Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache

Corpus Linguistics and Interdisciplinary Perspectives on Language

Bd. / Vol. 1

Marek Konopka / Jacqueline Kubczak
Christian Mair / František Štícha
Ulrich H. Waßner (Hgg.)

Grammatik und Korpora 2009

Dritte Internationale Konferenz

Grammar & Corpora 2009

Third International Conference

Mannheim, 22.-24.09.2009

narr
VERLAG

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

Gedruckt auf chlorfrei gebleichtem und säurefreiem Werkdruckpapier.

Internet: www.narr.de
E-Mail: info@narr.de

Layout: Tröster, Mannheim
Printed in Germany

ISSN 2191-9577
ISBN 978-3-8233-6648-5

Inhalt / Contents

III. Methodologie korpuslinguistischer Grammatikforschung / Methodologies of corpus-linguistic Grammar Research

IV. Einblicke in die aktuelle Forschung / Insights into current studies

Vorwort

Mit dem vorliegenden Band werden Beiträge zur Dritten Internationalen Konferenz *Grammatik und Korpora* (*Grammar & Corpora 3*) dokumentiert, die vom 22. bis zum 24. September 2009 an der Universität Mannheim stattfand und vom Institut für Deutsche Sprache (IDS) in Mannheim[1] in Zusammenarbeit mit der Albert-Ludwigs-Universität Freiburg[2] ausgerichtet wurde.[3] Die Organisation der Konferenz wurde von der Deutschen Forschungsgemeinschaft finanziell unterstützt.

Die Konferenzreihe *Grammar & Corpora* wurde 2005 von Frantisek Štícha (Prag) ins Leben gerufen, und so fanden die beiden ersten Konferenzen[4] in der Tschechischen Republik statt. Während die erste von korpusgrammatischen Projekten aus der Bohemistik bestimmt war, kamen schon auf der zweiten neben slawistischen Beiträgen stärker solche zum Zuge, die weitere europäische Sprachen, besonders Englisch und Deutsch, fokussierten. Als der Austragungsort der Konferenz bei ihrer Drittauflage nach Mannheim wanderte, fand schließlich eine regelrechte „Westerweiterung" statt. Nicht nur, dass korpusgestützte Grammatikforschung aus den Bereichen Germanistik und Anglistik noch ausgiebiger berücksichtigt werden konnte, jetzt spielte auch noch eine weitere Philologie eine markante Rolle, und zwar die Romanistik. Insgesamt wuchs die Zahl der genauer thematisierten Einzelsprachen auf zehn an. Entsprechend internationalisierte sich auch die Referentenschar weiter, in der jetzt Vertreterinnen und Vertreter aus 18 Ländern zu finden waren.

---

1 Vertreten durch Marek Konopka, Jacqueline Kubczak und Ulrich H. Waßner, Mitarbeiter des Projekts „Grammatische Variation im standardnahen Deutsch", vgl. http://www.ids-mannheim.de/gra/korpusgrammatik.html.

2 Vertreten durch Christian Mair, Englisches Seminar / Freiburg Institute for Advanced Studies (FRIAS).

3 Vgl. Brunner, Annelen / Hein, Katrin / Hennig, Sophie (2009): Bericht von der Dritten Internationalen Konferenz „Grammatik und Korpora", Mannheim, 22.-24.9.2009. In: Sprachreport 4 / 2009: 26-29.

4 Zu den Beiträgen vgl. Štícha, František / Šimandl, Josef (Hg.) (2007): Grammatika a korpus / Grammar & Corpora 2005. Praha: Ústav pro jazyk český, AV ČR; sowie Štícha, František / Fried, Mirjam (Hg.) (2008): Grammar & Corpora / Grammatika a korpus 2007. Praha: Academia.

Die Entwicklung, die die Konferenzreihe nahm, ist nicht unabhängig zu sehen von der Etablierung und dem rasanten Fortschritt der korpusorientierten Grammatikforschung. Sie spiegelt auch die Tatsache wider, dass in der sich auf digitale Korpora stützenden Linguistik neben einzelsprachlichen Schwerpunkten schon immer übereinzelsprachliche Aspekte von besonderer Bedeutung waren. Dies rührt ursprünglich daher, dass man im Bemühen, zum einen möglichst große Datenbestände anzulegen und recherchierbar zu machen und zum anderen ständig die Möglichkeiten für die automatische Auswertung von Sprachdaten zu optimieren, darauf angewiesen ist, immer wieder Anschluss an die weltweite technologische Entwicklung zu suchen. Die Entwicklungen im methodologischen Bereich sind folglich oft nicht einzelsprachgebunden, was übereinzelsprachliche Perspektiven fördert und nicht zuletzt auch kontrastiven oder komparativen Studien zugutekommt.

Die beiden Seiten korpusorientierter Grammatikforschung vor Augen – die oft primär einzelsprachbezogenen grammatischen Untersuchungen und die eher übereinzelsprachlichen methodologischen Aspekte –, entschlossen sich die Organisatoren, die Konferenz größtenteils in zwei entsprechenden Arbeitsgruppen abzuhalten. Dies sollte allerdings keine Trennung in voneinander isolierte Abteilungen bedeuten, was schon dadurch deutlich wurde, dass den Arbeitsgruppensitzungen die nicht auf einen Bereich fixierten Plenarvorträge vorangestellt wurden und dass auch eine – beide Bereiche übergreifende – Postersession hinzukam. Ziel einer solchen Strukturierung der Konferenz war es also vielmehr, Vertreter der beiden Bereiche zusammenzubringen und gebührend zu Wort kommen zu lassen – in der Erwartung, dass dadurch der kritische Dialog in konstruktiver Weise vorangetrieben wird. Auch auf die verschiedenen Traditionen und unterschiedlichen empirischen Schwerpunktsetzungen der Grammatikforschung in den Einzelphilologien blickend, hofften die Organisatoren, mit der international angelegten Konferenz einen Umschlagplatz für Ideen zu bieten, wo Unterschiede produktiv genutzt werden können, indem philologische Tradition und empirisch fundierte Beschreibung, theoretische Linguistik, Korpuslinguistik und Computerlinguistik über die Grenzen der untersuchten Sprachen hinweg in eine Diskussion miteinander treten.

Die Anlage der Konferenz schlägt sich in der Struktur des vorliegenden Bandes nieder. Er ist in die Kapitel I. „Plenarvorträge", II. „Korpusgestützte Grammatikforschung", III. „Methodologie korpuslinguistischer Grammatikforschung" und IV. „Einblicke in die aktuelle Forschung" aufgeteilt, wobei das

erste und das letzte Kapitel thematisch unspezifisch bleiben und letzteres die auf Posterpräsentationen zurückgehenden bzw. vom Umfang her etwas kürzeren Beiträge gruppiert. Den Herausgebern ist bewusst, dass kein Versuch, die Beiträge innerhalb der Kapitel in einer bestimmten Weise anzuordnen, bei der Fülle der behandelten Aspekte den Kreuz-und-Quer-Bezügen zwischen den Aufsätzen wirklich gerecht werden kann. Sie entschieden sich daher für eine Reihenfolge, deren einzige Aufgabe es ist, die Auffindbarkeit des näher Zusammenhängenden für den punktuell interessierten Leser zu erleichtern: Mit Ausnahme von Kapitel III sind die Beiträge innerhalb der Kapitel nach behandelten bzw. zur Veranschaulichung der Thesen benutzten Sprachen geordnet, wobei die durch die Anzahl der jeweiligen Beiträge naheliegende Reihenfolge Deutsch – Englisch – slawische Sprachen – romanische Sprachen gewählt wurde. Da in Kapitel III allgemeine methodologische Probleme und weniger die Einzelsprachgrammatiken im Vordergrund standen, musste hier ein anderes Ordnungsprinzip eingeführt werden: Eher programmatische bzw. theoretische Beiträge wurden vorangestellt, und für die weiteren Beiträge wurde – nach dem schwerpunktmäßig behandelten grammatischen Bereich und traditionellen Vorstellungen folgend – die Reihenfolge von der Graphematik/Phonologie über Morphologie und Syntax bis hin zu ebenenübergreifenden Fragestellungen gewählt.

Aus unserer Sicht als Organisatoren und Herausgeber hat die Tagung die weiter oben angedeuteten Grundannahmen eindrucksvoll bestätigt: (1) Die systematische Nutzung von Korpora hat die Grammatikforschung in den letzten Jahren wirkungsvoll vorangebracht und birgt großes Potenzial für weitere Fortschritte, und (2) die weithin übliche Fokussierung auf die jeweils einzelsprachliche korpuslinguistische Tradition ist schädlich und muss im kontinuierlichen Dialog über die Fachgrenzen hinweg überwunden werden. Darüber hinaus gab es aber auch unerwartete und gerade deshalb wertvolle Lernerfahrungen, die durchaus Anlass für weiteres Nachdenken sein sollten.

Korpora sind nützlich und haben die Grammatikschreibung auf vielfältige Weise befruchtet. Sie vermitteln adäquatere Einsichten in den tatsächlichen Gebrauch einer Sprache und ermöglichen es manchmal überhaupt erst, Phänomenen und Regularitäten auf die Spur zu kommen, die in den Grammatikbüchern bisher zu wenig oder noch gar keine Beachtung gefunden haben und vor allem auch nicht normativ reguliert sind. Das ist unbestritten. Dennoch bleibt ihr Nutzen im Einzelnen noch genauer zu definieren. Korpusdaten sind

solchen Daten, die aus anderen Quellen gewonnen werden – etwa in Elizitationsexperimenten oder durch Befragung der muttersprachlichen Intuition –, immer dann überlegen, wenn es um die statistische Verteilung von Varianten in Texten oder um die Einbettung von grammatischen Strukturen in authentische Verwendungskontexte geht. An einem einfachen Beispiel illustriert, das sich an den Beitrag von Julia Richling im vorliegenden Band anlehnt: Wie häufig man im heutigen Deutsch in E-Mail-Kommunikation *ich geb* anstelle von *ich gebe* schreibt, erfährt man nicht durch Befragung seiner eigenen Intuition und auch nicht durch Befragung anderer Sprecherinnen und Sprecher, sondern aus einschlägigen Korpora.

In welchem Umfang aber authentische Korpusdaten in unsere Referenzgrammatiken, Wörterbücher und Lehrmaterialien Eingang finden müssen, ist eine sehr viel schwierigere Frage. Niemandem würde etwa einfallen, den folgenden Satz aus einem englischen Korpus (F-LOB)[5]

> [...] we demonstrated a modified sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) method for visualization of factor VIII heavy chain (FVIII HC) polypeptides.

als Illustrationsbeispiel für den einfachen transitiven Satz des Musters S-V-O anzuführen – dazu ist das Objekt *a modified sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) method for visualization of factor VIII heavy chain (FVIII HC) polypeptides* einfach zu lang und auch zu unverständlich. Wohl seinen Sinn hätte ein solches Beispiel aber natürlich im fachsprachlichen Unterricht für Naturwissenschaftler.

Ein anderes Problem macht uns das nächste Beispiel bewusst. Es entstammt einem transkribierten Radio-Interview aus einem anderen englischen Korpus (ICE-GB)[6]. Ein Choreograph beantwortet die Frage, worin sich aus seiner Sicht die Arbeit mit Behinderten von der mit professionellen Tänzern unterscheidet.

---

[5] F-LOB, „Freiburg update of the LOB (= Lancaster-Oslo / Bergen) Corpus" – entstanden Anfang der 1990er Jahre in Freiburg im Rahmen eines Projekts zur systematischen korpusgestützten Untersuchung von aktuell ablaufenden Sprachwandelprozessen im heutigen Englisch. Siehe auch http://icame.uib.no/newcd.htm.

[6] ICE-GB ist die britische Komponente des International Corpus of English, einer Sammlung von Vergleichskorpora zu derzeit zehn nationalen Standardvarietäten des Englischen. Zu ICE-GB vgl. http://www.ucl.ac.uk/english-usage/projects/ice-gb/, zum Projekt insgesamt: http://ice-corpora.net/ice/index.htm.

> Well I think I think it's interesting because there's limitations put on what you're able to do and it's finding ways around those limitations
>
> And I think when you're <,> choreographing <,> uhm with very able-bodied people particularly because I'm used to <,> working <;> with very able-bodied people who are capable of doing more or less anything you ask <,> then uh you don't have those restrictions you uhm <,> whereas within this particular class you- there are limitations
>
> And it makes it fa- in a way far <,,> easier I'd say to be able to create because <,> you can't go off at- in all tangents or in c- every direction

Das Beispiel steht für Glanz und Elend der korpuslinguistischen Behandlung von gesprochener Spontansprache. Es zeigt sehr schön die spezifische grammatische Komplexität der Sprechsprache, erlaubt andererseits aber keinerlei Rückschlüsse auf die Zusammenhänge zwischen grammatischer Struktur und Prosodie – wohl die zentrale Frage in jeder grammatischen Analyse gesprochener Sprache.

Während die Größe der zur Verfügung stehenden geschriebensprachlichen Korpora und gleichzeitig die Ansprüche an deren für erforderlich gehaltenen Umfang exponential steigt – wo man vor 25 Jahren mit einer Million Wörtern zufrieden war, hat man heute oft das Hundert- und Tausendfache zur Verfügung –, sind die Korpora gesprochener Spontansprache nach wie vor eher klein und stehen, was fast noch schlimmer ist, sehr oft nur in Form mäßig annotierter orthographischer Transkriptionen zur Verfügung. Hier besteht eine Forschungslücke, die noch gefüllt werden muss, wenn die Arbeiten zur Grammatik gesprochener Sprache die Qualität der korpusgestützten Forschung zur geschriebenen Sprache erreichen sollen.

Was den von uns angestrebten wissenschaftlichen Dialog über die Fachgrenzen hinweg betrifft, wurde uns vor Augen geführt, dass dieser zwar allseits gewünscht und begrüßt wird, in der praktischen Umsetzung jedoch einer Strategie bedarf, die wir als *intelligente Mehrsprachigkeit im akademischen Bereich* bezeichnen möchten. Als Tagungssprachen setzten wir Deutsch, eine wichtige Objektsprache und die Sprache des Gastgeberlandes, sowie Englisch als die führende Wissenschaftssprache in der Linguistik fest. Gerade die Philologien widersetzen sich mit Recht einer sprachlichen Monokultur – und sind dennoch für die optimale Verbreitung ihrer Ergebnisse häufig auf das Englische als Wissenschaftssprache angewiesen. Ein Vorschlag für eine intelligente Mehrsprachigkeit in der Wissenschaft muss dieses Paradox anerkennen. Wir

plädieren dafür, es nicht dadurch aufzulösen, dass Wissenschaftlerinnen und Wissenschaftler auf Kosten ihrer Muttersprache oder der untersuchten Objektsprache auf das Englische verpflichtet werden, sondern dass sie das Englische als *zusätzliches* Medium zur Verfügung gestellt bekommen – etwa in Form von Übersetzungsangeboten bei Konferenzen und nachfolgenden Publikationen. Auf diese Weise wird mehrsprachige Linguistik auch in einem zunehmend einsprachigen globalen Wissenschaftssystem möglich sein und gedeihen.

Eine Fortführung des wissenschaftlichen Gedankenaustausches zu den genannten Fragekomplexen – sowohl des korpuslinguistischen im engeren Sinn als auch des weiteren sprachpolitischen – ist notwendig, und eine Fortsetzung der Reihe daher höchst wünschenswert. In Anbetracht des oben Gesagten erschiene für die korpusorientierte Darstellung grammatischer Variabilität im standardsprachlichen und nicht-standardsprachlichen Gebrauch im Rahmen von *Grammar & Corpora 4* ein romanischsprachiges Land als Austragungsort besonders geeignet.

Zurück jedoch zur Gegenwart und zu gebührenden Worten des Dankes, die wir gerne aussprechen. Dieser Tagungsband hätte ohne die tatkräftige und auch inhaltlich anregende Hilfe vieler Personen nicht und vor allem nicht in der vorliegenden Form zustande kommen können. Es ist den Herausgebern deswegen ein ehrliches Anliegen, den Autoren der Beiträge sowie insbesondere den im Folgenden genannten Personen für ihre Unterstützung zu danken. So stand uns das Programmkomitee der Konferenz zur Seite, in dem Holger Keibel, Mannheim, Stefan Pfänder, Freiburg, und Gisela Zifonun, Mannheim, mitwirkten. Claire Holfelder, Speyer, hat viele der englischsprachigen Beiträge mit muttersprachlicher und inhaltlicher Kompetenz redigiert. Bei allen editorischen Tätigkeiten wurden wir intensiv und kompetent unterstützt von unseren studentischen Hilfskräften Ulrike Stölzel (Mannheim), Anastasia Cobet und Susanne Gundermann (Freiburg). In der Publikationsstelle des IDS kümmerten sich Sonja Tröster und Norbert Volz um die Endredaktion sowie um das Layout und die Erstellung der Druckvorlage. Nicht zuletzt hat der Direktor des IDS, Ludwig M. Eichinger, auf vielfältige Weise die Publikation dieses Tagungsbandes unterstützt. Ihnen allen gilt unser herzlicher Dank!

Die Herausgeber

Preface

The present proceedings contain a collection of papers presented at the Third International Conference *Grammar & Corpora*, which took place at the University of Mannheim, Germany, from 22 to 24 September 2009, and which was hosted by the Institut für Deutsche Sprache (IDS, Institute for the German Language), Mannheim,[7] in co-operation with the Albert Ludwig University of Freiburg.[8] The conference was sponsored by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation).

The conference series *Grammar & Corpora* was inaugurated in 2005 by Frantisek Štícha (Prague), and hence the first two conferences[9] took place in the Czech Republic. While the first of them was largely devoted to corpus-grammatical projects from the field of Czech linguistics, papers focusing on other European languages, particularly English and German, were already much in evidence at the subsequent gathering. When Mannheim was chosen as the venue of the third conference, a veritable "westward expansion" occurred. Not only was corpus-based grammar research from the fields of English and German philology incorporated more extensively, but, with the addition of papers devoted to French and other Romance languages, another important gap in coverage was filled. In total, the number of individual languages addressed in detail increased to ten, and accordingly, the group of contributors became more international as well, with representatives now originating from 18 countries.

The evolution of the conference series cannot be seen independent of the establishment and rapid development of corpus-oriented grammar research. It also reflects the fact that, apart from language-specific concerns, cross-linguis-

---

7 Represented by Marek Konopka, Jacqueline Kubczak, and Ulrich H. Waßner, staff members of the project "Grammatische Variation im standardnahen Deutsch" ("Grammatical variation in near-standard German"), cf. www.ids-mannheim.de/gra/korpusgrammatik.html. For a brief report on the conference see Brunner, Annelen / Hein, Katrin / Hennig, Sophie (2009): Bericht von der Dritten Internationalen Konferenz "Grammatik und Korpora", Mannheim, 22.-24. 9. 2009. In: Sprachreport 4 / 2009: 26-29.

8 Represented by Christian Mair, English Department / Freiburg Institute for Advanced Studies (FRIAS).

9 As for the contributions cf. Štícha, František / Šimandl, Josef (eds.) (2007): Grammatika a korpus / Grammar & Corpora 2005. Praha: Ústav pro jazyk český, AV ČR; and Štícha, František / Fried, Mirjam (eds.) (2008): Grammar & Corpora / Grammatika a korpus 2007. Praha: Academia.

tic aspects have always been of great significance in linguistic research that is based on digital corpora. This is of course chiefly due to the fact that, in seeking to create as large data bases as possible and in striving to optimize the possibilities for automatic analysis of linguistic data, one needs to keep up with the worldwide technological state of the art. Consequently, methodological developments in corpus-based research are frequently not tied to individual languages, and this, ultimately, favors cross-linguistic perspectives and, above all, contrastive or comparative studies.

Bearing in mind these two facets of corpus-based grammar research – the oftentimes language-specific grammatical investigations on the one hand and the cross-linguistically relevant methodological aspects on the other –, the organizers decided to set up the conference in two corresponding work groups. This was not meant to constitute a separation into two isolated sections, which was already apparent from the fact that the work groups' meetings were preceded by plenary talks which were not committed to one particular section and were accompanied by a poster session also encompassing both sections. Thus, the purpose of the conference design was rather to bring together representatives of the two (sub)fields and let them have their due say, assuming that in this way a critical but constructive dialog would be furthered. Also in view of the diverse traditions and different empirical focal points of research within the individual philologies, the organizers hoped to make this international conference a marketplace of ideas, where differences would turn out to be productive, with philological tradition and empirical description, linguistic theory, corpus linguistics, and computational linguistics embarking on discussions beyond the boundaries of the single languages under scrutiny.

The design of the conference is reflected in the structure of the present proceedings, which are divided into the Chapters I. "Plenary talks", II. "Corpus-based grammar research", III. "Methodologies of corpus-linguistic grammar research", and IV. "Insights into current research". The first and the last chapter remain thematically unspecific, with the latter grouping together papers originating in the poster session and some shorter papers read at the conference. The editors are aware that, due to the multiplicity of aspects addressed, no attempt at ordering the contributions within the chapters in a particular manner could truly do justice to the cross-references between the articles. Therefore, they opted for an order which simply serves to facilitate finding the more closely related ones for the selective reader. Except for Chapter III, the contri-

butions are ordered according to the languages treated or used for the illustration of theoretical arguments. Here, the sequence German – English – Slavic languages – Romance languages has been chosen since this suggested itself on the basis of the respective number of contributions. Given that Chapter III is rather concerned with general methodological problems than with language-specific grammars, a distinct organizing principle had to be introduced here: More programmatic or theoretical contributions were placed first, and other papers followed – according to their focal area of grammatical research and following the traditional conception – from graphematics/phonology via morphology and syntax to cross-thematic issues.

From our perspective as organizers and editors, the conference has impressively confirmed the two above-sketched assumptions: (1) The systematic use of corpora has effectively advanced grammar research during the past few years and bears great potential for further progress, and (2) the widely practiced restriction to language-specific corpus-linguistic traditions is detrimental and needs to be overcome through continuous transdisciplinary dialog. On the other hand – and we do not want to conceal that either – there have also been unexpected and therefore all the more valuable learning experiences, which should by all means serve as a starting point for further contemplation.

Corpora are useful and have stimulated grammar writing in manifold ways. They provide insights into actual language use and sometimes even make us aware of phenomena and regularities which have gone unnoticed in the grammatical tradition and have never been subject to prescriptive regulation. This is undisputed. Nonetheless, their particular usefulness remains to be defined more precisely in each individual case. Corpus data have an advantage over data obtained from other sources – such as elicitation experiments or native speaker interviews –, whenever it is the statistical distribution of variants in texts or the embedding of grammatical structures in authentic usage contexts that is at issue. Illustrating this with a simple example that is based on Julia Richling's article in the volume at hand: We will not learn about the frequency of the spelling variant *ich geb* instead of *ich gebe* in present-day German e-mail communication by consulting our own intuition or by asking other speakers but only by turning to pertinent corpora.

However, it is a much more difficult question to what extent authentic corpus data should enter our reference grammars, dictionaries, and teaching materi-

als. It would not occur to anyone, for instance, to cite the following sentence from an English corpus (F-LOB)[10] in order to illustrate the simple transitive sentence pattern S–V–O:

> [...] we demonstrated a modified sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) method for visualization of factor VIII heavy chain (FVIII HC) polypeptides.

For that purpose, the object *a modified sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) method for visualization of factor VIII heavy chain (FVIII HC) polypeptides* is simply too long and too unintelligible as well. Yet, such an example would very well make sense in a scientific-English class using technical jargon.

The next example brings to mind a further problem. It is taken from the transcription of a radio interview from another English corpus (ICE-GB)[11]. A choreographer is telling us how he thinks his work with disabled people is different from his work with professional dancers:

> Well I think I think it's interesting because there's limitations put on what you're able to do and it's finding ways around those limitations
>
> And I think when you're <,> choreographing <,> uhm with very able-bodied people particularly because I'm used to <,> working <;> with very able-bodied people who are capable of doing more or less anything you ask <,> then uh you don't have those restrictions you uhm <,> whereas within this particular class you- there are limitations
>
> And it makes it fa- in a way far <,,> easier I'd say to be able to create because <,> you can't go off at- in all tangents or in c- every direction

This excerpt exemplifies the misery and the splendor of the corpus-linguistic treatment of spontaneous speech. On the one hand, it very nicely illustrates the specific grammatical complexity of spoken language, but on the other hand, it does not allow for any conclusions about the correlations between grammatical structure and prosody – probably the central concern of any grammatical analysis of spoken language.

[10] F-LOB, the Freiburg update of the LOB (= Lancaster-Oslo / Bergen) Corpus, was compiled in the 1990s in Freiburg in the frame of a project on the corpus-based real-time study of ongoing morphosyntactic change in present-day English. See http://icame.uib.no/newcd.htm for further details.

[11] ICE-GB is the British component of the International Corpus of English, an expanding collection of comparable one-million-word corpora documenting at present ten national varieties of Standard English. Cf. www.ucl.ac.uk/english-usage/projects/ice-gb/ for ICE-GB, and http://ice-corpora.net/ice/index.htm for the project as a whole.

While the benchmark for the size of written corpora rises exponentially – where one used to be satisfied with a million words 25 years ago, a hundred or even a thousand times that amount of text is now common –, our corpora of spontaneous speech still remain fairly small and, what is almost worse, are often only available in the form of moderately annotated orthographic transcriptions only. This constitutes a research gap that must be closed if studies dealing with spoken language grammar are to measure up to the quality of corpus-based research on written language.

With regard to the transdisciplinary scientific dialog we were aspiring to, we realized that it is indeed desired and welcomed on all sides, but its practical implementation requires a strategy which we would like to refer to as *intelligent multilingualism in academia*. We decided that the conference languages should be German, an important object language and the language of the host country, and English, the leading academic language in the field of linguistics. It is precisely the language sciences that must take a stand against a linguistic monoculture – but nevertheless, they are often dependent on English as the language of science for the sake of an optimal dissemination of their findings. Our plea for an intelligent multilingualism in academia acknowledges this paradox and advocates the position that it should not be solved by obliging individual researchers to use English at the expense of their native tongue or the object language under scrutiny but by providing them with the English language as an additional medium – for example by offering translation support at conferences and for the subsequent publication. In this way, a multilingual linguistics will be viable and thrive even in a global scientific system that is increasingly monolingual.

It is necessary to continue the scholarly exchange of ideas concerning this range of issues – both those relating to corpus-linguistics in the narrower sense and those regarding language policies in a wider sense –, and hence a continuation of the *Grammar & Corpora* conference series is very desirable. Bearing in mind the above-said, "Charting grammatical variability in standard and non-standard usage" would be an excellent topic for *Grammar & Corpora 4*, and a Romance language-speaking country the appropriate venue.

But back to the present and to due words of thanks. These conference proceedings could not have been accomplished without the active and enthusiastic support of many people. First of all, the editors would like to thank the authors of the contributions for meeting tight publication deadlines. In addition, we

were assisted by the program committee of the conference, which included Holger Keibel, Mannheim, Stefan Pfänder, Freiburg, and Gisela Zifonun, Mannheim. Claire Holfelder, Speyer, helped edit many of the English articles with both native speaker proficiency and thematic competence. In all other editorial matters, we were supported substantially and competently by our student assistants Ulrike Stölzel (Mannheim), Anastasia Cobet and Susanne Gundermann (Freiburg). At the Publication Office of the IDS, Sonja Tröster and Norbert Volz took care of the final editing, layouting, and typesetting. Last but not least, the director of the IDS, Ludwig M. Eichinger supported this publication in many ways. It is to all these that we would like to express our sincere gratitude.

The editors

I. Plenarvorträge / Keynote Speeches

BRUNO STRECKER

Korpusgrammatik zwischen reiner Statistik und „intelligenter“ Grammatikografie

Abstract

Von Grammatikern erwartet man Auskunft darüber, wie man zu reden und zu schreiben hat, eine Erwartung, die sich auf die Annahme stützt, es stehe grundsätzlich immer schon fest, was in Sprachen wie etwa dem Deutschen als korrekt gelten kann. Tatsächlich kann jedoch nicht einmal davon ausgegangen werden, dass es so etwas wie das Deutsche als eindeutig bestimmten Gegenstand gibt. Alles, was als Deutsch zu fassen ist, sind ungezählte schriftliche und – sofern aufgezeichnet – mündliche Äußerungen. Bis vor wenigen Jahren waren diese Daten praktisch nur unzureichend zu nutzen, weshalb Grammatikern wenig anderes übrig blieb, als auf der schmalen Basis durch Introspektion gewonnener Daten Simulationen eines allgemeinen Sprachgebrauchs zu entwickeln. Mit der Verfügung über riesige Korpora maschinenlesbarer Texte haben sich die Voraussetzungen für die Untersuchung grammatischer Strukturen entscheidend verändert. Für die Grammatikforschung ergaben sich damit neue Perspektiven: zum einen ein radikaler Bruch mit der Tradition grammatischer Analysen, der weitgehend auf eine statistische Auswertung von Kookkurrenzen setzt, zum andern – weniger radikal, mehr traditionsverbunden – die Möglichkeit, konventionell kompetenzgestützt erarbeitete Regelhypothesen anhand von Daten zu validieren, wie sie in sehr großen Textkorpora vorliegen und dem, was als Deutsch gelten kann, so nah kommen, wie dies irgend erreichbar ist, da sie durchweg in dem Bemühen zustande kamen, sich korrekt auszudrücken.

1. Was von Grammatikern erwartet wird

Von Grammatikern erwartet man vor allem, dass sie Auskunft darüber geben können, wie man reden und schreiben soll, wenn man Fehler vermeiden will.[1] Bei Muttersprachlern beschränkt sich dies – das zeigen zahllose Anfragen bei diversen Sprachberatungsstellen – in aller Regel auf Entscheidungen in Zwei-

[1] Selten ist mir diese Erwartung klarer und eindringlicher begegnet als bei einer von mir geleiteten Podiumsdiskussion anlässlich der Jahrestagung 2008 des Instituts für Deutsche Sprache. Eine – mir nicht namentlich bekannte – Dame aus dem Publikum erklärte, nachdem sie einen Anspruch „der Leute“ auf Richtlinien unterstellt hatte, dass wir – die Grammatiker – zu bestimmen hätten, was „Sache ist“, und dies selbst dann, wenn nach Lage der Forschung nichts als ausgemacht korrekt gelten kann.

felsfällen. Wer Deutsch als Zweit- oder Fremdsprache erlernen will oder gar muss, geht meist noch weiter und erwartet von einer Grammatik, dass sie die Sprache insgesamt in klaren, eindeutigen Regeln erfasst.

In beiden Fällen sind die Erwartungen durchaus verständlich: Wer hierzulande die Schule besucht hat, verlässt diese in der festen Überzeugung, in Sachen deutsche Sprache sei immer und überall klar geregelt, was als korrekt zu gelten hat und was nicht. Und wer sich der Mühe zu unterziehen hat, zusätzlich noch Deutsch zu lernen, wünscht sich wenigstens eindeutige und möglichst einfache Regeln.

Einen Grammatiker bringen solche Erwartungen in eine verzwickte Lage, denn, was er auch tut, er wird manchen gegen sich aufbringen:

- Akzeptiert er als ehrlicher Forscher die Rolle nicht, die man ihm geradezu aufdrängen will, gefährdet er sich und letztlich sogar seinen ganzen Berufsstand, da er partout nicht leisten will, was alle Welt – Kollegen natürlich ausgenommen – von ihm erwartet.
- Lässt er sich auf das Spiel ein, werden ihn kluge Leute mit einigem Recht fragen, wie er denn dazu komme, sich in einer Angelegenheit von derart allgemeinem Interesse als Richter aufzuspielen.

Die Chancen, sich aus dieser Doppelbindung zu befreien, stehen für Grammatiker nicht allzu gut, denn das Dilemma ergibt sich stets neu aus einer fatalen Grundannahme hinsichtlich ihres Forschungsgegenstands: Alle Welt redet von Sprachen so, als seien sie an sich klar umrissene Systeme, die freilich steter sorgsamer Pflege bedürfen, damit sie nicht durch unbedachten Gebrauch verunreinigt werden. Dass diese Grundannahme selbst problematisch sein könnte, kommt nur wenigen in den Blick.

Wie verhält es sich wirklich mit Gegenständen dieser Art, etwa mit der deutschen Sprache? Unter den Dingen, die als deutsch bezeichnet werden, gibt es tatsächlich manches, das als eindeutig bestimmt gelten kann, so etwa das deutsche Staatsgebiet, die deutsche Staatsangehörigkeit, die deutsche Fußballnationalmannschaft. Schwieriger wird es schon, wenn von deutscher Baukunst, deutschem Denken oder deutscher Lebensart die Rede ist. Gefragt, was dies denn sei, wird man bestenfalls mit Beispielen antworten können und sich im Übrigen auf dieselbe Weise aus der Affäre ziehen wie Augustin, der auf die Frage, was die Zeit sei, antwortet:

> Was ist also die Zeit? Wenn mich keiner danach fragt, weiß ich es, wenn ich es Fragenden erklären will, weiß ich es nicht ...[2]

Will man sich darüber klar werden, um welcher Art Gegenstand es sich bei der deutschen Sprache handelt, sollte man genauer betrachten, was etwa den Unterschied zwischen dem deutschen Staatsgebiet und deutscher Lebensart ausmacht. Von beiden wird gesagt, sie seien deutsch, doch, was das eine klar umrissen erscheinen lässt, fehlt gerade beim zweiten, nämlich Grenzen, die sich nicht einfach natürlich ergeben haben, sondern ausdrücklich gezogen wurden.

Wie die Dinge liegen, gleicht das Deutsche mehr deutscher Lebensart denn dem deutschen Staatsgebiet, und dies ließe sich allenfalls ändern, wenn die Gesetzgeber deutschsprachiger Länder für den Geltungsbereich ihres Rechts detaillierte und umfassende Regelungen für den Sprachgebrauch erlassen und diese, wenn erforderlich, auch bei Strafandrohung durchsetzen. Doch, was so entstehen könnte, wäre dann wesentlich anderer Natur als das, was uns als unsere Sprache in einem Prozess überkommen ist, den man mit Adam Smith (1776) als das Wirken einer unsichtbaren Hand bezeichnen kann.

2. Das Deutsche – was ist das?

Wenn das Deutsche etwas in der Art der deutschen Lebensart oder des deutschen Geistes ist, wie kann es dann überhaupt Gegenstand einer Grammatik sein? Nun, auf dieselbe Weise, in der auch die deutsche Lebensart Gegenstand der Erforschung sein kann: Man sammelt und analysiert möglichst viele Erscheinungen, von denen man annimmt, dass sie als deutsch durchgehen, oder man kramt – so man sich selbst zu den Sprechern des Deutschen rechnet – in seinen Erinnerungen, um nach und nach erst Einzelereignisse, dann Gebräuchliches, Musterhaftes herauszuarbeiten.

Bevor ich genauer darauf eingehe, in welcher Form beides zu realisieren ist und was damit jeweils zu erreichen ist, noch zwei Bemerkungen:

- In jedem Fall kann festgehalten werden, dass so etwas wie *das* Deutsche selbst nicht Gegenstand der Forschung sein kann, denn das Deutsche ist, um mit Kant zu sprechen, nichts als eine Hypostasierung – böse Menschen würden sagen, eine *idée fixe*. Wir wären deshalb nicht schlecht beraten,

[2] „Quid est ergo tempus? si nemo ex me quaerat, scio; si quaerenti explicare velim, nescio." (Aurelius Augustinus Confessiones XI, 14).

wenn wir gar nicht erst von *dem* Deutschen reden wollten. Als bloße *façon de parler* mag das ja angehen, aber selbst dann ist man ständig in Gefahr, im Sinn einer Vergegenständlichung missverstanden zu werden.

– Kritische Leser könnten vermuten, damit stünde man kurz vor einem viziösen Zirkel: Wenn zwar das Deutsche in seiner Totalität nicht zu erforschen sein soll, aber doch von Fall zu Fall zu entscheiden bleibt, ob etwas als deutsch gelten kann oder nicht, dann scheint das Grundproblem nur verschoben, nicht aufgehoben.

Die Vermutung ist verständlich, aber sie verkennt, dass mit den fallweisen Entscheidungen das Problem in eine Form gebracht wird, in der es doch noch eine praktikable Lösung finden kann, denn einen Text oder eine mündliche Äußerung als deutsch oder nicht deutsch zu klassifizieren, erfordert keine prinzipiell anderen Fähigkeiten, als etwa Gegenstände nach ihrer Farbe als rot oder nicht rot zu sortieren. Man braucht dafür keine allgemeine Idee davon zu bemühen, was denn nun „Röte" eigentlich sei. Wieso sollte man dergleichen im Fall von Texten brauchen? Tatsächlich kann es sogar gelingen, Texte nach Sprachen zu klassifizieren, die man überhaupt nicht versteht, solange man nur einige ihrer Besonderheiten kennt, etwa so: Viele Wörter mit *ö* und *ü*, am Ende eines Wortes oft *-ler* oder *-lar*, ein *i* ohne I-Punkt – das dürfte Türkisch sein; ab und an eine Tilde auf einem *n*, das wird wohl Spanisch sein. Wollte man prinzipiell in Frage stellen, dass wir in der Lage sind, ein Prädikat wie „ist deutsch" korrekt anzuwenden, ohne gleich schon zu wissen, was *das* Deutsche ist, dann hätten wir wenig Grund anzunehmen, dass wir *irgendein* Prädikat korrekt verwenden können, und könnten ebenso gut gleich alles Reden einstellen.

Kann man erst einmal davon ausgehen, dass sich feststellen lässt, ob ein gegebener Text oder Gesprächsbeitrag deutsch gehalten ist, steht im Prinzip alles, was so geschrieben, gedruckt oder aufgezeichnet wurde, als Ausgangsmaterial für eine Erforschung von Regularitäten und Konventionen zur Verfügung, die in der Kommunikation unter Deutschsprechenden eine Rolle spielen können. Und das ist weit, weit mehr als selbst kompetenteste Sprachteilhaber in einem langen Leben je zur Kenntnis nehmen könnten.

3. Kompetenz und/oder Korpus

Man mag darüber streiten und man hat darüber gestritten, ob Grammatiken ernstlich dazu beitragen können, Menschen in die Lage zu versetzen, erfolgreich mit Mitgliedern einer Sprachgemeinschaft zu kommunizieren, in der sie nicht aufgewachsen sind, denn schließlich kamen jene selbst weitestgehend ohne Grammatik zu ihrer Sprachkompetenz. Eine interessante Funktion scheinen mir Grammatiken jedoch in jedem Fall haben zu können: Sie können, Landkarten gleich, Orientierungshilfen geben. Entsprechend sollte ihre Leistung, ganz wie jene von Landkarten, danach bewertet werden, als wie hilfreich sie sich erweisen. Gelingt es den Nutzern einer Grammatik, mit deren Hilfe komplexe Äußerungen besser zu verstehen und über ihre zunächst gegebenen Kenntnisse hinaus, selbst akzeptable Gesprächsbeiträge oder Texte zu verfassen, dann leistet eine Grammatik, was man von ihr erwarten darf, und dann ist letztlich völlig gleichgültig, wie sie zustande kam.

Nur, wie kommt man zu einer solchen Grammatik? Wie bereits angesprochen, bieten sich grundsätzlich zwei Vorgehensweisen an:

a) Betrachtet man sich als kompetentes Mitglied der Sprachgemeinschaft, kann man damit beginnen, sich nach und nach daran zu erinnern, wie und mit welchen Mitteln man in dieser Gemeinschaft Aufgaben löst, die sich bei dem Versuch stellen, sich mit anderen zu verständigen. Und auf der Grundlage solcher Erinnerung kann man dann versuchen, in Form von Grammatikregeln eine Art Simulation der eigenen Praxis zu entwickeln. Ob Simulationen, die auf diese Weise zustande kommen, in irgendeiner Weise die mentalen Fähigkeiten kompetenter Sprachteilhaber nachbilden oder nicht, ist m.E. eine müßige Frage, denn sie kann nicht und muss auch nicht beantwortet werden. Kriterium für die Bewertungen solcher Simulationen kann einzig und allein sein, ob sie jemand, der sich an ihnen orientiert, in die Lage versetzen, an Kommunikationen zumindest insoweit erfolgreich teilzuhaben, dass seine Äußerungen akzeptiert werden.

b) Man kann nach der Wittgenstein'schen Maxime handeln,[3] nicht gleich zu denken, sondern erst einmal zu schauen und dazu möglichst viele Daten zu sammeln, die von kompetenten Sprechern und Schreibern *grosso modo* als relevante Sprachdaten erkannt werden. Man muss dabei nicht unbedingt selbst in der Lage sein, das Gesammelte zu verstehen. Auf der Basis der Datensammlung kann man sich dann daran machen, mittels statistischer Verfahren Musterhaftes und Wiederkehrendes herauszuarbeiten.

[3] Ludwig Wittgenstein (1953: § 66).

Noch bis vor wenigen Jahrzehnten konnte man Vorgehensweise (b) mehr als rein theoretische Möglichkeit betrachten denn als echte Alternative. Nicht dass es an Masse gefehlt hätte, wohl aber an effizienten Verfahren, die Datenmassen auszuwerten. Wirklich gangbar war nur der Weg über vorgängigen, mehr oder weniger natürlichen Spracherwerb und darauf aufbauende Introspektion, ein Verfahren, das man neudeutsch-technisch als kompetenzgestützt bezeichnen könnte.

So ganz bei Null musste dabei freilich schon lange, sehr lange niemand mehr beginnen, denn man konnte sich auf das stützen, was Generationen von Grammatikern bereits auf den Weg gebracht hatten. Und man hat sich darauf gestützt, manchmal auch auf Biegen und Brechen, etwa dann, wenn Kategorisierungen, die an griechischem und lateinischem Material entwickelt worden waren, auf Deutsches und gar Chinesisches angewandt wurden. Alles in allem kann man jedoch sagen, dass Grammatiker auf diese Weise mit Bienenfleiß Beachtliches erarbeitet haben. Der allgegenwärtige Streit der verschiedenen Schulen von Grammatiktheoretikern ist längst ein Streit auf sehr hohem Niveau.

Mit dem Aufkommen von Verfahren, Sprachdokumente in elektronischer Form zu speichern und mittels immer raffinierterer statistischer Algorithmen zu analysieren, ist den klassischen „intelligenten" Verfahren echte Konkurrenz erwachsen. Auch wenn mancher es nicht wahr haben mag: Die „dummen" Automaten können riesige Textkorpora durchforsten, dabei – ohne Rückgriff auf tradierte grammatische Kategorien – Strukturen und Muster auffinden, ja vielleicht bald schon Grammatiken weitgehend sich selbst finden lassen.[4] Kupietz / Keibel (2009a) sprechen in diesem Zusammenhang von Emergenz, ein Konzept, das zugleich hervorragend zu Vorstellungen davon passt, wie sich Konventionen und Regularitäten im Bereich der Kommunikation ohne bewusst planende, ordnende Eingriffe ergeben haben könnten.[5]

Heißt dies, dass bald schon Informatik und Statistik die klassische Grammatikforschung überflüssig machen werden? Für manche Anwendungsbereiche mag dies tatsächlich zutreffen, allerdings nicht unbedingt für jene, für die Grammatiken immer schon geschrieben wurden. Es wäre jedoch vermessen, wollte ich hier Prognosen wagen, denn dafür verstehe ich viel zu wenig von Informatik und Statistik.

---

4 So ganz ohne intelligente Eingriffe wird es dazu freilich nicht kommen, doch die Intelligenz mehr informatisch-mathematischer denn grammatischer Natur sein.

5 Keller (2003, 2009), Strecker (1987), Ullman-Margalit (1977).

Man muss hier freilich keinen Gegensatz sehen. Man kann die neuen technischen Möglichkeiten vielmehr auch als Chance begreifen, die klassische Grammatikforschung doch noch zu einem Zweig der empirischen Wissenschaften zu entwickeln, indem man die immer schon vorhandenen Ansätze zur empirischen Validierung von Hypothesen über Regularitäten um geeignete Recherchen in maschinenlesbaren Textkorpora erweitert, wie wir sie etwa am Institut für Deutsche Sprache mit dem Deutschen Referenzkorpus (DeReKo) zur Verfügung haben.

Grammatiker haben schon immer versucht, Bestätigungen für ihre Regelhypothesen zu finden, etwa indem sie das Urteil anderer Sprachteilhaber darüber einholten oder in Texten von anerkannten Autoren Textpassagen suchten, die ihre Regelhypothesen exemplifizierten oder zu exemplifizieren schienen. Ihr Horizont blieb dabei freilich verständlicherweise eher beschränkt. Erst mit dem Zugang zu riesigen maschinenlesbaren Textkorpora wie dem DeReKo und sicher auch mit der Nutzung von Internet-Suchmaschinen wie Google bot sich die Chance, den eigenen Horizont in ungeahnter Weise zu erweitern.

Was durch Recherchen in Korpora aufzufinden ist, hat zwar auf den ersten Blick einen gewaltigen Nachteil gegenüber einer Befragung kompetenter Sprachteilhaber, doch genauer besehen kann sich dies sogar als Vorteil erweisen. Der vermeintliche Nachteil besteht darin, dass Korpusrecherchen nicht ohne weiteres punktgenaue Antworten liefern, während befragte Sprachteilhaber direkt antworten können. Ein Beispiel: Man kann Probanden fragen, ob sie folgende Äußerung für korrekt formuliert halten:

> Wie er dem Mann sein Gesicht, der dort gestanden sei, gesehen hat, ist er sich nicht mehr ganz so sicher gewesen, ob er alles richtig gemacht hat, weil der hat ganz unglücklich ausgesehen.

Viele Probanden – zumal solche mir gehobener Schulbildung – würden vermutlich sehr schnell antworten, das alles sei gar nicht korrekt, man könne nicht sagen: *dem Mann sein Gesicht*, und schon gar nicht, dass dieser dort gestanden „sei", usw.

Hat sich die Sache damit erledigt? Ich denke nicht, und zwar aus einer ganzen Reihe von Gründen:

- Was, wenn dieser Satz nicht von mir zu Illustrationszwecken erfunden, sondern tatsächlich von jemandem vorgebracht worden wäre – in gutem Glauben, sich so korrekt auszudrücken? Nach meiner Erfahrung könnte

dies sehr wohl so sein. Wieso sollte dann das Urteil anderer, die sich nicht mit mehr Recht als Mitglieder derselben Sprachgemeinschaft betrachten können, mehr Gewicht haben als seine spontane Äußerung?

– Vielleicht würde sich der Sprecher ja sogar selbst korrigieren, wenn er kritisiert wurde. Doch bedeutet dies, dass er sich zunächst fehlerhaft geäußert hätte? Nicht unbedingt, denn bei seinem Urteil ließ er sich möglicherweise nicht von dem leiten, was er und zahllose andere üblicherweise tun, sondern von Normen, die man ihm im Unterricht eingetrichtert hat. Es lassen sich unschwer Beispiele dafür finden, dass Menschen ihren eigenen Sprachgebrauch nicht überblicken. Ein durchaus typisches Beispiel: Als ich eine sehr normbewusste Bekannte darauf hinwies, sie habe soeben gesagt: „So etwas habe ich noch nie gesehen gehabt", behauptet sie steif und fest, dergleichen käme ihr nie über die Lippen. Erst als weitere Anwesende meine Beobachtung bestätigten, gab sie klein bei, immer noch, ohne so recht zu glauben, dass dies alles seine Richtigkeit habe.

– Man sollte Äußerungen in ihrer natürlichen Umgebung betrachten. Als Beispielsätze werden sie uneigentlich verwendet. Sie stehen gewissermaßen voll im Schlaglicht und können deshalb Kritik auf sich ziehen, die bei einer Verwendung in einem alltäglichen Kontext gar nicht aufgekommen wäre.

– Wenn Formulierungen wie *dem Mann sein Gesicht* jahrhundertelang erfolgreich gebraucht wurden und noch werden,[6] und dies trotz massiver Kritik seitens selbst ernannter Sprachpfleger, dann kann es sich dabei wohl kaum um etwas handeln, was nicht zum tradierten Bestand deutscher Ausdrucksformen gehört.

Stützt man sich bei der Validierung von Regelhypothesen auf Textkorpora, wird man in aller Regel keine punktgenaue Antwort zu einem bestimmten Beispiel bekommen, denn hierfür ist die Menge möglicher Äußerungen bei Weitem zu groß. Es kann jedoch gelingen, eine aussagekräftige Antwort auf *die* Frage zu erhalten, die man *eigentlich* stellen wollte – oder sollte, denn Beispiele sind ja immer nur als eben solche gedacht. Eigentlicher Gegenstand der Validierung ist nicht das konkrete Beispiel, sondern die Regelhypothese nach der es akzeptabel oder inakzeptabel sein sollte.

In Textkorpora immenser Größe wird man auf der Ebene von Phrasen mit hoher Wahrscheinlichkeit reichlich Belege für die Strukturen finden, deren Vorkommen man überprüfen will. Ein entscheidender Vorteil dieses Vorge-

[6] Siehe hierzu auch Paul (1916, Bd. 3, § 241) sowie Behaghel (1923, Bd. 1, § 448f.).

hens ist, dass man dabei auf Äußerungen trifft, die durchweg in der Absicht vorgebracht wurden, sich korrekt zu äußern, denn üblicherweise drücken die Leute sich so aus, wie man sich ihrer Meinung nach ausdrücken sollte.

Vor allem aber kann man festhalten: Schon die heute verfügbaren maschinenlesbaren Textkorpora exemplifizieren reicher und breiter als alles zuvor Gekannte, was Mitglieder einer Sprachgemeinschaft für richtig gehalten haben. Und wenn sich dabei nicht das einheitliche Bild ergibt, das mancher gern hätte, so wird man doch feststellen müssen, dass man nie näher an das herankommen wird, was den Sprachgebrauch der Gemeinschaft ausmacht, als eben auf diese Weise. Nur eine Kleinigkeit stört noch: Man kommt in vielen Fällen nicht so ohne weiteres an die Informationen heran, die man für eine Validierung braucht. Doch das sind überwiegend technische Schwierigkeiten, die letztlich in Griff zu bekommen sein sollten. Mehr dazu später.

Aber ist es nicht so, werden Kritiker einwenden, dass in Korpora jede Menge Fehlerhaftes und vor allen auch nicht Einschlägiges enthalten sein kann, wenn man nicht doch schon bei ihrer Zusammenstellung eine sorgfältige Auswahl trifft? Das bringt uns zu der Frage, wie ein Referenzkorpus für eine Grammatik aufgebaut sein sollte, die anhand eines Textkorpus validiert, diversifiziert oder gar von Grund auf entwickelt werden soll.

4. Was ein Referenzkorpus bieten sollte und was es nicht bieten kann

Am Beginn des Aufbaus jedes Textkorpus stehen Entscheidungen an: Was soll aufgenommen werden und was nicht. Dabei sind zwei – durchaus schwerwiegende – Fehler zu vermeiden: allzu lasche Kriterien und – mehr noch – allzu strenge Kriterien anzuwenden.

Allzu lasch wäre etwa ein Kriterium, nach dem alles aufzunehmen wäre, was mit lateinischen Buchstaben verfasst ist. Vielleicht wäre ein solcher Missgriff später über eine aufwändige statistische Auswertung des Materials zu beheben, die geeignet wäre, die Texte über strukturelle und lexikalische Besonderheiten zu sortieren, doch das brächte nur eine zusätzliche Komplikation mit sich, ohne einem dann wenigstens die Entscheidung zu ersparen, feststellen zu müssen, ob ein Text denn nun etwa deutsch, englisch oder italienisch gehalten ist.

Allzu streng wiederum wären die Kriterien, wenn man jeden einzelnen Text zuvor von Experten zertifizieren lassen wollte. Ganz abgesehen davon, dass

man auf diese Weise auf Jahre hinaus kein signifikant großes Korpus zusammenbrächte, man hätte damit genau das zunichte gemacht, was den Sinn eines Referenzkorpus ausmacht, nämlich ganze Klassen von Texten, so etwa ganze Jahrgänge von Zeitungen, Zeitschriften und Buchreihen pauschal als gleichsprachig aufzunehmen, und sich so an die Vorgabe zu halten, vom wirklichen Sprachleben auszugehen und nicht doch wieder schon vorab zu wissen, was überhaupt sein darf.

Auch wenn man sich entschieden hat, keine allzu strengen Kriterien anzulegen, bleibt das Problem, auf Texten welcher Art das Referenzkorpus aufgebaut sein sollte. Eine Überlegung ist, ein ausgewogenes Korpus aus Texten verschiedenster Lebens- und Erfahrungsbereiche zusammenzustellen, so dass jeder dieser Bereiche entsprechend seiner Bedeutung im gesellschaftlichen Leben repräsentiert wäre. Dagegen spricht m.E. zweierlei:

a) Es wird aus praktischen wie theoretischen Gründen nie gelingen, echte Ausgewogenheit zu erreichen, denn niemand weiß, welche Bereiche im Leben der Mitglieder einer – selbst nur unscharf zu bestimmenden – Sprachgemeinschaft überhaupt eine Rolle spielen und von welcher Bedeutung sie dabei sind.

b) Nichts gewährleistet, dass eine Gewichtung von Erfahrungsbereichen, die unter soziologischen Gesichtspunkten vorzunehmen wäre, den Erfahrungen entspricht, die Sprachteilhaber mit Texten machen.

Angemessen scheint mir ein Vorgehen, wie es dem Aufbau von DeReKo zugrunde liegt.[7] Kupietz / Keibel (2009b: 53) charakterisieren dieses so:

> Unlike other well-known corpora, like, e.g. the *British National Corpus* (*BNC*) or the core corpus of the *Digital Dictionary of the 20th Century German Language*, the DEREKO archive itself does not intend to be balanced in any way. The rationale behind this is that the term balanced – just as much as the term representative – can only be defined with respect to some given statistical population. The resource itself should not dictate a specific population, nor should it define which properties of the population are of particular relevance. Instead, these issues should, as far as possible, be decided by the individual researcher depending on their general research interests and the specific question they seek to answer.

Ganz praktisch gesehen bedeutet dies: Nimm, was du kriegen kannst. Solange man nur schriftliches Material und dabei nicht gerade ausschließlich moderne Lyrik oder Texte aus einem alltagsfernen Forschungsbereich auswählt, ist al-

7 http://www.ids-mannheim.de/kl/projekte/korpora/.

lein von Interesse, ein möglichst großes Korpus aufzubauen, denn, man muss es sagen, dabei kommt einem zugute, dass allein schon die Schriftform die Bandbreite möglicher Variationen deutlich beschränkt, da spätestens seit der Erfindung des Buchdrucks Schriftliches zunehmend standardisiert wurde. Natürlich schadet es nichts, wenn man dabei versucht, Texte verschiedener Provenienz zu berücksichtigen, aber man sollte unter keinen Umständen nur Texte auswählen, die den Ansprüchen an einen vermeintlichen Standard genügen, denn das müsste dem Versuch einer Validierung grammatischer Regeln anhand von Korpusdaten jeden Sinn nehmen.

Vielleicht noch eine Anmerkung in Sachen Ausgewogenheit: Ein Korpus wie das DeReKo, in das sehr viel Tageszeitungen aufgenommen wurden und weiterhin aufgenommen werden, gerät möglicherweise ganz von selbst so ausgewogen, wie dies ein Korpus nur sein kann, denn was könnte ausgewogener sein als das, was Tageszeitungen zu bieten haben, die vom Kleintierzüchtertreffen bis zur Literaturkritik so ziemlich über alles berichten, was Alltag und Festtage einer Sprachgemeinschaft ausmachen.

Wer daran interessiert ist, das Sprachleben in seiner ganzen Breite zu erfassen, wird natürlich Wert darauf legen, nicht nur schriftliche Texte zur Verfügung zu haben, sondern auch mündliche Äußerungen zu erfassen. Das ist zwar durchaus verständlich, m.E. jedoch derzeit nicht auf demselben Niveau zu realisieren wie im Bereich der Schrift: Zum einen hat man hier mit fast unüberschaubar großer Variation zu rechnen, zum andern behindern sehr massive rechtliche Probleme die Sammeltätigkeit. Während man es bei Texten mit Produkten zu tun hat, die überwiegend zur Publikation bestimmt waren, so dass meist nur noch Verwertungsrechte zu klären sind, haben mündliche Äußerungen – von eher untypischen öffentlichen Auftritten abgesehen – privaten, sprich vertraulichen Charakter.

Zu den rechtlichen Problemen im Umgang mit Daten mündlicher Kommunikation kommt deren in Vergleich zu Texten weit stärkere Situationsbindung sowie, zumindest derzeit, noch das Problem, dass geeignete Werkzeuge fehlen, um riesige Mengen von Sounddaten oder gar Videodaten maschinell zu durchsuchen. All dies schien uns in der Grammatikabteilung des Instituts für Deutsche Sprache Grund genug, uns bei unserem eben beginnenden korpusgrammatischen Projekt erst einmal auf schriftliche Daten zu beschränken. Dieses Projekt möchte ich jetzt kurz skizzieren, um einen Eindruck davon zu vermitteln, wie man sich eine korpusgestützte Grammatik vorstellen könnte, die ei-

nerseits die neuen technischen Möglichkeiten nutzt und andererseits doch der Tradition verpflichtet bleibt und mithin die Erkenntnisse von Generationen „intelligenter" Grammatiker zu nutzen sucht.

5. Korpusgestützte Validierung von Grammatikregeln und Analyse von Variation und Zweifelsfällen

Das Projekt trägt den Namen *Grammatische Variation im standardnahen Deutsch* (*Vorstudien zu einer Korpusgrammatik*), und nach allem, was ich bislang vorgebracht habe, wird es kaum überraschen, dass ich damit nicht restlos glücklich bin. Die Bezeichnung stammt aus einer Zeit, in der wir mit unseren Überlegungen noch am Anfang standen, und aus praktischen Gründen, die mehr mit Laufzeiten von Arbeitsplänen als mit Forschung zu tun haben, schien es besser, die einmal eingeführte Bezeichnung beizubehalten und ihr eine neue Lesart zu geben, nach der, was als standardnahes Deutsch gelten könnte, ein zentrales, wenn nicht das zentrale Ergebnis unserer korpusgestützten Forschungen sein soll. Das heißt: „Standard" verstehen wir nicht im Sinn einer Norm, und zwar weder im Sinn einer vorgefundenen noch einer durchzusetzenden Norm. Wenn etwas als Standard betrachtet werden soll, dann muss sich dies erst auf der Grundlage intensiver Recherchen in den Textkorpora über Feststellungen zur Frequenz von Ausdrucksformen erweisen.

Zu den Zielen unseres Projekts:

Durch ein dezidiert korpuslinguistisches Herangehen soll bei der Untersuchung grammatischer Variation erreicht werden:

- ein höherer Grad an Detailtreue, der auch die Aufdeckung bisher nicht erfasster Muster und Strukturen möglich macht,
- eine genaue Ermittlung der Frequenz und Distribution von Phänomenen (insbesondere bei grammatischen Alternativen).

Methodisch und im deskriptiven Herangehen orientiert sich das Projekt an der *Grammatik der deutschen Sprache* (*GDS*)[8] und anderen grammatischen Projekten des IDS, wobei – wie schon in der *Systematischen Grammatik* (*Grammis*)[9] – das robuste Format einer oberflächenorientierten Konstituentenstrukturgrammatik mit syntaktischen Funktionen und der empirische Gehalt ge-

[8] Zifonun / Hoffmann / Strecker et al. (1997).

[9] Online zu finden unter http://hypermedia.ids-mannheim.de/pls/public/sysgram.ansicht.

stärkt werden. Im Bereich der Empirie richtet sich das Projekt nach korpuslinguistischen Ansätzen aus. Auf der Basis der morphosyntaktisch annotierten IDS-Korpora geschriebener Sprache in ihren zukünftigen Ausbaustufen wird ein (virtuelles) Projektkorpus eingerichtet, das Texte aus dem gesamten deutschen Sprachraum in hinreichend großer Zahl umfasst.

Den Ausgangspunkt für die Korpusanalysen bilden die Ergebnisse der *GDS*, der *Grammatik in Fragen und Antworten*,[10] der Konnektorenprojekte[11] und des Valenzwörterbuchs *VALBU* (Schumacher et al. 2004). Dort vorgenommene Einschätzungen und Erklärungen zu Phänomenbereichen, die sich dort als besonders variationsreich herausgestellt haben, sollen anhand von Korpusanalysen überprüft, vertieft und, wenn erforderlich, auch modifiziert werden. In Frage kommen dabei unter anderem Bereiche wie:

- Flexion (z.B. *niemand – niemanden/em, sämtliche einschlägige/en Informationsquellen*)
- Wortbildung (z.B. *Interessebekundung – Interessenbekundung – Interessensbekundung*)
- Komparation (z.B. *rot – röter/roter, weitestgehend – weitgehendst*)
- Rektion von Verben und Präpositionen (z.B. *ich versichere Sie meiner Unterstützung – ich versichere Ihnen meine Unterstützung, wegen dem Geld – wegen des Geldes*)
- Valenz von Verben, Adjektiven und Substantiven (z.B. *sich um Objektivität bemühen – sich redlich bemühen, fähig zu – fähig, ein Mittel für – ein Mittel gegen*)
- Kongruenz / Korrespondenz (z.B. *Für Rückfragen stehen / steht Ihnen die Kundenberatung oder Herr Krause zur Verfügung.*)
- Bildung von Verbalperiphrasen (z.B. *gemacht gehabt hat, schriebe – würde schreiben*)
- Passivierbarkeit (z.B. *Jetzt wird sich gewaschen! Gedanken werden gehabt.*)
- Gebrauch der Tempora (z.B. *sie ging/ist gegangen*)
- Modus in der indirekten Rede (z.B. *Er sagt, er ist/sei/wäre zufrieden.*)
- Realisierungsmöglichkeiten für verschiedene Konstituenten (z.B. *der Wunsch zu gewinnen – der Wunsch, man möge gewinnen – der Wunsch, dass man gewinnt*)

[10] http://hypermedia.ids-mannheim.de/pls/public/fragen.ansicht.

[11] http://www.ids-mannheim.de/gra/konnektoren/.

- Wortstellung (z.B. *hat erklären lassen wollen – hat wollen erklären lassen – erklären hat wollen lassen*)
- Ellipsen (z.B. *Man glaubt, er sei fromm. Ist er nicht.*)
- Textgrammatik / Kohärenz (*Mit dem Nachhilfeschüler komme ich ganz gut zurecht, während / aber / jedoch mit dem Bruder ist es ziemlich schwierig. – Mit dem Nachhilfeschüler komme ich ganz gut zurecht, während es mit dem Bruder ziemlich schwierig ist.*)

6. Zwei exemplarische Studien

Um einen Eindruck davon zu vermitteln, was wir dabei im Einzelnen vorhaben und wie wir vorgehen wollen, hier noch exemplarisch zwei Studien, für die zwar noch nicht alle künftig verfügbaren korpuslinguistischen Verfahren genutzt werden konnten, die m.E. jedoch die Grundideen ganz gut veranschaulichen können.

Anfang dieses / diesen Jahres

Heißt es nun *Anfang dieses* oder *Anfang diesen Jahres*? So oder ähnlich lautet eine der häufigsten Anfragen bei Sprachberatungsstellen. Man liest und hört beides oft genug, um verunsichert zu sein, ob man mit dem eigenen Sprachgebrauch hier richtig liegt. Was Fragen dieser Art auch für den Theoretiker interessant macht, ist der Umstand, dass sich hieran exemplarisch zeigen lässt, wie hilflos eine nur kompetenzgestützte Grammatik ist, wenn sie auf solche Zweifelsfälle trifft. Die Berufung darauf, dass im Deutschen der Genitiv von *diese*, *dieses*, *dieser* im Neutrum und Maskulinum *dieses* lauten müsse, verkennt völlig ihre eigene Bodenlosigkeit. Was hier weiter helfen kann, sind allein Recherchen in großen Textkorpora, denn nur so kann man sich überhaupt erst einmal Klarheit darüber verschaffen, was hier, jenseits sporadischer Beobachtungen, vorliegt, und vielleicht sogar eine Erklärung dafür finden, wieso sich hier ein Zweifelsfall ergeben konnte.

Hier, was eine Suche nach *dieses Jahres* und *diesen Jahres* in DeReKo ergab:

dieses Jahres:	121 837
diesen Jahres:	12 650

Was lässt sich daraus schließen? Nun, zum einen, dass *dieses Jahres* in DeReKo etwa zehn mal so häufig verwendet wird wie *diesen Jahres* und dass dies zwar recht eindeutig, jedoch nicht so eindeutig ist, als dass man die alternative Form einfach als dummen Fehler abtun könnte.

Interessant ist, was sich zeigt, wenn man das Problem verallgemeinert und untersucht, wie häufig *dieses* und *diesen* vor anderen Nomina als Genitivform auftritt. Da man den beiden Wortformen natürlich nicht ansieht, ob es sich um Genitivformen handelt oder nicht, muss man sich entweder auf die Ergebnisse eines Taggers verlassen oder sich ein Suchmuster ausdenken, mit Hilfe dessen man die Suche so weit einschränken kann, dass nicht übermäßig viel Handarbeit mehr erforderlich ist.

Da ich Taggern bei derart problematischen Entscheidungen nicht über den Weg traue, habe ich die zweite Möglichkeit vorgezogen und im IDS-Recherchesystem COSMAS II diese exemplarisch zu verstehenden Suchmuster verwendet:

> *dieses* +w2 WODER *Namens Verfassers Autors Tages Wochenendes Abends Jahrzehnts Jahrhunderts Jahrtausends Jahrgangs Monats Semesters Quartals Zeitraums Vorgangs Amtes Kindes Mannes Landes Buches Typs Inhalts*

Die Suche ergab:

Namens	1 239	*Amtes*	674
Verfassers	3	*Kindes*	240
Autors	502	*Mannes*	1 997
Tages	2 608	*Landes*	6 248
Wochenendes	607	*Buches*	2 469
Abends	2 818	*Typs*	1 950
Jahrzehnts	1 254	*Inhalts*	139
Jahrhunderts	9 412	*Wesens*	47
Jahrtausends	691	*Dorfes*	108
Jahrgangs	309	*Politikers*	74
Monats	3 983	*Staatsmannes*	7
Semesters	107	*Autos*	483
Quartals	34	*Wagens*	410
Zeitraums	468		
Vorgangs	357	Summe	38 978

diesen +w2 WODER *Namens Verfassers Autors Tages Wochenendes Abends Jahrzehnts Jahrhunderts Jahrtausends Jahrgangs Monats Semesters Quartals Zeitraums Vorgangs Amtes Kindes Mannes Landes Buches Typs Inhalts*

Das Ergebnis:

Namens	37	*Amtes*	0
Verfassers	0	*Kindes*	0
Autors	0	*Mannes*	0
Tages	11	*Landes*	2
Wochenendes	0	*Buches*	1
Abends	4	*Typs*	79
Jahrzehnts	5	*Inhalts*	29
Jahrhunderts	28	*Wesens*	0
Jahrtausends	1	*Dorfes*	0
Jahrgangs	2	*Politikers*	0
Monats	261	*Staatsmannes*	0
Semesters	2	[*Autos*	157
Quartals	0	jedoch alles Plurale]	
Zeitraums	0		
Vorgangs	0	Summe	619

Hier fällt auf:

Anders als vor *Jahres* tritt die Form *diesen* – mit wenigen Ausnahmen – vor anderen Nomina nicht oder so gut wie gar nicht auf. Interessant sind vor allem die Ausnahmen *Monats*, *Typs* und *Inhalts.* Im Fall von *Monats* kann man vermuten, die Form *diesen* werde in Analogie zu *Jahres* gewählt, womit natürlich noch nicht erklärt ist, wieso sie dort verstärkt zu beobachten ist. Ganz aus dem Rahmen scheinen die Formen *diesen Typs* und *diesen Inhalts* zu fallen, denn hier kann schwerlich eine semantisch motivierte Analogie vorliegen.

Auf der Suche nach einer Erklärung für diese Erscheinungen, die Systematikern so gar nicht ins Konzept passen wollen, können wieder Recherchen in Textkorpora weiterhelfen. Da die infrage stehenden Formen vor allem bei *Jahres*, *Inhalts* und *Typs* signifikant häufig auftreten, könnte es sich lohnen zu betrachten, in welchen Kontexten diese Nomina sonst als Genitivformen auftreten, wenn nicht gerade *dieses* oder *diesen* ihnen vorangehen.

Unter den Ausdrücken, die *Jahres* in den Texten des DeReKo vorangehen, fanden sich bei einer durch Zufallsauswahl auf 100 000 Belege eingeschränkten Recherche zu über 43% die folgenden Adjektive, denen kein bestimmter Artikel voranging:

vergangenen	17 588
nächsten	6 966
letzten	5 152
kommenden	4 408
vorigen	2 018
laufenden	940
heurigen	444
gleichen	254
folgenden	121
darauf folgenden	24
verflossenen	20
vorangegangenen	20
selbigen	2

Keines dieser Adjektive findet sich auch nur annähernd so häufig etwa vor *Mannes*, *Kindes*, *Landes*. An *Jahres* heran kommt hier, wie eben auch bei vorangehendem *diesen* nur *Monats*, womit ebenso ein Zeitraum zu bezeichnen ist.

Im Falle von *Inhalts* und *Typs* trifft dies freilich nicht zu, doch treten auch bei diesen, ganz wie bei *Jahres* und *Monats*, sehr viele Adjektivattribute auf, denen kein Artikel vorangeht. Hier, was in DeReKo vor *Inhalts* und *Typs* an solchen Attributen zu finden war:

Fundstellen für *Inhalts* insgesamt:	4 605
mit Adjektivattribut ohne vorangehenden Artikel:	2 110
Fundstellen für *Typs* insgesamt:	14 038
mit Adjektivattribut ohne vorangehenden Artikel:	3 019

Geht man davon aus, dass alles, was uns an Sprache regelhaft vorkommt, letztlich auf Analogien beruht, dann kann es kaum noch überraschen, wenn man in Verbindung mit *Jahres*, *Monats*, *Inhalts* anstelle des von Puristen bevorzugten *dieses* immer häufiger auf *diesen* trifft, denn *diesen* lässt sich sehr gut in die Liste

der hier aufgeführten Adjektivattribute einreihen und kann dann auch ganz wie sie als Qualifikator aufgefasst werden. Man sieht: Anhand einer Recherche in großen Textmengen wurden Zusammenhänge deutlich, die einem bei rein kompetenzgestützten Vorgehen gar nicht zugänglich wären.

Die Recherchen konnten dabei noch ganz ohne Rückgriff auf grammatische Klassifikationen durchgeführt werden. Wenn hier von Nomina und Adjektivattributen die Rede war, dann war das nur als bequeme Abkürzung für ein Publikum von Experten zu verstehen. Für die Suchen selbst war nicht mehr erforderlich als grundsätzlich eine Volltextsuche mit regulären Ausdrücken leisten könnte. Im Fall meines nächsten Beispiels kommt man schnell an Grenzen, wenn man nur dieses Werkzeug zur Verfügung hat.

Du sagtest, es stünde mir so gut!

Evelyn Hamann hat mit diesem Satz in einem Sketch von Loriot die halbe Nation zum Lachen gebracht – ein untrügliches Zeichen dafür, dass hier mit *stünde* eine Ausdrucksform vorliegt, die zwar nicht für falsch, doch für etwas schrullig gehalten wird. Und genau solche Formen wird man produzieren, wenn man sich an allzu simple Maximen hält, wie sie heute noch in manchen Sprachlehrbüchern in Sachen indirekte Redewiedergabe zu finden sind. Grammatiken mit wissenschaftlichem Anspruch gehen an diese Thematik sicher weit differenzierter heran, doch bislang hat sich m.W. niemand der Mühe unterzogen, vorurteilslos in großen Textkorpora nachzuprüfen, wie denn nun tatsächlich geredet und geschrieben wird. Aber vielleicht liegt das auch daran, dass Nachschauen hierbei alles andere als einfach ist.

Hat man Zugriff auf riesige Textmengen, wie sie etwa im DeReKo vorliegen, sind erst einmal die nötigen Voraussetzungen gegeben, denn, soviel lässt sich pauschal feststellen: Beispiele für indirekte Redewiedergabe finden sich darin mit Garantie zuhauf, denn diese Form sprachlichen Handelns gehört zum Alltagsgeschäft menschlicher Kommunikation. Doch damit ist man erst einmal wie Michelangelo, als er den Marmorblock beschafft hatte, aus dem sein David werden sollte. Ohne „intelligente“ Sprachanalyse helfen einem hier weder raffinierte Suchalgorithmen noch ausgeklügelte Statistik weiter, denn allein damit ließe sich noch nicht einmal das Problem bestimmten.

Wie also vorgehen? Da jede Redewiedergabe – direkt oder indirekt – unverzichtbar mit der Verwendung eines Verbs verbunden ist, mit dem auf den wiederzugebenden Sprechakt Bezug genommen werden kann, empfiehlt es sich, erst einmal herauszufinden, welche Verben dafür infrage kommen. Will man

sich dabei nicht ganz auf seine Phantasie verlassen, kann man dazu die von Belica (2001-2007) entwickelte Kookkurrenzdatenbank nutzen. Wie man dabei dann vorgehen kann, soll hier jetzt nicht im Einzelnen ausgeführt werden, nur so viel: Damit lässt sich sehr effizient schon bald eine Liste einschlägiger Verben zusammenstellen, die ziemlich erschöpfend sein dürfte. Damit hat man die wichtigsten Suchbegriffe beisammen.

Als Nächstes ist zu bestimmen, wie Textpassagen aufgebaut sein können, mit denen Rede indirekt wiedergegeben werden soll. Die Grundtypen sind exemplarisch schnell bestimmt, denn allzu viel Verschiedenes kommt hier nicht infrage. Hier ein Beispiel für die wichtigsten Typen:[12]

> Sie sagte / erklärte / erläuterte / betonte, dass sie so kurz vor dem Examen keine Zeit für solche Späße hat / habe / hätte.
>
> Dass sie so kurz vor dem Examen keine Zeit für solche Späße hat / habe / hätte, sagte / erklärte / erläuterte / betonte sie.
>
> Sie sagte / erklärte / erläuterte / betonte, so kurz vor dem Examen hat / habe / hätte sie keine Zeit für solche Späße.
>
> So kurz vor dem Examen, sagte / erklärte / erläuterte / betonte sie, hat / habe / hätte sie keine Zeit für solche Späße.
>
> So kurz vor dem Examen hat / habe / hätte sie keine Zeit für solche Späße, sagte / erklärte/erläuterte/betonte sie.

Die Beispiele lassen bewusst offen, ob nun die Form *hat*, *habe*, *hätte* oder *haben würde* angemessen wäre, denn ob dies im allgemeinen Sprachgebrauch tatsächlich zu entscheiden ist und wenn ja, anhand welcher Kriterien, soll ja gerade Gegenstand der Korpusrecherche sein.

Man kann davon ausgehen, dass in DeReKo hunderttausende von Sätzen dieser Art zu finden sind. Die Schwierigkeit ist nur, wie man an sie herankommt, denn die derzeit verfügbaren Rechercheverfahren sind weit, sehr weit davon entfernt, Suchen vom Typ „etwas in dieser Art" zu unterstützen.

In den Recherchen, die ich im Rahmen unseres Projekts *Grammatik in Fragen und Antworten* zu indirekter Redewiedergabe angestellt habe, musste ich mich, mangels besserer technischer Möglichkeiten und Kenntnisse, auf exemplari-

[12] Aus rein praktischen Gründen habe ich hier darauf verzichtet, auch noch die umgangssprachlich recht verbreitete periphrastische Form *haben würde* zu berücksichtigen. Die exemplarisch gemeinte Recherche wäre sonst um ein Vielfaches komplexer und komplizierter ausgefallen.

sche Suchen beschränken, bei denen als finite Verbformen im *dass*-Satz meist nur Formen von *sein* und *haben* sowie von *wollen*, *können*, *dürfen*, *müssen*, *sollen* berücksichtigt wurden. Die Suchausdrücke sahen dabei beispielsweise so aus:

Dass [^,]* ist, resümier[a-z]*

\<betont [a-z]* [a-z]*, dass [^,]* habe[n]\>

\<erklärte, dass [^,]* haben k[a,ö]nn

Informatiker werden sich jetzt vielleicht ein Lachen kaum verkneifen können, aber bereits anhand von Suchen dieser Art waren interessante Feststellungen zu machen. So zeigten sich etwa beachtliche Unterschiede im Auftreten von Indikativ, Konjunktiv Präteritum und Konjunktiv Präsens nach den präteritalen Formen dieser Sprechaktverben:

Verb	Indikativ	Konj. Prät.	Konj. Präs.
andeutete	**30,62**	6,51	62,87
anmerkte	8,57	8,31	83,12
antwortete	11,79	7,18	81,03
äußerte	18,21	4,09	77,7
bedauerte	18,64	1,91	79,45
behauptete	4,28	13,71	82,01
beklagte	7,81	3,82	88,37
bemängelte	9,73	3,99	86,28
bestätigte	24,26	2,08	73,66
bestritt	5,43	4,23	**90,34**
beteuerte	5,02	5,40	89,58
betonte	8,12	3,71	88,17
bezweifelte	24,79	7,14	68,07
darlegte	20,00	8,00	72,00
erklärte	8,56	6,02	85,42
erläuterte	13,71	4,35	81,94
erwiderte	8,42	9,47	82,11
erzählte	14,50	9,70	75,80
forderte	15,55	13,96	70,49

Verb	Indikativ	Konj. Prät.	Konj. Präs.
fragte ob	6,22	**22,66**	71,12
fragte warum	21,37	9,54	69,09
fragte wo	17,36	10,18	72,46
fragte wer	14,04	11,24	74,72
fragte wie	15,70	17,41	66,89
führt an/aus	8,09	3,88	88,03
gab zu bedenken	8,22	4,45	87,33
kritisierte	11,25	2,41	86,34
kündigte an	18,28	2,45	79,27
sagte	7,92	6,35	85,73
schrieb	16,53	6,35	77,12
unterstellte	8,33	8,33	83,34
verkündete	18,07	2,80	79,13
verlautete	14,41	2,96	82,63
versprach	18,69	9,73	71,55
warnte	10,38	5,46	84,16

Angaben in Prozent

Während die Konjunktiv-Präsens-Form bei Sprechaktverben in der 3. Person Präteritum trotz merklicher Häufigkeitsschwankungen als Standardform gelten kann, ergab sich bei Sprechaktverben für das Präsens und Präsensperfekt bei meinen exemplarischen Recherchen ein anderes Bild:

Verb	Indikativ	Konj. Präs.	Konj. Prät.
bedauert	**63,4**	36,37	0,23
bemängelt	35,64	63,27	1,09
bestätigt	59,74	39,48	0,78
erklärt	34,23	**63,85**	1,92
fragt	52,25	43,73	4,02
sagt	47,5	50,99	1,51
unterstellt	37,5	53,13	**9,37**
verspricht	69,35	29,57	1,08

Angaben in Prozent

Bemerkenswert auch, welche Abhängigkeiten zwischen der Personalform des Sprechaktverbs und der Häufigkeit der verschiedenen Verbmodi im *dass*-Satz festzustellen waren:

Liegt das Sprechaktverb in der 1. Person Präteritum vor, folgen im anschließenden *dass*-Satz, wie die exemplarischen Recherchen zeigen, Indikativ-Formen und Konjunktiv-Präsens-Formen mit etwa gleicher Häufigkeit. Konjunktiv-Präteritum-Formen machen etwa 8,4% aller Formen aus.

In etwa bestätigt fand sich eine gängige Annahme zum Gebrauch der Konjunktiv-Präteritum-Form als Ersatzform für aufgrund des Formensynkretismus als solche nicht erkennbare pluralische Konjunktiv-Präsens-Formen: Nach *sagte*, *erklärte*, *behauptete*, *meinte* fanden sich in den folgenden *dass*-Sätzen über 75% Konjunktiv-Präteritum-Formen – verglichen mit den Verhältnissen bei singularischen Formen eine geradezu überwältigende Menge.

Das alles ergab sich, wie gesagt, bei nur exemplarischen Recherchen, aber es macht Laune, die Untersuchung auf der Basis der inzwischen vollständig getaggten Korpora und besseren Recherchewerkzeuge wieder aufzunehmen, um zu einem wesentlich differenzierten Bild zu kommen, das auch Interdependenzen zwischen den verschieden Faktoren berücksichtigten kann, die exemplarisch nur jeweils unter der Annahme *ceteris paribus* zu betrachten waren.

Literatur

Behaghel, Otto (1923): Deutsche Syntax. Bd. 1. Heidelberg: Carl Winters Universitätsbuchhandlung.

Belica, Cyril (2001-2007): Kookkurrenzdatenbank CCDB – V3.2 – Eine korpuslinguistische Denk- und Experimentierplattform für die Erforschung und theoretische Begründung von systemisch-strukturellen Eigenschaften von Kohäsionsrelationen zwischen den Konstituenten des Sprachgebrauchs. Institut für Deutsche Sprache. Mannheim. Internet: http://corpora.ids-mannheim.de/ccdb/.

Keller, Rudi (2003): Sprachwandel. Von der unsichtbaren Hand der Sprache. 3. Aufl. Tübingen: Francke.

Keller, Rudi (2009): Konventionen, Regeln, Normen. Zum ontologischen Status natürlicher Sprachen. In: Konopka / Strecker (Hg.), 9-22.

Konopka, Marek / Strecker, Bruno (2009) (Hg.): Deutsche Grammatik – Regeln, Normen Sprachgebrauch. Berlin / New York: de Gruyter.

Kupietz, Marc / Keibel, Holger (2009a): Gebrauchsbasierte Grammatik: Statistische Regelhaftigkeit. In: Konopka / Strecker (Hg.), 33-52.

Kupietz, Marc/Keibel, Holger (2009b): The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In: Working Papers in Corpus-based Linguistics and Language Education 3. Tokyo: Tokyo University of Foreign Studies (TUFS), 53-59.

Paul, Hermann (1916): Deutsche Grammatik. Bd. 3. Halle: Niemeyer.

Schumacher, Helmut / Kubczak, Jacqueline / Schmidt, Renate / de Ruiter, Vera (2004): VALBU – Valenzwörterbuch deutscher Verben. Tübingen: Narr.

Smith, Adam (1776): An inquiry into the nature and causes of the wealth of nations. Edinburgh: Strahan / Cadell.

Strecker, Bruno (1987): Strategien des kommunikativen Handelns. Düsseldorf: Schwann.

Wittgenstein, Ludwig (1953): Philosophische Untersuchungen. Oxford: Blackwell.

Ullmann-Margalit, Edna (1977): The emergence of norms. Oxford: Clarendon Press.

Zifonun, Gisela / Hoffman, Ludger / Strecker, Bruno et. al (1997): Grammatik der deutschen Sprache. 3 Bde. Berlin / New York: de Gruyter.

DOUGLAS BIBER / BETHANY GRAY

Is conversation more grammatically complex than academic writing?

Abstract

Conversation is usually considered to be grammatically simple, while academic writing is often claimed to be structurally complex, associated primarily with a greater use of dependent clauses. Our goal in the present paper is to challenge these stereotypes, based on the results of large-scale corpus investigations. We argue that both conversation and professional academic writing are grammatically complex but that their complexities are dramatically different. Surprisingly, the traditional view that complexity is realized through extensive clausal embedding leads to the conclusion that conversation is more complex than academic writing. In contrast, written academic discourse is actually much more 'compressed' than elaborated, and the complexities of academic writing are realized mostly as phrasal embedding rather than embedded clauses.

1. Introduction

Grammatical complexity is often linked with elaboration and clausal embedding in linguistic theory. A 'simple' clause has only a subject, verb, and object or complement. A 'simple' noun phrase has a determiner and head noun. Additions to these structures represent elaboration, resulting in 'complex' grammar. In particular, there is widespread agreement that embedded clauses are an important type of grammatical complexity (often contrasted with 'simple' clauses; see e.g., Huddleston 1984: 378, Willis 2003: 192, Purpura 2004: 91, Carter / McCarthy 2006: 489).

Conversation has long been described as grammatically simple in these terms. Conversational participants share time and place, and they normally also share extensive personal background knowledge. As a result, pronouns and vague expressions are common, and referring expressions generally do not need to be elaborated in conversation. Because of these factors, conversational grammar is assumed to be generally not complex, employing "simple and short clauses, with little elaborate embedding" (Hughes 1996: 33).

In contrast, academic writing is claimed to be structurally complex, shown by longer sentences, longer 't-units' (a main clause plus all associated dependent clauses), "longer and more complex clauses with embedded phrases and clauses" (Hughes 1996: 34), and generally a greater use of subordinate clauses (see, e.g., O'Donnell / Griffin / Norris 1967, O'Donnell 1974, Kroll 1977, Chafe 1982, Brown / Yule 1983).

These stereotypical portrayals of conversation and academic writing reflect the most salient characteristics of both. For example, some of the most noticeable characteristics of conversation are the hesitations, false starts, and short non-clausal utterances, because none of these features are normally appropriate in formal writing. The following conversational excerpt illustrates these characteristics:

Text Excerpt 1: Conversation

Non-clausal utterances are marked in **bold**

Barry: *I went to the Institute of Terror.*
Wendy: *You went to where?*
Barry: ***The Institute of Terror.***
[...]
Wendy: ***Oh.***
Barry: *It's pretty cool. You want to go? I've got free tickets.*
Wendy: *Is it – it's a – how long is it going to be open?*
Barry: ***Until the thirty first.***
Wendy: ***Cool.*** *It's, it's an, it's actually pretty scary and stuff?*
Barry: *I wouldn't go so far as to say it's really scary*
Wendy: *But it's cool.* [laugh]
Barry: ***Yeah.*** [...] *I'll go with you. I wouldn't pay for you or anything*
Wendy: [laugh]
Barry: *But I'll go with you.*
Wendy: *It's expensive, isn't it? It's like five bucks.*
Barry: ***Yeah****, this one's six.*
Wendy: ***The one down here?*** *And you have free tickets?*
Barry: ***Well, yeah.*** [...]
Wendy: ***Wow. Cool.***

This conversation additionally illustrates the reliance on short, simple clauses, such as *where's that?*, *it's pretty cool*, *I've got free tickets*, *I'll go with you*, and *It's like five bucks*. If most conversation included only the grammatical features illustrated in Text Excerpt 1, we would be justified in making the generalization that conversation was generally not grammatically complex (as measured by the traditional criteria).

In contrast, one of the most noticeable characteristics of academic writing is that sentences tend to be long, and readers usually attribute that fact to the presence of numerous embedded clauses. Text Excerpt 2 illustrates this style of discourse:

Text Excerpt 2: Academic writing: Philosophy textbook

Embedded clauses marked with [ ]

> [*Even if propositional attitude accounts succeeded in their own terms*], *they would not explain most of* [*what should be explained by a theory of emotion*]. *Propositional attitude theories are often presented* [*as if they were a simple consequence of the idea* [*that emotions involve the occurrence of mental states* [*which represent states of affairs in the world* (*states with "content"*)]]].
>
> [...]
>
> [*What is distinctive about the propositional attitude theory*] *is the interpretation* [*it gives to the words thought and belief*]. *The mainstream philosophical tradition* [*in which Lyons is located*] *assumes* [*that our everyday understanding of these notions is adequate for a theory of emotion*].

Here again, if most written academic texts incorporated this same dense use of embedded clauses, we would be justified in making the generalization that academic writing was highly complex as measured by that criterion.

However, consideration of a single text excerpt from a register does not provide an adequate basis for such conclusions. Rather, this is exactly the kind of research question that corpus-based research can contribute to (see, e.g., Biber / Conrad / Reppen 1998, McEnery / Tono / Xiao 2006). By basing analyses on large, representative collections of texts, it is possible to discover patterns of use that are generalizable to a register, rather than more specific patterns that characterize only particular texts. Further, corpus-based methods usually entail quantitative analysis, permitting description of the extent to which a

linguistic pattern is typical of a register. Both of these analytical characteristics are important here. First, corpus research shows that the general 'complexity' characteristics of conversation and academic writing are quite different from those that are especially salient in individual texts. And second, corpus research shows that both conversation and academic writing use grammatical complexity features to some extent; the major difference between them is in the quantitative extent to which they rely on different sets of features.

Our goal in the present paper is to challenge the stereotypes described above, based on the results of large-scale corpus investigations. We argue that both conversation and professional academic writing are grammatically complex – but their complexities are dramatically different. Surprisingly, if we adopt the traditional view that complexity is realized through extensive clausal embedding, the evidence presented below would lead us to conclude that conversation is more complex than academic writing. In contrast, written academic discourse is actually much more 'compressed' than elaborated, and the complexities of academic writing are realized mostly as phrasal embedding rather than embedded clauses.

The following sections present the results of large-scale corpus analyses that document these patterns of use. Section 2 introduces the corpora and linguistic features used for the analyses. Then, the analyses themselves are discussed in Section 3, which surveys the synchronic patterns of use for features associated with structural elaboration versus compression. In conclusion, we briefly discuss functional motivations for these patterns of use.

2. Corpus and grammatical features used for the analysis

We employ corpus-based analysis to describe the typical discourse styles of conversation and academic writing, investigating the extent to which both registers employ grammatical devices associated with structural elaboration. Previous corpus-based studies have documented the different complexities of spoken and written registers. For example, multi-dimensional studies of register variation (e.g., Biber 1988, 1992, 2006) have shown repeatedly that certain dependent clause types (e.g., *because*-clauses and *WH*-clauses) are more strongly associated with speech than writing. The *Longman Grammar of Spo-*

ken and Written English (Biber et al. 1999) provides more detailed descriptions of the grammatical features that are common in conversation versus those that are common in academic writing.

Building on this previous research, the present study focuses on the grammatical devices in English that are associated with structural elaboration. The descriptions below contrast the patterns of use in conversation to those in professional academic writing, based on analysis of a large corpus of texts for each of these two registers.

The conversation subcorpus is taken from the *Longman Spoken and Written Corpus* (see Biber et al. 1999: 24-35). The subcorpus includes 723 text files and c. 4.2 million words of American English conversation. These are conversations collected by participants who agreed to carry tape recorders for a two-week period. The corpus thus represents one of the largest collections of natural face-to-face conversations available.

We constructed a corpus of academic research articles (c. 3 million words), sampled from four general disciplines: science / medicine, education, social science (psychology), and humanities (history). We collected texts from three 20-year intervals (1965, 1985, 2005) to enable the description of short-term historical change. However, for the purposes of the present study, we consider these as a single group (429 texts, c. 2.9 million words), contrasted with conversation.

The corpora were grammatically annotated ('tagged') using software developed for the *Longman Grammar of Spoken and Written English* and earlier corpus studies of register variation (e.g., Biber 1995). Then, more specialized computer programs were developed for detailed linguistic analyses of specific types of structural elaboration.

Table 1 lists the types of dependent clauses that we considered for our analysis of structural elaboration. These dependent clauses can serve three major syntactic functions: complement clauses, which usually function as the direct object of a verb; adverbial clauses, which modify the main verb; and post-nominal relative clauses, which modify a head noun. In addition, dependent clauses can be finite (with tense overtly marked) or non-finite.

Grammatical feature	Examples
Finite complement clauses	*I wonder* ***how he is today.*** *I thought* ***that was just too funny.***
Non-finite complement clauses	*We'd love* ***to come.*** *They talk about* ***building more.***
Finite adverbial clauses	*She won't narc on me,* ***because she prides herself on being a gangster.*** *You can have it* ***if you want.***
Finite relative clauses	*A method* ***that would satisfy the above conditions*** *...* *a repressor substance* ***which prevents the initiation*** *...*
Non-finite relative clauses	*the assumptions* ***given above*** *...* *initiatives* ***involving local authorities*** *...*

Table 1: Selected grammatical features associated with structural elaboration

We also considered grammatical devices that result in a 'compressed' rather than 'elaborated' discourse style, illustrated in Table 2. These are all phrases rather than dependent clauses, used to modify a head noun. Attributive adjectives and pre-modifying nouns occur before the head noun ('pre-modifiers'), while prepositional phrases occur after the head noun ('post-modifiers').

Grammatical feature	Examples
Attributive adjective (adjective as noun pre-modifier)	*a* ***large*** *number,* ***unusual*** *circumstances*
Noun as noun pre-modifier	***human*** *actions,* ***membrane*** *structure*
Prepositional phrase as noun post-modifier	*the scores* ***for male and female target students in the class*** *the mechanism* ***for penetration of protein through the ovariole wall***

Table 2: Selected grammatical features associated with structural compression

Most of these features could be identified accurately using automatic computer programs. However, prepositional phrases required hand coding to determine when the phrase was functioning as a noun modifier versus adverbial. This analysis was based on a sub-sample of tokens (every fourth occurrence) from a sub-sample of the corpus (48 conversations and 41 academic research articles). The counts for all linguistic features were converted to a 'normed' rate of occurrence (per 1 000 words) for each text (see Biber / Conrad / Reppen 1998: 263-264).

3. Structural elaboration and compression in conversation versus academic writing

As noted above, researchers have usually focused on dependent clauses (or subordinate clauses) as the primary measure of grammatical complexity or structural elaboration. What they have less often noticed is that there is extensive clausal embedding in conversation. In particular, complement clauses (also called 'nominal clauses') are very common, especially *that*-clauses and *WH*-clauses. Complement clauses normally fill a direct object slot, making it possible for a relatively short utterance to have multiple levels of embedding. For example, the following short utterance has two embedded complement clauses:

You know [*you could get* [*what you wanted*]]

Unlike adverbial clauses and relative clauses, complement clauses are not optional structures; rather, they take the place of a required noun phrase. In conversation, the complement clause usually occurs with a transitive verb (e.g., *think*, *know*, or *want*): the complement clause substitutes for the noun phrase as the direct object of the verb. As a result, these structures can contain multiple levels of structural embedding. For example, the following relatively short sentence from conversation has four embedded complement clauses, each occurring as the object of the preceding main verb:

But I don't think [*we would want* [*to have it* [*sound like* [*it's coming from us*]]]].

Adverbial clauses are optional rather than obligatory clause elements. However, these clause types are also commonly found in conversation, as in:

She married him [*because Clinton's father died before Clinton was born*]
[*If anybody wakes me up early*] *they die*

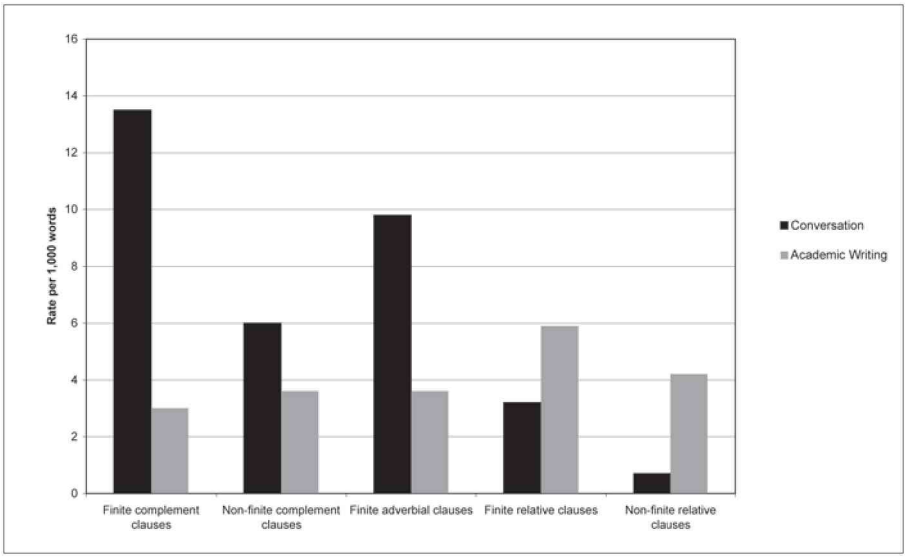


Figure 1: Common dependent clause types in conversation vs. academic writing

Our corpus investigations show that the structures illustrated above represent strong general differences between academic writing and conversation: that they are much more frequent in conversation than in academic writing. Thus, Figure 1 shows that both complement clauses and adverbial clauses are much more frequent in conversation than in academic writing. These differences are strongest for finite clauses (e.g., *that*-clauses and *WH*-clauses functioning as complement clauses; *because*-clauses and *if*-clauses functioning as adverbial clauses). However, the same general pattern holds for non-finite complement clauses (*to*-clauses and *ing*-clauses). In contrast, relative clauses are more frequent in academic writing than in conversation (especially non-finite relative clauses, such as *the concept of society* ***proposed here***).

Text Excerpt 3 illustrates the pervasive use of embedded clauses in conversation. Unlike Text Excerpt 1 above, this conversational excerpt shows how certain kinds of dependent clauses can occur with extreme density in normal conversational interactions. For the most part, these kinds of structural elaboration do not feel complex, and they certainly do not inhibit normal communication. However, they are clearly 'complex' according to the definition of embedded clauses added on to simple clauses. Overall, Figure 1 shows that there are around twice as many dependent clauses in conversation as in academic writing. Thus, if we limited our comparison to these features, we would be forced to conclude that conversation is more complex than academic writing.

Text Excerpt 3: Conversation

Dependent clauses are marked in **bold**

Gayle: *And Dorothy said* ***Bob's getting terrible with, with the smoking****. Uh, he's really getting defiant about it* ***because there are so many restaurants where you can't smoke*** *and he just gets really mad and won't go to them.*

[...]

Peter: *Well they, they had a party. I forget* ***what it was.*** *They had it at a friend's house. I can't remember* ***why it wasn't at their house*** *any way. And they had bought a bottle of Bailey's* ***because they knew I liked Bailey's.***

[...]

Gayle: I *can't remember* ***who it was.*** *One of us kids.*

[...]

Peter: *Oh. I'll tell you* ***I think the biggest change in me is since I had my heart surgery.***

Gayle: *Really? Yeah I guess my, I mean* ***I know my surgery was a good thing but***

Peter: <?> *It makes you* ***think****. You realize* ***it can happen to you.***

The obvious question at this point is to ask why academic research writing seems grammatically complex. That is, given that dependent clauses are generally more frequent in conversation than in writing, we need to account for the perception that academic texts are hard to process. Part of this perception is caused by difficult subject matter and complex vocabulary. However, there are also grammatical features that make a major contribution to this complexity. In particular, the structural elaboration of academic writing is realized mostly as phrases without verbs. For example, consider the following sentence from a Biology research article:

The knowledge of tissue distribution of each novel molecular species is the first step toward the understanding of its possible function.

This sentence consists of only a single main clause, with the main verb *is*. There are no dependent clauses in this sentence. The sentence is relatively long because there are multiple prepositional phrases:

of tissue distribution

of each novel molecular species

toward the understanding

of its possible function

In addition, many of the noun phrases include extra nouns or adjectives as pre-modifiers before the head noun:

tissue distribution
novel molecular species
possible function

In their main clause syntax, sentences from academic writing tend to be very simple. Thus, consider the following sentence from a Psychology research article:

This may indeed be ***part*** [***of*** *the reason* [***for*** *the statistical link* [***between*** *schizophrenia and membership* [***in*** *the lower socioeconomic classes*]]]].

Similar to the example from biology above, the clausal syntactic structure of this sentence is extremely simple, with only one main verb phrase:

X may be Y (*This may be part*)

All of the elaboration here results from prepositional phrases added on to noun phrases. Thus, unlike conversation, academic writing does **not** frequently employ dependent clauses for structural elaboration. Rather, we find a more 'compressed' style, employing embedded phrases rather than fuller dependent clauses.

As Figure 2 shows, academic writing relies heavily on non-clausal phrases instead of dependent clauses to add information. Most of these phrases occur embedded in noun phrases. Many of these structures are adjectives modifying a head noun (e.g., *theoretical orientation*) or nouns pre-modifying a head noun (e.g., *system perspective*). But the most striking difference from conversation is for the use of prepositional phrases as noun post-modifiers. Many of these are *of*-phrases (e.g., *an interpretation* ***of the general form of mitochondria***), but other prepositions are also commonly used for this function (e.g., *the complex relations* ***between three components***; *understanding rational approach* ***to politics***). Prepositional phrases used as adverbials (e.g., ***From the systems perspective****, these stages are marked by...*) are also more common in academic writing than in conversation, but the difference is much less strong.

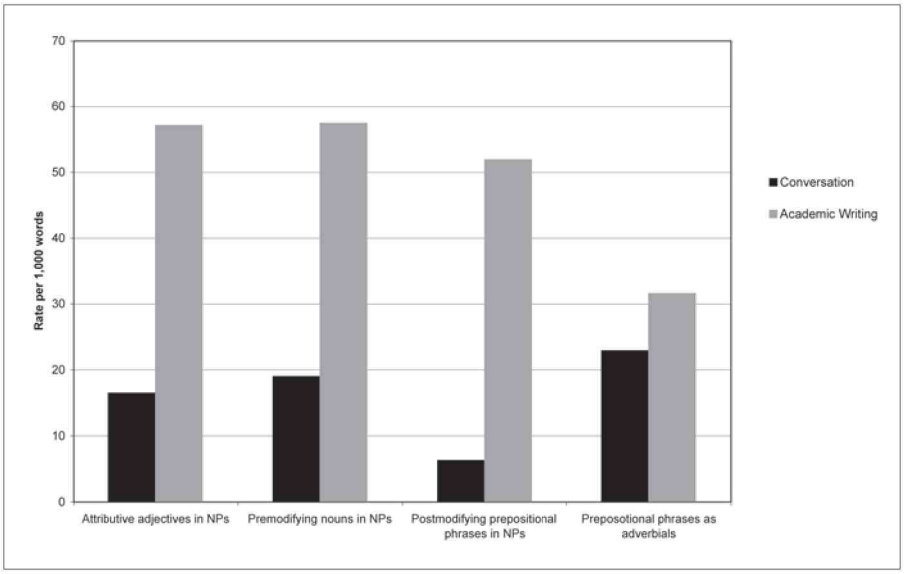


Figure 2: Common dependent phrasal types in conversation vs academic writing

It is not the case that there are no dependent clauses in academic writing. Rather, as Figure 1 above shows, dependent clauses are relatively frequent in academic writing, especially noun modifiers and non-finite clauses. Text Excerpt 4 illustrates these patterns.

Text Excerpt 4: Academic research article

Main Verbs are underlined; main verbs in dependent clauses marked in **bold**

A number of important themes have emerged from previous research ***exploring*** *the links among gender, interaction, and collaborative learning. First, in mixed-gender interactions boys tend to* ***dominate*** *apparatus, teacher attention, and peer discussion within the classroom (for a review, see Howe 1997). Second, the comparative context and, in particular, task are important in* ***determining*** *how children* ***engage*** *in collaborative interaction (Holmes-Lonergan 2003). Third, a child's gender and that of his or her conversation partner appear to* ***affect*** *the dynamics of conversation but not necessarily the answer that children* ***agree*** *on (Leman 2002): Specifically, boys tend to* ***show*** *greater resistance to girls' arguments, but ultimately all children tend to* ***opt*** *for the developmentally more advanced answer after interaction, regardless of whether a girl or a boy* ***has put*** *these arguments forward.*

A further consideration for studies of gender and interaction is the influence of both speakers and partners gender on conversation. Leaper (1991) examined both these types of gender effect on communication between pairs of children in two different age groups (5 and 7 years) who ***were engaged*** in play with a puppet.

However, this same text excerpt also illustrates the more important grammatical pattern that sharply distinguishes between academic writing and conversational discourse: the heavy reliance on phrasal rather than clausal elaboration. Thus, Text Excerpt 4 is repeated as Text Excerpt 5 below, highlighting the complex noun phrases with phrasal rather than clausal modifiers. A quick glance at this excerpt shows that the majority of this text is composed of such structures.

Text Excerpt 5: Academic research article [repeated from Text 4]

Complex noun phrases with no clausal embedding are marked in **bold**

A number of important themes *have emerged from* ***previous research*** *exploring* ***the links among gender, interaction, and collaborative learning.*** *First, in* ***mixed-gender interactions*** *boys tend to dominate* ***apparatus, teacher attention, and peer discussion within the classroom*** *(for a review, see Howe, 1997). Second,* ***the comparative context and, in particular, task*** *are important in determining how children engage in* ***collaborative interaction*** *(Holmes-Lonergan, 2003). Third,* ***a child's gender and that of his or her conversation partner*** *appear to affect* ***the dynamics of conversation but not necessarily the answer*** *that children agree on (Leman, 2002): Specifically, boys tend to show* ***greater resistance to girls' arguments****, but ultimately all children tend to opt for* ***the developmentally more advanced answer after interaction****, regardless of whether a girl or a boy has put these arguments forward.*

A further consideration for studies of gender and interaction *is* ***the influence of both speakers and partners gender on conversation****. Leaper (1991) examined* ***both these types of gender effect on communication between pairs of children in two different age groups (5 and 7 years)*** *who were engaged in* ***play with a puppet.***

In fact, relatively few noun phrases in Text Excerpt 5 are simple noun phrases. The majority of noun phrases contain some manner of phrasal complexity. Furthermore, noun phrases in academic prose often have multiple levels of phrasal embedding within a single noun phrase. For example, the following complex noun phrase contains two prepositional phrases functioning as noun postmodifiers (head noun of phrase in bold, prepositional postmodifiers bracketed):

the ***influence*** [*of both speakers and partners gender*] [*on conversation*]

Thus, despite stereotypical beliefs about the complexity of academic writing stemming from subordinate clauses, it appears that one of the more distinctive complex structures of academic prose have been largely overlooked: phrasal modification.

4. Conclusion

In summary, the stereotype that writing is more elaborated than speech is not supported by corpus evidence. In fact, using traditional measures of elaboration – considering the use of dependent clauses – we would conclude that the opposite was the case: that conversation is more complex and elaborated than academic writing. However, that conclusion would also be an over-simplification, because it does not fully capture the characteristics of either conversation or academic writing.

However, the elaboration of conversation is very restricted in nature. As noted above, most of the dependent clauses in conversation are integrated into the clause structure: complement clauses normally fill an object slot controlled by a transitive verb. As such, these dependent clauses are not 'elaborating' in the same way that adverbial clauses and relative clauses are. In addition, the structural patterns in conversation are very restricted lexically. For example, although there are over 200 different verbs that can control a *that* complement clause (e.g., *assume*, *ensure*, *feel*, *hear*, *imply*, *indicate*, *propose*, *realize*, *suggest*), only three verbs account for c. 70% of all occurrences of this clause type in conversation: *think* (35%), *say* (20%), *know* (13%) (see Biber et al. 1999: 667-670). The lexical restriction is even stronger with *to* complement clauses, where c. 50% of all occurrences are controlled by the verb *want* (see ibid.: 710-714). Thus, the overall frequency of dependent clauses in conversation is largely due to a few high frequency lexico-grammatical patterns.

On the other hand, the lack of elaboration in academic writing is in part an artifact of inadequate measures, rather than an accurate characterization of academic writing. That is, elaboration has normally been analyzed by considering the extent to which dependent clauses are used in a text. By that measure, we would conclude that academic writing is actually less elaborated than conversation. However, that measure misses the most important structural characteristic

of academic written discourse: the reliance on phrasal rather than clausal elaboration. Most sentences in academic prose are elaborated in the sense that they have optional *phrasal* modifiers, especially nominal pre-modifiers (adjectives or nouns) and nominal postmodifiers (e.g., prepositional phrases).

These phrasal modifiers are elaborating because they are optional, providing extra information. At the same time, though, these structures are condensed or compressed: the opposite of elaborated. That is, phrasal modifiers are alternatives to fuller, elaborated expressions that use clausal modifiers (e.g., *the effect of gender* can be paraphrased with a relative clause, as in *the effect* ***which is caused by gender***).

There are good reasons why compressed, phrasal expressions are preferred over elaborated clausal expressions in academic writing: they are more economical; they allow for faster, more efficient reading; and they are equally comprehensible to the expert reader despite the fact that some explicit meaning is lost when fuller clauses are reduced to phrasal structures. In contrast, conversation relies on a relatively small set of very productive verbs controlling complement clauses to convey information, with much less reliance on complex noun phrases.

Thus, academic writing is dramatically different from speech but not in the ways that conform to the stereotypes of complexity created through the use of embedded dependent clauses. Rather, academic writing has developed a unique style, characterized especially by the reliance on nominal / phrasal rather than clausal structures. Consequently, perhaps the question should not be which register is more or less complex, but instead, in what respects are conversation and academic writing each complex in their own distinctive ways?

References

Biber, Douglas (1988): Variation across speech and writing. Cambridge: Cambridge University Press.

Biber, Douglas (1992): On the complexity of discourse complexity: A multidimensional analysis. In: Discourse Processes 15: 133-163.

Biber, Douglas (1995): Dimensions of register variation: A cross-linguistic comparison. Cambridge: Cambridge University Press.

Biber, Douglas (2006): University language: A corpus-based study of spoken and written registers. Amsterdam: John Benjamins.

Biber, Douglas / Conrad, Susan / Reppen, Randi (1998): Corpus linguistics: Investigating language structure and use. Cambridge: Cambridge University Press.

Biber, Douglas / Johansson, Stig / Leech, Geoffrey / Conrad, Susan / Finegan, Edward (1999): Longman grammar of spoken and written English. London: Longman.

Brown, Gillian / Yule, George (1983): Discourse analysis. Cambridge: Cambridge University Press.

Carter, Ronald / McCarthy, Michael (2006): Cambridge Grammar of English. Cambridge: Cambridge University Press.

Chafe, Wallace (1982): Integration and involvement in speaking, writing, and oral literature. In: Spoken and written language: Exploring orality and literacy. Ed. by Deborah Tannen. Norwood, NJ: Ablex, 35-54.

Huddleston, Rodney (1984): Introduction to the grammar of English. Cambridge: Cambridge University Press.

Hughes, Rebecca (1996): English in speech and writing: Investigating language and literature. London: Routledge.

Kroll, Barbara (1977): Ways communicators encode propositions in spoken and written English: A look at subordination and coordination. In: Discourse across time and space (SCOPIL no. 5, eds. E.O. Keenan and T. Bennett). Los Angeles: University of Southern California, 69-108.

McEnery, Tony / Xiao, Richard / Tono, Yukio (2006): Corpus-based language studies: An advanced resource book. London: Routledge.

O'Donnell, Roy C. (1974): Syntactic differences between speech and writing. In: American Speech 49: 102-110.

O'Donnell, Roy C. / Griffin, William J. / Norris, Raymond (1967): A transformational analysis of oral and written grammatical structures in the language of children in grades three, five, and seven. In: Journal of Educational Research 61: 36-39.

Purpura, James (2004): Assessing Grammar. Cambridge: Cambridge University Press.

Willis, David (2003): Rules, Patterns and Words: Grammar and Lexis in English Language Teaching. Cambridge: Cambridge University Press.

MIRJAM FRIED

Grammatical analysis and corpus evidence

Abstract

This study explores the interdependence of qualitative and quantitative analysis in articulating empirically plausible and theoretically coherent generalizations about grammatical structure. I will show that the use of large electronic corpora is indispensable to the grammarian's work, serving as a rich source of semantic and contextual information, which turns out to be crucial in categorizing and explaining grammatical forms. These general concerns are illustrated by the patterns of use of Czech relative clauses (RC) with the non-declinable relativizer *co*, by taking a set of existing claims about these RCs and testing their accuracy on corpus material. The relevant analytic categories revolve around the referential type of the relativized noun, the interaction between relativization and deixis, and the semantic relationship between the relativized noun and the proposition expressed by the RC. The analysis demonstrates that some of the existing claims are fully invalid in the face of regularly attested semantic distinctions, while others are more or less on the right track but often not comprehensive or precise enough to capture the full richness of the facts.

1. Introduction

One of the central challenges in articulating empirically grounded generalizations about grammatical patterning is the task of maintaining balance between two sources of pressure: the need to identify the inventory of relatively stable, predictably recurrent patterns that we can collectively refer to as 'grammar', while respecting and capturing the inherently dynamic, variable nature of grammatical structure. It has been increasingly noted and argued in functionally and cognitively oriented research that descriptive and explanatory adequacy in grammatical descriptions requires reference to meaning and to patterns of usage; such an approach, in turn, calls for systematic attention to a sufficiently representative body of authentic linguistic material. The goal of this paper is to show that the use of large electronic corpora is indispensable to the grammarian's work, primarily as a rich source of semantic and contextual information, which is highly relevant in categorizing and explaining grammatical forms. This is in contrast to the traditional methods and approaches, in which grammatical descriptions generally take the form of static, discrete

'rules' that are often formulated on the basis of introspection. Even when textual evidence is taken into account, it is used rather unsystematically and selectively, and any quantificational claims based on such evidence have, at best, very limited informational value. The present study argues for combining qualitative and quantitative evidence as two interdependent dimensions of grammatical analysis which aims at both descriptive and explanatory adequacy.

The theoretical and methodological issues will be illustrated on one particular syntactic form in Czech, concerning the use and classification of relative clauses (RC) with the non-declinable relativizer *co*, shown in (1); the relativizer is often accompanied by a resumptive pronoun, in (1a) exemplified by the personal pronoun *ho* 'him'. The absolutive RCs constitute a relativization strategy that is formally distinct from agreeing RCs, introduced by a fully declinable agreeing relative pronoun *který* 'which'; the examples in (2) are constructed agreeing variants of (1).[1]

(1) a. *Ten člověk, co jste ho za mnou kdysi poslal,*
that man CO AUX.2PL 3SG.ACC.M after me once sent[2]
{*viděl jste ho ještě někdy potom?*}
"The man [CO] you sent [him] to me a while back, {did you ever see him again later}?"

b. *ta paní, co u nás bydlí, je moc hezká*
that woman CO at us lives is much pretty
"the woman who lives with us is very pretty"

(2) a. *Ten člověk, kterého jste za mnou kdysi poslal,*
that man which.ACC.SG.M AUX.2PL after me once sent
{*viděl jste ho ještě někdy potom?*}
"The man who[m] you sent to me a while back, {did you ever see him again later}?"

b. *ta paní, která u nás bydlí, je moc hezká*
that woman which.NOM.SG.F at us lives is much pretty
"the woman who lives with us is very pretty"

---

[1] Unless otherwise noted, the examples all come from the SYN2000 corpus of written Czech.

[2] Abbreviations used in the glosses: AUX 'auxiliary', SG/PL 'singular/plural', NOM 'nominative', ACC 'accusative', DAT 'dative', GEN 'genitive', INS 'instrumental', M 'masculine', F 'feminine', NEG 'negation', PRES 'present', FUT 'future', IMP 'imperative', PST 'past', RF 'reflexive'.

While the agreeing RCs are fairly well understood, the absolutive RCs have so far attracted only sporadic attention among Czech linguists, although some partial studies of their properties and distribution do exist (Zubatý 1918; Poldauf 1955; Svoboda 1967, 1972; Lešnerová / Oliva 2003) and reference grammars or other comprehensive grammatical works may briefly mention them (Trávníček 1951, Kopečný 1962, Šmilauer 1972, *Mluvnice češtiny* 1987, Grepl / Karlík 1998). As a first step toward a more comprehensive examination of the absolutive RCs, this study will take a subset of existing claims about them and the 'rules' for their form, interpretation, and distribution as presented in the Czech grammatical literature, and test their accuracy on corpus material. The relevant analytic categories, with implications for relativization strategies beyond the Czech facts, will revolve around the referential type of the relativized noun (henceforth referred to as the head N), the interaction between relativization and deixis, and the semantic relationship between the head N and the proposition expressed by the RC. The analysis, which takes into account frequency-based quantitative patterns of usage, will demonstrate that some of the existing claims are either fully invalid, or too general to capture relevant semantic distinctions, while others are more or less on target but often too inflexible to truly capture the attested facts. In general, the point of the present work will be to introduce corpus evidence into the task of analyzing the absolutive RCs (or *co*-RCs) in their full, empirically documented complexity.

Thus, on the basis of corpus material, the present study argues for a more dynamic approach to grammatical analysis, one in which grammatical generalizations can be structured in cognitively and communicatively coherent networks of related grammatical patterns. The networks simultaneously provide a tool for (i) identifying points of potential fluctuations within the usage of a particular form and (ii) tracking incipient shifts between the form and / or function of a given grammatical pattern.

2. Background – relative clauses with the relative pronoun *který*

Relative clauses marked by the agreeing relative pronoun *který* cover a broad functional and semantic spectrum. For the purposes of this study, I will take it for granted that we can, at a minimum, identify the interpretations exemplified in Table 1; this taxonomy is a synthesis of two existing and roughly compatible accounts of these clauses (Svoboda 1972: 109, Grepl / Karlík 1998: 184-196) that, taken together, provide a sufficient level of detail to be useful.

Table 1: Examples of RCs with relative pronoun *který*

I-A. Determinative restrictive	
1. Concept/category membership/ defining feature of head N	*Jsou lidé, které o tomhle nikdy nepřesvědčíš.* are people.NOM which.ACC.PL about this never NEG.convince.FUT.2SG "There are people who you'll never convince." (Svoboda 1972)
2. 'Kind of' specification	*Hledáme manažerku, která umí francouzsky.* seek.PRES.1PL manager.ACC.SG.F which.NOM.SG.F know.PRES.3SG French "We're looking for [a] manager who [can] speak French."
3. Identification	*Podej mi knihu, která leží tam na stolku.* hand.IMP.2SG 1SG.DAT book.ACC.SG.F which.NOM.SG.F lies there on table "Hand me [the] book that's over there on the table."
4. Characterization	*Včera jsem viděl film, který natočil.* yesterday AUX.1SG see.PST.SG.M film.ACC.SG.M which.ACC.SG.M made {*Forman ještě v Československu.*} (Grepl / Karlík 1998) "Yesterday I saw [a] movie that Forman made {when still [working] in Czechoslovakia.}"
I-B. Determinative non-restr.	{*ale nakonec mně bude chybět i*} *ten Zetka, kterým jsme ve třídě všichni opovrhovali.* that Z.NOM.SG.M which.INS.SG.M AUX.1PL in class all.NOM.PL.M look.down.PST.PL "{but in the end I'll be missing even} that [guy] Zetka, who the whole class looked down on"
II. Non-determinative (always non-restrictive)	
II-A. Explicative	{*ale nakonec mně bude chybět i*} *Zetka, kterým jsme ve třídě všichni opovrhovali.* Z.NOM.SG.M which.INS.SG.M AUX.1PL in class all.NOM.PL.M look.down.PST.PL.M "{but in the end I'll be missing even} Zetka, who the whole class looked down on"
II-B. Continuative	*Hledal asi hodinu poštovní schránku, kterou nenašel.* seek.PST.SG.M maybe hour mailbox.ACC.SG.F which.ACC.SG.F NEG.find.PST.SG.M "He[spent] about an hour looking for [a] mailbox, which he didn't find."

Functionally, the RCs form two major classes: + / – determinative (type I vs. II) and + / – restrictive (type I-A vs. the rest). The former captures the RC's status according to its (ir)relevance for identifying, or determining, the referent of the head N, while the latter establishes restrictiveness-based distinctions within the determinative patterns; non-determinative patterns are all non-restrictive. The determinative restrictive clauses come in several semantic flavors. The RC may determine the head N in terms of *category* membership by expressing some fundamental, defining features of the head N (ex. I-A-1) in Table 1; as a possible (and potentially non-existent) token of a *kind* (ex. I-A-2); as a concrete individual that is fully *identified* in a given context by the proposition expressed in the RC (ex. I-A-3); or as a concrete unique referent that is *characterized* as such by the RC but whose identity cannot be fully established in a given context (ex. I-A-4). The determinative non-restrictive RCs (type I-B) co-occur with head Ns that consist of a deictically anchored noun with unique reference (e.g., proper nouns); the obligatory presence of the demonstrative pronoun *ten / ta / to* "that.m / f / n" individuates the referent in context and contributes to its identifiability. Type I-B forms a minimal pair with non-determinative *explicative* RCs (type II-A), in which the head N is also a noun with unique reference but any presence of a demonstrative pronoun is prohibited; the job of these RCs is to provide further commentary about a referent that is already fully identified without the RC. Finally, *continuative* RCs express a proposition that is logically independent of the properties of the head N and is in a coordination relation to the main clause (type II-B).

The meanings and functions exemplified in Table 1 can be organized in a preliminary representational taxonomy sketched in Figure 1. It has been acknowledged (e.g., Svoboda 1972: 109, Grepl / Karlík 1998: 187) that it may not always be easy (or even possible) to categorically differentiate one type from another. Certain semantic overlaps and somewhat fluid transitions between parts of the taxonomy are apparent, particularly among the non-restrictive uses (in the diagram enclosed in the gray area), but potentially also in the characterization RCs since these do not allow explicit deixis and do not ensure full identification of the head N, in contrast to other restrictive RCs; their somewhat special relationship to the individuating function is indicated by the dotted line in Figure 1.

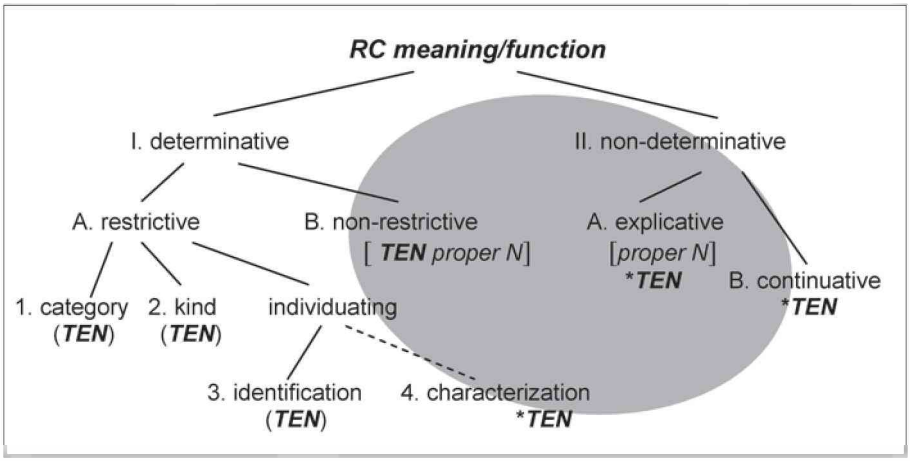


Figure 1: Functional and semantic classification of RCs with relative pronoun *který*

The properties of the agreeing RCs can thus be summarized as follows: (i) They can have both restrictive and non-restrictive interpretation and there is no obligatory marking associated with (non)restrictiveness, whether in written or spoken Czech. (ii) The relative pronoun agrees in number and gender with the relativized noun. (iii) Agreeing RCs are stylistically neutral (in terms of register, genre, text-type). (iv) Different semantic types appear to interact with deixis in different ways, showing distinct collocational preferences regarding the use of the demonstrative pronoun (in the diagram indicated by the pronoun *TEN* "that"); explicit deixis is possible in restrictive clauses of type I-A-1,2,3 (indicated by the parentheses), obligatory in the determinative non-restrictive clauses (type I-B), and prohibited everywhere else (indicated by the asterisk).

3. Relative clauses with absolutive relativizer *co*

Against the background of the agreeing RCs just described, the RCs with the absolutive relativizer *co* can be characterized as follows. As shown in the introductory examples in (1) and (2), the two types of RCs often (though not always) appear interchangeable. The absolutive relativizer is often accompanied by a resumptive (personal) pronoun, which agrees with the head N in number and gender and indicates the head N's grammatical function in the RC by being marked for the appropriate case (in the agreeing RCs, all three categories are expressed by the relative pronoun *který*). Czech grammatical literature al-

most uniformly describes the absolutive RCs as stylistically restricted and, specifically, as just a colloquial variant of the agreeing RCs.[3] However, beyond these general observations, the existing treatment leaves a lot unanswered about the defining features and the distribution of the absolutive relativizer and about the functions of these RCs. In this section, I will examine several specific claims that have been put forward about the *co*-RCs and confront them with what can be found in the *Czech National Corpus*.

The first two claims, somewhat interrelated, are rather general and have to do with the interaction of *co*-RCs with the determinative function and with deixis. The remaining questions to be addressed are more specific and concern concrete distributional constraints that follow from the first two claims. I will turn to the general issues first.

3.1 Absolutive RCs in non-determinative functions

Existing accounts mostly agree that the absolutive RCs do not occur in non-determinative uses, i.e., they cannot replace the agreeing relative pronoun *který* in the type II clauses of the taxonomy (e.g., Trávníček 1951: 1164, Svoboda 1972: 106, *Mluvnice češtiny* 1987: 528). The terminology may vary from author to author but the conceptual consensus is clear, although sometimes it is only implied by the examples used for the *co*-RCs rather than being explicitly stated. Some accounts do hedge their classification by noting that RCs "usually" (e.g., Šmilauer 1972: 262) express determination, but no further commentary is offered as to the conditions under which they may not serve this function. Overall, the functional and distributional constraints on *co*-RCs in Czech grammatical literature can be summarized as a prohibition on non-determinative contexts (type II), with a stronger version formulated by Svoboda (1972: 106), who states it as a prohibition on non-restrictive contexts (both type II and type I-B).

Let us now examine how these hypothesized distributional constraints hold up against corpus evidence. The material on which my observations are based consists of a randomly selected sample of 879 relevant tokens, each of which is coded for the functional / semantic type of RC (according to the taxonomy in

[3] This blanket statement turns out to be an exaggeration. The corpus shows that the textual distribution of *co*-RCs is more complex and any stylistic or genre-based conditioning of *co*-RC usage requires finer-grained semantic analysis. I will ignore this dimension here, leaving it to future research, in which both written and spoken data must be used.

Figure 1); animacy, grammatical number, and semantic type of the head N, and the grammatical function the relativized referent serves in the RC, as reflected in the case form of the resumptive pronoun. The latter is exemplified by the introductory examples: in (1a) we have a resumptive pronoun in the accusative since the head N is the direct object in the RC, while (2a) contains no resumptive pronoun since the head N is the subject of the RC. Some other case options are illustrated below: a dative in (3) and instrumental in (4).[4]

(3) *bojí* *se* *tlustých* *líných* *kocourů,* *co* *jim* *tam* *pan* *Bajer*
they.fear RF fat lazy tomcats CO 3PL.DAT there Mr. B.
nosí *jídlo*
carries food
"[they] are afraid of the fat, lazy tomcats [CO] Mr. Bajer brings food to [them]"

(4) {*Pomyslel jsem na doktora Fausta a*} *na* *ten* *starý* *problém,* *co* *jsem*
on that old problem CO AUX.1SG
se *jím* *zabýval* {*při nudnějších hodinách náboženství*}
RF 3SG.M.INS be.occupied.PST.SG.M
"{I thought of doctor Faust and} of that old problem [CO] I used to occupy myself with [it] {during somewhat boring lessons in theology}"

The sample also contains numerous tokens of two additional kinds of RCs that are not explicitly included in the *který*-taxonomy in Figure 1 and that require further commentary. One, exemplified in (5), is defined semantically by the type of its head N: the head is a relatively generic temporal expression (typically *doba* 'time period', but also various calendric units) that marks a point on a timeline. The RC identifies the temporal point as having particular properties, and the RC never refers back to the time expression through a resumptive pronoun, regardless of the grammatical function in the RC; I will label these RCs as 'temporal':

(5) a. *Od* *té* *doby,* *co* *jsem* *byl* *u* *vás* *doma*
from that time CO AUX.1SG be.PST.SG.M at you home
{*na návštěvě, už utekly asi čtyři roky*}
"{It's been about four years since} the time [CO] I visited you in your home"

[4] It is a well-known fact that the resumptive pronoun is prohibited in the subject function (2a), obligatory in all oblique functions (genitive, dative, instrumental, locative), and appears to be optional as a direct object, at least in some cases. The details of the distribution of the accusative pronoun are not well understood and require further study; I will not be concerned with this issue here.

b. *Tu neděli, co u nás byla paní Bohdalová, {padly všechny světové rekordy}*
that Sunday CO at us was Mrs. B.
"That Sunday [CO] Mrs. Bohdalová came by {all world records were broken}"

The other type, to my knowledge never mentioned in the existing accounts, is structurally and semantically more complex but for our purposes, it is sufficient to note that it is semantically distinct from all other RCs and must be, therefore, categorized as a separate subtype. I will refer to it as a 'quantifying' RC and a typical example is given in (6). Semantically, the RC is a particular – namely, quantifying – semantic subtype of the Explicative RC: the proposition expressed in the RC is presented as applicable to all possible members of the class denoted by the head N and the quantity is presented as exhaustive (notice, for example, the presence of the universal quantifier *všechny* "all" modifying the head N, as an explicit marker of the exhaustiveness). Moreover, the RC always contains a resumptive pronoun in the genitive of quantity (*jich* "of them").

(6) *Všechny politické strany, co jich máme, {zastávají skvělé*
all political parties CO 3PL.GEN have.PRES.1PL
a objevné myšlenky: prosperitu, pořádek, péči o potřebné, morálku, svobodu}.
"All the political parties – [the full number CO] we have [of them] – {advocate splendid, novel ideas: prosperity, order, care for the needy, morality, freedom}."

Both of these special subtypes of *co*-RCs also show signs of formulaicity, both structural and semantic, but those details need not concern us here. Their idiosyncrasies are discussed in Fried (in press).

If we just do a simple count of all the tokens in the sample, the determinative RCs outnumber the non-determinative ones at the rate of about 9:1, or, put differently, the non-determinative tokens represent around 10% of the total; this is shown in actual numbers at the top of Table 2. At this global level, then, the corpus distribution lends support to the intuition (however vaguely stated and left without evidence or argumentation) that *co*-RCs 'usually' express determination; put differently, the cautious formulation offered by some grammarians is closer to reality than any categorically stated prohibition on non-determinative functions (found in most accounts). However, it is very instructive to take a closer look at the distribution of the individual semantic types within each of the two broad categories, as summarized in Table 2. The columns represent the semantic types within the RC taxonomy, while the rows

itemize the RCs according to the grammatical function of the resumptive pronoun (including the null expression in the nominative) and the two special constructions exemplified above in (5-6).

	Total	I. Determinative					II. Non-determinative		
	789[5]	714					75		
		Cat.	Kind	Ident.	Charact.	Unique N	Explic.	Cont.	Other
	Totals	1	17	640	43	13	71	1	3
NOM	306	1	15	242	20	12	12	1	3
ACC	147		2	127	12	1	4		
DAT	24			18	3		3		
GEN	12			6	6				
INS	9			7	1		1		
LOC	5			4			1		
Quant.	50						50		
Temp.	236			236					

Table 2: Distribution of *co*-RCs according to function and semantic type[5]

First of all, it is evident that the totals are somewhat skewed if we include the two special and partially formulaic types, i.e., the quantifying and temporal RCs; they are each disproportionately frequent within their functional category. The temporal RCs are always of the Identification type and the quantitative RCs cannot be anything but a subtype of Explicative RCs. If we leave these two types out of the count, the relative frequency of non-determinative tokens goes up significantly from the overall 9:1 ratio; without the two special constructions, the occurrence of non-determinative RCs in the sample almost doubles, making up about 19% of all tokens. That alone undermines the traditionally held view that *co*-RCs are excluded from non-determinative functions.

At the same time, the distribution suggests an explanation for the traditional – and evidently inaccurate – analysis. The evidence that is typically offered in support of the analysis of *co*-RCs as purely determinative focuses solely on the Continuative type, and the corpus indeed confirms that this function is rather unexpected (although not completely unattested), as shown in Table 2. Instead, the vast majority of the non-determinative uses of *co*-RCs are found in the Explicative category whether we count the quantifying construction or

[5] This number includes only full nouns as the head N, not personal pronouns. Those will be discussed in Section 3.3.

not. An example of an Explicative *co*-RC is the second *co*-clause in (7): the proposition expressed in the RC simply elaborates on the description of Admiral Nelson, whose identity is already fully established without the RC. In contrast, the first *co*-RC in the example is a straightforward case of an Identification function, restricting the referential range of the phrase *ten pán* "the man".

(7) *Ten pán, co stojí nad vámi, je admirál Nelson, co porazil v roce*
that man CO stands above you is Admiral N. CO defeated in year
{1805 zlého Napoleona}
"The man that is standing above you is Admiral Nelson, who in 1805 defeated {the bad [guy] Napoleon}."

However, the existence of Explicative RCs, as a distinct semantic subtype of non-determinative relativization, has not been identified or acknowledged explicitly except in the analysis of *který*-RCs in Grepl / Karlík (1998) and its existence, let alone its specific properties, has not been considered anywhere in the context of *co*-RCs.

It is also worth noting that by far the most common usage of *co*-RCs centers on the Identification function, with Characterization being a distant second. This patterning is consistent with the traditional view that marking restrictiveness might be the core domain of the *co*-RCs, but yet again, the corpus provides tangible evidence that it is merely a tendency, however strong. Identification constitutes the focal point within a wider distributional range and thus cannot be presented in the form of a categorical 'rule' along the lines of, for example, Svoboda's (1972: 106) conclusions. In fact, one preliminary generalization we can draw from the correlations gathered in Table 2 is the following: there is a hierarchy of semantic preferences exhibited by the distribution of *co*-RCs in authentic discourse. Crucially, the hierarchy does not follow a clean determinative / non-determinative distinction as the traditional accounts suggest, but rather follows the finer semantic distinctions. The hierarchy appears to take the shape suggested in (8), which entails that the most common, typical candidate for the use of a *co*-RC is the context of identifying or otherwise describing specific individuals (around the middle portion of the taxonomy in Figure 1); the symbol '*' indicates that Continuative RCs are barely attested in the corpus. I will return to the significance of this hierarchy in Section 4.[6]

[6] Other interesting observations emerge from Table 2 as well, such as, for example, the overwhelming preference of *co*-RCs in which the head Ns serve the subject function inside the RC. Due to space limitations, I have to leave this aspect of the distribution aside for now.

(8) **Hierarchy of semantic preferences:**
identification (type I-A-3) > characterization (I-A-4) > explicative (II-A) > kind-of (I-A-2) > non-restr. determinative (I-B) > category (type I-A-1) > *continuative (II-B)

3.2 Correlation with deixis

To the extent that deixis has been addressed at all in the context of relativization, it has been noted that *co*-RCs are predominantly deictic (Svoboda 1967: 10, 1972: 105-106): their primary function is to point – in space, time, or discourse – to specific entities, thereby uniquely identifying (or individuating) the referent of the head N. A concrete manifestation of this relationship is the collocation of the head N with the demonstrative pronoun *TEN* "that", as illustrated in the introductory examples in (1);[7] this collocational pattern is also hypothesized to be the historical origin of the *co*-RCs (Svoboda 1967: 10). However, actual corpus data call for a substantially more nuanced analysis. For the sake of expediency, I will refer to the head Ns that are modified by *TEN* as 'deictic' and the head Ns without a demonstrative as 'non-deictic'.

First of all, the full sample splits down the middle between the deictic (396) and non-deictic (393) tokens; this alone contradicts Svoboda's assertion quite robustly. Moreover, the assumed dominance of deictic contexts in the distribution of *co*-RCs becomes even less convincing when we consider correlations between deixis and other criteria concerning the nature of the head Ns, namely, animacy and number. All three parameters – deixis, animacy, and number – are known to correlate with differences in degrees of referentiality or individuation and it is therefore relevant to examine how they interact in the context of the RCs as well, since they all can be expected to bear on the question of determination and restrictiveness.

Before we address these correlations, though, let us note that there is one domain in which the dominance of deictically marked head Ns appears to be confirmed. In the temporal RCs, the collocation with *TEN* is more than twice as likely as the use of a bare N: out of the total of 236 tokens, a demonstrative phrase as the head NP occurs in 159 cases, in contrast to 77 cases without *TEN*. Considering that the temporal usage of the *co*-RCs is often the only one that the

[7] The use of capital letters (*TEN*) is a typographical indication that I am only referring to a lexeme, without making explicit reference to its morphological shape, particularly the formal differences in gender and number.

existing accounts consider as accepted in the literary language (Trávníček 1951: 1165) and therefore on a par with *který*-RCs, it is not surprising that the temporal clauses may simply be taken as the only (or at least the primary) example of *co*-RCs. Nevertheless, even here the use of *TEN* is far from obligatory.

Let us now turn to the correlation between deixis and animacy. In order to remove any bias contributed by the two special constructions, i.e., the temporal RCs, which are overwhelmingly deictic and necessarily with inanimate head Ns only, and the quantifying RCs, which are necessarily non-deictic, I will exclude those tokens from the counts in the rest of this section. Deixis also seems irrelevant in the oblique grammatical functions (DAT, GEN, INS, LOC). The actual token frequencies of these forms are included in the counts and shown in Table 3, but their contribution to the analysis is marginal. After all, these case forms are rather rare to begin with compared to the direct cases (NOM, ACC), as we saw in Table 2.

Total	**Non-deictic**			**Deictic**		
503	Total	266 (=53%)		Total	237 (=47%)	
		animate	**inanimate**		**animate**	**inanimate**
		46%	**54%**		**56%**	44%
NOM	173	**61%**	39%	133	**78%**	22%
ACC	73	12%	**88%**	74	19%	**81%**
DAT	10	8	2	14	11	3
GEN	4	–	4	8	4	4
INS	4	–	4	5	–	5
LOC	2	–	2	3	–	3

Table 3: Deixis and animacy

The general pattern captured in the top portion of Table 3 shows two things: (i) the overall distribution favors non-deictic over deictic context, albeit not in a dramatic way (53% over 47%), and (ii) there is a general asymmetry between inanimate and animate head Ns: inanimate Ns are more frequent in the non-deictic contexts, while animate Ns outnumber inanimate Ns in the deictic contexts. The relative frequencies do not provide an overwhelming contrast but are sufficiently suggestive of the potential correlation between animacy and an explicitly marked determination. This potential comes into relief when considered in relation to the grammatical functions played by the head N in the RC.

Subjects (NOM) and indirect objects (DAT) generally attract animate referents more than inanimates, but it is interesting that in the nominative, the likeli-

hood of animate head Ns increases significantly in the deictic contexts (78%) compared to the non-deictic contexts (61%). This is particularly striking in light of the fact that in the actual number of tokens, animate head Ns are about equally distributed across deictic and non-deictic contexts (104 vs. 106, respectively). This asymmetry is not contradicted by the dative pattern and is further confirmed by the accusative pattern, where the vast majority of head Ns are inanimate entities (again, not surprisingly) but their distribution in deictic vs. non-deictic contexts displays a comparable correlation between animacy and deixis. Their distribution with respect to deixis is about even (64 tokens in non-deictic contexts vs. 60 in deictic ones), but inanimate non-deictic contexts are somewhat more likely (88%) than deictic ones (81%).

Given these patterns, we may explore further the hypothesis that the use of the demonstrative pronoun has to do with the degree of referentiality of the head N, rather than being an inherent property of the *co*-RCs. To test this possibility further, we can probe the distribution of grammatical number (singular vs. plural) as another relevant parameter. Overall, singular head Ns are more frequent than plural Ns (59% vs. 41%, respectively) and this distributional asymmetry becomes even more pronounced when we track the correlations with deixis and animacy. The relative frequencies are summarized in Table 4. We can see that there is about the same number of singular tokens in both non-deictic and deictic contexts (151 vs. 149), but the likelihood of a singular head N goes up in deictic contexts; the ratio singular : plural in the sample is roughly 5:3 (63% over 37%) in favor of the singular, while the difference between singular and plural in non-deictic contexts is less pronounced (57% over 43%).

Total	**Non-deictic**			**Deictic – *TEN***		
503	Total	266 (=53%)		Total	237 (=47%)	
		animate	inanimate		animate	inanimate
		123	143		133	104
Sg.	151 (57%)	51%	**61%**	**149** (63%)	**65%**	55%
Pl.	115 (43%)	49%	39%	**88** (37%)	35%	45%

Table 4: Distribution of the demonstrative *TEN* relative to animacy and number

The distributions in our sample thus suggest that the presence of the demonstrative cannot be attributed to the *co*-RCs as their inherent feature but, rather,

depends on the properties of the head N, particularly number and animacy. The prototypical constellation that attracts deixis appears to be a singular animate N. It also follows from the frequencies, though, that number ranks higher than animacy in determining preferential co-occurrence with *TEN*. We can propose a hierarchy of deictic contexts (i.e., the structure [*TEN* N, *co*]) as follows; note that animacy plays a role in the singular NPs but does not seem to make any difference in the plural (the numbers are percentages of a given configuration in [*TEN* N] occurrences):

(9) head N = Anim. sg > Inanim. sg > (Anim. pl, Inanim. pl)
37% 23% 20% 20%

The correlations in (9) are consistent with treating the usage of *co*-RCs as an issue of individuation or high referentiality, rather than simply deixis. The preferred head N tends to be a highly individuated / referential entity, at the expense of less individuated / referential ones.

3.3 Head Ns with unique reference

Perhaps the least explored domain within the proposed taxonomy of Czech relativization are the segments in the middle, at the hypothesized boundary between determinative non-restrictive clauses (I-B) and the non-determinative Explicative clauses (II-A). Both of these segments involve the same type of head N (nominals with unique reference) and the crucial difference between them is the obligatory presence of *TEN* with the former and obligatory absence of *TEN* with the latter. It is the demonstrative that contributes the determinative function (I-B), thereby invoking an interpretation that is based on some sort of contrast, whether explicitly stated or just implied; in the absence of the demonstrative (II-A), no contrastive reading is available. We thus obtain the interpretive distinction between (7) above and (10) below. While in (7), the communicative objective is to offer further commentary about Admiral Nelson, in (10), the speaker's goal is to establish the identity of a guy named Vantoch:

(10) *nejste vy ten Vantoch, co se se mnou v Jevíčku prával,*
aren't 1PL.NOM that V. CO RF with me in J. fight.PST.SG.M
{*když jsme byli kluci*?}
"are you that [guy] Vantoch, who used to have fights with me in Jevíčko, {when we were little boys}?"

On the one hand, the use of the person's last name suggests unique reference and thus complete identification. The context, however, places this person in contrast to other schoolmates among which the speaker is trying to single out just one, by offering a description that might set Vantoch apart from other potential candidates. We cannot classify the reading as restrictive (there is only one person named Vantoch that the speaker went to school with), but the demonstrative creates a distinctly different setting from the bare noun structures illustrated in (7); in (10), the RC is relevant for the head N's precise identification.

If we take seriously the blanket prohibition on non-determinative usage, discussed in Section 3.2, we should expect no attestations of the kind in either (7) or (10); Svoboda (1967: 7, 1972: 106) states this condition directly. In reality, we find both, as has already been noted and quantified in Table 1, although not in any overwhelming numbers (25 tokens in the sample). The existing accounts thus overstate the case by making a categorical judgment, but the basic insight about the limited compatibility of *co*-RCs with unique-reference Ns is on the right track. The deictic usage (I-B, example 10) is essentially consistent with the patterning discussed in the preceding section in that the majority of tokens involve proper nouns denoting human individuals (i.e., animate singular). As expected, the deictic usage of unique-reference Ns outnumbers the non-deictic usage, but only at the ratio of about 3:2, which indicates that non-deictic usage (specifically the Explicative) is not only possible but is not even all that exotic within the domain of unique-reference head Ns. Overall, then, the corpus contradicts both of the two general claims: the requirement of deixis on the head N as an inherent feature of *co*-RCs and the expectation that *co*-RCs cannot serve non-determinative functions or co-occur with *TEN*.

Sorting out the issue of unique reference also extends to one particular subtype of head nominals, namely, personal pronouns. The full range of such pronouns can occur in the agreeing RCs with the relative pronoun *který* 'which', e.g., the structures *já, který* "I who", *my, kteří* "we who", etc. It follows from the assumptions about *co*-RCs being necessarily determinative that *co*-RCs cannot be headed by pronouns that necessarily mark unique reference, such as *já* "I" and *ty* "you-sg." (yielding **já, co* "I who" or **ty, co* "you-sg. who"). The reasoning, explicated in Svoboda (1967: 6) goes as follows: the speaker and the hearer are fully and uniquely identified by the pronoun itself and cannot, therefore, be modified ('determined') by an RC whose semantic range is limited to indicating restrictiveness or at least determination. If we

take the speaker as an example, it should not be possible to restrict reference to a specific ego in contrast to the same ego. The presence / absence of unique-reference personal pronouns thus, again, speaks to the issue of (non)restrictiveness and (non)determinativeness.

The sample contains 90 unambiguous tokens of personal pronouns as the head N of *co*-RCs and the distribution of the 1st and 2nd pers. sg. confirms Svoboda's insight that the singular pronouns *já* "I" and *ty* "you-sg." are incompatible with the determinative function: they are very rare in the sample (altogether only six tokens among all the personal pronouns) and only one of those, found in a dialog of a theatrical play and shown in (11), can be classified as helping establish the referent's identity.

(11)	{FANKA:	*To sou voni, milostpane?*			
	LOUPEŽNÍK:	*Ne, to jsem já, Fany.*}			
	FANKA:	*Kterej já?*			
	LOUPEŽNÍK:	*Já, co*	*tu*	*byl*	*ráno.*
		I [CO]	here	was	morning
	FANKA:	*Ten zabitej? To už běhaj?*			
	{FANKA:	'Is that you, sir?			
	LOUPEŽNÍK:	'No, Fanny, it's me.'}			
	FANKA:	'Which me?'			
	LOUPEŽNÍK:	'I who was here this morning.'			
	FANKA:	'The dead one? You're on your feet again?'			

The utterance in line 3 explicitly presents a setting that presupposes multiple referents, by posing the question "which [one] I?"; but the full context is also conducive to this shift since Fanka is evidently faced with the task of choosing between two distinct individuals: one that she expects and addresses in the first line (her master), and another, who shows up, unexpectedly, instead (the master's daughter's young admirer). Note also that the contrastive context, necessary for the determinative reading, is not marked explicitly by anything in the sentence itself (e.g., by using a demonstrative, as was the case in (10) above) but merely follows from the broader context. This observation further supports the generalization that the determinative function of *co*-RCs need not be encoded directly as an inherent feature of these RCs.

Aside from this clearly shifted reading, however, the remaining tokens, exemplified in (12) below, are all cases of non-determinative usage. They all fit the Explicative category, in simply adding an informative comment about the

speaker or hearer, with no identificational relevance; in (12) the RC actually suggests the flavor of a *because*-clause: not just "... I who needs it more" but "... I, since I need it more".

(12) *proč jsem nevyhrála já, co to víc potřebuji?*
why AUX.1SG NEG.win.PST.SG.F 1SG.NOM CO it more need.PRES.1SG
"Why wasn't the winner me, who needs it more?"

Overall, the corpus confirms that determinative readings are quite marginal with the 1st and 2nd pers. sg. pronouns, but does not substantiate any absolute prohibition on non-determinative usage of the *co*-RCs. When these pronouns do appear they of course have to be non-determinative, which follows from their inherent nature as unique-reference nominals.

4. Functional and semantic range of absolutive relativization in Czech

Based on the attested frequencies in the corpus sample, certain properties emerge that can be seen as *prototypically* associated with *co*-RCs; they are listed in (13):

(13) **Prototypical features of *co*-RCc**

Function:	determinative restrictive
Semantics:	individuation of head referent
Syntax:	relativized N is the subject in RC
Head N:	concrete, animate, singular entity

The functional and semantic features may appear, on the whole, to conform to the traditionally posited constraints. There is one important difference, though: the corpus shows them to be mere tendencies to start with and a closer look at the specific semantic subtypes helps us piece together a much more nuanced picture that leads to a deeper understanding of the nature of this relativization strategy.

Let us now recall the proposed taxonomy of relativization in Figure 1, which reflects the current state of knowledge in Czech grammatical literature and which we took as the starting point for our analysis. However, rather than a strict taxonomy with discrete boundaries, we can view the diagram as delimiting a particular functional or conceptual space (in the spirit of typological se-

mantic maps, e.g., Croft / Shyldkrot / Kemmer 1987; Haspelmath 1997, 2003; Croft 2001 or constructional maps that have been proposed for the purpose of capturing grammatical patterning in a single language, e.g., Fried 2005, 2009) within which attested meanings of RCs can be coherently organized. Such a space presupposes fluid transitions between individual nodes, which is also more consistent with the often observed difficulty in classifying individual tokens as categorically belonging to one type or another, particularly across the gray domain in the middle.

Figure 1 represents the space that is fully covered by the agreeing *který*-RCs and if we were to incorporate the existing accounts of the *co*-RCs, it would amount to essentially admitting *co*-RCs as coinciding with the determinative node (type I) and all its subtypes (with some disagreement left open concerning subtype I-B) and as being excluded from the non-determinative node (type II) and its subtypes. However, if we map the corpus distribution onto this space, we can not only establish points of similarity and dissimilarity in relation to the agreeing *který*-relativization, but can also begin to articulate an empirically grounded and descriptively much more accurate account of the *co*-RCs as a distinct grammatical pattern. Figure 2 summarizes our findings in a preliminary representation of the relevant conceptual space; the dashed-line oval delimits the core domain in which *co*-RCs are attested.

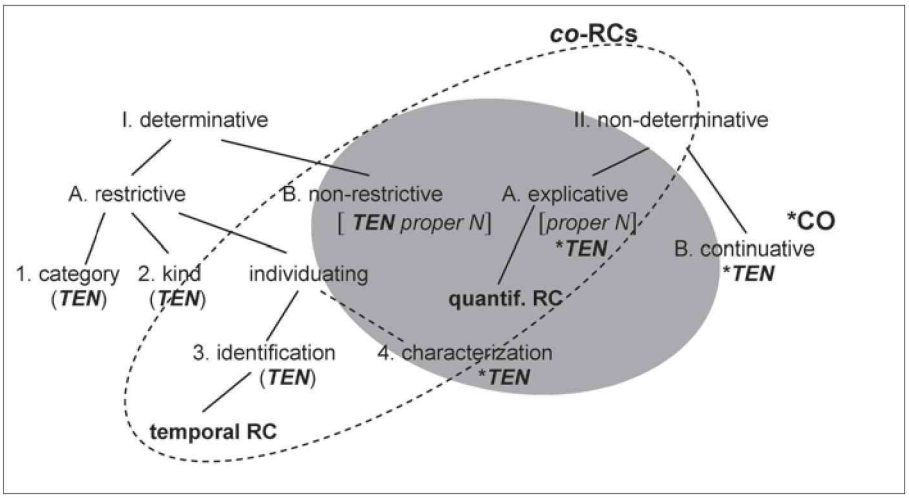


Figure 2: Distribution of absolutive RCs in the corpus

The map in Figure 2 captures several important (and newly established) facts about the nature of the *co*-RCs. In general, their primary function is centered on individuating the referent of the head N in a presupposed contrastive context; the individuation may lead (and overwhelming does so) to pure identificational meaning, but need not (yielding a less definite characterization meaning instead). This general functional preference remains very strong and manifests itself in several ways:

i) The focal point of the functional range is the Identification function (type I-A-3), within which the temporal RCs occupy a prominent position of a highly entrenched, semantically distinct, and formally partially formulaic subtype of identificational RCs.

ii) In contrast, *co*-RCs are very rare in the remaining, less central restrictive functions (determining kinds and categories or class membership). This may not be a very surprising outcome; establishing generic reference or class membership involves diminished individuation and/or a lower degree of referentiality, which is conceptually incompatible with the preferentially individuating function of *co*-RCs.

iii) To the extent that *co*-RCs extend out of the identificational range, they cover primarily the domain of unique-reference head Ns, whether determinative or non-determinative. The latter, moreover, includes a special construction (quantifying RCs) which represents a formally and semantically distinct, well-entrenched, and partially formulaic subtype of the Explicative RCs. The pull toward the unique-reference head Ns can again be motivated by the general affinity toward individuation: these head Ns, by definition, mark highly individuated, highly referential entities.

iv) The individuating character of the *co*-RCs correlates with certain semantic properties of the head Ns: the corpus demonstrates a preference for singular animate entities. Consequently, and contrary to certain existing analyses, these RCs are much less dependent on the presence of demonstrative pronouns to ensure an individuating interpretation. The most we can conclude about explicitly marked deixis is that inherently highly individuating nouns (singular, animate) show stronger co-occurrence patterns with deixis than less individuated ones (plural, inanimate).

v) Finally, we can hypothesize that the *co*-RCs, by pointing to highly individuated entities, form a relatively tight conceptual unit with their head Ns, which they either help identify or at least add some contextually salient information about them. This hypothesis can easily accommodate the Explicative non-determinative usage as well: the relationship between the RC and the head N in this pattern is reminiscent of a subordinating relation – one in which the embedded clause bears signs of conceptual dependence on the head N. In contrast, the same conceptual closeness cannot be expected in the essentially coordinating relation characteristic of the Continuative readings of relativization patterns (type II-B) since the relationship between the head N and the RC is very loose here; the two clauses (main clause and RC) express two conceptually independent propositions, just like other, formally explicit, coordinating structures.

It is perhaps also worth noting that the space in which the *co*-RCs most commonly operate coincides with the 'gray' area in the middle of the map, namely, the part that is generally considered the fuzziest domain, with the least distinct boundaries, which tend to be most dependent on actual discourse context. I hope to have shown that the fuzziness may become much less intractable with the use of corpus data, through which semantic and contextual aspects of grammatical patterning can be readily available and aid in accurate analysis.

5. Conclusions

The goal of this study was to explore the potential of integrating qualitative analysis with frequency-based evidence provided by an electronic corpus, confronted with grammatical descriptions that have been formulated without the use of any large corpora. The observations and results reported in the case study concerning Czech absolutive RCs should be taken as no more than the very first step in a more thorough investigation of this particular grammatical pattern, which has not yet received a truly systematic and comprehensive treatment. However, certain partial generalizations emerge, including potential implications for the study of RCs beyond just the Czech patterns.

In order to fully understand the use and distribution of the *co*-RCs, also in contrast to the formally agreeing relativization pattern, we must take into account finer semantic and contextual distinctions than traditionally applied.

These RCs appear to form a distinct cluster of relativization functions and meanings that all have to do with individuating the head N. The corpus material suggests that while the core domain of the *co*-RCs resides in identification functions, the clauses have spread into other, non-restrictive and non-determinative functions well beyond what traditional analyses admit possible. At the same time, the spread into the non-determinative territory is not random, but follows a conceptually coherent path within a relativization network that organizes all the attested functions and meanings of Czech relative clauses. The path, moreover, suggests a particular direction in the development of RCs, namely, gradual erosion of restrictiveness as a linguistically explicitly marked distinction. While the Czech *co*-RCs can be considered preferentially (though by no means universally) restrictive, the corpus reveals quite clearly that the absolutive relativizer *co* by itself cannot be taken as a reliable marker of restrictiveness. The diachronic dimension of this spread and the details of the *co*-RC development from the hypothesized deictic origins to what the synchronic corpus documents will require much more research. Nevertheless, even this preliminary analysis has some value for broader theoretical and typological studies concerning the status of restrictiveness as a relevant notion in classifying the inventory of RCs. The Czech facts appear to confirm the cross-linguistic observation that restrictiveness is not a highly salient linguistic category that requires explicit marking and, therefore, should not be used as a fundamentally important criterion for analyzing RCs. Instead, the salient notions, which would deserve further testing in other languages as well, seem to include the referential type of the relativized noun, the interaction between relativization and deixis, and the semantic relationship between the head N and the proposition expressed by the RC.

Finally, the present work also shows that the use of large electronic corpora enriches grammatical descriptions in several respects. Corpus material serves as an important source of semantic and contextual information, which turns out to be crucial in categorizing and explaining grammatical forms; forces us to acknowledge and directly address the dynamic nature of language; helps identify specific usage-based factors that affect variability in linguistic 'rules' and categorization; and offers greater reliability of quantificational evidence, provided we exercise a necessary dose of skepticism about its infallibility and apply adequate controls. It is clear that on the basis of corpus evidence, we can arrive not only at sufficiently dynamic, multi-faceted, and, hence, more accu-

rate generalizations about a given form itself, but also capture subtle shifts in its distribution, depending on specific, well-defined criteria. The use of corpus material has the potential of bringing the grammarian's work to a new and more realistic level of analysis.

References

Croft, William (2001): Radical Construction Grammar. Syntactic theory in typological perspective. Oxford: Oxford University Press.

Croft, William / Shyldkrot, Hava Bat-Zeev / Kemmer, Susanne (1987): Diachronic semantic processes in the middle voice. In: Giacalone Ramat, A. / Carruba, Onofrio / Bernini, Giuliano (eds.): Papers from the Seventh International Conference on Historical Linguistics. Amsterdam / Philadelphia: Benjamins, 179-192.

Fried, Mirjam (2005): Constructing grammatical meaning: isomorphism and polysemy in Czech reflexivization. In: Studies in language 31, 4: 721-764.

Fried, Mirjam (2009): Plain vs. situated possession in a network of grammatical constructions. In: McGregor, William (ed.): Expression of possession. Berlin: de Gruyter, 213-248.

Fried, Mirjam (in press): Vztažné věty s nesklonným *co*. In: Štícha, František (ed.): Kapitoly z české gramatiky. Prague: Academia.

Grepl, Miroslav / Karlík, Petr (1998): Skladba češtiny. Olomouc: Votobia.

Haspelmath, Martin (1997): Indefinite pronouns. Oxford: Clarendon Press.

Haspelmath, Martin (2003): The geometry of grammatical meaning: semantic maps and cross-linguistic comparison. In: Tomasello, Michael (ed.): The new psychology of language. Mahwah, NJ: Lawrence Erlbaum Publishers, 155-175.

Kopečný, František (1962): Základy české skladby. Praha.

Lešnerová, Šárka / Karel, Oliva (2003): Česká vztažná souvětí s nestandardní strukturou. In: Slovo a slovesnost LXIV, 4: 241-252.

Mluvnice češtiny III – Skladba (1987). Praha: Academia.

Poldauf, Ivan (1955): Vztažné věty v angličtině a v češtině. In: Sborník VŠP, Jazyk a literatura II: 159-194.

Svoboda, Karel (1967): Vztažné věty s nesklonným *co*. In: Naše řeč 50,1: 1-12.

Svoboda, Karel (1972): Souvětí spisovné češtiny. (= Acta Universitatis Carolinae, Philologica XLIII). Praha: Universita Karlova.

Šmilauer, Vladimír (1972): Nauka o českém jazyku. Praha: SPN.

Trávníček, František (1951): Mluvnice spisovné češtiny II – Skladba. Praha: Slovanské nakladatelství.

Zubatý, Josef (1918): Jenž, který, kdo, co atp. In: Naše řeč II, 37: 37-44.

Source of data

ČNK – SYN2000: Czech National Corpus (ČNK). Internet: http://www.korpus.cz. Ústav Českého národního korpusu FF UK, Praha 2000.

FRANÇOISE GADET

What can be learned about the grammar of French from corpora of French spoken outside France*

Abstract

This paper looks at some questions which were considered quite differently before corpora became the ordinary way to describe languages, focusing on the following points:

a) Our ultimate objective is to document how wide-reaching the appellation "French" can be, given the extent of variation found in the corpora: is it possible to document the whole variational span of "the French language", in what Chaudenson (2003: 182) would call "the limits of intra-linguistic variability of French"?

b) Are there grammatical phenomena which could be looked at differently and analysed using corpora?

c) Is it possible to generalise in an explanatory perspective, and to determine something of the principles which lie behind the difference between standards and vernaculars?

d) The discussion of ordinary and non-standard data, mostly spoken (only occasionally written) will lead me to consider if it is possible to qualify *vernacular varieties* as such, in what Chambers called in 2000 "universal sources of the vernacular" and in 2003 "vernacular roots"; and what I shall choose to call here "vernacular resource" (see the concluding remarks in section 3).

The linguistic variation data which will be looked at in sections 1 and 2 are primarily diatopic, and occasionally diastratic: they mostly come from geographically "peripherical" French (mostly North American) and socially "marginal" French, which means here ways of speaking which have been subjected to / been the target of (albeit sometimes in a limited way) normative pressures, as is the case for ordinary spoken French, "popular French", youth language, child language, and different types of urban or rural vernaculars.

1. French as a material for corpora: Variation and ordinary data

The point of view adopted here will concern syntax and discourse as corpora do not require the same extension and the same inherent properties according to the different purposes they are built for. Our approach will be variational (and not variationist, according to the Labovian meaning of the word), trying to widen a system perspective to an intricate understanding of multilayer dimensions.

* Thanks are due to Henry Tyne for polishing my writing in English.

1.1 Objectives for large-scale corpora of French

French is, among the major well documented occidental languages, rather particular as far as corpora are concerned: they appeared late on in comparison with other major European languages (English of course, but also Italian or German).[1] The reason for this delay can be suggested to lie in the fact that French linguists have remained a little "normative" as far as ordinary and non-standard data are concerned, even if they pretend not to be. Or rather, they have been generally "ideology-of-the-standard-oriented", and heavily biased towards the written language or an idealised if not mythical form of spoken language; and other linguists working on French have frequently followed the French in terms of this attitude.

A consequence of this delay in the uptake of corpus-based approaches to French is that there is no equivalent of, say, the *Handbook of Varieties of English* (Kortmann et al. (eds.) (2004), see in particular the synopsis by Kortmann / Szmrecsanyi[2]), continuing an already long-established tradition (see several former studies, among which Cheshire (ed.) (1991), rather sociolinguistically-oriented); and studies of French lag far behind those of English, both for variational diatopic descriptions and theoretical generalisations or explanations.[3]

The large diversity in French spoken beyond France is to be considered in relation to the consequences of two of its historical peculiarities:

- The colonial expansion along the 17th, 18th and 19th centuries (transplantation outside Europe, mainly in America and in Africa);

[1] For state-of-the-art (now dated) but very concise presentations of corpora of spoken English, Italian, Portuguese, Spanish and German, see *RFLA* (1996).

[2] See also Szmrecsanyi / Kortmann (2009), continuing Kortmann / Szmrecsanyi (2004). Specialists of the areas concerned have been asked to rate a catalogue of 76 well established (by variationist, variational, dialectological and creolist research) non standard features, concerning 11 "core areas" of morpho-syntax. They were classified into: a) "pervasive if not obligatory", b) "exists but not frequently", and c) "does not exist or is not documented". This was done for 60 L1 and L2 almost non standard varieties of English worldwide. I personally consider that to date it would be totally excluded to try to build such a grid for French.

[3] Albeit Chambers (2000) claims that sociolinguistics remained a bit "provincial" or "insular" in what concerns generalisations "across language borders", such is probably much more the case for the sociolinguistics of French, which often seems to simply apply to French problematics designed for English. Traditions of description are also quite different, current grammatical studies on English being much more often variation-oriented (Sylviane Granger p.c.), even if, of course, specialists of different areas do exist.

- The effects of normative pressures in France itself after the 17th century, the time of standardisation.

The complex history of French worldwide has given rise to a large diversity of linguistic vernaculars all around the world (in what, following Trudgill 1986, one could call "colonial French", i.e. every territory beyond the original European ones): the diversity of French today is the result of different factors, events and processes, often complex. They concern particularly history (e.g. secondary diasporas), the diversity of types of contacts, with different languages of different types (the most frequently involved being English), the relationships with standardisation (see for example Erfurt 2008) and the vitality (several situations being obsolescent or "close to death"). French is therefore interesting for general linguistics, for the great diversity its vernaculars offer, fanning out as far as creole languages (French-based creoles being the most widespread following English-based ones, and belonging to two very different geographic areas). There is a great diversity of ways of speaking French, and this is what gives it some importance, much more than its number of speakers, which remains modest: with its roughly 90 million L1 speakers, together with about 20 million L2 speakers, French is usually ranked the 11th or 12th most spoken language in the world; English, by comparison, is probably the second (after Chinese) or the third (after Spanish). French is, therefore, even if it is far behind English, the only world-language whose diversity can be compared to English (see Gadet / Ludwig / Pfänder 2009 for a tentative typology referring linguistic data to ecological situations).[4]

One might have expected this complex situation to give rise to a certain willingness to collect large-scale corpora. This is indeed what has happened, but only rather slowly and parcimoneously (Cappeau / Gadet 2007). In comparison to hexagonal French, corpora of French outside France started to be gathered earlier and are still more numerous, as linguists interested in ordinary oral varieties of French simply had no other way than gathering corpora to document the grammatical specificities of a particular variety, especially for those ways of speaking for which there is no tradition of description and no or very little literary production. For North American French the gatherings started from the

[4] The reflections presented in this paper owe much to my ongoing participation in two projects: the realisation of a Reference Grammar of French, for which I am in charge of the variation data; and the project CIEL_F (*Corpus International Ecologique de la Langue Française* – see end of section 1 for more details).

end of the 1960s (among the first ones, the Montreal corpus 1971 – see Thibault / Vincent (1990), also for its continuation in 1984, and Vincent (2008) for a second continuation in 1995) and a bit later for different African French (beginning of the 1980s – see the journal *Le Français en Afrique*).

At the time when these "colonial" gatherings started, research institutions in France were setting a massive programme in motion concerning written data (*Frantext*). Thus the collection of spoken material started later on in France, at the end of the 1970s, and was first the mere outcome of individual initiatives, like Claire Blanche-Benveniste and the *GARS* research team at Aix-en-Provence University (starting in 1977 – see Blanche-Benveniste (2000)). It has more or less been the same in other European countries (at the beginning of the 1980s for Michel Francard and the constitution of a database by *VaLiBel* group in Francophone Belgium).

The main objectives of these different gatherings differ in terms of their theoretical positioning: for example, syntactic for *GARS* and mostly sociolinguistic for *VaLiBel* (to document variation). If we except those projects and the *Français fondamental* in the 1950s, which is more a collection of recordings than a corpus, the first corpora of spoken hexagonal French were initiated outside France, at the end of the 1960s. They were gathered following a British initiative of teachers of French for "Corpus d'Orléans", or in Germany – see Gülich (1970), who says today (p.c.) that it is through listening to spoken corpora that she established her research object, as she became sensitive to the oral recurrence of some adverbs like *alors*, *puis* or *enfin* which she ultimately called "Gliederungssignale". As for the format, most of those corpora follow a model of interviews.

The absence in Europe of French-speaking public policy concerning spoken corpora (particularly in France) contrasts with what happened in Quebec: the corpora collected from the 1970s onwards, especially the oldest of them, Sankoff et al. (1976), were gathered following a political impulsion which led ultimately in the 1980s to the defining of "français standard d'ici" (See Boisvert / Laurendeau 1988, Sankoff et al. 1976, Thibault / Vincent 1990, Vincent 2008). Nevertheless, none of these corpora from the 1970s and 80s is nowadays easily available.

1.2 External aspects of the worldwide diversity of French

If it is possible to consider that there is a "French-speaking world" from a linguistic point of view (which is the assumption of all corpora gathering projects), there are important differences in what is to be learned from French spoken in America (mostly L1, mostly minority language except in Quebec, sometimes obsolescent or at least not very widely used) and in Africa (mostly L2, in some countries being only an official language meeting formal usages, in others becoming part of the local landscape, endogeneous and gaining new emergent processes).[5] The crucial difference is probably not so much between native and non-native use (whose interest is partly overestimated through an *Ideology of the standard* attitude), but rather the whole ecological organisation. A major characteristic of francophony studies, contrasting sharply with anglophony appears to be a lack of relationships between "africanists" and "americanists" (see on the contrary the scope of a journal like *English Worldwide* for English, with no equivalent for French to date).

In this paper, the examples will mostly be taken in North American vernaculars (Canada and the US, especially Louisiana), which constitute a coherent set: at first historically, and, as it is possible to suppose, with structural impacts. An additional reason is that it is the area in which the first and most far-reaching descriptions have been carried out since the 1970s and even with returns to the same field (e.g. in Montreal and in Ontario). A factor of differentiation between the different French could be the time when the language was exported: before or after the 17th century and standardisation in France itself (while trying not to overestimate the outcomes of standardisation and of the action of prescriptive grammarians on the way ordinary people usually *talk*). The historical way of looking at this question is to try to reconstruct what kind of language was spoken by sailors and settlers: Who were they? Where did they come from? What were their usual ways of speaking? etc. ... These facts are of interest to historians as well as to linguists especially concerning the difference of outcomes according to whether the area was settled in first or second diaspora, the conditions of settlement, and the history of co-habitation of settlers with people speaking another language, especially English. See in particular

[5] See *inter alia* Queffélec (2008) for a historical perspective reaching from colonial French to the post-colonial African French of today in a very clear presentation of the major trends. He sorts the countries into two groups. The first group includes the majority of "Francophone" African countries, e.g. Niger, Centrafrique or Rwanda, the second one consists only of Senegal, Cameroon, Congo, Ivory Coast and Gabon.

several papers by Morin, among which (2002) with an important bibliography, but unfortunately only concerning pronunciation; Chaudenson (2003) for general considerations – *passim*, and more particularly (2003: 145-155); and articles in Valdman / Auger / Piston-Hatlen (2005) which all also give historical elements for North American situations.

1.3 A tentative typology of studies in variational syntax of French

Four major types of syntactical studies seem to emerge from diatopic corpora studies:

- The description of a whole area (see e.g. Seutin 1975 on rural French in Quebec, close to dialectology and seeking pre-defined grammatical phenomena; or different articles in Valdman / Auger / Piston-Hatlen 2005);
- Studies of a specific phenomenon in a specific area (e.g. Arrighi 2005 on prepositions in New Brunswick, or different articles in Lefebvre (ed.) 1982 for Montreal, on interrogatives, dislocations, *que*, clitics, WH-words, adverbial use of adjectives…);
- The comparison of a syntactic phenomenon in different points of a given area (e.g. Neumann-Holzschuh / Wiesmath 2006 for Acadian French in Canada and Louisiana on non-finite verbal forms);
- The comparison / contrast of a grammatical phenomenon in different areas, historically related or not (e.g. Vinet 2001 contrasting Quebec French vs. Swiss French, on interrogatives or the use of the pronoun *ça* – but not relying particularly on corpora):

(1) Ta mère *est-tu* là? (Quebec, Vinet 2001: 27)
Is your mother here?[6]

(2) Les frites, on *ça* prépare en tout *ça* mettant dans une friteuse (Switzerland, Vinet 2001: 117)
Chips, you prepare them by putting all that in a deep fryer

[6] For our examples, it would have been better to give non-standard equivalents in English. However, due to the difficulty of translating ordinary, vernacular and non-standard data, I just propose approximative glosses (more or less standardised). It is clear that this way of translating constitutes a problem for how the data is considered (we could say one of the first problems: see further considerations concerning the description of non-standard data, particularly in foot notes), but translating more precisely would require too many explanations which would be superfluous for many readers as well as for the purpose of this paper. My transcriptions always reproduce those of the original authors, for orthography as well as for punctuation. From now on, the grammatical facts under study will be italicised in the examples.

What to my knowledge has never or seldom been done, however, is the study of a linguistic phenomenon throughout the whole of the "French-speaking world". Furthermore, there is a fifth type of study which had seemed even more neglected until now: the study of a same grammatical phenomenon in all types of ordinary French, geographically *peripheral* as well as socially *marginal*, which would allow us to consider the variational scope of the whole of the French language.[7]

Here are some examples of these grammatical questions which could thus be considered and reinterpreted in taking non-standard data into consideration: presence / absence of the negative particule *ne*, presence / absence of *que* introducing a complement clause, as in (3), nature of the *que* as in (4), dislocation and word order as in (5), indirect interrogatives as in (6) and (7):

(3) je crois pas les Français *ils le font* (Parisian suburb, Gadet corpus)
I don't think French do such a thing

(4) donc après j'ai travaillé dans un magasin Hechter Newmann + *que* ça me plaisait pas du tout (Clermond-Ferrand, CRFP, PRI-003)
afterwards, I worked in a Hechter Newmann shop and I didn't like it at all

(5) tu sais / le chien de la voisine / mort / ils l'ont trouvé (Parisian suburb, Gadet corpus)
you know / the neighbour's dog / dead / they found him

(6) On s'inquiète un peu *quoi-ce que* les gens fait (New Brunswick, Arrighi corpus 2005: 173, text 13, t 2)
we are a little worried about what people do

(7) elle connait *elle veut quoi* (La Réunion, Ledegen 2007)
she knows what she wants

[7] Linguists used to work with non-standard phenomena know how problematic such a proposition is: for example, an "ideology-of-the-standard" perspective could lead to the interpretation of all syntactic phenomena as being conceived with respect to their distance from standard grammatical categories (different types of gap theories). For critics of this stereotyped position, see *inter alia* Ploog (2002) on some phenomena of Ivorian African French, whose instability is particularly accurate to address this type of questioning; and Cappeau / Gadet (in press) on apparently only technical choices concerning corpora of spoken French. For the numerous ties between vernacular and historical facts, see Brunot / Bruneau (1949).

Many more questions to do with variation could be formulated: when a locus of syntactic variation is identified, the question has to be asked as to its possible diatopic or diastratic interpretations. It appears that variation data do not occur in all parts of the grammar of French, and the most concerned areas are the following: pronouns, prepositions and verbs (tense, modality, aspect), negation, interrogation, relatives, and word order (see Gadet in press); a list which it is interesting to compare with that of Kortmann / Szmrecsanyi (2004) for English.

1.4 A detailed example: auxiliaries *avoir* and *être*

We shall now consider one of the more frequently described phenomena concerning tenses and conjugations. Different descriptions of ordinary French attest to a tendency to build compound tenses with the auxiliary *avoir* instead of the two-auxiliary system of standard French (*être* and *avoir*), as in (8) to (10):

(8) ils ont bati un tas de maisons depuis moi *je m'ai marié* (Louisiana, Stäbler 1995: 49)
they've built a lot of houses since I married

(9) *j'ai resté* soixante-dix jours au lit (Corpaix corpus, quoted in Blanche-Benveniste 1977: 105)
I stayed in bed for seventy days

(10) j'ai venu, j'ai monté, je m'ai fait mal, j'ai tombé, je m'ai acheté un costume, j'm'ai foutu la gueule en bas (Bauche 1920: 105)
I came, I went up, I hurt myself, I fell down, I bought me a suit, I fell arse-over-tit

(11) dans un sens quand elle *a* parti, ça m'a fait plus de peine que quand mon père *est* parti (Montreal, Sankoff / Cedergren corpus 1971, quoted in Sankoff / Thibault 1977: 94)
in a sense, when she left I was more worried than when my father left

The result depends on the type of verbs (transitive, intransitive or pronominal). However, in order to see whether we are dealing with issues of language structure or issues of language contact, three types of argument can be considered:

a) The phenomenon is also attested elsewhere, and foremost in ordinary hexagonal French (see (9), and the older (10)), and, as Blanche-Ben-

veniste (1977) writes, "it appears that the syntactic variations noticed in Montreal are also found in our corpora from Aix and Marseilles" (100, my translation of the original French);

b) examples like (11) show much variability, even in the same communicative situation and from the same speaker;

c) corpora fall into two groups from the point of view of auxiliaries:

1) Corpora attesting a categorical usage of *avoir* (or nearly categorical): for example New Brunswick (Péronnet corpus 1985 – see Péronnet 1989), Louisiana (Stäbler 1995), Prince Edward Island (King corpus 1987). There appear to be very few exceptions to this levelling, and one can wonder if they have to do with the same verbs everywhere: King / Nadasdi (2005), comparing different varieties of Acadian French, indicate *naître* and *mourir* as exceptions – *to be born* and *to die*, and it is also the case in Louisiana:

(12) l'enfant est né sur la berge (Stäbler 1995: 112)[8]
the child was born on the riverbank

(13) il avait soixante-dix ans ou plus quand il est mort (ibid.: 126)
he was seventy or more when he died

2) Corpora showing more variability or instability: for example, Montreal (Montreal, Sankhoff / Cedergren corpus 1971 – see example (10)), Vermont (Russo / Roberts 1999), Western Canada (Manitoba, Hallion 2000), Ottawa-Hull (Poplack 1989).

At first sight, it is not easy to determine what would be the ecological differences between the first and the second group of corpora. Since intensity of contact with English does not seem to be a sufficient factor, is there another hidden principle? Or is it merely a structural factor concerning French, as the simplification to only one auxiliary is not unknown in places where there is no contact or where contact is not dominant; and the phenomenon is not recent either (Hallion 2000: 360 for an enumeration of North American examples). The Swiss grammarian Henri Frei already signaled an opposition in 1929, reproduced here in (14) which, according to him, only takes place with under-

[8] Louise Péronnet (1989) indicates that she never finds *il est né*, but instead *il a venu au monde* (literally: he came into the world).

educated lower class speakers ("dans le bas peuple, complètement inculte", Frei 1929: 86) – it is well documented that it is not the case: see Blanche-Benveniste (1977) for a corpus of French children:

(14) il est mort *vs.* il a mouru (or *mouri* in New Brunswick, *cf.* Péronnet 1989)
he is dead vs. *he died*

1.5 Conclusion to Section 1

A third hypothesis for explaining differences between corpora emanating from the same place concerns their inherent qualities, or the way they were gathered. This very serious question will not be considered further here: what are the effects, on the inherent properties of a corpus, of the way it was gathered? With what possible grammatical consequences? Therefore, what are the biases of the many different modalities of data collection? (see also Vincent 2008).

More corpora are, then, still badly needed and a further diversification of the circumstances of data collection appears to be necessary: they need to be collected on a more systematic basis, diversifying data in terms of, for example, the following:

- Places of collection, taking into consideration the local history and ecology;
- Types of speakers: what are their fundamental differentiating characteristics? Certainly not the same everywhere, and certainly not (or not only) according to socio-demographic criteria (see *inter alia* Mendoza-Denton 2002 for the dangers of working with essentialist identities);
- Activities and the domain they belong to (e.g. talking with family and friends *vs.* talking in work situations);
- Different situations, taking into consideration more than one unidimensional axis like formal / informal, proximity / distance, planned / unplanned, public / private ...;
- Different discursive genres and interactions, including spoken / written;
- Emergent identities (beyond socio-demographic pre-categorisation).

Such a diversification is what the CIEL_F project is currently doing, documenting the same interactional pattern data from different areas of francophony around the world (www.ciel-f.net).

Our second section will then proceed with a presentation of some empirical data, keeping in mind more general challenging questions, concerning in particular the effects of a broadening of the data considered on the way of formulating syntactic problems.

2. Reevaluations of received knowledge about French grammar through corpus investigation

Some questions will be now considered through ordinary variation data. We will look at these in three sub-sections:

a) Are there functional organisations of forms which have not been seen (not "seeable"?) from a standard point of view or from the only perspective of hexagonal French?

b) Can some of the observed variation phenomena be "explained" as being mere interferences with the contact language?

c) How far are varieties isolated entities, or can variation facts be regarded as being on a continuum, and varieties as complex nodes of factors?

Corpora help query the received conception of *variety* as a specific combination of linguistic features or variants.

2.1 Some functional organisations of forms

2.1.1 Pronouns

The Quebecian grammarian Jean-Marcel Léard has identified five morphological features which differentiate what he calls the Quebecian variety of French from the hexagonal variety (Léard 1995, based on the Estrie corpus for Quebecian facts):

(15) *a(l)* instead of *elle* [*she*]

(16) [i] or [j] instead of *lui* [*him*]

(17) [i] instead of *elles* [*they feminine*]

(18) *-autres* added to all plural forms (*nous-autres*, [*we*], *eux-autres* [*they*])

(19) *eux-autres* instead of *elles* (no gender differentiation, as between *ils* et *elles*, for *they*)

Most of these changes are not completely unknown in European French: (15) is found, even though it is probably not the most frequent form (it is still said to be "popular"), (16) is quite frequent and even systematic in some ordinary spoken varieties, especially for some verbs like *dire* ("to say"), (17) is well documented for the past (e.g. Bauche 1920) and still exists, (18) seems to be local (especially in Belgium and the North of France), and (19) is the only form which could be said to be unknown or only heard sporadically in Europe.

What then could be said to be Quebecian here? For Léard, the crucial opposition lies between "autonomous" and "non-autonomous" positions. It is well known that the difference in grammatical functions for pronouns does not systematically lead to differences of forms in hexagonal French (see *nous* and *elle*, to be found in all positions, vs. *je / moi* or *il / lui*, which differ according to the syntactic position, subject or object). Quebec French, then, would appear to generalise and regularise an opposition which is only sporadic in standard French.

However, can this be said specifically "Quebecian"? In fact, all these forms are found elsewhere in North American varieties. See Fox (2005) for a corpus from Massachussetts:

(20) j'ai des tantes puis j'ai des cousines qui sont un peu agées mais ils aiment pas parler français avec moi [...] so c'est un peu difficile de parler avec eux-autres parce qu'ils aiment pas répondre (Fox 2005: 46)

I have aunts and cousins who are a little older but they don't like to talk French with me [...] *so it is a little difficult to talk with them because they don't like to answer*

2.1.2 Tense: the two futures

Grammatical tradition often gives us variant forms for which it is not easy to identify differences. This is the case for the two futures, periphrastic and simple, whose meanings are regularly said to be different, as shown by this SNCF (French national railway company) announcement, (21), contrasting with (21'):

(21) le train 7206 entrera en gare voie F (first announcement)

(21') le train 7206 va entrer en gare voie F (second announcement)[9]

Train number 7206 will arrive / is going to arrive (= is arriving) at platform F

[9] Thanks are due to Paul Cappeau, a regular train user, for this example.

(21') is uttered at the very moment of arrival of the train. However, can such a distinction be generalised? Some examples found in Canadian corpora, like (22) or (23), which of course are not very frequent, allow a different, pragmatic interpretation:

(22) j'me pose la question? Si *j'vas l'reprendre* ou si *j'le reprendrai pas* (Quebec City, Deshaies corpus (her transcription), Deshaies / Laforge 1981: 28)
I wonder whether I shall take it again or if I shall not

(23) un jour je me dis *je vais le faire* / le lendemain, je me dis non *je ferai pas* ça (Quebec, Estrie corpus, Léard 1995)
one day I think I am going to do it the day after I say I won't

There appears to be a link between simple future and negative sequence on one hand, periphrastic future and affirmative sequence on the other hand. Such a hypothesis has also been presented for European French far less frequently (Vet 1993). Does it mean, then, that the phenomenon is not yet very visible in European French (but could perhaps be shown through statistics)? Thanks to older attestations from other areas of North American French, we can extend the localisation: see, for example, Conwell / Juilland (1963) for Louisiana, Seutin (1975) for Ile-aux-Coudres in rural Quebec, LeBlanc (2007) for Ottawa-Hull (relying on the Poplack corpus).

2.1.3 Final infinitive

Other interesting examples concern localisation given to some forms, which can be reevaluated through a corpus-based approach. For example, a "final infinitive", exemplified in (24) is known by grammarians as non-standard, and only roughly evaluated (sometimes as "Belgian", sometimes as "popular", sometimes as "dialectal" ...). A localisation which is too narrow has to be revised, as a similar form (albeit superficially) is also found in Louisiana, as in at least one of the possible analyses of (25):

(24) il m'a toujours battue et c'était dur *pour moi l'aimer* (Paris, popular, Bauche 1920: 110)
he always beat me and it was difficult for me to love him

(25) il a fait faite un gros plancher *pour nous-autres danser dessus* (Louisiana, Stäbler 1995: 181)
he had a wooden floor built for us to dance on

As the structure is not that frequent (and there are probably syntactic differences between (24) and (25)), it is a little early to give a list of varieties of French in which such infinitives are attested (once more, we do not have sufficient corpus data at our disposal). However, we know for sure that it cannot be (or be only) a Belgian form as there is no attested historical relation between Belgium and Louisiana.

2.1.4 Prepositions

Corpora may enable us to oppose counterevidence to items of "received knowledge", like: "In French, the prepositions *sur*, *sous* and *dans* become *dessus*, *dessous* and *dedans* in stranded position." Thanks to some Acadian corpora, this assertion can be said to be at least partly false:

(26) elle mettait du câble pis elle amarrait ça *sus* (New Brunswick, Arrighi corpus 2005: 166, text 5, t 2)
she put wires and then she tied that on

(27) non à ce temps-là dire moi j'étais trop jeune tu sais on pensait point . de regarder qu'est-ce qu'on avait *dans* (New Brunswick, Arrighi corpus 2005: 455, text 23, t 2)
no at that time I was too young you know . we didn't think at looking at what we had inside

2.2 Deciding whether a feature is an interference or not

It is not always easy to decide whether a phenomenon is better understood as an interference with the other language or as an indigeneous process of reanalysis and restructuration in French, in particular as English and French share several typological characteristics (see Mougeon / Nadasdi / Rehner 2005, Gadet / Jones 2008, Neumann-Holzschuh 2009). Corpora are very useful for refining judgements, offering evidence for or against interference, allowing a revision of hasty "explanations" in terms of contact, which shows that an evaluation is a complex process.

The most explicit argumentation on this point was given by Mougeon / Nadasi / Rehner (2005) (relying on the corpus presented in Mougeon / Beniak 1991): their four-step argumentation relies on a process of two internal arguments and two external ones:

a) Is there an equivalent feature in the other language?

b) Can the innovative feature be attributed to internally-motivated processes (e.g. can it be understood as an analogical regularisation)?

   What happens in other varieties of the language (varieties being in intense vs. low contact, with the same language, with another language, or no contact at all)?

c) What is the distribution of the innovation within the speech community (what types of speakers are concerned)?

d) Obviously, such a methodology in its whole can only be applied to a specific community (as Mougeon / Nadasdi / Rehner 2005 do for Ontario), not on a large-scale picture as we try to do here.

2.2.1 Subjunctive

Let us first consider the case of the dwindling use of subjunctive in North American varieties as in (28) or (29), which, it is currently said, could be influenced by (if not a copy of) indifferentiation between indicative and subjunctive in English:

(28) faut que tu *prends* la gaffe pour la monter à bord (Newfoundland, Brasseur corpus)
you must take the boathook to take it on board

(29) pour pas que ça va dedans le riz blanc (Louisiana, Stäbler 1995: 54)
so that it doesn't go into the white rice

The decline in use of subjunctive is certainly open to an explanation in terms of simplification. Is such a phenomenon typical of all North American varieties, or only of obsolescent ones? In fact, neither explanation would appear to work as the phenomenon is also found in several (the majority? all?) peripherical and marginal varieties, as in example (30), which reproduces (3), and (31):

(30) je crois pas les Français ils le *font*
I don't think the French do it

(31) sans qu'on *sait* pourquoi (Switzerland, awkward writer, Frei 1929)
without us knowing why

The fourth step of distribution within the community does not apply here, and the conclusion would be more of internal re-organisation than of contact-induced change. However, contact with English can of course be seen as a reinforcing factor.

2.2.2 Simplification of conjugation

Some examples, where there is an English equivalent, could be interpreted as simplifications of the conjugation, as (31) and (32):

(32) les petits garcons *peut* venir (Louisiana, Stäbler 1995: 115)
small boys can come

(33) dans l'ancien temps . . une vieille fille là . . . tout le monde disait ça faisait pitié mais asteure je trouve *les femmes mariées fait* pitié (New Brunswick, Arrighi corpus 2005: 194, text 18, t 2)
formerly we used to say spinsters were a pitiful sight but now I feel married women are a pitiful sight

These forms constitute a regularisation in French conjugation, a phenomenon known as "default singular"[10] which appears to occur in several languages (and for Chambers 2009 could be a candidate for a universal feature). These forms are not usually described in marginal varieties (except perhaps in child language), and one has to wonder if they are mostly ways of speaking used by "semi-speakers". Therefore, it is impossible to interpret them without knowing who the utterers are and at what frequency and in which circumstances they were produced (which means they have to be contextualised as fully as possible).

2.2.3 Prepositions

The same argumentation can apply to prepositions: *avec pas* (*de*) (= *sans*) is often looked at as a calque of English *without*, as in (34):

(34) C'est dur *avec pas* de compas (Newfoundland, Brasseur corpus)
it is difficult without a compass

[10] This term is either a bad denomination or a typical example of the way non-standard phenomena are regularly considered in terms of distance from standard grammar categories (gap, difference). In the present case, it would certainly be better not to characterise this form as a singular, but rather a number neutralisation, or a restructuration of the modalities of number marking. We shall return to a discussion of these terms in part 3.

It could certainly be argued that *avec pas de* is a regularisation, therefore a type of simplification. However, we know it is also to be found in geographical areas where there is no contact, even if it is not very frequent: in the 503 pages of transcription of the Falkert corpus, it is only to be found three times:

(35) l'hiver le monde a resté avec pas de pain ni rien du tout (Magdalen Islands, Falkert corpus – see Falkert 2007: 242)
in winter everybody went without bread, nothing

(36) il est venu avec pas un sou (Paris, popular, Chaudenson / Mougeon / Beniak 1993)
he came without any money

See also the same type of argumentation in Dubois/Noetzel/Salmon (2005), concerning several local prepositions in Louisiana.

2.2.4 Preposition stranding

Preposition stranding is known to be possible in French with *après*, *pour*, *contre*, *avec*, *sans*; and with three prepositions whose form is modified, *dessus* (*sur*), *dessous* (*sous*) and *dedans* (*dans*). But never with *à* and *de*, says the doxa. Examples of stranding in hexagonal French are given in (37) and (38):

(37) y a un truc prévu pour (Paris)
there's a thing (especially) made for the job

(38) le soleil me vient contre (Marseilles, Valli 1995)
the sun is shining against me

According to King / Roberge (1990), there is at least one variety of French in the world which can also strand *à* and *de*, and which freely allows stranding in positions avoided in Standard French, like passivisation, interrogatives and relatives. It is Prince Edward Island French, an anglicized obsolescent variety (which does not imply that there are no more perfectly fluent speakers), where the following examples were found:

(39) qui-ce tu vas à Ottawa à travers de? (King / Roberge 1990, relying on King corpus 1987)
whom did you go to Ottawa through?

(40) qui-ce tu as fait le gâteau pour? (King / Roberge 1990, relying on King corpus 1987)
whom did you do the cake for?

(41) le ciment a été marché dedans (ibid.)
the cement was walked in

(42) la fille que j'ai donné la job à (ibid.)
the girl I gave the job to

(43) Robert a été beaucoup parlé de au meeting (ibid.)
Robert was spoken about a lot at the meeting

Data from Prince Edward Island French therefore stand as counterevidence to the two following received wisdom assertions in French grammar: "prepositions *à* and *de* are not strandable"; and "preposition stranding is much more constrained in French than in English". It remains to be demonstrated how this fact is to be understood, if we agree that stranded *à* and *de* became part of the French language transmitted to children in Prince Edward Island. Will it be considered as a direct structural borrowing from English, or as a reanalysis of the syntactic properties of the prepositions *à* and *de*? Still other questions concern the areas in North America where such structures will be found. As sporadic attestations, from Nova Scotia or from Western Canada seem to show that they are not totally ignored elsewhere (see King / Roberge 1990), even if less frequent than in Prince Edward Island, we can wonder what their extension really is. See (44), from the New-Brunswick Boudreau / Perrot corpus 2000:

(44) la seule solution que le monde peut penser à c'est la guerre
the only solution the world can think of is war

2.3 A glance at a received wisdom in sociolinguistics: the notion of *variety*. Discrete varieties or continuum of features?

We shall end this second section by revisiting the notion of *variety* whose drawbacks from a linguistic point of view are well known: see Gadet (2008). Gadet shows that this notion, although easily accountable for from a social-historical point of view or from the emic point of view of the agent, becomes difficult, if not impossible, to define from a linguistic point of view especially in the way a variety would differ from all other varieties of the same language.

We will thus take a glance at finely grained dialectal variations within what is generally looked at as the "same area",[11] which we will discuss through Acadian features. Neumann-Holzschuh / Wiesmath (2006) show, for what they call the "Acadian continuum", that different factors or phenomena will not always be congruent. Acadian areas all seem to be roughly conservative, but some features can be innovative (e.g. conditional after *si* and *si que* as in (45)). Neumann-Holzschuh / Wiesmath (2006) considered six features, regarded as typically Acadian, in five different Acadian areas: Louisiana, New Brunswick, Newfoundland, Nova Scotia East and Nova Scotia West:[12]

a) *quoi* in interrogatives (46)
b) morphology of person 6 (47)
c) morphology of person 1 (48)
d) imperfect subjunctive (49)
e) use of simple past tense ("passé simple" – e.g. (50))
f) *point* in negation (51)

(45) d'abord si qu'on se mettait ensemble icitte pis qu'on fait une demande / pis *si qu'on voudrait* avoir une clinique (Nova Scotia, Petras corpus, Petras 2008: 45)
if we all come together then we make a request then we would like to have a clinic

(46) *quoi* t'espères (Louisiana, Stäbler 1995: 151)
what are you waiting for?

(47) ils parl*ont* (New Brunswick – Standard French *ils parlent*)
they speak

(48) je parl*ons* (New Brunswick – Standard French *je parle*)
I speak

(49) faulait quelqu'un *restit* / FEEDer les vaches / et les poules (Nova Scotia, Petras corpus, Petras 2008: 97)
Somebody had to stay to feed cows and chicken

[11] Such a formulation supposes that it would at least be possible to specify what a "same area" is, which raises several problems, among which the question of point of view, etic or emic, from which agents and experts could diverge. See Gadet / Ludwig / Pfänder (2009) for discussions around the terms *geographical area* and *communicative area*.

[12] The diversity among the five main areas where French is still spoken in Nova Scotia is well known, related to differences in history and in local ecology (see Flikeid 1989).

(50) Pis je *furent* danser. Je *dansirent* bien. Après ça, je *venurent* BACK[13] (Nova Scotia, Hennemann corpus)
then I went dancing. I danced a lot. Then I came back

(51) j'avais *point* assez de lait (Nova Scotia, Petras 2008: 308)
I had not enough milk

They then propose the following table which they call "scale of Acadianicity":

	Lou	NB	NF	NSe	NSw
quoi in interrogative	+/0	+	+	+	+
ils … ont	+/0	+	+	+	+
je … ons	0	0/+	+	+	+
Imperfect subjunctive	0	+/0	+/0	+	+
Simple past tense	0	0	0	+	+
Negation in (*ne*) … *point*	0	0	0	0	+

Where + = present, 0 = absent or not documented, and +/0 = found but not systematic. Of course, these signs cannot avoid drastic simplification.

As can be seen, no two sets (to avoid the term "varieties") behave exactly in the same way. Now, what can be said to be more typically "Acadian"? The diversity of phenomena, for an area which is not that big, and which is generally considered as cohesive (albeit the diversity of histories in different areas) leads us to wonder what the term "area" means and what can be considered prototypical.

This diversity is also one of the lessons of corpora if they are diversified enough. How far do varieties exist, definable from a linguistic point of view? Can *variety* be considered a linguistic concept? Most studies rely on the hypothesis that linguistic varieties do exist, and do differ linguistically one from the other. Corpora lead to challenging questions about the definition of *varieties*, if such things do exist. Can they be defined in terms other than socio-historical, what Gadet (2008) argues to be impossible?

[13] The convention for transcribing bilingual corpora uses capitals for words from the other language. It would be necessary to comment the 3rd person plural forms with 1st person singular pronoun, which we will not do here.

2.4 Conclusion to Section 2

Why do language users as well as linguists (who are, after all, also users of language) seem to be so eager to allocate variants, for example according to varieties, in the emic point of view of agents as well as searching for "explanations" through correlations? Empirical research has to consider what these different varieties of French vernaculars (and others which have not been studied here) have in common and in how they differ from standard versions of the language. Thus some questions are better formulated thanks to large-scale corpora:

- Is it possible (and interesting) to distinguish between the specific parts taken by structural, socio-historical, or cognitive processes? Or shall we talk of complex multi-causation (for which the term "ecological" is particularly suitable)?
- To what extent is there a relation within a "same language" between grammatical areas in which variation is found, and areas sensitive to contact phenomena (between intra-linguistic variability and inter-linguistic variability)?

Now, if we want to avoid the dangers of "insularity" that are often underlined by critics of sociolinguistics (see Chambers 2000 quoted in footnote 3), it is time to consider possible generalisations and explicative principles behind our observations on vernaculars of French spoken outside France.

3. Concluding remarks: Vernacular resources

If we accept the idea that features encountered in peripheral and/or in vernacular varieties are not randomly distributed, and that it is only through the study of large collections of diversified data that they will be further investigated, some generalisations have to be thought of concerning the relationships between standard and vernacular features. As is well known, vernacular data are the most difficult to collect in ecological circumstances.

3.1 Functional and formal features of vernaculars (of French and in general)

The major differences between standards and vernaculars are traditionally mostly looked at as being above all a matter of socio-cultural status and thus of function and usage. However, this does not mean that ecological factors will

not have formal and structural repercussions. Some phenomena (e.g. *avoir* as generalised auxiliary, the negation in *pas*, or *que* as a general introducer of subordination) will be found in different varieties without historical or geographical relationships, and their ubiquity is unlikely to be due to sociolinguistic diffusion (in particular because they entered the language at a time where communication between distant locations was impossible or very rare).

Therefore, the question of a generalisation must explore other avenues. In how far is it possible to extend the principles guiding the phenomena described here and concerning French to different vernaculars of French (if not all), and perhaps to vernaculars in general (of all languages)? Is it possible to isolate, in the study of data from different vernaculars considered in parts 1 and 2, those features that are specific:

a) to all varieties of French (structural point of view),
b) to Romance languages or to Indo-European languages (typological point of view),
c) to vernacular varieties of different if not all languages (point of view of universals), and
d) to spoken ordinary varieties of all languages?[14] (sociolinguistic and interactional point of view).

In order to formulate such generalisations on the first point (which is the main topic of this article), I will not retain the term *francoversals* offered by Kortmann / Szmrecsanyi (2004), Szmrecsanyi / Kortmann (2009) alongside *angloversals*, because I feel it is:

a) more the product of a perspective of the quest for universals, whatever the level, than a balanced approach taking into consideration structural features and typology as well as socio-history and ecology of the languages, and
b) too much in the line of designing reflections on other languages upon reflections on English.

[14] Vernaculars are spoken varieties *par excellence*, and we can suppose that there are some consequences for a language to being mostly spoken and seldom or never written (thus spoken mainly in ordinary circumstances, tied to orality). They would therefore necessarily be influenced by the actual processes of mostly face-to-face interaction in speech production and reception, as a consequence of ordinary uses of the language by ordinary speakers in everyday circumstances. For the same type of questioning, but from a mostly syntactical point of view, see Deulofeu (1983) who discusses the first three points of view.

The history / ecology of the migration of Englishes worldwide is clearly not the same as that of French, which could have formal consequences. I would rather be in the line of reflection of Chaudenson, designed specifically for French and French-based creoles, concerning "français zéro" ("to determine empirical and statistical points, areas and limits of variation in French" – Chaudenson 2003: 185 [my translation]) and what can be learned from what he calls "marginal French" (belonging to different kinds – see also Chaudenson / Mougeon / Beniak 1993).

Another question is whether it is possible to go beyond considerations of frequencies and / or probabilities. According to Szmrecsanyi / Kortmann (2009), no feature appears in more than 80% of the Englishes: they can thus be said to be frequent, but not universals. Is it also the case for French? We would be keen to hypothesize that some features like *ne* deletion or *que* as a discursive resource will show very high frequency in all vernaculars (if not categoricity). But the question remains open if, when considered in different situations and in different linguistic networks, they can still be said to be "the same" (see for example sequences involving *que*).

3.2 Vernaculars as free *loci* for "natural processes"

A recurrent theme in folklinguistics is that vernaculars can be characterised using the terms *simple*, *simplified*, *simplicity* or *simplification*.[15] These terms are difficult to define, and Ferguson / DeBose (1977) take *simplification* as a technical term comprising two components: increase in morpho-phonemic *regularity* (i.e. *regularisation* or *analogy*) and regular correspondance between content and expression – i.e. *transparency*).

As there are no normative pressures or norm awareness (or a relative lack of overt norms) in vernaculars, all kinds of naturally regularising trends have a chance to come to the fore, making the system more transparent (and therefore easier for outsiders to learn it). In "ordinary" varieties, the forces able to counterbalance "natural processes" are fewer and less powerful (see Stein 1997). These processes are easy to conceive in terms of phonological considerations (in particular prosodic),[16] and it can be asked whether it is possible to

[15] See Gadet (1991, 2003) for French. *Simplicity* is a recurrent theme in many works to do with change, taken into account by different traditions under different perspectives, like neutralisation, regularisation, iconicity, analogy, markedness, optimality, etc.

[16] It would be more accurate to say that *naturalness* is often interpreted as having manifestations in body investissment (or "incorporation", in Bourdieu's terms – see Bourdieu 1991). Once more, the question can be raised as to what can be deemed "natural" in language, beyond the phonic level.

enlarge them to grammatical processes (e.g. Chambers 2003). Chaudenson calls these "auto-regulation" processes which can be observed in all languages as soon as the normative pressure is partly or completely relaxed (and that was already the idea of *advanced language*, "français avancé" in Frei 1929).

A recent volume (Filppula / Klemola / Paulasto (eds.) 2009) proposes to discuss the notion of "vernacular universals", and reworks the idea of comparing processes shared by pidgins, child language and vernaculars. For Chambers, what they have in common cannot be learned features or processes, and he proposes to regard them as "primitives". Szmrecsanyi / Kortmann (2009), through quantitative bases, are led to retain two components which concern 38% of variability: on one side the axis of morpho-syntactic *complexity / simplicity*, and on the other the degree of *analycity*. 38% is of course far from enough to comfort the idea of "vernacular universals".

For Trudgill (2009) the search for vernacular universals can only be a failure, and he suggests looking at non-standard features the other way round: the point would not be that much if "universals of vernaculars" can be distinguished in opposition to what standards universally impose upon languages, in their way of inhibiting change and distinguishing themselves from the corresponding vernaculars[17] (see Berruto 1983 and Kroch 1978). This point of view can be related to the concept of *distinction* by Bourdieu ("distinctiveness" – among others, see Bourdieu (1991) for an English translation). "Prestige speakers seem to mark themselves off as distinct from the common people." (Kroch 1978: 30): according to him, the point is thus not that underprivileged speakers particularly simplify their ways of speaking, but rather that privileged speakers do complicate, retain archaic or literary differentiations, or inhibit natural innovations. Kroch is mostly concerned by the phonic level, whereas Trudgill (2002) tries to enlarge the perspective to all linguistic levels including grammar. Once more, a restriction lies in the fact that looking at non-standard varieties in light of the standard can appear as supposing that non-standard features are some kinds of equivalents to standard features (see also footnotes 7 and 10).

[17] It could be just a side effect of the ideology of the standard that linguists generally seem to take for granted that vernaculars depart from standards and not the other way round. Nevertheless, the second proposition would be historically more accurate, as standardisation and standards, to do with the process of the constitution of modern nation-states, are recent processes and products in different European languages. A good example taken by Trudgill (2009) concerns multiple negation: for him it is not multiple negation which is a feature of vernaculars, but the absence of multiple negation which is a feature of non-vernaculars (i.e. standards).

3.3 Koineisation in particular circumstances

Trudgill (1986) summarised what happens when languages gain new territories in colonies. At first, they undergo *mixing* (either between dialects or languages from speakers coming from different places, or with languages spoken in the area). As time passes, *focusing* begins to take place for speakers within a new identity, and the variants begin to be an object of *reduction*. When a new dialect-formation takes place through *koineisation*, it comprises a process of *levelling* (loss of marked and / or minority variants), and one of *simplification:* even minority forms may survive if they are linguistically simpler. In 1986, Trudgill interpreted this process through the psycho-social notion of *accommodation*, where later (2002 and on) he conceives it in the more sociological term of *networks* (the outcomes of their characteristics, in particular what type of ties they exhibit – weak or strong, as well as closeknit or open).

Lodge (2004) has illustrated this model using French. Studying the history of French in France as a dissemination out of Paris, he shows the key role of contact between speakers coming to the capital from everywhere within and outside the country, thus arriving with different mother tongues: they arrived in Paris more particularly during the "industrial phase" of history, from the 18th century onwards, disseminating along the 19th and 20th centuries. And such processes are probably not going to slow down in times of globalisation.

Large cities thus appear to be particularly important places for linguistic processes: the place where different kinds of people meet, where there is contact. Urbanisation regularly gives birth to *koineisation*[18] ("linguistic melting pot") erasing the more untypical or differentiating features (regional and / or atypical features), on a phonological as well as grammatical level: it is therefore a particular type of *levelling*, taking the shape of *simplification* (regularisation, transparency, analogy).[19] Lodge thus follows Trudgill, for whom the wider the community, the more simplified the language. People have to communicate in spite of their different origins as they have different ways of speaking and, more importantly for syntax and discourse, have little basic knowledge and

[18] This is also the term used by Chaudenson (2003) to refer to the first stage of harmonisation between the new settlers' ways of speaking, having led ultimately either to colonial French in North America or to creoles in islands.

[19] It will be useful here to refer to linguistic reflections in terms of *markedness* (marked / unmarked). See Ludwig (2001) for a historical and conceptual synthesis on this topic.

few backgrounds in common. Therefore, the levelling processes take place even outside phonology, and concern also grammar and discourse facts, in more elaborated ways.

The same types of processes happen in all metropoles around the world, and such koineisation constitutes a general sociolinguistic process. The functional generalisation with formal consequences is therefore that languages have to become simpler in cities and especially in capital cities. Manessy (1992) made a demonstration of these types of processes in African cities, which could be synthetised as "what cities and urban areas do to languages when they are used mostly as lingua francas between people having different mother tongues and different ways of speaking".

3.4 Conclusion to part 3: Towards a new way of considering variation phenomena and vernaculars

French then appears, just like all world languages, to be a locus of variation, which can be fully recognised only through a generalisation of corpora as diversified as possible. At the grammatical level (and especially with syntax), this variation appears to be limited to some specific grammatical areas or points. Diatopic variation is known to be, especially for widely spoken languages like English or French, the widest locus of variation. However, one can still wonder how far other orders (like diastratic explorations), would allow to exhibit different linguistic facts belonging to other variational areas, or the same data organised in different ways with different constraints, or quite other variation data.

This is another story, which would have to take into consideration the outcomes of globalisation (see e.g. Parisian suburbs,[20] where it is necessary to take contact phenomena into consideration). We will no doubt find different emergent phenomena, not having been attested elsewhere (at least until now) in the Francophone world.

[20] Such will be the objective of a third big project I am involved in (see footnote 4 for the two others): an ANR/ESRC project (see Gadet/Gardner-Chloros 2009) concerning a study of change in English and in French, comparing London and Paris according to the outcomes of migrant contact languages spoken by young people of both English and French. For the linguistic effects of contact in a Canadian situation where English and French are implied, see Perrot (2005) for a study of the way of speaking French called *chiac* in Moncton, a city in New Brunswick.

Appendix: Corpora alluded to

I do not pretend here to offer an inventory of corpora of spoken French worldwide: the corpora listed here are only those alluded to in the text, a small part of the existing corpora of spoken French. In particular, I did not refer to three of the major corpora available on the web for French in France: PFC (http://www.projet-pfc.net/?accueil:intro)[21], CLAPI (http://clapi.univ-lyon2.fr), and CFPP2000 (http://ed268.univ-paris3.fr/syled/ressources/Corpus-Parole-Paris-PIII/index.html). Selection has been even still more drastic for North American corpora, as they are more numerous, albeit scarcely disposable on the web (the only Canadian exception being the recently born Sherbrooke corpus, http://pages.usherbrooke.ca/cfpq/index.php). A much larger list, for part a product of an ongoing inquiry, will be published soon on the DGLFLF website (see Cappeau / Gadet 2007 for a first presentation).

The dates given are only roughly indicative, some of them referring to the time of the fieldwork, others to the time of publication of the corpus, and still others (especially when the corpus is not open to public access) to one of the published papers whose analyses rely on examples from the corpus. I try to cite corpora under their most frequently given name, and some of the names are just to be taken as ways of reference.

Canada

Newfoundland: Brasseur 2001

Magdalen Islands: Falkert 2007

Nova Scotia: Flikeid 1989, Petras 2004 (see 2008), Hennemann 2005 (see 2007)

Prince Edward Island: King 1987 (see www.yorku.ca/rking)

New Brunswick: Péronnet 1985 (see 1989), Wiesmath 2000 (see 2006), Arrighi 2005, Boudreau / Perrot 2000 (see CRLA@umoncton.ca)

Quebec: Montreal 1971 (see Sankoff et al. 1976), Montreal 1984 (see Thibault / Vincent 1990, Vincent 2008), Montreal 1978 (Centre-sud, see Lefebvre (ed.) 1982), l'Ile-aux-Coudres (see Seutin 1975), Deshaies 1981 (Quebec City), Corpus de l'Estrie 1972 (Beauchemin-Martel, see Boisvert / Laurendeau 1988), Ottawa-Hull 1982 (see Poplack 1989, www.linguistics.uottawa.ca/faculty/poplack.html).

[21] All sites last visited 10 / 2010.

Ontario: Corpus Mougeon / Beniak (comparing minority and majority francophone areas – http://www.yorku.ca/rmougeon/frenchv.htm; see Mougeon / Beniak 1991)

Western Canada: Hallion 1995 (Manitoba, see Hallion 2000)

United States

New England: Russo / Roberts 1999 (Vermont), Fox 2005 (Massachussetts)

Louisiana: Conwell / Juilland 1963, Stäbler 1995

Others (Europe)

France: Corpus d'Orléans (ESLO, www.univ-orleans.fr/eslo/), Corpaix (GARS, see Blanche-Benveniste 2000); CRFP (= Corpus de référence du français parlé, see Equipe Delic 2004); Parisian popular French (see Gadet); corpus de Picardie (see Coveney 2002).

Belgium: VaLiBel (Francard, http://valibel.fltr.ucl.ac.be)

Older attestations:[22] Bauche 1920 (Paris, popular), Frei 1929 (Switzerland, mostly written).

[22] These texts can hardly be called "corpora" in the now received meaning of the term, as is also the case for all historical texts (e.g. Conwell / Juilland (1963) for Louisiana or Seutin (1975) for Quebec – who observed more than 100 inhabitants of a small island). Nevertheless, they have to be taken as such, as they constitute the only possible mode of documentation allowing to go back that far into the past of the spoken language.

References

Arrighi, Laurence (2005): Etude morphosyntaxique du français parlé en Acadie. Une approche de la variation et du changement linguistique en français. Ph.D Diss. Univ. d'Avignon.

Bauche, Henri (1920): Le langage populaire. Paris: Payot.

Berruto, Gaetano (1983): L'italiano popolare e la semplificazione linguistica. In: Vox Romanica 42: 38-79.

Blanche-Benveniste, Claire (1977): L'un chasse l'autre, le domaine des auxiliaires. In: Recherches sur le Français Parlé 1: 100-148.

Blanche-Benveniste, Claire (2000): Corpus de français parlé. In: Bilger, Mireille (ed.): Corpus – Méthodologie et applications linguistiques. Paris: Honoré Champion & Presses Universitaires de Perpignan, 15-25.

Boisvert, Lionel / Laurendeau, Paul (1988): Répertoire des corpus québécois de langue orale. In: Revue québécoise de linguistique appliquée 17, 2: 241-259.

Bourdieu, Pierre (1991): Language and symbolic power. Harvard: Harvard University Press.

Brasseur, Patrice (2001): Dictionnaire des régionalismes du français de Terre-Neuve. Tübingen: Niemeyer.

Brasseur, Patrice / Falkert, Anika (eds.) (2005): Français d'Amérique: approches morphosyntaxiques. Paris: L'Harmattan.

Brunot, Ferdinand / Bruneau, Charles (1949): Précis de grammaire historique de la langue française. Paris: Masson & Cie.

Cappeau, Paul / Gadet, Françoise (2007): Où en sont les corpus sur les français parlés? In: Revue Française de Linguistique Appliquée XII, 1: 129-133.

Cappeau, Paul / Gadet, Françoise (in press): Transcire, ponctuer, découper l'oral: bien plus que de simples choix techniques. In: Cahiers de linguistique.

Chambers, Jack (2000): Universal sources of the vernacular. In: Sociolinguistica 14: 11-15.

Chambers, Jack (2003): Sociolinguistic theory: linguistic variation and its social significance. 2. ed. Oxford: Blackwell.

Chambers, Jack (2009): Cognition and the linguistic continuum from vernacular to standard. In: Filppula / Klemola / Paulasto (eds.), 19-32.

Chaudenson, Robert (2003): La créolisation: théorie, applications, implications. Paris: L'Harmattan.

Chaudenson, Robert / Mougeon, Raymond / Beniak, Edouard (1993): Vers une approche panlectale de la variation du français. Paris: Didier-Erudition.

Cheshire, Jenny (ed.) (1991): English around the world: sociolinguistic perspectives. Cambridge: Cambridge University Press.

Cheshire, Jenny / Stein, Dieter (eds.) (1997): Taming the vernacular. From dialect to written standard language. London / New York: Longman.

Conwell, Marilyn / Juilland, Alphonse (1963): Louisiana French Grammar. The Hague: Mouton & Co.

Coveney, Aidan (2002): Variability in spoken French. A sociolinguistic study of interrogation and negation. 2nd ed. with a postface. Exeter: Elm Bank Publications.

Deshaies, Denise / Laforge, Eve (1981): Le futur simple et le futur proche dans le français parlé de la ville de Québec. In: Langues et linguistique 7: 23-37.

Deulofeu, José (1983): L'étude des langues parlées et la typologie des langues. In: Recherches sur le Français Parlé 5: 103-123.

Dubois, Sylvie / Noetzel, Sybille / Salmon, Carole (2005): Les innovations en français cadien: interférences ou changements motivés de façon interne au système? In: Brasseur / Falkert (eds.), 27-38.

Equipe Delic (2004): Présentation du corpus de référence du français parlé. In: Recherches sur le français parlé 18: 11-42.

Erfurt, Jürgen (2008): Le français du XXe siècle. Variétés linguistiques et processus de standardisation. In: Erfurt, Jürgen / Budach, Gabriele (eds.): Standardisation et déstandardisation. Frankfurt a.M.: Peter Lang, 13-34.

Falkert, Anika (2007): Le français acadien des Iles de la Madeleine. Unpubl. Ph.D. Diss., Univ. of Avignon and Regensburg.

Ferguson, Charles / DeBose, Charles (1977): Simplified registers, broken languages and pidginization. In: Valdman, Albert (ed.): Pidgin and Creole linguistics. Bloomington: Indiana University Press, 99-125.

Filppula, Markku / Klemola, Juhani / Paulasto, Heli (eds.) (2009): Vernacular universals and language contacts. Evidence from varieties of English and beyond. New York / London: Routledge.

Flikeid, Karen (1989): Moitié anglais, moitié français: emprunts et alternance de langues dans les communautés acadiennes de la Nouvelle-Ecosse. In: Laforge, Lorne / Péronnet, Louise (eds.): Revue québécoise de linguistique théorique et appliquée 8-2. Québec: Association québécoise de linguistique, 177-228.

Fox, Cynthia (2005): La variation syntaxique dans le français de Woonsocket: esquisse d'une grammaire du franco-américain. In: Brasseur / Falkert (eds.), 39-48.

Le français en Afrique. Internet: http://www.unice.fr/ILF/ofcaf (last visited: 08/2010).

Frei, Henri (1929): La grammaire des fautes. Genève: Republications Slatkine.

Gadet, Françoise (1991): Simple, le français populaire? In: LINX 25: 63-78.

Gadet, Françoise (2003): La relative française, difficile et complexe. In: Kriegel, Sibylle (ed.): Grammaticalisation et réanalyse. Approches de la variation créole et française. Paris: Editions du CNRS, 251-268.

Gadet, Françoise (2008): Les français 'marginaux' dans une perspective dialinguistique. In: Baronian, Luc / Martineau, France (eds.): Le français d'un continent à l'autre. Québec: Presses de l'Université Laval, 171-191.

Gadet, Françoise (in press): Sociolinguiste dans une grammaire: la variation dans une grammaire du français. Actes du 25e CILPR. Presses de l'Université d'Innsbruck.

Gadet, Françoise / Gardner-Chloros, Penelope (2009): Multi-cultural London English and multi-cultural Parisian French. ANR-ESRC unpubl. Project 2010-2013. London / Paris.

Gadet, Françoise / Jones, Mari (2008): Variation, contact and convergence in French spoken outside France. In: Journal of Language Contact, 238-248. Internet: http://www.jlc-journal.org (last visited: 08/2010).

Gadet, Françoise / Ludwig, Ralph / Pfänder, Stefan (2009): Francophonie et typologie des situations. In: Cahiers de linguistique 34, 1: 143-162.

Gülich, Elisabeth (1970): Makrosyntax der Gliederungssignale im gesprochenen Französisch. München: Fink.

Hallion, Sandrine (2000): Étude du français parlé au Manitoba. Unpubl. Ph.D. Diss., Univ. de Provence.

Hennemann, Julia (2007): Remarques à propos du système prépositionnel de l'acadien en Nouvelle-Ecosse. In: LINX 57: 79-90.

King, Ruth / Nadasdi, Terry (2005): Deux auxiliaires qui voulaient mourir en français acadien. In: Brasseur / Falkert (eds.), 103-111.

King, Ruth / Roberge, Yves (1990): Preposition stranding in Prince Edward Island French. In: Probus 2: 351-369.

Kortmann, Berndt / Szmrecsanyi, Benedikt (2004): Global synopsis. Morphological and syntactic variation in English. In: Kortmann et al. (eds.), 1142-1202.

Kortmann, Berndt et al. (eds.) (2004): A handbook of varieties of English. Berlin / New York: de Gruyter.

Kroch, Anthony (1978): Toward a theory of social dialect variation. In: Language in Society 7, 1: 17-36.

Léard, Jean-Marcel (1995): La grammaire québécoise d'aujourd'hui. Comprendre les québécismes. Montréal: Guérin universitaire.

LeBlanc, Carmen (2007): Le futur périphrastique dans le français parlé: une question d'habitude. Ph.D. Diss., Univ. Ottawa.

Ledegen, Gudrun (2007): L'interrogative indirecte in situ à la Réunion: 'elle connaît elle veut quoi'. In: Le français parlé au 21e siècle: normes et variations géographiques et sociales. Actes du colloque à l'Université d'Oxford. Paris: L'Harmattan, 177-200.

Lefèbvre, Claire (ed.) (1982): La syntaxe comparée du français standard et populaire. Montréal / Québec: Office de la langue française, 2 volumes.

Lodge, R. Anthony (2004): A sociolinguistic history of Parisian French. Cambridge: Cambridge University Press.

Ludwig, Ralph (2001): Markiertheit. In: Haspelmath, Martin / König, Ekkehard / Oesterreicher, Wulf / Raible, Wolfgang (eds.): Language typology and language universals. An international Handbook = Sprachtypologie und sprachliche Universalien. Ein internationales Handbuch. (= Handbooks of linguistics and communication science – Handbücher zur Sprach- und Kommunikationswissenschaft 20). Berlin / New York: de Gruyter.

Manessy, Gabriel (1992): Modes de structuration des parlers urbains. In: Des langues et des villes. Paris: Didier-Erudition, 7-27.

Mendoza-Denton, Norma (2002): Language and identity. In: Chambers, Jack K. et al. (eds.): The handbook of language variation and change. Oxford: Blackwell Publishing, 475-499.

Morin, Yves-Charles (2002): Les premiers immigrants et la prononciation du français au Québec. In: Revue Québécoise de Linguistique 31, 1: 39-78.

Mougeon, Raymond / Beniak, Edouard (1991): Linguistic consequences of language contact and restriction: The case of French in Ontario. Oxford: Oxford University Press.

Mougeon, Raymond / Nadasdi, Terry / Rehner, Katherine (2005): Contact-induced linguistic innovations on the continuum of language use: the case of French in Ontario. In: Bilingualism: Language and Cognition 8, 2: 99-115.

Neumann-Holzschuh, Ingrid (2009): Contact-induced structural change in Acadian and Louisiana French, Mechanisms and motivations. In: Langage & Société 129: 47-68.

Neumann-Holzschuh, Ingrid / Wiesmath, Raphaele (2006): Les parlers acadiens: un continuum discontinu. In: Revue Canadienne de Linguistique Appliquée 9, 2: 233-249.

Péronnet, Louise (1989): Le parler acadien du sud-est du Nouveau-Brunswick. Eléments grammaticaux et lexicaux. New York / Bern: Peter Lang.

Perrot, Marie-Eve (2005): Le chiac de Moncton: description synchronique et tendances évolutives. In: Valdman / Auger / Piston-Hatlen, 307-326.

Petras, Cristina Anca (2008): Les emprunts et la dynamique linguistique. Unpubl. Ph.D. Thesis, Univ. Iasu.

Ploog, Katja (2002): L'approche syntaxique des dynamiques langagières: non-standard et variation. In: Cahiers de grammaire 27: 77-96.

Poplack, Shanna (1989): The care and handling of a mega-corpus. In: Fasold, Roger / Schiffrin, Deborah (eds.): Language change and variation. Amsterdam: Benjamins, 411-451.

Queffelec, Ambroise (2008): L'évolution du français en Afrique noire, pistes de recherche. In: Holter, Karen / Skattum, Ingse (eds.): La francophonie aujourd'hui, réflexions critiques. Paris: L'Harmattan, 63-76.

RFLA (= Revue française de linguistique appliquée) (1996): Corpus: de leur constitution à leur exploitation. Vol. 1-2. Internet: http://www.rfla-journal.org (last visited: 08/2010).

Russo, Marijke / Roberts, Julie (1999): Linguistic change in endangered dialects: The case of alternation between avoir and être in Vermont French. In: Language Variation and Change 11: 67-85.

Sankoff, David et al. (1976): Méthodes d'échantillonnage et utilisation de l'ordinateur dans l'étude de la variation grammaticale. In: Cahier de linguistique 6: 85-125.

Sankoff, Gillian / Thibault, Pierrette (1977): L'alternance entre les auxiliaires avoir et être en français parlé à Montréal. In: Langue française 34: 81-108.

Seutin, Emile (1975): Description grammaticale du parler de l'Ile-aux-Coudres. Montréal: Presses de l'Université de Montréal.

Stäbler, Cynthia (1995): La vie dans le temps et asteur. Ein Korpus von Gesprächen mit Cadiens in Louisiana. Tübingen: Narr.

Stein, Dieter (1997): Syntax and varieties. In: Cheshire / Stein (eds.), 35-50.

Szmrecsanyi, Benedikt / Kortmann, Bernd (2009): Vernacular universals and angloversals in a typological perspective. In: Filppula / Klemola / Paulasto (eds.), 33-53.

Thibault, Pierrette / Vincent, Diane (1990): Un corpus de français parlé: Montréal 84, historique, méthode et perspectives de recherche. Québec: Presses de l'Université Laval.

Trudgill, Peter (1986): Dialects in contact. Oxford: Basil Blackwell.

Trudgill, Peter (2002): Linguistic and social typology. In: Chambers, Jack / Schilling-Estes, Natalie / Trudgill, Peter (eds.): Handbook of Linguistic Variation and Change. London: Routledge, 707-728.

Trudgill, Peter (2009): Vernacular universals and the sociolinguistic typology of English dialects. In: Filppula / Klemola / Paulasto (eds.), 304-322.

Valdman, Albert / Auger, Julie / Piston-Hatlen, Deborah (2005): Le français en Amérique du Nord, état présent. Québec: Presses de l'Université Laval.

Valli, André (1995): Notes sur la variation linguistique en français. In: Recherches sur le Français Parlé 13: 91-109.

Vet, Co (1993): Conditions d'emploi et interprétation des temps futurs en français. In: Verbum 4: 71-84.

Vincent, Diane (2008): Corpus, banques de données, collections d'exemples. Réflexions et expériences. In: Cahiers de Linguistique 33, 2: 81-96.

Vinet, Marie-Thérèse (2001): D'un français à l'autre: la syntaxe de la micro-variation. Montréal: Fides.

Wiesmath, Raphaèle (2006): Le français acadien. Analyse syntaxique d'un corpus oral recueilli au Nouveau-Brunswick/Canada. Paris: L'Harmattan.

II. Korpusgestützte Grammatikforschung / Corpus-based Grammar Research

CHRISTA DÜRSCHEID / STEPHAN ELSPASS / ARNE ZIEGLER

Grammatische Variabilität im Gebrauchsstandard: das Projekt „Variantengrammatik des Standarddeutschen“

Abstract

Im Beitrag werden die Methodologie und die Ziele eines Projekts vorgestellt, das anstrebt, auf der Grundlage eines breiten Korpus von Texten aus allen Ländern und Regionen des zusammenhängenden deutschen Sprachgebiets die Variation in der Grammatik der geschriebenen deutschen Standardsprache zu erfassen, in einem Handbuch zu dokumentieren und damit eine Basis sowohl für Grammatiken als auch für weitergehende grammatische Untersuchungen zu schaffen. Nach einleitenden Bemerkungen zum Projekt und zu der Frage, in welcher Relation die geplante „Variantengrammatik des Standarddeutschen“ zum bereits erhältlichen „Variantenwörterbuch des Deutschen“ von Ammon et al. (2004) steht, folgt ein Forschungsüberblick zur grammatischen Variation in der Standardsprache. Dann werden Beispiele für grammatische Variabilität in verschiedenen Phänomenbereichen gegeben, und es wird anhand von zwei Fallbeispielen gezeigt, wie eine grammatische Beschreibung dieser Phänomene aussehen kann. Um Angaben zur arealen Distribution grammatischer Varianten machen zu können, wird den Analysen ein Korpus zugrunde gelegt, das sich auf den *geschriebenen* Standard beschränkt und darunter den Sprachgebrauch in der Presse fasst. Das Korpus, das als Basis für die Erstellung der geplanten Variantengrammatik dient, wird im Beitrag kurz vorgestellt, außerdem wird erläutert, welche Zielsetzungen mit einer solchen Grammatik verbunden sind.

1. Zum geplanten Projekt

Die areale Variation in der Grammatik der deutschen Standardsprache hat in der Grammatikographie – trotz einer immer stärkeren Orientierung an Textkorpora – bislang kaum Beachtung gefunden. Dieser Typus von Variation ist aber nicht etwas außerhalb der Standardsprache Anzusiedelndes, sondern Realität innerhalb der deutschen Standardsprache. Entsprechende Varianten können daher auch nicht pauschal als sozial markiert angesehen oder einer „Grammatik der gesprochenen Sprache“ zugeschlagen werden. Areale Unterschiede im Gebrauchsstandard umfassen u.a. die Wortstrukturierung, die Phrasenstruktur und die Rektion. So finden sich in deutschsprachigen Zeitungen der Gegenwart – je nach Land oder Region – morphologische und morphosyntaktische Varianten wie *Zugmitte / Zugsmitte*; *Störenfried / Störefried*;

die Parks / die Pärke / die Parke oder *Das Wetter ändert / Das Wetter ändert sich.* Auch syntaktische Varianten wie *Bereits sind die Ämter besetzt / Die Ämter sind bereits besetzt* oder *Gut, gibt es Bauern / Gut, dass es Bauern gibt*, welche die Regularitäten der Vorfeldbesetzung und die Verbstellung im Nebensatz betreffen, zählen dazu. Das Nichterkennen bzw. Nichtanerkennen dieser Variation führt in verschiedenen Bereichen zu Problemen: in der Grammatikschreibung zum vorschnellen Ausschluss von grammatischen Varianten (und damit auch von möglichen Grammatikalisierungswegen), die nicht im Blickfeld der Grammatikographen liegen, und in sprachnormvermittelnden Instanzen wie Schule und Universität zu ungerechtfertigten Markierungen von regionalen oder nationalen Varianten des Standarddeutschen als 'Fehler'.

Um hier einen Kontrapunkt zu setzen, sollen in dem geplanten Projekt „Variantengrammatik des Standarddeutschen" Phänomene der eben genannten Art erfasst und kodifiziert werden. Dabei handelt es sich um eine Zusammenarbeit von Sprachwissenschaftlern aus Deutschland (Stephan Elspaß), Österreich (Arne Ziegler) und der Schweiz (Christa Dürscheid). Angestrebt wird die systematische Aufbereitung standardsprachlicher Varianten, die sich im deutschen Sprachgebiet auf überregionaler Ebene finden, die aber nicht dialektal sind und sich nicht dem lexikalischen Bereich zuordnen lassen. Ein wichtiges Referenzwerk für das geplante Projekt ist das Variantenwörterbuch (Ammon et al. 2004), das die Standardvarietäten des Deutschen in Österreich, der Schweiz und Deutschland sowie Liechtenstein, Südtirol, Ostbelgien und Luxemburg berücksichtigt, seinen Schwerpunkt aber auf die lexikalische Ebene legt. Zwar findet sich im Variantenwörterbuch vereinzelt auch die Beschreibung von Phänomenen, die auf der Ebene der Grammatik zu verorten sind (z.B. zum *am*-Infinitiv in Sequenzen wie *Er ist am schlafen*). Solche Einträge sind wertvoll, sie lassen jedoch keine umfassende Rekonstruktion einer Variantengrammatik zu. Außerdem liegt dem Variantenwörterbuch keine systematische Datenerhebung zugrunde; die von den beteiligten Forschergruppen vorgebrachten Einschätzungen zum Vorkommen eines Phänomens wurden nicht an einem großen Textkorpus überprüft. Dies wird dagegen in dem geplanten Projekt, das auf einem breiten Korpus von standardsprachlichen Texten basieren wird (vgl. Kap. 4), der Fall sein.

Ziel des Projekts ist also die Erstellung eines Korpus, das nach Abschluss der Projektarbeit öffentlich zugänglich gemacht werden soll. Ein weiteres wichtiges Ziel ist die Konzeption und Publikation einer Variantengrammatik in Form eines Handbuchs. Diese Grammatik soll, analog zum Zweifelsfälle-Du-

den (Duden 2007), alphabetisch geordnete Artikel mit drei Typen von Lemmata enthalten: Erläuterungen zu einzelnen Wortformen (z.B. zum Gebrauch von *trotzdem* als Konjunktion) und zu grammatischen Termini (z.B. zum Terminus 'Fugenelement') sowie umfassende Überblicksartikel (z.B. zu den Prinzipien der Reflexivierung in der deutschen Standardsprache, genauer: in den Standardvarietäten des Deutschen). Zu jedem Lemma werden Textbeispiele angeführt, korpusbasierte Angaben zur arealen Distribution gemacht und (auch für Laien nachvollziehbare) theorieübergreifende Erläuterungen gegeben. Dabei kann es durchaus sein, dass dialektale Sprachgebrauchsmuster als Erklärungshorizont für die Analyse einzelner Phänomene des Standarddeutschen dienen. Die Beschreibung dialektal-grammatischer Phänomene steht aber nicht im Fokus des Projekts, die wird bereits an anderer Stelle geleistet (vgl. Glaser 2003). Auch ist wichtig zu betonen, dass syntaktische Muster, die aus der zeitlichen Bedingtheit der gesprochenen Sprache (Synchronizität, Linearität, Flüchtigkeit) und den damit verbundenen Produktions- und Rezeptionseigenschaften resultieren, keinen Eingang in unsere Grammatik finden werden. Solche Muster werden in der Duden-Grammatik in einem separaten Kapitel zur gesprochenen Sprache ausführlich beschrieben. Dazu gehören u.a. Referenz-Aussage-Strukturen (z.B. *den ausweis den brauch ich dann auch noch*) und Apokoinu-Konstruktionen (z.B. *des is was furchtbares is des*, Beispiele aus Duden 2009: 1165-1244). Konstruktionen dieser Art unterliegen den spezifischen Produktionsbedingungen der „On line-Syntax" (vgl. Auer 2000). Und auch wenn sie im Geschriebenen vorkommen (z.B. in Texten konzeptioneller Mündlichkeit), gehören sie nicht zu dem, was unter Standardsprache verstanden wird.

Daran schließt sich unmittelbar eine Frage an: Welchen Begriff von 'Standardsprache' legen wir überhaupt zugrunde? Bekanntlich gibt es zahlreiche Definitionen, die hier nicht referiert werden können (vgl. zu einer instruktiven Übersicht Dovalil 2006: 59-63). Eine Definition ist nach Ammon (1995: 74) die, dass die (nationale oder areale) Standardvarietät die in den Schulen unterrichtete, kodifizierte Sprache ist, wobei Ammon an anderer Stelle (ebd.: 78) aber selbst betont, dass es teilweise unklar sei, welche Veröffentlichungen zum Sprachkodex gehören und welche nicht, und dass Lehrer „sich faktisch keineswegs immer streng am Sprachkodex" orientieren. Im Folgenden wollen wir den Blick von den Kodizes denn auch auf den Sprachgebrauch selbst richten, also auf die grammatischen Strukturen, die im geschriebenen Standard tatsächlich vorkommen. Ausgangspunkt ist damit das, was Ammon (ebd.: 88) als „Gebrauchsstandard" bezeichnet und „in semantischer Opposition zum Ter-

minus *kodifizierter Standard*" [Hervorh. i. Orig.] sieht. Ammon hält dazu fest: „Diese Kennzeichnung [...] impliziert, daß die fraglichen Varianten nicht durch den Sprachkodex im oben erläuterten Sinn als standardsprachlich ausgewiesen sind." (Ammon 1995: 88). Im Zentrum steht also nicht, was bereits in den Grammatiken erfasst ist (dazu braucht es kein neues Projekt), sondern das, was im Sprachgebrauch regelhaft vorkommt. Das kann sich mit den Einträgen in den Grammatikkodizes decken, es kann aber auch darüber hinaus gehen. So kann es durchaus sein, dass im Sprachgebrauch grammatische Strukturen verwendet werden, die in den Kodizes nicht oder nur abwehrend erfasst sind (z.B. *wegen* + Nominalgruppe im Dativ).

Wir folgen mit unserer Auffassung, den Sprachgebrauch selbst ins Zentrum zu stellen, Peter Eisenberg, der sich im Zusammenhang mit seiner Neubearbeitung des Zweifelsfälle-Dudens für die konsequente Orientierung am Sprachgebrauch ausspricht und betont, wie wichtig es sei, Aussagen zum Sprachgebrauch auf korpusbasierte Recherchen abzustützen (vgl. Eisenberg 2007). Als geeignete Basis für die empirische Fundierung von Feststellungen über den Standard sieht Eisenberg den Sprachgebrauch in der überregionalen Presse in Deutschland an. Er schreibt: „Süddeutsche Zeitung, Frankfurter Rundschau, Die Zeit, Frankfurter Allgemeine Zeitung und einige andere setzen in Deutschland das Maß" (ebd.: 217). Dem Benutzer seines Wörterbuchs solle ein „Angebot" gemacht werden, aus verschiedenen gebräuchlichen Varianten diejenige auszuwählen, mit der „er sich im geschriebenen Standard unauffällig bewegen kann" (ebd.: 226). Nun wird man sich aber auch in der Schweiz und in Österreich bzw. regional in Deutschland mit bestimmten Varianten, die in Deutschland bzw. in bestimmten Regionen in Deutschland auffällig sind, durchaus unauffällig im geschriebenen Standard bewegen können.[1] Da der Schwerpunkt im geplanten Projekt auf der diatopischen Variation liegt, weiten wir daher den Standardbegriff auf den Sprachgebrauch in der schweizerischen und österreichischen Presse aus. Des Weiteren betrachten wir die grammatische Variation innerhalb der betreffenden Länder und berücksichtigen daher vor allem den Sprachgebrauch der jeweiligen regionalen Presse. So kommen wir nicht in die Situation, entscheiden zu müssen – und diese Entscheidung wäre notwendigerweise in hohem Maße subjektiv –, welche Zeitungen das Maß setzen; wir wählen vielmehr Zeitungen, die es regional gibt und die regional eine bestimmte Reichweite haben.

[1] Eisenbergs Festlegung würde streng genommen bedeuten, dass der Sportteil der *Frankfurter Rundschau* standardsprachlicher ist als etwa die politischen Kommentare in der *Rhein-Zeitung*, der *Augsburger Allgemeinen* oder der *Vorarlberger Nachrichten*. Das erscheint uns nicht plausibel.

Diatopische Variation, das sei an dieser Stelle noch eigens betont, macht selbstverständlich nicht an den Landesgrenzen Halt. Wir gehen also nicht davon aus, dass wir in Zeitungstexten grammatische Varianten finden werden, die z.B. exklusiv an die deutsche, schweizerische oder österreichische Standardvarietät geknüpft sind. Ohnehin gilt nicht in allen Fällen, dass eine Standardvarietät im ganzen Land im Gebrauch ist. Dies sieht man an der Schweiz, wo nur 17 der 26 Kantone einsprachig (deutsch) sind. Aus diesen Gründen sind denn auch Bezeichnungen wie „plurinationale Variation" bzw. „Plurinationalität", die sich in den Arbeiten von Ammon (1995 und öfter) finden, problematisch. Geeigneter scheint uns das pluriareale Konzept, das in der sprachlichen Heterogenität des Deutschen ein typologisches, nicht an Staatsgrenzen gebundenes Merkmal sieht (vgl. etwa Scheuringer 1996, Reiffenstein 2001).

2. Grammatische Variabilität in der Grammatikschreibung und in der Forschung

Im Folgenden stellen wir Arbeiten vor, in denen die grammatische Variation des Standarddeutschen bereits Berücksichtigung findet. Dabei handelt es sich um grammatische Einzelstudien. Eine Überblicksdarstellung, die als Referenzwerk dienen könnte, gibt es nicht. Geordnet sind die Ausführungen daher nach den drei Hauptvarietäten des Standarddeutschen in der Schweiz, Österreich und Deutschland (eingedenk der Vorbehalte, die eine solch nationale Zuordnung mit sich bringt, siehe oben).

Eine Monographie, die eine systematische Beschreibung der Grammatik des Schweizer Standarddeutschen anstrebt, ist die vor vier Jahrzehnten erschienene Arbeit von Kaiser (1969 / 1970). Der Autor führt in diesem zweibändigen Werk eine Vielzahl von grammatischen Phänomenen an, wobei hier aber syntaktische Aspekte im engeren Sinne, etwa Regularitäten der Satzstrukturierung oder der Wortstellung, kaum Beachtung finden. Auch sind die Vorannahmen, die der Studie zugrunde liegen (und sich etwa in der Titelgebung „Die Besonderheiten der deutschen Schriftsprache in der Schweiz" widerspiegeln), nicht mehr aktuell. Ein grammatisches Referenzwerk zur Standardsprache in der Deutschschweiz ist dieses Buch somit nicht. Das gilt auch für die in einem Schweizer Verlag publizierte Grammatik „Richtiges Deutsch" von Heuer / Gallmann / Flückiger (2008), in der Einzelphänomene zur Grammatik im Schweizer Standarddeutsch erfasst sind, dies aber nur selektiv geschieht, da der Schwerpunkt auf der grammatischen Beschreibung des Gemeindeutschen

liegt. Weiter gibt es in der Forschungsliteratur eine kleinere Zahl von Studien zu Einzelaspekten. Diese stammen von Gelhaus (1972) zur Rektion bestimmter Präpositionen, von Rohrer (1973) zum Konjunktivgebrauch und von Stirnemann (1980) zur Syntax des im Schulunterricht verwendeten Deutsch. Vergleichbare Arbeiten aus jüngerer Zeit existieren kaum, von Ausnahmen wie Rüttimann (2002) – einer nicht veröffentlichten Lizenziatsarbeit, die sich mit Fragen der Kasusrektion bei Präpositionen befasst – und Dürscheid / Hefti (2006) abgesehen.

Zum österreichischen Standarddeutsch liegen ebenfalls einige Arbeiten vor, jedoch befassen sich diese in erster Linie mit lexikalischen Besonderheiten und nur selten mit grammatischen Phänomenen. Ein eindrucksvolles Bild dieser Situation vermittelt der im Jahre 2006 veröffentlichte Sammelband mit dem Titel „Zehn Jahre Forschung zum Österreichischen Deutsch: 1995-2005 – eine Bilanz" (Muhr / Sellner (Hg.) 2006). Unter den 17 Beiträgen findet sich lediglich ein Aufsatz, der sich mit einer grammatischen Besonderheit, nämlich dem Gebrauch des Genitivs, auseinandersetzt (Sellner 2006). Auch in der Online-Bibliographie zum „österreichischen Deutsch", die bis zum Jahr 2005 / 2006 reicht, erscheinen unter den 498 erfassten Arbeiten nur vier unter dem Stichwort „Grammatik".[2] Selbst die kurze Einführung „Österreichisches Deutsch" von Ebner (2008) behandelt das Thema „Grammatik" nur auf knapp drei von 48 Seiten. Eine neuere Arbeit (Zeman 2009) enthält ein 45-seitiges Kapitel zum Thema „Grammatische Merkmale des österreichischen Deutsch", darin finden sich aber neben einem kurzen Abschnitt zum Satzbau vor allem Erläuterungen zum Wortschatz und erstaunlicherweise auch zur Aussprache. In einem Beitrag von Muhr (1995) dagegen sind verschiedene grammatische Phänomene berücksichtigt – in erster Linie freilich derart, dass sie schlicht aufgelistet werden. Problematisch sind hier zudem die empirische Basis, die Methode sowie auch die grammatische Erklärung und Differenzierung der Daten. Zu den Artikeln, die grammatische Variation berücksichtigen, gehört auch Tatzreiter (1988), der sich vor allem morphologischen Aspekten widmet (etwa im Bereich Genus). Daneben sei ein Aufsatz von Stubkjær (1997) erwähnt, der sich ebenfalls mit einem morphologischen Problem, dem Präsensparadigma starker Verben, befasst. Ein Beitrag, der syntaktisch-strukturelle Probleme im österreichischen Deutsch untersucht, stammt von Patocka (1997). Jedoch hat Patocka nicht die geschriebene Standardsprache im Blick,

[2] Vgl. http://www-oedt.kfunigraz.ac.at/OEDTRADIO/content/05-Mat/blioedt2.html#Grammatik%20des%20OD (Stand: 12 / 2010).

sondern nimmt eine eher dialektologische Perspektive ein. Und schließlich ist auch Wiesinger (2006) zu nennen, der sich in seinem Beitrag zum österreichischen Amtsdeutsch u.a. der Syntax widmet.

Kommen wir schließlich zum deutschen Standarddeutsch: Dieses gilt in der Grammatikographie meist – wie selbstverständlich – als der 'unmarkierte' Fall. Dementsprechend gibt es wenige Arbeiten, die sich mit grammatischer Standardvarianz im Deutschen befassen (z.B. Elter 2005). Die bisherigen Markierungen in den Grammatiken heben meist Besonderheiten im Süden Deutschlands hervor, seltener werden Gebrauchsvarianten einzelner Regionen oberhalb der Main-Linie genannt. Früh hat schon Götz (1995) in ihrem Aufsatz „Regionale grammatische Varianten des Standarddeutschen" auf entsprechende Mängel in der Grammatikschreibung hingewiesen. Zwar würden in den Grammatiken einzelne Fälle von Variation erfasst, doch geschehe dies sehr unsystematisch, außerdem fänden sich z.T. ungenaue Angaben zur regionalen Verbreitung. Dass die Grammatik des deutschen Deutsch 'oberhalb' der Dialekte aber keineswegs einheitlich ist, ergibt sich schon aus der Arbeit von Henn-Memmesheimer (1986) zu syntaktischen „Nonstandardmustern", in der es u.a. um Syntagmen geht, die zwischen Standard und so genanntem „Substandard" oszillieren (z.B. *Stücker drei / vier*). Hinweise auf mögliche Standardvariation haben sich in den letzten Jahren schließlich aus den Karten des *Atlas zur deutschen Alltagssprache* (*AdA*) ergeben, die tatsächlich vorkommende grammatische Varianten und ihre Verbreitung in der – mehr oder weniger standardnahen – Alltagssprache dokumentieren. Doch trotz dieser Ansätze (und dazu zählen auch die Arbeiten zu den Unterschieden zwischen BRD- und DDR-Deutsch) kann festgestellt werden, dass die Existenz eines deutschen Standarddeutsch und seine interne grammatische Variation ein blinder Fleck in der Grammatikographie des Deutschen ist (vgl. dazu auch Elspaß 2010).

3. Phänomenbereiche

An dieser Stelle sollen, nur stichwortartig, einige Beispiele für grammatische Variation im Deutschen aufgelistet werden, um dem Leser einen Eindruck von der Vielfalt der Phänomene zu geben. Diese Phänomene sind bereits in den einschlägigen Grammatiken des Deutschen (siehe Literaturverzeichnis) erfasst, jedoch ohne verlässliche Angaben zu ihrer arealen Distribution. Das soll in der geplanten Variantengrammatik anders werden. Hier die Auswahl:

- Flexion: starke vs. schwache Verbform (*speisen / spies*), Variation in der Pluralbildung (*Pärke / Parks*) und im Genus (*das / der Abszess*);
- Wortbildung: + / – substantivische Doppelform (*Entscheid / Entscheidung*; *Tapezier / Tapezierer*); + / – Einfügen eines Fugenelements (*Klasslehrer / Klassenlehrer*), + / – Erhalt des Ortsnamens in der Derivation (*Aachen > Aachener*; *Pfäffikon > Pfäffiker*); + / – Suffix bei Adverbien (*durchweg(s)*, *durchgehend(s)*, *weiter(s)*);
- Kasusrektion: Genitiv- vs. Akkusativ-Adverbialkasus (*aller / alle zehn Minuten*); Dativ- vs. Akkusativobjekt (*jdm. / jdn. anrufen*); präpositionales Objekt vs. direktes Objekt (*auf etwas vergessen / etwas vergessen*);
- Verbsyntax: + / – trennbares Verb (*anvertrauen*, *anerkennen*, *widerspiegeln*, *aberkennen*); + / – Perfektbildung mit *sein* bei Bewegungsverben und den Positionsverben *sitzen*, *stehen*, *liegen*;
- Einzelfälle: Artikelgebrauch bei geografischen Namen (z.B. *das Tirol*); *was für ein* vs. *welcher*; *trotzdem* in der Funktion von *obwohl*.

Um verlässliche Aussagen zum Vorkommen der jeweiligen Varianten machen zu können, werden die für das Projekt ausgewählten Pressetexte (vgl. dazu Kap. 4) mit einer geeigneten Korpusanalyse-Software durchsucht. Auch sollen weitere Phänomene in die Suche einbezogen werden, die in den bisherigen Grammatiken noch gar nicht erfasst sind. Diese werden induktiv, auf der Basis datengeleiteter Musteranalysen ermittelt, die auf statistischen Vergleichen von Frequenzen komplexer n-Gramme gründen. Zwei Beispiele für solche Phänomene, die in den Grammatiken bisher noch keine Erwähnung finden, seien im Folgenden gegeben. Sie gehören in den Bereich der Verbstellung und der Reflexivierung. Detaillierte Informationen zu ihrer arealen Distribution, wie sie für das Handbuch vorgesehen sind, können hier nicht gegeben werden; darüber kann erst die Korpusrecherche Aufschluss geben. Es sei aber an dieser Stelle bereits gesagt, dass Phänomen 1 (Verberststellung im Nebensatz) dem Schweizer Standarddeutsch zugeordnet wird und Phänomen 2 (Gebrauch des Reflexivpronomens) seinen Ausgangspunkt bei Daten aus dem österreichischen Standarddeutsch nimmt.

Kommen wir zu Phänomen 1, zur Verberststellung im Nebensatz (vgl. dazu ausführlich Dürscheid / Hefti 2006). Ein Beispiel hierfür ist der Satz *Schön, haben Sie sich für die Rigi entschieden*, der in einem Begleitbrief der Rigibahnen zum Versand von Geschenkgutscheinen zu lesen war. Ein solches Satzmuster

ist in der Deutschschweiz in der Mundart sehr verbreitet (vgl. dazu den Aufsatz von Lötscher (1997) mit dem Titel „Guet, sind Si doo"), es handelt sich also zweifellos um eine Interferenz. Die gemeindeutsche Variante dazu ist das Satzgefüge *Es ist schön, dass Sie sich für die Rigi entschieden haben*, das aus einem Kopulasatz mit einem emotional bewertendem Prädikatsnomen (vgl. auch *schade*, *gut*) und einem *dass*-Nebensatz besteht. In der Konstruktion *Schön, haben Sie sich* [...] handelt es sich zwar auch um eine Verbindung aus Haupt- und Nebensatz, doch enthält in diesem Fall der übergeordnete Satz nur das Prädikatsnomen. Außerdem wird der Nebensatz nicht mit der Konjunktion *dass* angeschlossen, was zur Folge hat, dass das finite Verb diese Position einnehmen kann und das Subjekt nach dem finiten Verb platziert wird. Eine solche Struktur ist allerdings nur bei Nachstellung des Nebensatzes möglich (vgl. die Ungrammatikalität des Satzes **Haben Sie sich für die Rigi entschieden, schön*). Und sie steht zwar in Analogie zu einem Verbzweit-Satz mit Kommentaradverb (vgl. *Erfreulicherweise haben Sie sich für die Rigi entschieden*), doch kann das Prädikatsnomen, anders als das Kommentaradverb, hier nur am Satzanfang platziert werden (vgl. *Sie haben sich *schön / erfreulicherweise für die Rigi entschieden*). In dieser Konstruktion besetzt es aber – dies sei eigens betont – nicht das Vorfeld, sondern stellt den einzig verbliebenen Teil eines übergeordneten Satzes dar (daher auch die Kommasetzung).

Konstruktionen dieser Art weisen Parallelen zu Satzgefügen auf, in denen der übergeordnete Satz nicht eine Kopula enthält, sondern ein emotional bewertendes Prädikat (wie z.B. *froh sein, sich freuen*). Das zeigt der Internetbeleg *Wir freuen uns, haben Sie sich zurück gemeldet*. Ein weiteres Beispiel hierfür ist der Satz *Lola und Nuria sind froh, können sie miteinander Spanisch reden* (entnommen aus dem Zürcher *Tages-Anzeiger* vom 12.10.2009: 18). Auch hier handelt es sich um ein Satzgefüge, in dem der Nebensatz mit Verberststellung auftritt (analog zu dem konstruierten Beispiel *Gut, können sie miteinander Spanisch reden*), und auch hier handelt es sich um einen faktitiven, nicht um einen konditionalen Nebensatz. Letzteres ist wichtig zu betonen, denn in konditionalen Nebensätzen kann das finite Verb im Gemeindeutschen durchaus am Anfang stehen (vgl. *Ich wäre froh, käme sie auf einen Besuch vorbei*). So wird im Zweifelsfälle-Duden (Duden 2007: 519) eigens erwähnt, dass eine solche Verbstellung im konditionalen Nebensatz möglich ist.[3] Dass sie unter bestimmten Bedingungen auch im faktitiven Nebensatz vorkommt, findet dagegen keine Erwähnung.

[3] Allerdings werden hier nur Beispiele gegeben, in denen der konditionale Nebensatz vorangestellt ist (z.B. *Versagen die Bremsen, dann ist alles verloren*), nicht aber nachgestellt.

Bei Phänomen 2 geht es um den Gebrauch des Reflexivums in Äußerungen des Typs *Denn manchmal erwartet sich der Kunde im Urlaub Ruhe und Frieden*. Konstruktionen dieser Art sind in der Standardvarietät in Österreich in zahlreichen lexikalischen Variationen im Hinblick auf das beteiligte Verb durchgängig und hochfrequent belegt und können somit als weitgehend konventionalisiert gelten. Dies gilt insbesondere für die Konstruktion *erwarten + sich*. Korpusrecherchen zeigen, dass in diesem Fall bereits gegenwärtig von einem stabilen Grad der Lexikalisierung in der Standardsprache in Österreich auszugehen ist. Die Grammatiken des Deutschen verzeichnen den hier dargestellten Phänomenbereich allerdings entweder gar nicht oder subsumieren ihn unter den reflexiven Gebrauch transitiver Verben nach dem Muster *Sie kämmt sich*. Analoge Beispiele zu dem angeführten sind daher in den meisten Grammatiken ebenfalls nicht zu finden (vgl. u.a. Engel 1996, Zifonun / Hoffmann / Strecker et al. 1997, Eroms 2000, Helbig / Buscha 2001, Hentschel / Weydt 2003). Zwar gibt es umfängliche Abhandlungen zu Reflexivpronomina und Reflexivität im Deutschen (vgl. u.a. Brinker 1969, Kunze 1997, Ágel 1997, Gunkel / Müller / Zifonun (Hg.) 2003, Zifonun 2004), aber die reflexiven Konstruktionen des dargestellten Typs werden meist nicht gesondert thematisiert oder gar als diatopisch markierte standardsprachliche Variante deklariert. Dem Zweifelsfälle-Duden sind Varianten im Gebrauch des Reflexivpronomens vollkommen unbekannt (vgl. Duden 2007: 820f.), und auch im jüngst erschienenen Lexikon „Wortarten des Deutschen" findet sich im Artikel zum Reflexivum ebenso wenig ein Hinweis in diese Richtung wie in Arbeiten zum so genannten *österreichischen Deutsch* (vgl. Siemund 2007: 707ff.). In diesen wird zwar eine im Vergleich zum bundesdeutschen Sprachgebrauch abweichende und verstärkte Verwendung der Reflexiva konstatiert, die Deskription erschöpft sich aber in Empfehlungen wie „Mit dem Reflexivpronomen 'sich' nicht geizen!" (vgl. Sedlaczek 2004).[4]

Neben *erwarten + sich*, wo das Reflexiv bereits obligatorisch auftritt, sind zahlreiche Konstruktionen belegt, in denen das Reflexivum als Adjunkt fungiert und damit fakultativ ist (vgl. z.B. *Nehmen Sie sich gegenüber Vorgesetzten ein Blatt vor den Mund*; entnommen aus dem *Kurier* vom 14.12.1997: 22). Während in den Fällen, in denen das Reflexiv als direktes Objekt aufscheint (z.B. *Sie kämmt sich*), die Argumentstruktur des transitiven Verbs und die phorische Funktion des Reflexivpronomens deutlich zum Ausdruck kommen, übernimmt das Reflexiv in den angeführten Konstruktionen offenbar eine an-

[4] Siehe auch unter: http://www.das-oesterreichische-deutsch.at (Stand: 12 / 2010).

dere Funktion, denn schließlich ist das direkte Objekt explizit genannt und keinesfalls koreferent mit dem Subjekt. Da die Möglichkeiten der Interpretation der Befunde zu Phänomen 2 bereits an anderer Stelle ausführlich diskutiert worden sind (vgl. Ziegler 2010), sollen sie hier nur kursorisch angeführt werden:

Zum einen spricht die Verwendung des Reflexivpronomens, das keinen eindeutigen semantischen Bezug zum direkten Objekt aufweist, dafür, dass hier die im Verb ausgedrückte Handlung modifiziert werden soll. Insofern würde es sich schlicht um eine Tendenz zur Reflexivierung transitiver Verben in der Standardvarietät in Österreich handeln, die am Beispiel *erwarten* + *Reflexiv* bereits ein fortgeschrittenes Stadium der Lexikalisierung erreicht hat, während in anderen Fällen noch ein früheres Stadium reflektiert wird. Zum anderen kann aber auch der Auffassung gefolgt werden, dass das Reflexivpronomen *sich* polyfunktional ist, und zwar insofern es als Reflexiv- und Medialmarker fungiert. Dies bedeutet, dass zwischen Reflexivität und Medialität im Deutschen formal nicht unterschieden wird (vgl. Ágel 2000: 151). Daraus wäre zu schlussfolgern, dass die vermeintlichen Reflexivkonstruktionen keine sind, sondern vielmehr Medialkonstruktionen darstellen. Folgt man dieser Argumentation, dann wäre das Pronomen *sich* in den angeführten Beispielen eine grammatikalisierte – wenn auch morphologisch maximal unterspezifizierte – Form einer diathetischen Markierung zur Anzeige des Mediums im Deutschen, d.h. die Belege würden systematisch eine mediale Diathese reflektieren. Ein Blick in die Sprachgeschichte bekräftigt diese These. So ist etwa im Deutschen Wörterbuch im Eintrag zum Lemma *sich* zu lesen:

> schon im got. tritt nicht nur sis, sondern auch sik zu intransitiven verben, deren begriff auf diese weise inniger mit dem subjekt verbunden wird [...]. diese verwendung des reflexivpronomens kann natürlich auch beim transitivum eintreten, dann entsteht ein richtiges medium. (DWB 16: 711)

Wie auch immer die hier dargestellten Fallbeispiele grammatischer Variabilität der Standardsprache analysiert werden müssen, es bleibt als Zwischenfazit festzuhalten, dass sich durch die hohe Frequenz der aufgezeigten Konstruktionsmuster regionale Besonderheiten im geschriebenen Gebrauchsstandard des Deutschen manifestieren. Die grammatischen Varianten werden nicht *ausnahmsweise*, sondern von einer mehr oder weniger großen Zahl von Sprechern / Schreibern akzeptiert. Sie zeigen dabei eine systematische Regelhaftigkeit in ihrer strukturellen Ausformung. Die Beispiele verdeutlichen ebenfalls,

dass unser Projekt dazu geeignet ist, die grammatiktheoretische Diskussion anzuregen und somit auch in dieser Hinsicht durchaus fruchtbar und vielversprechend scheint. Und überdies ist zu konstatieren: Die dargestellten Varianten finden keine Berücksichtigung in vorliegenden Grammatiken.

4. Korpuszusammenstellung

Das den Analysen zugrunde liegende Korpus soll Texte deutscher Standardsprache der Gegenwart umfassen, wobei wir uns, wie in Kapitel 1 dargelegt, zunächst auf den *geschriebenen* Standard beschränken und darunter im Wesentlichen den Sprachgebrauch in der Presse ansehen. Eine Ausweitung auf die Untersuchung der gesprochenen Standardsprache würde dagegen auch Phänomene erfassen, die der Online-Syntax (siehe oben) geschuldet sind und daher nicht in unseren Untersuchungsbereich fallen. Außerdem würde sie eine andere Methodologie und insbesondere – wegen notwendiger Transkriptionsarbeiten in großem Stil – einen viel höheren Aufwand erforderlich machen. Der Fokus des Projekts auf die diatopische Variation in der Standardsprache rechtfertigt die Konzentration auf regionale Zeitungen, und da besonders auf regionale Meldungen. Es ist vorgesehen, 57 online publizierte Zeitungen aus den deutschsprachigen Ländern nach einem bestimmten Regionenschlüssel heranzuziehen. Dazu wird eine Einteilung des Gesamtgebiets in 17 Sektoren vorgenommen; diese orientiert sich an der regionalen Gliederung im Variantenwörterbuch (Ammon et al. 2004: XXXIVff.), in der Deutschland in sechs und Österreich in vier Sektoren eingeteilt wird und Liechtenstein, Luxemburg und Südtirol als eigene Gebiete gewertet werden. Einzig bei der Einteilung der deutschsprachigen Schweiz (in einen (nord-)westlichen, einen (nord-)östlichen und einen südlichen Teil) weichen wir von den Vorgaben des Variantenwörterbuchs, das die Schweiz als homogenen Block behandelt, ab, da neuere Untersuchungen zur Binnendifferenzierung der dialektalen Syntax in der Schweiz (z.B. Glaser 2003, Bucheli Berger 2005) eine solche Untergliederung auch für die geschriebene Standardsprache als lohnend erscheinen lassen.

Pro Sektor wird eine bestimmte Anzahl von Tageszeitungen herangezogen: Für die kleineren deutschsprachigen Länder und Gebiete können ein bis zwei Zeitungen als hinreichend betrachtet werden.[5] Für die Sektoren in der Schweiz und in Österreich werden zunächst je zwei Zeitungen berücksichtigt, für die deut-

[5] Für Ost-Belgien ließ sich überhaupt nur eine einzige deutschsprachige Tageszeitung ermitteln.

schen Sektoren hingegen – wegen der größeren Ausmaße und höheren Bevölkerungszahlen – je sechs Zeitungen. Bei der Auswahl der Zeitungen für Deutschland soll zusätzlich auf eine ausgeglichene Distribution innerhalb der Sektoren geachtet werden. Dies geschieht dadurch, dass die sechs Zeitungen möglichst gleichmäßig auf die in diesem Sektor erfassten Bundesländer verteilt sind. Damit soll eine „Unausgewogenheit des Korpus" vermieden werden, wie sie in der Zusammenstellung des Zeitungskorpus für das Variantenwörterbuch auftrat (vgl. Kleiner 2006: 114). Pro Zeitung wird schließlich ein Textumfang von fünf Millionen Wortformen angestrebt, so dass das Gesamtkorpus des arealen Gebrauchsstandards des Deutschen ca. 285 Millionen Wortformen umfasst und die Teilkorpora ungefähr gleichmäßig über das Sprachgebiet distribuiert und ungefähr gleich groß sind. Selbstverständlich müssen in Bezug auf die Texte in regionalen Zeitungen Agenturmeldungen, Werbebanner, Anzeigen etc. herausgefiltert werden, da diese als vorgefertigte Texte von den Zeitungsredaktionen übernommen werden und daher eben kaum Rückschlüsse auf den regionalen Gebrauchsstandard zulassen.

5. Ausblick

Nachdem in den vorangehenden Kapiteln die Eckdaten des geplanten Projekts vorgestellt, ausgewählte Phänomenbereiche diskutiert und die Methoden der Datenerhebung dargelegt wurden, soll nun zusammengefasst werden, welche Zielsetzungen wir mit unserem Projekt verbinden. Fünf Punkte seien hier resümierend genannt:

1) Wie bereits erwähnt, soll die Variation in der Grammatik der geschriebenen deutschen Standardsprache umfassend dokumentiert werden. Eben dies wurde von den bisherigen grammatischen Überblicksdarstellungen nicht geleistet.

2) Das geplante Handbuch soll nicht nur ein Grundlagenwerk in der Variationslinguistik darstellen, es soll auch für interessierte Laien zugänglich und benutzbar sein. Gedacht ist hier an Personen, die aus beruflichen oder privaten Gründen Auskunft über die Normgemäßheit bestimmter Varianten begehren (z.B. Lehrer, Lektoren, Übersetzer).

3) Von der Darstellung grammatischer Varianz und ihrer theoretischen und empirischen Fundierung in einem Handbuch soll auch die Grammatikographie des Deutschen profitieren. Sie kann sich bis dato auf keine einzige umfassende empirische Untersuchung der Variation in der Standardgram-

matik stützen. Folglich sind die – wenigen – Angaben in den Grammatiken zum einen mehr oder weniger impressionistisch, zum anderen aber auch lückenhaft.

4) Durch die Dokumentation grammatischer Varianten in einem Handbuch soll langfristig ein Beitrag dazu geleistet werden, dass die arealen Ausprägungen des Standarddeutschen als gleichwertig anerkannt werden. In dieser Hinsicht hat das Projekt eine zentrale sprachpolitische Bedeutung.

5) Ein nach Regionen in den deutschsprachigen Ländern ausgewogenes Korpus als Basis für grammatische Untersuchungen liegt bisher nicht vor. In korpuslinguistischer Hinsicht beschreitet das geplante Projekt in der deutschsprachigen Grammatikforschung damit völliges Neuland.

Es bleibt zu hoffen, dass die geplante „Variantengrammatik des Standarddeutschen" dem Leser ebenso nützlich sein wird, wie es das Variantenwörterbuch bereits ist – und zusätzlich noch einen Mehrwert aufweist, da die Variantengrammatik korpusbasiert sein wird. Außerdem kann sie ein nützliches Pendant zu einer Korpusgrammatik werden, die derzeit am Institut für Deutsche Sprache (IDS) in Planung ist. Ziel dieses Projekts ist „die korpusgestützte Erforschung der Variation im standardnahen Deutsch (einschließlich der Variation im Standard selbst), die längerfristig eine Grundlage für die Erstellung einer Grammatik des Deutschen bilden soll" (vgl. http://www.ids-mannheim.de/gra/korpusgrammatik.html Stand: 12/2010). Während in unserem Projekt der Fokus auf der Beschreibung diatopischer Variation liegt und hierfür die Erstellung eines neuen, ausgewogenen Korpus erforderlich ist, wird es dort um die Beschreibung jeglicher („standardnaher") Variation in der Grammatik gehen, was mit den bereits am IDS vorhandenen Korpora geschriebener Sprache geleistet werden soll. Insofern ergänzen sich die beiden Projekte in ihrer jeweiligen Zielsetzung. Was sie gemeinsam haben, ist ihre Hinwendung zur Korpuslinguistik. Darin zeigt sich eine Entwicklung, die in der Grammatikforschung erfreulicherweise immer mehr zunimmt.

Literatur

AdA = Atlas zur deutschen Alltagssprache. Bearb. v. Stephan Elspaß und Robert Möller (2003ff.). Internet: http://www.uni-augsburg.de/ada (Stand: 12 / 2010).

Ágel, Vilmos (1997): Reflexiv-Passiv, das (im Deutschen) keines ist. Überlegungen zu Reflexivität, Medialität, Passiv und Subjekt. In: Dürscheid, Christa / Ramers, Karl Heinz / Schwarz, Monika (Hg.): Sprache im Fokus. Tübingen: Niemeyer, 147-187.

Ágel, Vilmos (2000): Valenztheorie. Tübingen: Narr.

Ammon, Ulrich (1995): Die deutsche Sprache in Deutschland, Österreich und der Schweiz. Das Problem der nationalen Varietäten. Berlin / New York: de Gruyter.

Ammon, Ulrich / Bickel, Hans / Ebner, Jakob et al. (2004): Variantenwörterbuch des Deutschen. Die deutsche Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol. Berlin / New York: de Gruyter.

Auer, Peter (2000): On line-Syntax – oder: was es bedeuten könnte, die Zeitlichkeit der mündlichen Sprache ernst zu nehmen. In: Sprache und Literatur 85: 43-56.

Brinker, Klaus (1969): Zum Problem der angeblich passivnahen Reflexivkonstruktionen in der deutschen Gegenwartssprache. In: Muttersprache. Zeitschrift zur Pflege und Erforschung der deutschen Sprache 79: 1-11.

Bucheli Berger, Claudia (2005): Passiv im Schweizerdeutschen. In: Christen, Helen (Hg.): Dialekt/ologie an der Jahrtausendwende. Dialect/ology at the turn of the millennium. Linguistik online 24, 3: 49-77. Internet: http://www.linguistik-online.de/24_05/bucheli.pdf (Stand: 12 / 2010).

Dovalil, Vít (2006): Sprachnormenwandel im geschriebenen Deutsch an der Schwelle zum 21. Jahrhundert. Die Entwicklung in ausgesuchten Bereichen der Grammatik. Frankfurt a.M. u.a.: Peter Lang.

Duden (2007): Richtiges und gutes Deutsch. Wörterbuch der sprachlichen Zweifelsfälle. (= Duden 9). 6., vollst. überarb. Aufl. Mannheim / Leipzig / Wien / Zürich: Dudenverlag.

Duden (2009): Duden – Die Grammatik. Unentbehrlich für richtiges Deutsch. (= Duden 4). 8., überarb. Aufl. Mannheim / Leipzig / Wien / Zürich: Dudenverlag.

Dürscheid, Christa / Hefti, Inga (2006): Syntaktische Merkmale des Schweizer Standarddeutsch. Theoretische und empirische Aspekte. In: Dürscheid, Christa / Businger, Martin (Hg.): Schweizer Standarddeutsch. Beiträge zur Varietätenlinguistik. Tübingen: Narr, 131-161.

DWB = Grimm, Jacob / Grimm, Wilhelm (1854-1960): Das Deutsche Wörterbuch. 16 Bde. (in 32 Teilbänden). Leipzig: S. Hirzel [Quellenverzeichnis 1971].

Ebner, Jakob (2008): Österreichisches Deutsch. Eine Einführung. Mannheim / Leipzig / Wien / Zürich: Dudenverlag.

Eisenberg, Peter (2007): Sprachliches Wissen im Wörterbuch der Zweifelsfälle. Über die Rekonstruktion einer Gebrauchsnorm. In: Aptum. Zeitschrift für Sprachkritik und Sprachkultur 3: 209-228.

Elspaß, Stephan (2010): Regional standard variation in and out of grammarians' focus. In: Lenz, Alexandra N. / Plewnia, Albrecht (Hg.): Grammar between norm and variation. (= VarioLingua 40). Frankfurt a.M. u.a.: Peter Lang, 127-144.

Elter, Irmgard (2005): Genitiv versus Dativ. Die Rektion der Präpositionen *wegen*, *während*, *trotz*, *statt* und *dank* in der aktuellen Zeitungssprache. In: Schwitalla, Johannes / Wegstein, Werner (Hg.): Korpuslinguistik deutsch: synchron – diachron – kontrastiv. Würzburger Kolloquium 2003. Tübingen: Niemeyer, 125-135.

Engel, Ulrich (1996): Deutsche Grammatik. 3., korr. Aufl. Heidelberg: Julius Groos Verlag.

Eroms, Hans-Werner (2000): Syntax der deutschen Sprache. Berlin / New York: de Gruyter.

Gelhaus, Hermann (1972): Vorstudien zu einer kontrastiven Beschreibung der schweizerdeutschen Schriftsprache der Gegenwart. Die Rektion der Präpositionen trotz, während und wegen. Unt. Mitarb. v. Roger Frey und Otfried Heyne. (= Europäische Hochschulschriften I: Deutsche Literatur und Germanistik 58). Bern / Frankfurt a.M. u.a.: Peter Lang.

Glaser, Elvira (2003): Schweizerdeutsche Syntax. Phänomene und Entwicklungen. In: Dittli, Beat / Häcki Buhofer, Annelies / Haas, Walter (Hg.): Gömmer MiGro? Veränderungen und Entwicklungen im heutigen Schweizer Deutschen. (= Germanistica Friburgensia 18). Freiburg, Schweiz: Universitätsverlag, 39-66.

Götz, Ursula (1995): Regionale grammatische Varianten des Standarddeutschen. In: Sprachwissenschaft 20: 222-238.

Gunkel, Lutz / Müller, Gereon / Zifonun, Gisela (Hg.) (2003): Arbeiten zur Reflexivierung. (= Linguistische Arbeiten 481). Tübingen: Niemeyer.

Helbig, Gerhard / Buscha, Joachim (2001): Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht. Neubearb. Berlin / München: Langenscheidt.

Henn-Memmesheimer, Beate (1986): Nonstandardmuster. Ihre Beschreibung in der Syntax und das Problem ihrer Arealität. (= Reihe Germanistische Linguistik 66). Tübingen: Niemeyer.

Hentschel, Elke / Weydt, Harald (2003): Handbuch der deutschen Grammatik. 3., völl. neu bearb. Aufl. Berlin / New York: de Gruyter.

Heuer, Walter / Gallmann, Peter / Flückiger, Max (2008): Richtiges Deutsch. Vollständige Grammatik und Rechtschreiblehre unter Berücksichtigung der aktuellen Rechtschreibreform. 28., überarb. Aufl. Zürich: NZZ Libro.

Kaiser, Stephan (1969/70): Die Besonderheiten der deutschen Schriftsprache in der Schweiz. 2 Bde. (= Duden-Beiträge. Sonderreihe: Die Besonderheiten der deutschen Schriftsprache im Ausland). Mannheim/Zürich: Duden Verlag.

Kleiner, Stefan (2006): Rezension 'Variantenwörterbuch des Deutschen'. Die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol. In: Zeitschrift für Dialektologie und Linguistik 73: 112-116.

Kunze, Jürgen (1997): Typen der reflexiven Verbverwendung im Deutschen und ihre Herkunft. In: Zeitschrift für Sprachwissenschaft 16: 83-180.

Lötscher, Andreas (1997): „Guet sind Si doo" – Verbstellungsprobleme bei Ergänzungssätzen im Schweizerdeutschen. In: Ruoff, Arno/Löffelad, Peter (Hg.): Syntax und Stilistik der Alltagsprache. Beiträge der 12. Arbeitstagung zur alemannischen Dialektologie. (= Idiomatica 18). Tübingen: Niemeyer, 85-95.

Muhr, Rudolf (1995): Grammatische und pragmatische Merkmale des Österreichischen Deutsch. In: Muhr, Rudolf/Schrodt, Richard/Wiesinger, Peter (Hg.): Österreichisches Deutsch. Linguistische, sozialpsychologische und sprachpolitische Aspekte einer nationalen Variante des Deutschen. Wien: Hölder-Pichler-Tempsky, 208-235.

Muhr, Rudolf/Sellner, Manfred (Hg.) (2006): Zehn Jahre Forschung zum Österreichischen Deutsch: 1995-2005 – eine Bilanz. Wien/Frankfurt a.M. u.a.: Peter Lang.

Patocka, Franz (1997): Syntaktische Austriazismen in der Verbstellung? In: Eichner, Heiner/Ernst, Peter/Katsikas, Sergios (Hg.): Sprachnormung und Sprachplanung. Festschrift für Otto Back zum 70. Geburtstag. 2., verb. Aufl. Wien: Edition Praesens, 51-60.

Reiffenstein, Ingo (2001): Das Problem der nationalen Varietäten. Rezensionsaufsatz zu Ulrich Ammon: Die deutsche Sprache in Deutschland, Österreich und der Schweiz. Das Problem der nationalen Varietäten. Berlin/New York 1995. In: Zeitschrift für deutsche Philologie 120: 78-89.

Rohrer, Carl (1973): Der Konjunktiv im gesprochenen Schweizer Hochdeutschen. Analyse von Radiogesprächen. (= Studia Linguistica Alemannica 3). Frauenfeld/Stuttgart: Huber.

Rüttimann, Martina (2002): Vergleichende Studie zur Kasusrektion bei den Präpositionen „wegen", „während" und „trotz" in drei deutschsprachigen Schweizer Zeitungen. Unveröff. Lizenziatsarbeit Univ. Zürich.

Scheuringer, Hermann (1996): Das Deutsche als pluriareale Sprache: Ein Beitrag gegen staatlich begrenzte Horizonte in der Diskussion um die deutsche Sprache in Österreich. In: Die Unterrichtspraxis/Teaching German 29, 2: 147-153.

Sedlaczek, Robert (2004): Das österreichische Deutsch. Wie wir uns von unserem großen Nachbarn unterscheiden. Ein illustriertes Handbuch. Wien: Ueberreuter.

Sellner, Manfred (2006): 'Trotz', 'wegen' und 'während' im Österreichischen Deutsch. Eine Pilotstudie. In: Muhr / Sellner (Hg.), 49-64.

Siemund, Peter (2007): Reflexivum. In: Hoffman, Ludger (Hg.): Deutsche Wortarten. Berlin / New York: de Gruyter, 707-725.

Stirnemann, Knut (1980): Zur Syntax des gesprochenen Schweizer Hochdeutschen. Eine Untersuchung zur Sprache des Deutschunterrichts an der Luzerner Kantonsschule. (= Studia Linguistica Alemannica 7). Frauenfeld / Stuttgart: Huber.

Stubkjær, Flemming Talbo (1997): Das Präsensparadigma der starken Verben im Österreichischen Deutsch. In: Muhr, Rudolf / Schrodt, Richard (Hg.): Österreichisches Deutsch und andere nationale Varietäten plurizentrischer Sprachen in Europa. Empirische Analysen I. Wien: öbv & hpt, 199-210.

Tatzreiter, Herbert (1988): Besonderheiten in der Morphologie der deutschen Sprache in Österreich. In: Wiesinger, Peter (Hg.): Das österreichische Deutsch. Wien u.a.: Böhlau, 71-98.

Wiesinger, Peter (2006): Das österreichische Amtsdeutsch. Eine Studie zur Syntax, Stil und Lexik der österreichischen Rechts- und Verwaltungssprache der Gegenwart. In: Wiesinger, Peter: Das österreichische Deutsch in Gegenwart und Geschichte. (= Austria: Forschung und Wissenschaft – Literatur 2). Wien u.a.: LIT.

Zeman, Dalibor (2009): Überlegungen zur deutschen Sprache in Österreich. Linguistische, sprachpolitische und soziolinguistische Aspekte der österreichischen Varietät. Hamburg: Verlag Dr. Kovač.

Ziegler, Arne (2010): „Er erwartet sich nur das Beste ...". Reflexivierungstendenz und Ausbau des Verbalparadigmas in der österreichischen Standardsprache. Zu einer Variantengrammatik des Deutschen. In: Bittner, Dagmar / Gaeta, Livio (Hg.): Kodierungstechniken im Wandel: Das Zusammenspiel von Analytik und Synthese im Gegenwartsdeutschen. (= Linguistik – Impulse & Tendenzen 34). Berlin / New York: de Gruyter, 67-81.

Zifonun, Gisela (2004): Reflexivierung in der Nominalphrase. In: Lindemann, Beate / Letnes, Ole (Hg.): Diathese, Modalität, Deutsch als Fremdsprache. Festschrift für Oddleif Leirbukt zum 65. Geburtstag. Tübingen: Stauffenburg, 135-152.

Zifonun, Gisela / Hoffmann, Ludger / Strecker, Bruno et al. (1997): Grammatik der deutschen Sprache. 3 Bde. (= Schriften des Instituts für Deutsche Sprache 7.1-7.3). Berlin / New York: de Gruyter.

REINHARD FIEHLER

Korpusbasierte Analyse von Univerbierungsprozessen

Abstract

Vorgestellt werden Ziele und erste Ergebnisse des Projektes „Univerbierung" am Institut für Deutsche Sprache. Das Projekt untersucht in verschiedenen Korpora, ob sich Prozesse der Univerbierung quantitativ belegen lassen. In Form von Univerbierungsprofilen sollen Univerbierungsverläufe dargestellt werden, d.h. die quantitativen Veränderungen, die zeitlich im Verhältnis der Getrennt- und Zusammenschreibungen eintreten (Kap. 1 und 2). Zugleich wird dabei methodologisch reflektiert, ob und inwieweit diese Korpora für solche Untersuchungen geeignet sind (Kap. 3). Exemplarisch werden einige Univerbierungsprofile vorgestellt (Kap. 4). Es handelt sich zum einen um Beispiele, bei denen sich die Normlage im Zuge der Rechtschreibreform nicht geändert hat, und zum anderen um solche, bei denen sie im Untersuchungszeitraum (1985-2008) verändert wurde. Die Untersuchungen zielen in der Perspektive darauf ab, Faktoren herauszuarbeiten, die Univerbierungsprozesse fördern bzw. hemmen, und aufzuklären, was Schreiber(-innen) als *ein* Wort gilt. Dies kann dazu beitragen, empirisch gestützt Komponenten des Wortkonzepts zu ermitteln (Kap. 5).

1. Univerbierung

Unter Univerbierung wird hier mit Bußmann (1983: 563) Folgendes verstanden:

> In der Wortbildung Vorgang und Ergebnis des Zusammenwachsens mehrgliedriger syntaktischer Konstruktionen zu einem Wort, z.B. *ob + schon* zu *obschon*.

Univerbierung ist damit ein Prozess, der zur Bildung neuer Lexeme führt (Lexikalisierung). Zugleich ist sie in vielen Fällen auch ein Prozess der Grammatikalisierung, wenn das neue Lexem einer anderen grammatischen Kategorie angehört als die Ausgangsbestandteile (*auf Grund* → *aufgrund*: Präpositionalphrase → Präposition).

Löst man die Metapher des Zusammenwachsens auf, so ist Univerbierung zum einen ein mentales Phänomen: Sprachliche Elemente werden nicht mehr als getrennt empfunden bzw. verstanden, sondern als zusammengehörig. Dies manifestiert sich nach außen und für andere erkennbar in der Zusammenschreibung (grafische Univerbierung) von Elementen, die bisher getrennt geschrieben wurden. In dieser Hinsicht betrachtet ist Univerbierung damit zum

anderen ein schriftsprachliches Phänomen, dessen Realisierung und Erkennbarkeit die Trennung von Wortformen durch Spatien zur Voraussetzung hat. Werden zwei Wörter, die bisher mehrheitlich durch ein Spatium getrennt wurden, nicht mehr getrennt-, sondern zusammengeschrieben, so liegt Univerbierung vor.[1] In der Spatiensetzung kommt zum Ausdruck, was Schreiber als *ein* Wort empfinden. Univerbierung reflektiert, dass bestimmte Wortfolgen im Zuge des Sprachwandels nicht mehr als solche, sondern zunehmend als *ein* Wort empfunden werden. Dabei unterliegen nicht alle Wortfolgen einer Tendenz zur Univerbierung, sondern nur bestimmte, d.h. es muss Faktoren geben, die die Univerbierung zweier Wortformen befördern, und solche, die ihr entgegenstehen.

Univerbierung ist ein diachroner Prozess, er führt von einer mehrheitlichen Getrennt- zu einer mehrheitlichen Zusammenschreibung.[2] Synchron betrachtet ist sein Resultat eine systematische Varianz: Die Koexistenz von getrennt- wie zusammengeschriebenen Formen (in einem sich zeitlich verändernden quantitativen Verhältnis). Univerbierung kann in Bezug auf einzelne Personen, auf Textsorten oder in Hinblick auf eine Sprach- bzw. Schreibgemeinschaft als Ganzes betrachtet werden. Univerbierung ist zudem ein fortwährender historischer Prozess, der immer neue Wortfolgen erfasst.

> [...] dass im Deutschen seit längerer Zeit die Tendenz zu beobachten ist, bei bestimmten Typen von Wortgruppen, die eine relativ abgegrenzte, einheitliche Gegebenheit der objektiven Realität (Gegenstand, Eigenschaft, Vorgang, Beziehung) benennen und damit eine den Einwortlexemen ähnliche Nominationsfunktion übernehmen, auch die geschriebenen Formative denen von Einwortlexemen anzugleichen und sie in eine geschlossene graphische Wortform zu überführen (graphische Univerbierung) (Nerius et al. 2000: 163)

Viele in der Schriftsprachgeschichte entstandene Univerbierungen werden heute nicht oder kaum mehr als solche erkannt (*seinerzeit*, *zufrieden*, *immerhin, durchaus* etc.). Analoge Prozesse finden sich nicht nur in der deutschen Schriftsprache, sondern auch in anderen Sprachen.

Getrennt- und Zusammenschreibung unterliegen in vielen Bereichen einer Normierung durch die kodifizierte Rechtschreibung. Die faktische Schreibung ist jedoch nicht allein durch die Norm bestimmt, sondern eine Resultante aus

---

[1] Ehemals nominale Bestandteile werden bei der Univerbierung – falls vorhanden – unter Vermeidung von Binnenmajuskeln kleingeschrieben.

[2] Zu dem der Univerbierung komplementären Prozess der Desintegration, der Getrenntschreibung von Bestandteilen eines bisher überwiegend zusammengeschriebenen Wortes (z.B. *Besucher Parkplatz*), vgl. Scherer (2010).

drei Faktoren: dem intuitiven Konzept (Sprachgefühl) des Schreibers, was *ein* Wort ist, der Normkenntnis des Schreibers und dem Willen des Schreibers, die Norm zu beachten.

Das primäre Interesse des Projekts gilt den faktischen Schreibungen, so wie sie in verschiedenen Korpora vorfindbar sind.[3] Die Normierung der Getrennt- und Zusammenschreibung ebenso wie Veränderungen dieser Normierung (z.B. im Rahmen der Rechtschreibreform) sind nur insoweit von Interesse, wie sie als – sicherlich zentrale – Faktoren diese Schreibungen mitbeeinflussen.

2. Ziele des Projekts

Motivation für die Beschäftigung mit Univerbierungsprozessen im Rahmen des Projekts „Univerbierung (KG-U)"[4] am Institut für Deutsche Sprache war zum einen die Erfahrung, dass sich die Veränderungen der Normen im Bereich der Getrennt- und Zusammenschreibung im Zuge der Rechtschreibreform, die ohne hinreichende Berücksichtigung faktischer Sprachwandelprozesse erfolgten, als problematisch erwiesen haben (vgl. Günther 1997, Bredel/Günther 2000), zum anderen aber auch eine punktuelle Beobachtung wie die, dass die Univerbierung *ausversehen* für mich völlig überraschend und von niemandem (außer den Schreiber(-innen)) beachtet offensichtlich dabei ist, Platz zu greifen: So sind bei Google Web mehr als ein Viertel der Schreibungen von *aus_Versehen*[5] univerbiert. Eine Vergleichsanalyse von Google Web, dem Deutschen Referenzkorpus (DeReKo) und dem Korpus des Digitalen Wörterbuchs der Deutschen Sprache (DWDS) ergab folgendes Bild:

- Google Web (2.9.09)

ausversehen:	196 000 Vorkommen	(26,7%)
aus Versehen:	734 000 Vorkommen	

- DeReKo (2.9.09)

ausversehen:	15 Vorkommen	(0,25%)
aus Versehen:	6 085 Vorkommen	

[3] Dieser Aspekt ist in der Literatur bisher wenig beachtet worden. Univerbierung wird dort vor allem unter den Gesichtspunkten Normierung (z.B. Herberg/Baudusch1989), Theorie des Wortes (z.B. Gallmann 1999, Weinberger 2005), Wortbildung (z.B. Eisenberg 2004, Fuhrhop 2007), Systemhaftigkeit (z.B. Jacobs 2005, 2007) und Sprachgeschichte (z.B. Nerius et al. 2000) behandelt.

[4] Mitarbeiter(-innen): Reinhard Fiehler, Kerstin Güthert. Das Projekt ist eine Vorstudie für das geplante Projekt „Grammatische Variation im standardnahen Deutsch (Korpusgrammatik)".

[5] Zur Kennzeichnung, dass sowohl auf die univerbierten wie auch auf die nicht univerbierten Schreibungen referiert wird, wird das Spatium mit einem Unterstrich versehen.

- DWDS (2.9.09)

ausversehen:	0 Vorkommen	(0%)
aus Versehen:	235 Vorkommen	

Diese gravierenden Unterschiede verweisen deutlich auf die Bedeutsamkeit, die Korpora und ihrer je spezifischen Zusammensetzung für solche Untersuchungen zukommt.

Aufgrund dieser Vorüberlegungen ergaben sich für das Projekt folgende Zielsetzungen:

- Sichtung und Prüfung deutschsprachiger Korpora auf ihre Eignung für die quantitative Untersuchung von Sprachwandelprozessen;
- Grammatische Kategorisierung von Univerbierungen und Bildung von Fallklassen;
- Korpusbasierte quantitative Beschreibung von Univerbierungsverläufen (Erarbeitung von Univerbierungsprofilen);
- Systematisierung univerbierungsfördernder und -hemmender Faktoren;
- Analyse des Einflusses von Normänderungen (im Zuge der Rechtschreibreform) auf Univerbierungsverläufe;
- Erarbeitung von Empfehlungen für die Normierung der Getrennt- und Zusammenschreibung auf empirischer Basis;
- Entwicklung eines empirisch basierten Wortkonzepts auf der Grundlage der univerbierungsfördernden und -hemmenden Faktoren.

3. Korpora

Im Rahmen des Ziels „Sichtung und Prüfung deutschsprachiger Korpora auf ihre Eignung für die quantitative Untersuchung von Sprachwandelprozessen" wurde am Beispiel von *statt_dessen* (und weiteren hier nicht dokumentierten Beispielen) eine Reihe frei zugänglicher Korpora getestet.[6] Bei den Korpora handelt es sich um das Deutsche Referenzkorpus (DeReKo) des IDS[7], das

[6] Für einen ähnlichen Korpusvergleich (allerdings zu anderen Zwecken) vgl. Stuyckens/Bröne (2009).

[7] Da im DeReKo ab 1985 die Anzahl der Wortformen pro Jahr sehr deutlich größer ist als im Zeitraum zuvor, liegt es nahe, den Zeitraum 1985-2008 als gesondertes Teilkorpus zu betrachten. Im Zeitraum vor 1985 gibt es lediglich in den Jahren 1949, 1954, 1959, 1964, 1969 und 1974 überdurchschnittliche Anzahlen von Wortformen pro Jahr. Diese sechs Messpunkte wurden als zweites DeReKo-Teilkorpus untersucht.

Digitale Wörterbuch der Deutschen Sprache (DWDS), das Korpus C4[8], das Schweizer Textkorpus (CHTK), die Datenbank gesprochenes Deutsch (DGD) des IDS, Google Web und Google Bücher[9]. In der folgenden Tabelle sind jeweils die Gesamtzahl der im betreffenden Korpus enthaltenen Wortformen, die Anzahl der Vorkommen von *statt_dessen*, der Anteil der Univerbierungen an der Gesamtzahl der Vorkommen und die Frequenz von *statt_dessen* (Vorkommen / Million Textwörter) angegeben (4.9.2009):

	Wortformen	Vorkommen *statt_dessen*	Anteil Univerbierungen	Frequenz Vork. / Mio
DeReKo 1949-1974	3,4 Mio	93	12 %	27,4
DeReKo 1985-2008	3219 Mio	153 050	63 %	47,6
DWDS	100 Mio	1338	19 %	13,4
C4	45,8 Mio	406	14 %	8,9
CHTK	20 Mio	116	22 %	5,8
DGD	5,8 Mio	4	0 %	0,7
Google Web	?	5 900 000	77 %	?

Die Untersuchungsergebnisse zu *statt_dessen* in den einzelnen Korpora möchte ich an dieser Stelle nicht im Detail diskutieren. Betrachtet man die Ergebnisse im Überblick, so wird deutlich, dass die untersuchten Korpora in jeder Hinsicht gewaltige Unterschiede aufweisen. Wie auch immer die unterschiedlichen Resultate im Einzelnen zu erklären sind, machen sie doch überaus deutlich, dass man sich bei Untersuchungen der angestrebten Art nicht auf ein Korpus wird beschränken können.

Geht man von einer statistisch relevanten Untergröße von 1 000 Vorkommen aus, so reduziert sich für *quantitative Untersuchungen* das Feld der Korpora auf Google Web, DeReKo 1985-2008 und DWDS. Für quantitative Untersuchungen von *zeitlichen Veränderungen* kommen nur DeReKo 1985-2008 und DWDS infrage. Google Web lässt (mit Ausnahme der Abfrage der Vorkommen im

[8] Gemeinsames Korpus des DWDS, des Austrian Academy Corpus (AAC), des Korpus Südtirol und des Schweizer Textkorpus (CHTK).

[9] Google Bücher wird im Weiteren wegen der noch unausgereiften Recherchemöglichkeiten nicht berücksichtigt. Vgl. Brückner (2009).

letzten Jahr) keine zeitliche Differenzierung zu.[10] Auch wenn DeReKo in quantitativer Hinsicht keine Wünsche offenlässt, so besteht sein Nachteil doch darin, dass, wenn man hinreichend große Anzahlen von Wortformen und Vorkommen pro Jahr zur Bedingung macht, nur ein für Sprachwandelprozesse sehr kleiner Zeitraum von 23 Jahren (1985-2008) betrachtet werden kann. Ein weiterer Nachteil ist, dass DeReKo ganz überwiegend aus Zeitungstexten besteht und nur eine geringe Textsortendiversifikation aufweist. DWDS überspannt den wesentlich größeren Zeitraum von 100 Jahren (1900-2000), jedoch liegt seine Größe und entsprechend die Anzahl der Vorkommen für gegenwärtige Ansprüche an der unteren Grenze des quantitativ Erforderlichen.

4. Beispiele für Univerbierungsprofile

Im Folgenden werden nun exemplarisch einige Univerbierungsverläufe in DeReKo 1985-2008 im Detail betrachtet.

4.1 *statt_dessen*

Die Recherche für *statt_dessen* im DeReKo 1985-2008 ergibt folgende Ergebnisse (4.9.2009):

Bei 3219 Mio. Textwörtern finden sich 153050 Vorkommen. Davon zeigen 57021 Getrenntschreibung, 96029 (= 63%) Zusammenschreibung. Die Frequenz beträgt 47,6 Vorkommen / Million Textwörtern.

Fasst man die Jahresergebnisse zu 3-Jahres-Zeiträumen zusammen, so ergeben sich folgende Anteile von Zusammenschreibungen:

1985-87:	36%	1997-1999:	31%[11]
1988-90:	26%	2000-2002:	79%
1991-93:	20%	2003-2005:	80%
1994-96:	19%	2006-2008:	95%

Norm (bis 1998): getrennt

Norm (ab 1998): zusammen

[10] Anders ist dies bei Google Bücher, wo eine zeitliche Differenzierung über das Erscheinungsjahr erfolgen kann. Aber – wie gesagt – sind die Recherchemöglichkeiten dort im Moment noch nicht ausgereift.

[11] In den Zeitraum 1997-1999 fällt das Inkrafttreten der Rechtschreibreform 1998. Er ist damit gewissermaßen ein Scharnierzeitraum, wobei der Vergleich der Zeiträume davor und danach von besonderem Interesse ist.

Abbildung 1 zeigt diese Ergebnisse in grafischer Darstellung:

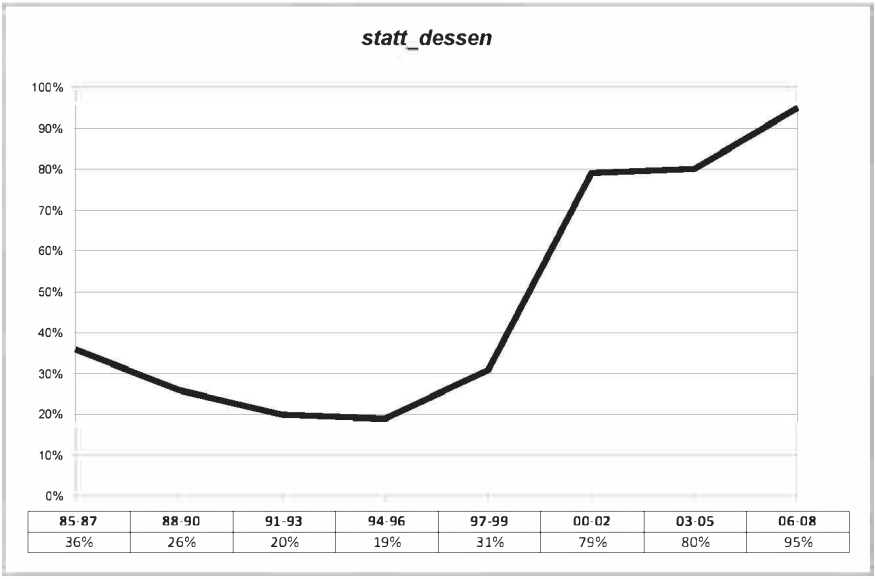


Abb. 1: Anteile der Univerbierungen von *statt_dessen* im DeReKo 1985-2008

In Zeitraum 1985-1996 ist eine kontinuierliche Abnahme der Anzahl der Zusammenschreibungen, also eine Entwicklung in Richtung auf die normentsprechende Getrenntschreibung, festzustellen. Der Anteil der Zusammenschreibungen bewegt sich dabei aber auch noch am Ende des Zeitraums auf vergleichsweise hohem Niveau.

Die Normänderung katalysiert dann im Zeitraum 2000-2008 die Univerbierung bis hin zu einer fast ausschließlichen Zusammenschreibung.[12]

4.2 *genauso_gut*

Die Recherche für *genauso_gut* im DeReKo 1985-2008 ergibt folgende Ergebnisse (4.9.2009):

[12] Inwieweit diese Entwicklung durch Rechtschreibkorrekturprogramme mitverursacht ist, bleibt zu prüfen. DeReKo ist ganz überwiegend ein Zeitungskorpus. Man kann davon ausgehen, dass in den Zeitungsredaktionen solche Programme verwendet werden. Im konkreten Fall ist festzuhalten, dass beispielsweise der Duden-Korrektor die Getrenntschreibung nicht als falsch ausweist.

Bei 3 219 Mio. Textwörtern finden sich 9 076 Vorkommen. Davon zeigen 6 672 Getrenntschreibung, 2 404 (= 26%) Zusammenschreibung. Die Frequenz beträgt 2,8 Vorkommen / Million Textwörtern.

Die Anteile der Zusammenschreibungen in den 3-Jahres-Zeiträumen betragen:

1985-1987:	57%	1997-1999:	49%
1988-1990:	60%	2000-2002:	10%
1991-1993:	63%	2003-2005:	9%
1994-1996:	60%	2006-2008:	2%

Norm (bis 1998): zusammen

Norm (ab 1998): getrennt

Abbildung 2 zeigt diese Ergebnisse in grafischer Darstellung:

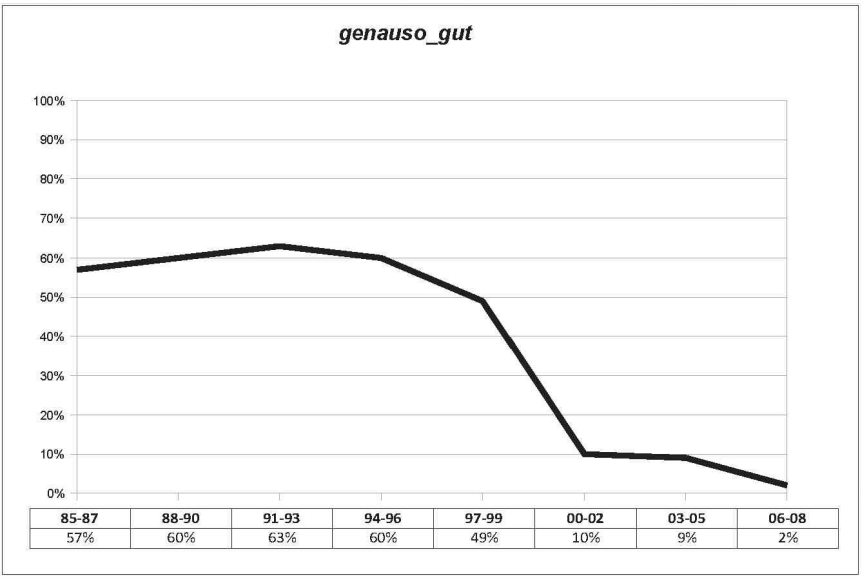


Abb. 2: Anteile der Univerbierungen von *genauso_gut* im DeReKo 1985-2008

Im Zeitraum 1985-1996 zeigt sich ein gleichbleibender Anteil (um 60%) der normentsprechenden Zusammenschreibung bei relativ hohem Anteil der Getrenntschreibungen. D.h. *genauso_gut* zeigt keine hohe Tendenz zur Univerbierung.

Von 1997-2008 katalysiert dann die veränderte Norm die Getrenntschreibung.[13]

4.3 *zu_Hause*

Die Recherche für *zu_Hause* im DeReKo 1985-2008 ergibt folgende Ergebnisse (4.9.2009):

Bei 3 219 Mio. Textwörtern finden sich 282 640 Vorkommen. Davon zeigen 223 690 Getrenntschreibung, 58 959 (= 21%) Zusammenschreibung. Die Frequenz beträgt 87,8 Vorkommen / Million Textwörtern.

Die Anteile der Zusammenschreibungen in den 3-Jahres-Zeiträumen betragen:

1985-1987:	16,05%	1997-1999:	19,24%
1988-1990:	16,44%	2000-2002:	20,63%
1991-1993:	14,45%	2003-2005:	20,38%
1994-1996:	19,72%	2006-2008:	23,69%

Norm (bis 1998): getrennt

Norm (ab 1998): getrennt + zusammen möglich

Abbildung 3 zeigt diese Ergebnisse in grafischer Darstellung:

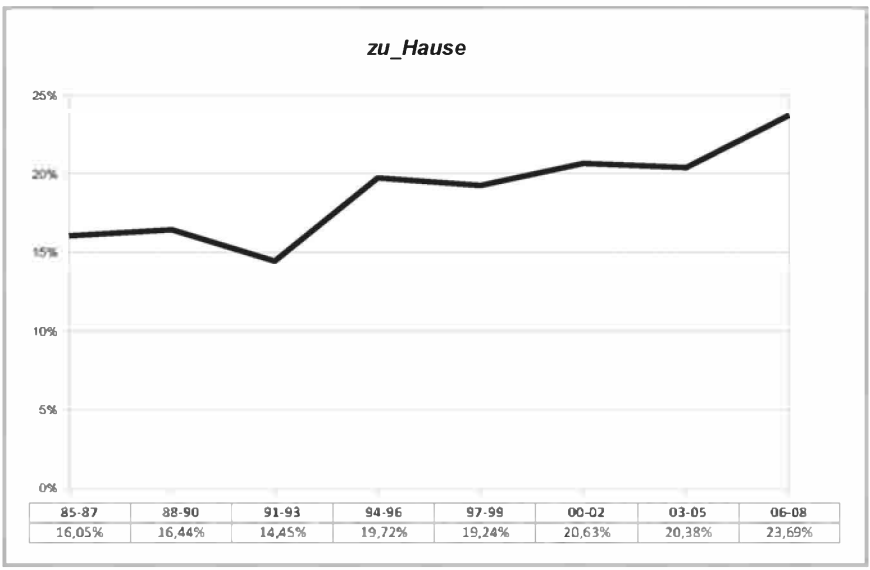


Abb. 3: Anteile der Univerbierungen von *zu_Hause* im DeReKo 1985-2008

[13] Der Duden-Korrektor markiert die Zusammenschreibung.

Über den ganzen Zeitraum 1985-2008 findet sich ein durchgehender leichter Anstieg der Zusammenschreibungen. Man kann dies als leichte Tendenz zur Univerbierung deuten. Die Normliberalisierung hat keinen erkennbaren Einfluss auf diese Tendenz.

Ein Korpusvergleich mit Google Web erbringt folgende Ergebnisse (7.9.09):

Google Web:

27,2 Mio. Vorkommen: Davon 15,3 Mio. Getrenntschreibungen und 11,9 Mio. (=44%) Zusammenschreibungen.

DeReKo 1985-2008:

282 640 Vorkommen: Davon 223 690 Getrenntschreibungen und 58 950 (= 21%) Zusammenschreibungen.

Sowohl die Beobachtung, dass die Anzahl der Vorkommen bei Google Web ca. um den Faktor 100 größer ist, wie auch die Feststellung, dass der Anteil der Univerbierungen deutlich über dem Wert von DeReKo liegt (hier mehr als doppelt so groß ist), sind Befunde, die sich der Tendenz nach auch bei weiteren Recherchen immer wieder belegen lassen.

4.4 *zu_Ende*

Die Recherche für *zu_Ende* im DeReKo 1985-2008 ergibt folgende Ergebnisse (4.9.2009):

Bei 3 219 Mio. Textwörtern finden sich 134 556 Vorkommen. Davon zeigen 132 887 Getrenntschreibung, 1 669 (= 1,24%) Zusammenschreibung. Die Frequenz beträgt 41,8 Vorkommen / Million Textwörtern.

Die Anteile der Zusammenschreibungen in den 3-Jahres-Zeiträumen betragen:

1985-1987:	4,74%	1997-1999:	1,58%
1988-1990:	4,78%	2000-2002:	0,59%
1991-1993:	2,04%	2003-2005:	0,98%
1994-1996:	1,42%	2006-2008:	0,73%

Norm (durchgehend): getrennt

Diagramm 4 zeigt diese Ergebnisse in grafischer Darstellung:

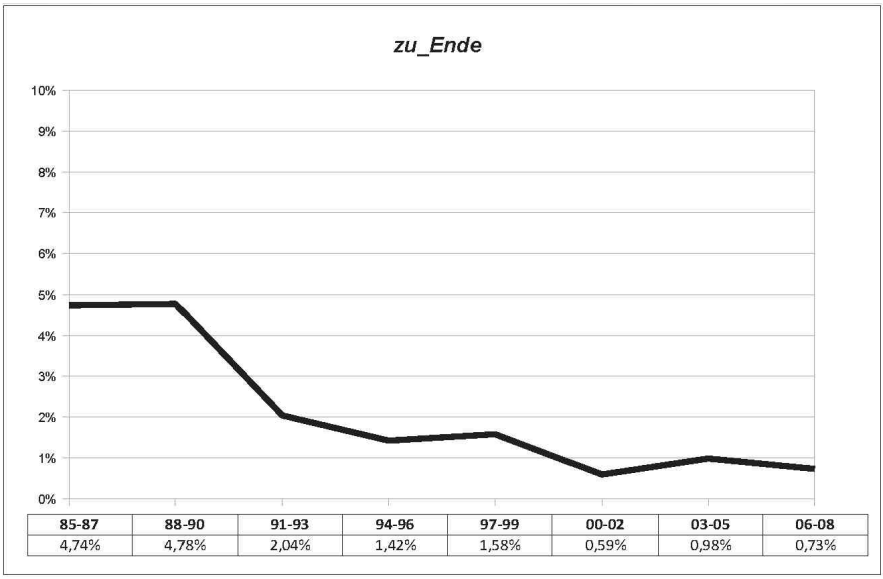


Abb. 4: Anteile der Univerbierungen von *zu_Ende* im DeReKo 1985-2008

Ein Korpusvergleich mit Google Web erbringt folgende Ergebnisse (7.9.2009):

Google Web:

5,287 Mio. Vorkommen: Davon 4,92 Mio. Getrenntschreibungen und 0,367 Mio. (= 6,94%) Zusammenschreibungen.

DeReKo 1985-2008:

134 556 Vorkommen: Davon 132 887 Getrenntschreibungen und 1 669 (= 1,24%) Zusammenschreibungen.

Der Anteil der Univerbierungen von *zu_Ende* ist bei Google Web zwar ca. 5-fach größer, absolut ist er aber in beiden Korpora gering.

Zu_Ende gehört ebenso wie das eben behandelte *zu_Hause* in das Feld der *zu_** Wortfolgen, es zeigt aber deutlich andere Ergebnisse als *zu_Hause*. Über den ganzen Zeitraum 1985-2008 ist eine durchgehende Anpassung auf hohem Niveau an die normentsprechende Getrenntschreibung festzustellen. *Zu_Ende* lässt keine Tendenz zur Univerbierung erkennen.[14] Der Unterschied in den Ergebnissen lässt u.U. darauf schließen, dass *zu_Hause* zunehmend nicht mehr als Präpositionalphrase verstanden wird, sondern als adverbiale Ortsangabe

[14] Der Duden-Korrektor markiert die Zusammenschreibung.

(analog zu bereits abgeschlossenen Univerbierungen wie *daheim* oder *hierzulande*), während *zu_Ende* nach wie vor als Präpositionalphrase mit einem Nomen als Kern gesehen wird.

4.5 *beieinander / beisammen_**

Im Folgenden werden nicht einzelne Wortfolgen betrachtet, sondern die Felder, die *beieinander_** und *beisammen_** jeweils mit den zweiten Bestandteilen *bleiben*, *haben*, *liegen*, *sitzen* und *stehen* bilden. Die folgende Tabelle zeigt die Ergebnisse, die die Recherche im DeReKo in allen Korpora des Archivs W ergibt, im Überblick (8.9.2009):

	*beieinander_*		*beisammen_*	
	zusammen	getrennt	zusammen	getrennt
bleiben	11 (11%)	90 (89%)	40 (29%)	96 (71%)
haben	7 (8%)	79 (92%)	31 (9%)	317 (91%)
liegen	301 (21%)	1164 (79%)	42 (24%)	133 (76%)
sitzen	49 (28%)	127 (72%)	207 (58%)	151 (42%)
stehen	70 (26%)	195 (74%)	37 (45%)	46 (55%)

Norm vor 1998: zusammen
Norm zwischen 1998 und 2006: *beieinander_*: getrennt
*beisammen_*: zusammen
Norm nach 2006: zusammen

Frappierendes Ergebnis ist, dass alle Wortfolgen dieser Felder (bis auf *beisammen_sitzen*) mehrheitlich getrennt geschrieben werden. Und dies, obwohl die Norm (vor 1998 und nach 2006 wieder) die Zusammenschreibung vorsieht. D.h. es ist zu vermuten, dass in diesen Wortfolgen Faktoren wirksam sind, die einer Univerbierung entgegenstehen und die die mehrheitliche Getrenntschreibung – auch gegen die Norm – befördern. Vermutlich ist es die hohe Silbenzahl, die die Zusammenschreibung erschwert.

beisammen_sitzen ist der einzige Fall in diesem Feld, bei dem die Zusammenschreibung überwiegt. Vermutlich ist dies auf die relativ frequente Existenz der Substantivierung *Beisammensitzen* (das nach den Belegen im DeReKo ganz überwiegend ein *gemütliches* ist) zurückzuführen, dessen Zusammenschreibung „abfärbt“.

Umgekehrt weisen *beieinander_haben* und *beisammen_haben* besonders hohe Anteile der Getrenntschreibung auf (92% bzw. 91%). Dies mag daran liegen, dass *haben* in dem Feld *bleiben*, *liegen*, *sitzen* und *stehen* semantisch eine Sonderstellung einnimmt. *haben* ist weniger auf Personen bezogen, als es die anderen Zweitbestandteile sind. Diese Sonderstellung kann unter Umständen die Tendenz zur Getrenntschreibung verstärken.

Die Analyse solcher Fälle, in denen in großen Anteilen bzw. sogar mehrheitlich gegen die Norm geschrieben wird, auf der einen Seite und die Untersuchung von quantitativen Unterschieden in Reihen (**_dessen*) bzw. Feldern (*zu_*; beieinander / beisammen_**) auf der anderen Seite können Aufschluss geben über Faktoren, die eine Univerbierung entweder befördern oder behindern.

5. Univerbierungsfördernde und -hemmende Faktoren

Im Folgenden sollen einige mögliche Kandidaten für univerbierungsfördernde und -hemmende Faktoren benannt werden. Sie sind exemplarisch gemeint. Beim jetzigen Stand der Arbeit können sie nicht anders als hypothetisch und unvollständig sein. Wenn möglich wird auf Beispiele aus den vorstehenden Analysen zurückgegriffen.

Beispiele für mögliche univerbierungsfördernde Faktoren:

- Existenz von Substantivierungen (*beisammen_sitzen*)
- Vorläuferuniverbierungen (*indessen*, *unterdessen* etc. für die Reihe **_dessen*)
- Hauptakzent auf 2, 1 als Auftakt[15] (*zu_Hause*)
- Hauptakzent auf 1 (*allzu_viel, vielmehr*)
- Hohe Kookkurrenz von 1 und 2 (*naja, um_Gottes_willen*)
- Ikonizität (*zusammenschreiben*)
- Auslaut 1 = Anlaut 2 (*warmmachen*)
- Intransparente Morphologie (*zuhauf, zulasten*)
- etc. etc.

Beispiele für mögliche univerbierungsbehindernde Faktoren:

- Hohe Silbenzahl von 1 (*beieinander_**)
- Semantische Andersartigkeit von 2 in einem 2-Feld (*beieinander_haben*)
- Semantische Eigenständigkeit von 1 und 2

[15] 1 meint den ersten Bestandteil der Wortfolge, 2 den zweiten.

- Unterschiedliche Bedeutung von Zusammenschreibung und Getrenntschreibung (*sitzen bleiben* vs. *sitzenbleiben*)
- Ikonizität (*getrennt schreiben*)
- etc. etc.

Die univerbierungsfördernden wie -behindernden Faktoren betreffen alle sprachsystematischen Ebenen. Die Entscheidungen, ob Schreiber(-innen) eine Wortfolge zusammen oder getrennt schreiben, sind – neben dem Normbezug – Resultante des Zusammenspiels dieser Faktoren. Sie sind Erscheinungsformen bzw. Ausdruck eines individuellen wie kollektiven Wortkonzepts. Die Explikation dieser Faktoren trägt dazu bei, zu operationalisieren, was Personen als *ein* Wort verstehen, und so ein empirisch fundiertes Wortkonzept zu erarbeiten.

Literatur

Bredel, Ursula / Günther, Hartmut (2000): Quer über das Feld das Kopfadjunkt. Bemerkungen zu Peter Gallmanns Aufsatz *Wortbegriff und Nomen-Verb-Verbindungen*. In: Zeitschrift für Sprachwissenschaft 19: 103-110.

Brückner, Dominik (2009): Die Google Buchsuche als Hilfsmittel für die Lexikographie. In: Sprachreport 25, 3: 26-31.

Bußmann, Hadumod (1983): Lexikon der Sprachwissenschaft. Stuttgart: Kröner.

Eisenberg, Peter (2004): Das Wort. Grundriss der deutschen Grammatik Bd. 1. 2. Aufl. Stuttgart / Weimar: Metzler.

Fuhrhop, Nanna (2007): Zwischen Wort und Syntagma. Zur grammatischen Fundierung der Getrennt- und Zusammenschreibung. Tübingen: Niemeyer.

Gallmann, Peter (1999): Wortbegriff und Nomen-Verb-Verbindungen. In: Zeitschrift für Sprachwissenschaft 18: 269-304.

Günther, Hartmut (1997): Alles Getrennte findet sich wieder – Zur Beurteilung der Neuregelung der deutschen Rechtschreibung. In: Eroms, Hans W. / Munske, Horst H. (Hg.): Die Rechtschreibreform: Pro und Kontra. Berlin: Erich Schmidt, 81–93.

Herberg, Dieter / Baudusch, Renate (1989): Getrennt oder zusammen? Ratgeber zu einem schwierigen Rechtschreibkapitel. Leipzig: Bibliographisches Institut.

Jacobs, Joachim (2005): Spatien. Zum System der Getrennt- und Zusammenschreibung im heutigen Deutsch. Berlin/New York: de Gruyter.

Jacobs, Joachim (2007): Vom (Un-)Sinn der Schreibvarianten. In: Zeitschrift für Sprachwissenschaft 26: 43-80.

Nerius, Dieter et al. (2000): Deutsche Orthographie. Von einem Autorenkollektiv unter der Leitung von Dieter Nerius. 3. Aufl. Mannheim / Leipzig / Wien / Zürich: Dudenverlag.

Scherer, Carmen (2010): Vom *Friseursalon* zum *Friseur Salon* – Desintegrationsprozesse in der Schreibung von Komposita. Ms.

Stuyckens, Geert / Bröne, Geert (2009): Brauchbarkeit von Korpora des geschriebenen Deutsch für DaF-Lehrende. Eine Fallstudie. In: Deutsch als Fremdsprache 46, 1: 3-9.

Weinberger, Elke (2005): *Kopf stehen* oder *kopfstehen*? Versuch einer grammatischen Sichtung des Grundproblems der Getrennt- und Zusammenschreibung im Bereich der Nomen-Verb-Verbindungen. Seminararb. Univ. Zürich. Internet: www.ds.uzh.ch/lehrstuhlduerscheid/docs/seminararb/weinberger-gzs-05.pdf (Stand: 03 / 2010).

HAGEN HIRSCHMANN

Eine für Korpora relevante Subklassifikation adverbieller Wortarten

Abstract

Die Wortart Adverb wird in vielen Wortartenklassifikationen als eine Art Restkategorie behandelt. Entsprechend wird auch im STTS das Wortartentag „ADV" für syntaktisch sehr unterschiedliche Wörter vergeben. Für viele linguistische Fragestellungen wird jedoch eine feinere Kategorisierung benötigt (so z.B. in Spracherwerbsstudien oder Arbeiten zu Modifikationsstrukturen, Konstituentenabfolgen usw.). Bislang kann nur in den wenigen syntaktisch annotierten Textkorpora systematisch nach adverbiellen Unterklassen gesucht werden.

In diesem Beitrag soll gezeigt werden, wie die heterogene Wortartenklasse Adverb, vertreten durch das STTS-Tag „ADV", so ausdifferenziert werden kann, dass die syntaktisch relevanten adverbiellen Kategorien auch ohne eine aufwändige syntaktische Annotation definiert und gefunden werden können.

Als syntaktische Klassifikationskriterien werden zum einen ein topologisch-distributionelles, zum anderen ein funktionales Kriterium herausgearbeitet und in einer hierarchischen Klassifikation miteinander verknüpft.

Unter topologisch-distributionellen Gesichtspunkten lassen sich genau drei adverbielle Wortklassen definieren: Adverb, Partikel, Modalpartikel. Hier werden die wesentlichen Eigenschaften dieser Wortarten zusammengetragen und es wird diskutiert, wie die Klassen unter funktionalen Kriterien weiter subklassifiziert werden können.

In einem Auswertungsteil wird exemplarisch gezeigt, welche Fragen sich mithilfe der vorgestellten Subklassifikation überhaupt erst bearbeiten lassen und dass sich auch komplexere syntaktische Strukturen mithilfe von Tokenannotationen wie den hier vorgeschlagenen finden lassen.

1. Einleitung

Diese Arbeit behandelt die korpuslinguistisch relevante syntaktische Ausdifferenzierung von adverbiellen Wörtern. „Adverbiell" meint hier zum einen allgemein der 'Restkategorie' Adverb zugehörig, zum anderen werden hiermit alle Wörter bezeichnet, die durch das im Stuttgart-Tübingen-Tagset (STTS; Schiller et al. 1999) enthaltene Wortartentag „ADV" repräsentiert werden. „Adverbielle" Einheiten grenzen sich als Worteinheiten von „adverbialen" Einheiten als syntaktisch-funktionale Einheiten ab. Letztere können Mehr-

wortsequenzen sein (wohingegen „adverbiell" ausschließlich auf Worteinheiten referiert). So sind z.B. *deshalb* und *aus diesem Grund* adverbial gesehen gleichwertige Elemente (sie können beispielsweise als kausale Adverbiale klassifiziert werden); *aus diesem Grund* enthält jedoch keine adverbiellen Wörter.

Korpuslinguistischen Arbeiten steht, so weit mir bekannt, keine Wortartenklassifikation zur Verfügung, die es erlaubt, angemessen fein definierte Klassen adverbieller Wörter (wie beispielsweise Modalpartikeln) zu untersuchen. Allgemein wird in Grammatiken die Wortart Adverb häufig als eine Art Restkategorie behandelt, obwohl es sich (bekannterweise) dabei nicht um eine homogene Wortklasse handelt. So verhält es sich auch im STTS. In dieser Arbeit soll gezeigt werden, dass eine feinere Klassifikation für viele linguistische Fragestellungen sinnvoll und notwendig ist.

Ich beziehe mich hier exemplarisch auf die Kategorie „ADV" im STTS-Tagset, weil es sich als ein Standard für das Deutsche durchgesetzt hat; die Ideen in dieser Arbeit lassen sich jedoch auf alle Klassifikationen beziehen, in denen die Klasse der Adverbien ähnlich grob definiert ist.

Das aus etwa 50 Wortartentags bestehende STTS-Tagset ist in vielen Bereichen recht differenziert. Beispielsweise wird im verbalen Bereich zwischen drei Verbtypen (Voll-, Hilfs- und Modalverb) und fünf flexionsmorphologischen Status (finit, infinit, partizipial, infinit mit eingeschlossenem *zu* und imperativisch) unterschieden, so dass sich im verbalen Bereich insgesamt ein Gesamtinventar von 15 Tags ergibt. Im adverbiellen Bereich gibt es auf den ersten Blick zwar auch unterschiedliche Tags, die fünf Partikel- und zwei Adverbklassen beschreiben, doch sind dies bis auf die Klasse „ADV" morphologische oder syntaktische Sonderklassen (beispielsweise Verbpartikeln, *am* bei Superlativen, *zu* vor Infinitiven, die Antwortpartikeln *ja* und *nein* usw.). Unter „ADV" werden Elemente zusammengefasst, die sich syntaktisch gesehen sehr unterschiedlich verhalten, wie in Kapitel 3 gezeigt wird.[1]

In vielen linguistischen Forschungsfragen sind allerdings nur bestimmte durch die Klasse „ADV" abgedeckte Worteinheiten bedeutsam, wie z.B. in Arbeiten zu doppelter Vorfeldbesetzung oder zu bestimmten Modifikationstypen.

---

[1] Automatische Tagger für das Deutsche (z.B. *TreeTagger*, Schmid 1994, oder *TnT*, Brants 2000) verwenden in der Regel das STTS und machen bei der Zuweisung des Tags „ADV" gemessen an den STTS-Guidelines (Schiller et al. 1999) kaum Fehler.

Andersherum formuliert, ist es kaum vorstellbar, dass eine Forschungsfrage auf die Gesamtklasse aller als „ADV" getaggten Wörter abzielt, da diese Klasse sehr uneinheitlich ist (siehe Kap. 3). Die meisten korpuslinguistischen Arbeiten in diesem Bereich befassen sich mit spezifischen adverbiellen Klassen – siehe z.B. die Arbeiten zu Modalpartikeln bei Lernenden des Deutschen als Fremdsprache von Möllering (2001) oder Vyatkina (2007), Arbeiten über die Grammatikalisierung bestimmter temporaler Adverbien zu Modalpartikeln (Pittner 2009) oder Beiträge zu bestimmten Adverbklassen in verschiedenen Registern (z.B. Rehbock 2009 zu Temporaladverbien in gesprochener Sprache). Alle diese Arbeiten müssen sich bei der Identifikation der behandelten adverbiellen Klassen aushelfen, indem sie lexematische Korpussuchen anstrengen, bei denen nach Oberflächenformen gesucht wird und homonyme Wortformen immer manuell desambiguiert werden müssen (Beispiele in Kap. 3). Das Problem bei solchen lexematischen Suchen ist, dass eine offene syntaktische (ebenso: semantische) Klasse wie z.B. die Modalpartikeln nicht lexematisch definiert bzw. abgedeckt werden kann.

Vor diesem Hintergrund soll hier herausgearbeitet werden, wie sich mit Blick auf die korpuslinguistische Arbeit eine syntaktische Subklassifikation des STTS-Tags „ADV" definieren lässt.

2. Die Repräsentation adverbieller Einheiten im STTS

Die Definition von „ADV" in den „Guidelines für das Tagging deutscher Textcorpora mit STTS" (Schiller et al. 1999) hat folgenden Wortlaut: „Als Adverbien werden nur reine, nicht von Adjektiven abgeleitete, nicht flektierbare Modifikatoren von Verben, Adjektiven, Adverbien und ganzen Sätzen verstanden" (ebd.: 56). Als ein wesentliches Kriterium gilt also ein flexionsmorphologisches (Unflektierbarkeit). Weiterhin werden in den „Guidelines" Pronominaladverbien (wie *damit* oder *darunter*) und Interrogativpronomen (*wo* oder *wann*) und die Negationspartikel *nicht* als gesonderte Klassen definiert.

In der Vergangenheit wurden unterschiedliche Vorschläge zur korpusrelevanten Einteilung von Wortarten im Allgemeinen gemacht (zu einem Vergleich unterschiedlicher Tagsets für das Deutsche siehe Rapp/Lezius 2001, für das Englische bzw. einen internationalen Standard vgl. Atwell 2008). Das STTS-Tagset hat sich als Standard für die meisten frei erhältlichen Korpora des Deutschen etabliert. Bei den Wortarteneinteilungen stehen nicht unbedingt rein

linguistische Kriterien im Vordergrund. In Verbindung mit der Frage nach der Trainierbarkeit von Taggern (automatischen Annotationswerkzeugen) gibt es oftmals pragmatische Entscheidungen bezüglich der Granularität und Beschaffenheit der Wortartenklassen, da sich bestimmte Klassen besser automatisch erkennen lassen als andere (vgl. Rapp / Lezius 2001: 8). Beispielsweise ist davon auszugehen, dass flexionsmorphologisch motivierte (Wortarten-)Klassen in flexionsreichen und syntaktisch weniger restringierten Sprachen wie dem Deutschen präziser zu definieren sind als syntaktisch motivierte Klassen (bestimmte Suffix-Strings weisen mit hoher Wahrscheinlichkeit auf bestimmte Wortartenklassen hin; anhand der Position eines Wortes kann aber nur begrenzt sein syntaktischer Status bestimmt werden). Dies kann mit ein Grund für die vorrangig flexionsmorphologische Kategorisierung von „ADV" und die daraus resultierende syntaktische 'Restkategorie' sein.

Das Wortartentag „ADV" ist also flexionsmorphologisch einheitlich definiert, syntaktisch jedoch unterspezifiziert. Die Unterspezifiziertheit soll hier aufgehoben werden, um Studien zur Syntax adverbieller Wörter eine bessere korpuslinguistische Grundlage zu geben.

Zunächst soll illustriert werden, welche Auswirkungen die STTS-spezifische Kategorisierung von „ADV" und anderen modifizierenden Wörtern hat. Dies sind vor allem zwei grundlegende 'Probleme':

a) Syntaktisch gleichwertige Einheiten werden unterschiedlichen Wortartenkategorien zugeordnet.

b) Die unterschiedlichen als „ADV" annotierten Einheiten repräsentieren syntaktisch eine äußerst heterogene Klasse.

Für Problem a) lassen sich die folgenden Beispiele anführen:

- Einheiten wie *verschieden* und *anders* erhalten unterschiedliche Wortartenzuweisungen (laut STTS „ADJD" und „ADV"), auch wenn sie syntaktisch äquivalent sind (beispielsweise in den Sätzen *Sie sind* ***verschieden*** und *Sie sind* ***anders***, in denen *verschieden* und *anders* dieselbe Art von syntaktischem Wort sind).
- Ähnlich verhält es sich bei dem Beispiel *Das hat sie* ***gekonnt*** *gemacht* und *Das hat sie* ***geschickt*** *gemacht*. Syntaktisch gesehen sind beide Elemente Adverbiale. Morphologisch wird hier jedoch unterschieden: *gekonnt* wird

auf einen Verbstamm (*können*) zurückgeführt, erhält demnach das STTS-Tag „VMPP" (Modalverb, partizipial), *geschickt* (auch wenn es oberflächlich gleich aussieht) wird (nachvollziehbarerweise) als lexikalisiertes Adjektiv analysiert und enthält entsprechend das STTS-Tag „ADJD".

Die Beispiele zeigen, dass – wie bereits betont – bei der Wortartenzuweisung (flexions-)morphologische Kriterien überwiegen. Dies ist unproblematisch, wenn die Nutzer dies wissen, also antizipieren können, welche Einheiten welche Wortartenzuweisung erhalten. 'Versagt' hat ein Wortartensystem dann, wenn die Kriterien zur Wortarteneinteilung uneinheitlich oder uneindeutig und nicht antizipierbar sind. Kriterien zur Wortartenklassifikation können traditionell wortbildungsmorphologisch, flexionsmorphologisch, syntaktisch oder semantisch motiviert sein. Die Kritik an sog. 'kriterienunreinen'[2] Klassifikationen ist alles andere als neu – so findet sich eine gerne zitierte Kritik bereits bei Sütterlin (1923: 97). Er führt aus, die landläufige Wortartenlehre (ursprünglich vor gut 2000 Jahren begründet durch den griechischen Grammatiker Dionysios Thrax) verwende dreierlei Klassifikationskriterien (morphologische, syntaktische und semantische), es dürfe aber nur eines 'gleichzeitig' angewendet werden. Eine aktuellere Darstellung der Problematik befindet sich beispielsweise in Knobloch / Schaeder (Hg.) (2005: 3ff.).

Man kann lange diskutieren, unter welchen Kriterien man Wortarteneinteilungen sinnvollerweise durchführen sollte bzw. welche Kriterien in hierarchischen Klassifikationen in welcher Reihenfolge angewendet werden sollten; letztendlich bleibt dies abhängig von einer konkreten Forschungsfrage oder einem bestimmten linguistischen Rahmenkonzept. (Zur Frage nach bestimmten Korpusannotationen in Abhängigkeit von Forschungsziel bzw. Forschungsfrage vgl. z.B. Garside / Leech / McEnery (Hg.) 1997.)

Die Heterogenität der Gesamtklasse „ADV" (Problem b)) lässt sich gut illustrieren an Suchbelegen von „ADV"-Kandidaten aus dem syntaktisch annotierten TiGer-Korpus (http://www.ims.uni-stuttgart.de/projekte/TIGER/) (das Korpus ist nach den oben genannten STTS-Guidelines (Schiller et al. 1999) wortartengetaggt).

[2] Dieser Begriff meint, dass mehr als ein Klassifikationskriterium in einem Klassifikationsschritt angewendet wird.

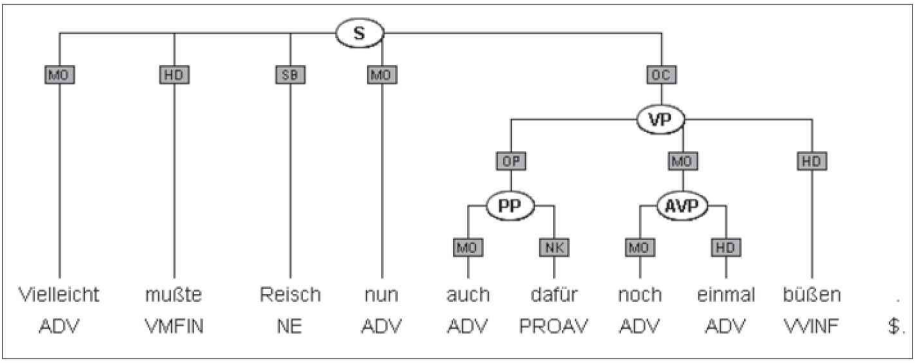


Abb. 1: TiGer Release 2007, Satz 9762
Beispiel für einen im TiGerschema annotierten Satz mit fünf „ADVs" unterschiedlichen Status

Abbildung 1 zeigt fünf als „ADV" getaggte Wörter. In der Analyse des TiGer-Schemas (im Internet unter www.dfki.de/~sbecker/articles/**tiger**_annot.pdf; Stand: 10/2009) haben diese unterschiedlichen syntaktischen Status bzw. unterschiedliche Bindungsrelationen: Zwei der Treffer sind Modifikatoren („MO") des gesamten Satzes, ein Treffer ist Modifikator in einer Präpositionalphrase (die in *dafür* enthaltene Präposition wird als Kopf der Phrase analysiert), die beiden letzten Treffer bilden gemeinsam eine Phrase, in der das erste Element (*noch*) Modifikator des Kopfes (des zweiten Elements *einmal*) ist. Nach dem TiGer-Schema liegen hier bei den fünf als „ADV" getaggten Wörtern also vier unterschiedliche syntaktische Typen vor, gemessen an der syntaktischen Anbindung (Relation zwischen „ADV" und Mutterknoten) und der syntaktischen Funktion (Kantenlabel).

Es ist leicht vorstellbar, dass für eine Forschungsfrage nur eine oder bestimmte dieser Klassen von „ADVs" interessant ist bzw. sind (beispielsweise nur „ADVs" wie *vielleicht*, die auf Satzebene modifizieren). In einem syntaktisch annotierten Korpus wie dem TiGerkorpus kann man, wenn man sich beispielsweise für Fokuspartikeln interessiert, alle Fälle von „ADV" definieren, die unmittelbare Modifikatoren in einer Nominalphrase oder Präpositionalphrase sind, und würde gemäß dem Treffer *auch* in Abbildung 1 fündig werden. Nun sind jedoch die meisten Textkorpora nicht syntaktisch annotiert.[3] In diesen Korpora lassen sich die syntaktischen Klassen nicht wie im TiGerkorpus ausdrücken. Da unter den frei nutzbaren Korpora für das Deutsche ausschließlich Zei-

[3] Da hier vollautomatische Taggingverfahren nicht präzise genug sind und die manuelle Arbeit zu ressourcen- bzw. kostenintensiv ist, ist nicht abzusehen, dass sich dies ändert.

tungskorpora syntaktisch annotiert sind, entsteht so für korpuslinguistische Untersuchungsbereiche, die auf andere Textsorten abzielen (z.B. alle in Kap. 2. genannten Referenzarbeiten), eine Lücke innerhalb der korpuslinguistischen Möglichkeiten.

Komplizierter wird es noch, wenn man sich für Adverbienketten interessiert (komplexe adverbielle oder adverbiale Strukturen). Sucht man nach Abfolgen von als „ADV" getaggten Einheiten – z.B. nach Zweierketten (Bigrammen) – so findet man ähnlich wie in Abbildung 1 unterschiedlichste syntaktische Klassen:

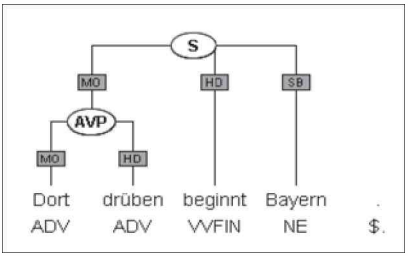


Abb. 2: TiGer Release 2007, Satz 28666
Beispielsatz für aufeinander folgende „ADVs" im TiGerkorpus, die eine Konstituente bilden

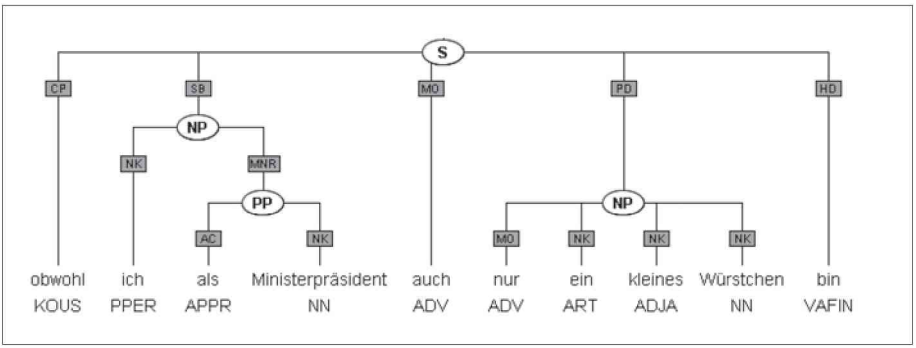


Abb. 3: TiGer Release 2007, Satz 1709
Beispielsatz für aufeinander folgende „ADVs" im TiGerkorpus, die zu unterschiedlichen Phrasen verschiedenen Status gehören

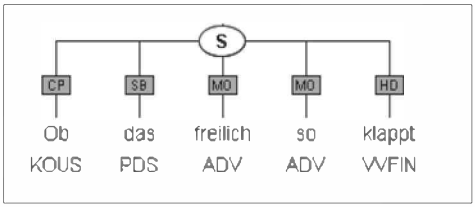


Abb. 4: TiGer Release 2007, Satz 61
Beispielsatz für aufeinander folgende „ADVs" im TiGerkorpus, die unabhängig voneinander auf Satzebene modifizieren

Die Abbildungen 2-4 mit Fundstellen von aufeinander folgenden „ADVs" zeigen, dass sich ganz unterschiedlich komplexe Strukturen dahinter verbergen können: Während in Abbildung 2 eine komplexe Adverbialphrase, bestehend aus zwei „ADVs" , vorliegt, gehören die aufeinander folgenden „ADVs" in den beiden anderen Suchergebnissen (Abb. 3 und 4) nicht zu derselben Phrase bzw. Satzkonstituente. Sie stehen also nur zufällig nebeneinander. Während nun in Abbildung 3 das erste „ADV" satzgebunden ist und das zweite zu einer Nominalphrase gehört, sind in Abbildung 4 beide „ADVs" Satzkonstituenten, die prinzipiell jeweils für sich permutierbar sind.

Auch hier wiederum hätte man ohne die syntaktische Annotation keine Möglichkeit, eine dieser drei Komplexitätsstufen zu finden. Beispielsweise wäre dies aber von Relevanz, wenn man nach Beispielen für doppelte Vorfeldbesetzung (zwei syntaktisch voneinander unabhängige Elemente vor dem finiten Verb im Hauptsatz) suchen wollte.

Viele korpusbasierte Arbeiten helfen sich aus diesem Grund, wie in Kapitel 1 beschrieben, durch die Beschränkung auf lexematische Prototypen. Wenn es allerdings das Ziel einer Untersuchung ist, eine bestimmte *syntaktische* Klasse repräsentativ zu untersuchen, ist dies streng genommen nicht zulässig. Zudem sind alle adverbiellen Einheiten potenziell homonym oder zumindest polyfunktional. Möchte man beispielsweise *doch* als Modalpartikel untersuchen, so besteht nicht nur das Problem, dass die Form auch Subjunktion sein kann. Dies wäre auch nicht problematisch, da an dieser Stelle moderne Tagger bereits verlässlich desambiguieren (vorausgesetzt, die Interpunktion im Text ist zuverlässig). *Doch* als „ADV" ist immer noch ambig, wie der syntaktische Unterschied von *doch* in den Sätzen *Doch kann das nicht wahr sein* und *Das kann doch nicht wahr sein* zeigt.

Im Folgenden gilt es also zu zeigen, wie die Vorkommen von „ADV" in STTS-annotierten Korpora so ausdifferenziert werden können, dass syntaktisch relevante Klassen gefunden werden können.

Da die Wortarteneinteilungen in den meisten Grammatiken Mischklassifikationen im oben dargestellten Sinne sind, bieten sich diese nicht als Grundlage einer konsequent syntaktischen Klassifikation an.

Hier werden nun diejenigen Kriterien aus der einschlägigen Grammatikforschung herausgearbeitet, welche sich zu einer rein syntaktischen Klassifikation adverbieller Einheiten verbinden lassen. Dabei steht vor allem auch eine gute Handhabbarkeit für Annotator(-inn)en (im Sinne von klaren Annotationsrichtlinien) im Vordergrund.

3. Syntaktische Ausdifferenzierung der Klasse „ADV"

Einen frühen Beitrag zur syntaktischen Ausdifferenzierung adverbieller Einheiten bietet Wladimir Admoni in „Der deutsche Sprachbau" (Admoni 1982), an dessen methodischer und terminologischer Tradition sich Helbig / Buscha (2001) orientieren. Hier werden die Klassen „Adverb", „Partikel" und „Modalwort" syntaktisch motiviert. Zum einen werden hierbei Wörter, die syntaktisch-funktional selbstständig, also funktions- oder satzgliedfähig sind, von solchen getrennt, die dies nicht sind. Diese Eigenschaft lässt sich einfach durch Topikalisierung testen (Vorfeldprobe) – sofern das Wort alleine im Vorfeld stehen kann, ist es (nach der Terminologie von Admoni 1982 und Helbig / Buscha 2001) Adverb oder Modalwort; sofern dies nicht der Fall ist, handelt es sich um eine Partikel. Die Stellungseigenschaft der Partikeln ist, dass sie fest an eine bestimmte Position innerhalb einer Mutterphrase gebunden sind und nur mit dieser gemeinsam permutiert werden können.

(1) *Das hat ihr **sehr** gefallen.*

(2) → ***Sehr** hat ihr das gefallen.*

→ *sehr* hat hier den Status eines Adverbs / Modalworts.

(3) *Das hat ihr **sehr** gut gefallen.*

(4) → ****Sehr** hat ihr das gut gefallen.*

→ *sehr* hat hier den Status einer Partikel.

An diesen Beispielen wird deutlich, dass sich ein Wort (wie es prinzipiell bei allen syntaktischen Klassifizierungen der Fall ist) auf seine syntaktischen Eigenschaften nur in einem aktuellen Verwendungskontext testen lässt. Das Phänomen, dass eine Wortform an unterschiedlichen syntaktischen Positionen auftritt, kann von Fall zu Fall mit Homonymie oder Polyfunktionalität beschrieben bzw. erklärt werden.

Bei der Unterscheidung von Adverbien und Modalwörtern müssen, Pittner (1999) und Helbig/Buscha (2001) folgend, andere Kriterien herangezogen werden als bei der Unterscheidung von Partikeln und diesen beiden Wortklassen, denn „Modalwörter unterscheiden sich morphologisch und in den Stellungseigenschaften nicht von den Adverbien" (Helbig/Buscha 2001, S. 430). Das Kriterium für die Unterscheidung ist der von dem Wort modifizierte Bereich (der Skopus). Beim Adverb ist dies (terminologisch konsequent) das Verb (bzw. die Verbalphrase, die je nach Syntaxmodell unterschiedlich beschrieben werden kann); beim Modalwort hingegen ist der Skopus der gesamte Satz, d.h. die Modifikation bezieht sich nicht unmittelbar und ausschließlich auf die Verbalhandlung, sondern auf einen Bereich außerhalb (in hierarchischen Syntaxmodellen: oberhalb) der VP.

(5) ..., S[*weil Jessica* ***vielleicht*** VP[***alleine*** *Bücher kauft*]].

Man kann dementsprechend mit „*Wie/Wo/Wann geschieht etwas?*" bzw. „*Wie/Wo/Wann tut X etwas?*" keine Modalwörter erfragen, sondern lediglich Adverbien, die auf die VP wirken. Auf die Frage „*Wie kauft Jessica Bücher?*" kann man in (5) demnach lediglich *alleine* erfragen, nicht jedoch *vielleicht*.

Diese beiden Modifikationsstrukturen kann man nicht nur mittels Frageprobe unterscheiden, sondern anhand verschiedener Testverfahren (z.B. auch durch die Überprüfung von Ersetzbarkeit u.a.), die in Pittner (1999) zusammengefasst werden (vgl. ebd.: 109).[4]

Vielleicht setzt also die gesamte Proposition in einen spezifischen (von der Sprechereinstellung abhängigen) Wahrscheinlichkeitsmodus, befindet sich also außerhalb der Verbalphrase, wohingegen *alleine* lediglich die Verbalhandlung modifiziert, also innerhalb der Verbalphrase liegt. Dieser Unterschied kann sowohl in semantischen als auch in syntaktischen Rahmenkonzepten be-

[4] Pittner benutzt in ihrer Arbeit die Termini „Modaladverbial" (entspricht hier „Adverb") und „Satzadverbial" (entspricht hier „Modalwort"), weil sich ihre Klassifikation auf syntaktische Einheiten bezieht, die wie hier Wörter, aber auch komplexere Phrasen oder Sätze sein können.

handelt werden und findet in mehreren grammatischen Modellen durch die Annahme und Darstellung voneinander abweichender Strukturen Berücksichtigung (in generativen Formaten beispielswiese durch VP- vs. IP- vs. CP-Attachment). Syntaktisch können solche Unterschiede sowohl in Phrasenstrukturmodellen als auch in Dependenzmodellen dargestellt werden.

Die bisher vorgestellten Klassen sind zwar rein syntaktisch, es sind aber unterschiedlich motivierte Klassen: Die Unterscheidung zwischen Adverbien / Modalwörtern und Partikeln geschieht unter distributionell-topologischen Gesichtspunkten, die Unterscheidung zwischen Adverbien und Modalwörtern kann als funktional motiviert beschrieben werden.

Nach diesen syntaktischen Klassifikationskriterien lassen sich auch die bisher noch nicht erwähnten Modalpartikeln (manchmal auch Abtönungspartikeln genannt) definieren. Sie sind topologisch eingeschränkt bis fixiert, können nicht topikalisiert werden, stehen aber nicht wie die Partikeln in Beziehung zu einer (lexikalischen) Mutterphrase. Beispiele sind

(6) *Das kann **ja** nicht wahr sein!*

(7) *Du bist mir **vielleicht** ein Charmeur!*

Sowohl *ja* als auch *vielleicht* können nicht in das Vorfeld permutiert werden, ebenso wenig können sie im Mittelfeld an eine andere Stelle gesetzt werden. Von der oben definierten Klasse der Partikeln unterscheidet sie die Eigenschaft, dass sie auch nicht in einer anderen (permutierbaren) Phrase integriert sind. Die beiden Klassen haben gemein, dass sie gewissermaßen die gesamte Satzphrase modifizieren.

Diese beiden hier vorgestellten syntaktischen Klassifikationskriterien dürfen ebenso wenig vermischt werden wie bspw. morphologische und syntaktische Klassifikationskriterien. Wie bereits in Kapitel 1 argumentiert wurde, können funktionale Klassen (wie kausales Adverbial oder Subjekt) gleichermaßen auf Wörter und Wortgruppen angewendet werden. Deswegen wäre es nicht sinnvoll, eine entsprechende Klassifikation nur auf Worteinheiten zu beziehen, sondern man würde größere Konstituenten mit einbeziehen. Die vorgestellten distributionellen Klassen hingegen sind wortspezifisch. Aus diesem Grund beschränkt sich die folgende Klassifikation hauptsächlich auf dieses Kriterium. Lediglich bei der Unterscheidung zwischen Adverbien und Modalwörtern wird der funktionale Aspekt berücksichtigt.

Analysiert man die adverbiellen Wörter strikt nach distributionellen Gesichtspunkten, so lassen sich die Vorkommen von „ADV“ zunächst in drei unterschiedliche Klassen einteilen:

a) Adverbien und Modalwörter

Sowohl Adverbien als auch Modalwörter können als Satzkonstituente im Vorfeld von Hauptsätzen auftreten. Die Unterscheidung zwischen Adverbien und Modalpartikeln kann unter distributionellen Gesichtspunkten nicht vorgenommen werden, weil die Wörter an denselben Positionen im Satz auftauchen können.

(8) ***Wahrscheinlich*** *war sie* ***dort*** *auch mit Peter schon einmal.*

(9) ***Dort*** *war sie* ***wahrscheinlich*** *auch mit Peter schon einmal.*

Sowohl *wahrscheinlich* als auch *dort* können alleine im Vorfeld des Satzes stehen oder auch variabel im Mittelfeld auftreten. Somit besitzen sie denselben distributionellen Rahmen. Voneinander unterschieden werden können sie erst auf einer Beschreibungsebene, die semantische, funktionale oder dependenzielle Aspekte berücksichtigt. Aus diesem Grund wird hier eine hierarchische Klassifikation erarbeitet, bei der auf die distributionelle Analyse (gemäß (8) und (9)) eine funktionale (gemäß (5)) folgt. Auf diese Weise kann *wahrscheinlich* in (8) und (9) der Status Modalwort und *dort* der Status Adverb zugeordnet werden.

b) Partikeln

Hierbei handelt es sich um die Partikelklassen, die unter den Termini „Grad-“, „Fokus-“, „Intensivierungs-“ oder „Steigerungspartikel“ behandelt werden. Manchmal werden diese differenzierend, manchmal austauschbar bzw. synonym verwendet (z.B. werden bei Hoffmann (Hg.) 2007 die „Gradpartikeln“ als eine Klasse von Wörtern behandelt (und dreifach ausdifferenziert), die in anderen Grammatiken oftmals als „Fokuspartikeln“ abgehandelt werden (vgl. z.B. Eisenberg 2004: 232f.).

Allgemein resultieren die unterschiedlichen Wortartenbeschreibungen in wissenschaftlichen Beiträgen aus verschiedenen Beschreibungskriterien und verschiedenen Terminologien. Oftmals sind, wie bereits erwähnt, die Einteilungs- bzw. Abgrenzungskriterien nicht klar formuliert; zudem werden unterschiedliche Termini für dieselbe Klasse verwendet. Zu dieser Problematik,

die man auch als Standardisierungsproblem interpretieren kann, und konkreten Diskrepanzen zwischen relevanten Beiträgen vgl. Eisenberg (2004: 231).

Für eine rein syntaktisch motivierte Klassifikation der Partikeln wird hier vorgeschlagen, analog zu der Klasse der Adverbien zuerst ausschließlich nach distributionellen Gesichtspunkten zu klassifizieren. Anschließend kann wie bei a) zusätzlich syntaktisch-funktional bzw. dependenz- oder skopusspezifisch unterschieden werden. Gemäß dem ersten Kriterium sind Partikeln topologisch abhängig von einer Mutterphrase, die sie modifizieren – sie sind in einer übergeordneten Phrase (Nominal- bzw. Determiner-, Adjektiv-, Präpositional oder Adverbialphrase) fixiert (meistens am linken Rand, selten am rechten) und können nur mit ihr zusammen permutiert werden.

(9) *Ihre Prüfung war* AP[***besonders / sehr / ziemlich / fast*** *gut*].

(10) AP[***Besonders / Sehr / Ziemlich / Fast*** *gut*] *war ihre Prüfung.*

Alle Elemente, die hier *gut* als Kopf der Adjektivphrase modifizieren, haben distributionell denselben Wert, wie auch immer sie semantisch subklassifiziert werden können: Sie müssen links von *gut* stehen und entsprechend mitverschoben werden, sofern *gut* beispielsweise topikalisiert wird. Einige (nur eine aus (9) bzw. (10)) können ebenso links in einer Präpositionalphrase stehen:

(11) PP[***Besonders / *Sehr / *Ziemlich / *Fast*** *über ihre Prüfung*] *freut sie sich.*

Partikeln können also nach ihrer Anbindung im aktuellen Satzkontext spezifiziert werden. Dies entspricht der Spezifizierung des Skopus; in (9) bzw. (10) ist der Skopus der Partikel eine Adjektivphrase, in (11) eine Präpositionalphrase. Zur Bestimmung des Fokus und Skopus bei Fokuspartikeln vgl. Dimroth / Klein (1996).

Leicht lässt sich der hochgradig ambige Charakter der adverbiellen Wörter zeigen, wenn man (9) und (10) mit folgendem Satz kontrastiert:

(12) *Sie freut sich ziemlich über ihre Prüfung.*

In (10) zeigt sich durch die gemeinsame Vorfeldposition von *ziemlich* und *gut*, dass *ziemlich* Modifikator von *gut* ist, also Partikelstatus hat. In (12) steht *ziemlich* links von der Präpositionalphrase, könnte hier also strukturell gesehen auch Partikel (einer PP) sein. Dies wird aber in (11) ausgeschlossen. *Ziemlich* besitzt in (12) also den Status einer Satzkonstituente bzw. eines Modalworts:

(13) *Sie freut sich* AdvP[*ziemlich*] PP[*über ihre Prüfung*].

c) Modalpartikeln

In der Literatur werden diese Elemente oftmals auch als 'Abtönungspartikeln' bezeichnet (was eine semantische Klasse impliziert). Den Elementen dieser Klasse ist syntaktisch gemein, dass sie nicht topikalisierbar sind, sondern im Mittelfeld meistens an einer Position nahe dem linken Rand (je nach Vorhandensein bestimmter anderer Konstituenten) auftreten müssen.

(14) *Sie kann sich das **ja** / **halt** / **wohl** / ... erlauben.*

(15) ***Ja** / ***Halt** / ***Wohl** / ... kann sie sich das erlauben.*

Aufgrund ihrer Nicht-Vorfeldfähigkeit können die Modalpartikeln nicht als satzgliedfähig / phrasenfähig angesehen werden. Aus diesem Grund werden sie oftmals gemeinsam mit den Partikeln bzw. als Unterklasse von ihnen behandelt. Hier wurde allerdings gezeigt, dass die als Partikel definierte Klasse topologisch abhängig von einem syntaktisch-lexikalischen Kopf ist. Modalpartikeln haben keinen solchen. Partikeln modifizieren 'untere' syntaktische Konstituenten (sie attribuieren AdvPs, APs, NPs bzw. DPs oder PPs), Modalpartikeln hingegen modifizieren wie die Modalwörter immer außerhalb der VP auf einer der „obersten" syntaktischen Ebenen (z.B. Satz; siehe (5)).

d) Konjunktionalpartikeln

Ein Sonderfall entsteht bei der Anwendung des STTS in deutschen Textkorpora dadurch, dass die koordinationsbeendenden Einheiten *usw.* und *etc.* bspw. von *TreeTagger* als Token bzw. Wörter behandelt werden. Folgerichtig müssen sie einen der vom STTS vorgesehenen Werte erhalten. Hierbei passt die Kategorie „ADV" unter allen Möglichkeiten noch am besten (auch in diesem Zusammenhang bestätigt sich die Klasse „Adverb" als Restkategorie). Dass nicht auch Einheiten wie *o.Ä.* als „ADV" getaggt werden, liegt daran, dass sie beim Tokenisieren in zwei Token getrennt werden (*o._Ä.*), denen dann plausible Tags zugeordnet werden können. Die graph(emat)ische Repräsentation von *usw.* und *etc.* verhindert dagegen eine weitere Zerlegung bzw. macht eine Behandlung dieser Einheiten als Wörter / Token plausibel. Das STTS kann, wie alle anderen Wortartensysteme, diese Einheiten nicht angemessen einordnen. Syntaktisch gesehen handelt es sich bei ihnen um einen Zweifelsfall von Wörtern, welche in ihren Eigenschaften zwischen Konjunktor und Konjunkt stehen. Diese Sonderfälle werden in keiner mir bekannten Grammatik behandelt.

Da sie keiner der bisher genannten Klassen zugeordnet werden können, werden sie einer Sonderklasse „Konjunktionalpartikel" (dies ist hier eher ein Arbeitsterminus) zugeordnet. Diese resultiert gemäß dem distributionellen Klassifikationskriterium aus der fixierten Stellung der Einheiten am rechten Rand von Koordinationen.

Für eine fundiertere Einordnung und Benennung bedürften diese Einheiten einer entsprechenden Analyse vor allem unter graphematischen und syntaktischen Gesichtspunkten.

Hiermit wären die distributionell und (hinsichtlich Adverbien) funktional unterscheidbaren „ADV"-Klassen genannt. In Kapitel 5 wird ein entsprechendes Annotationsschema für adverbielle Wörter vorgestellt.

Es gibt bezogen auf die besprochenen Klassen einige schwer zu klassifizierende Wörter. Viele Problemfälle entstehen durch die Tendenz einiger Modalwörter, selten im Vorfeld aufzutreten (*doch*, *auch*, *zwar* und andere). *Doch* als Modalwort konkurriert wie *aber* mit der Verwendung als Subjunktion, es kann jedoch auch die Vorfeldposition einnehmen.

(15) ***Doch*** *hat er sich trotzdem bemüht.* (Verwendung als Modalwort)

vs.

(16) ***Doch*** *er hat sich trotzdem bemüht.* (Verwendung als Subjunktion)

Aber ist noch schwieriger zu klassifizieren. Steht es satzinitial, so ist es immer Subjunktion (es steht links vom Vorfeld und kann nicht wie *doch* die Vorfeldposition einnehmen). Die Form kann jedoch auch in zwei unterschiedlichen Verwendungsweisen im Mittelfeld auftreten:

(17) *Sie war krank, hat die Prüfung* ***aber*** *abgelegt.*

(18) *Das hat sie sich* ***aber*** *schön zurechtgelegt!*

In (17) wird *aber* wie die adversativen Modalwörter *trotzdem, dennoch, jedoch* usw. verwendet. Es zeigt dieselben distributionellen Eigenschaften wie diese Wörter, außer dass es nicht vorfeldfähig ist (es scheint bei Topikalisierung 'automatisch' in die Subjunktionsposition links außerhalb des Vorfeldes zu geraten). So kann es als nicht vorfeldfähiges Modalwort klassifiziert werden (im Handbuch der deutschen Konnektoren wird es mit einer Reihe anderer Wörter wie *bereits* als nicht vorfeldfähiger Adverbkonnektor klassifiziert, vgl. Pasch et al. 2003: 4ff.).

In (18) zeigt *aber* eine dritte Verwendungsweise – die der Modalpartikel. Der Unterschied besteht in der Satzgliedwertigkeit (bzw. Phrasenhaftigkeit); *aber* in (17) ist innerhalb des Mittelfeldes permutierbar wie andere Modalwörter, in (18) besitzt es hingegen keinen Satzgliedwert, da es ausschließlich an einer bestimmten Mittelfeldposition in Exklamativsätzen auftreten kann (es besitzt hierbei auch nicht die adversative Bedeutung des Modalwortes).

Bei der Klassifikation von *auch* treten häufig zwei konkurrierende Lesarten auf:

(19) *Er wird morgen* ***auch*** *im Schwimmbad sein.*

In Sätzen wie (19) gibt es folgende Interpretationen für *auch*:

(20) ***Auch*** *im Schwimmbad wird er morgen sein.*

(21) ***Auch*** *er wird morgen im Schwimmbad sein.*

(22) ***Auch*** *wird er morgen im Schwimmbad sein.*

Gemäß diesen Lesarten wäre *auch* in (20) und (21) als Partikel (einmal mit PP-Skopus, einmal mit NP / DP-Skopus) und in (22) als Modalwort zu klassifizieren. In Fällen, in denen man aufgrund des Äußerungskontextes nicht desambiguieren kann,[5] muss man zu einer Defaultlösung greifen, sofern man Ambiguität nicht annotieren kann.

4. Annotationsrichtlinien

In Kapitel 3 wurde gezeigt, dass man für die syntaktische Ausdifferenzierung der flexionsmorphologisch motivierten Wortart „ADV“ des STTS zwei Bestimmungskriterien definieren kann: ein topologisch-distributionelles, welches nach (potenziellen) Stellungspositionen der Einheiten fragt, und ein syntaktisch-funktionales, welches nach dem Skopus bzw. nach der Bindungsrelation des Modifikators fragt. Das zweite Kriterium ist in der hier erarbeiteten Klassifikation nur relevant für die Unterscheidung zwischen Adverbien und Modalwörtern. (Es könnte auch verwendet werden, um beispielsweise zu spezifizieren, an welcher Art von Kopf eine gegebene Partikel adjungiert.)

Aus den in Kapitel 3 erarbeiteten „ADV“-Unterklassen, den entsprechenden syntaktischen Eigenschaften und genannten Tests lässt sich das folgende Annotationsschema ableiten.

[5] Beispielsweise würde Evidenz für (21) vorliegen, wenn in (19) der Satzakzent auf *auch* liegt

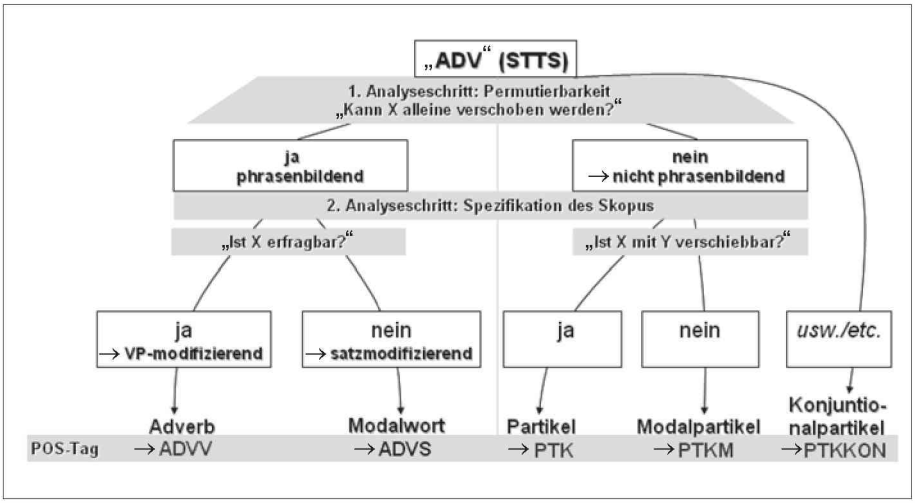


Abb. 5: Guidelines für die Annotation adverbieller Einheiten (nach STTS)

Gemäß den vorangegangenen Betrachtungen gibt es fünf zu vergebene Tags:

Ist ein „ADV“ alleine topikalisierbar, so wird es mithilfe der in Kapitel 3 genannten Tests entweder als VP-modizierend (→ „ADVV“) oder als satzmodifizieriend (→ „ADVS“) kategorisiert. Ist dies nicht der Fall, so wird geprüft, ob es einen Kopf modifiziert, mit dem es gemeinsam permutiert werden kann. In diesem Fall handelt es sich um eine Partikel (→ „PTK“). Trifft dies nicht zu, handelt es sich um eine Modalpartikel (→ „PTKM“). Da *usw.* und *etc.* eine geschlossene Klasse darstellen, kann hier formbasiert vorgegangen werden (distributionell gilt für diese Einheiten, dass sie am rechten Rand von Koordinationen auftreten).

Das vorgeschlagene Vorgehen soll an folgendem Satz illustriert werden:

(23) *Zwar können heute selbst Milchpreise ja immer weiter fallen , aber …*
ADV VMFIN ADV ADV NN ADV ADV ADV VVINF $, KON

Die Wortartenwerte in (23) wurden gemäß den STTS-Guidelines (Schiller et al. 1999) vergeben und entsprechen dem Output des *TreeTaggers*.

Gemäß den hier vorgeschlagenen Guidelines werden nun die unterschiedlichen „ADV“-Klassen der sechs „ADV“-Token ausdifferenziert:

(24) *Zwar können heute selbst Milchpreise ja immer weiter fallen , aber …*
ADVS VMFIN ADVV PTK NN PTKM PTK ADVV VVINF $, KON

Beispiel (24) zeigt, dass alle der „ADV"-Subklassen in (23) enthalten sind: Das im Vorfeld stehende (demnach satzgliedwertige) *zwar* ist ein Modalwort (nicht erfragbar), *heute* ist ein Adverb (topikalisierbar und erfragbar), *selbst* erhält den Wert für eine Partikel (mit *Milchpreise* zusammen permutierbar), *ja* für Modalpartikel (nicht topikalisierbar, mit keiner anderen Phrase permutierbar), *immer* für Partikel (mit *weiter* zusammen permutierbar) und *weiter* für Adverb (als Kopf von *immer weiter* topikalisierbar und erfragbar).

5. Auswertung eines Testkorpus und Fazit

Um exemplarische Suchen spezifischer adverbieller Kategorien illustrieren zu können, wurde ein Testkorpus von 45620 Token auf die beschriebene Weise bearbeitet. Das Korpus besteht aus Essaytexten, weil bei einer argumentativen Textsorte wie dieser davon auszugehen ist, dass viele unterschiedliche Modifikatoren verwendet werden.

In den annotierten Daten kann man nun die Verwendung der hier definierten Klassen quantitativ vergleichen. Die Verteilung der Klassen Adverb, Modalwort, Partikel und Modalpartikel (die Klasse Konjunktionalpartikel soll an dieser Stelle keine Gewichtung erhalten) lässt sich gemäß Abbildung 6 darstellen:

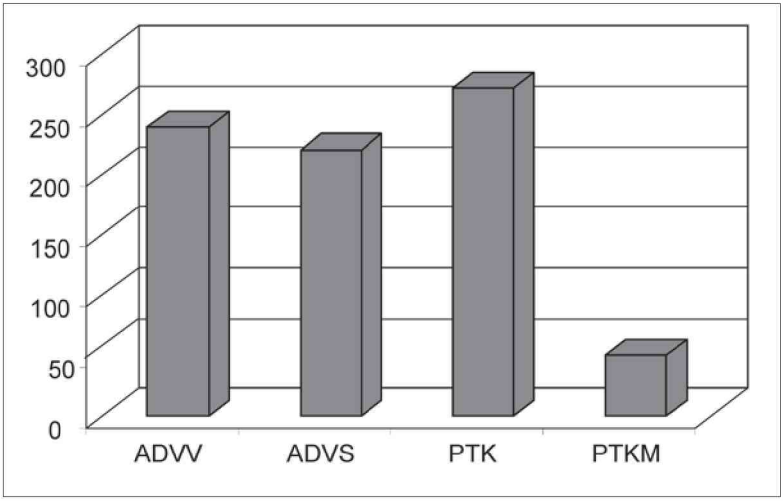


Abb. 6: Quantitative Verteilung von Adverbien (ADVV), Modalwörtern (ADVS), Partikeln (PTK) und Modalpartikeln (PTKM) im Testkorpus (Zahlen sind normalisiert auf 10000 Token)

Abbildung 6 zeigt, dass die Verwendung von Adverbien, Modalwörtern und Partikeln in dem untersuchten Essaykorpus recht ausgewogen ist. Erwartungsgemäß liegt die Anzahl der verwendeten Modalpartikeln deutlich unter der der anderen drei Klassen. Die Verteilung der vier Klassen kann als individuelles Profil der untersuchten Textsorte angesehen werden (in anderen Textsorten, Genres oder Registern sind andere Verteilungen zu erwarten). Wie allgemein bekannt ist, treten in gesprochenen Texten Modalpartikeln deutlich häufiger auf als in geschriebenen. Deshalb ist in diesem Kontext beispielsweise die Frage interessant, welche Modalpartikeln dennoch in geschriebenen Texten auftauchen (wie in dem untersuchten Testkorpus) und was die Faktoren für ihr Auftreten sind. Die Frage nach textsortenspezifischen Verwendungen bestimmter adverbieller Klassen und Lexeme ist ohne quantitative Studien wie den hier angedeuteten nicht zu beantworten.

Wortform	Vorkommen ADVV	Vorkommen ADVS	Vorkommen PTK	Vorkommen PTKM
aber	–	16,00	–	0,20
eigentlich	–	5,92	–	0,43
erst	0,88	–	3,29	–
immer	10,96	–	10,08	–
mehr	5,92	–	12,28	–
noch	6,58	–	14,03	–
nur	–	1,75	32,22	–
schon	4,60	–	7,45	2,85
selbst	1,32	–	5,25	–
so	11,92	2,38	17,97	–
sogar	–	1,10	3,29	–
überhaupt	–	0,31	2,19	2,41
viel	2,41	–	5,48	–

Tab. 1: Vorkommen von ambigen Wortformen als Adverbien (ADVV), Modalwörtern (ADVS), Partikeln (PTK) und Modalpartikeln (PTKM) (Zahlen sind normalisiert auf 10 000 Token)

Wertet man die hier vorgestellten syntaktischen Klassen auf Lexeme und ihre Frequenzen hin aus, so ist zum einen festzustellen, dass es klassenspezifische Lexeme gibt wie z.B. *morgen* als reines Adverb, *äußerst* als reine Partikel oder *ja* als reine Modalpartikel (die Antwortpartikel *ja* erhält das STTS-Tag „PTKANT", fällt demnach gar nicht erst in die ausdifferenzierte Klasse „ADV"). Auf der anderen Seite ist zu sehen, dass viele Wortformen in mehreren Kategorien auftreten. Man kann nun nach Verteilungen von Kategorien bei polyfunktionalen Wortformen suchen, wie Tabelle 1 illustriert.

Zum einen fallen an der Tabelle 1 diejenigen Formen auf, welche in drei der vier Verwendungsweisen vorkommen (*schon*, *so* und *überhaupt*). Beispiele aus dem Korpus, die die unterschiedlichen Verwendungen von *schon* illustrieren, sind folgende:

(25) *..., und* ***schon*** *hat er die Grenze erreicht.*

(26) *Die meisten haben* ***schon*** *in jungen Jahren mit der Polizei zu tun, ...*

(27) *... wer würde den eigenen Beruf* ***schon*** *als „weniger nützlich für die Gesellschaft" einstufen?*

Beispielsatz (25) zeigt die Verwendung von *schon* als Adverb, Satz (26) die Verwendung als Partikel, Satz (27) die als Modalpartikel.

Im Fall von *überhaupt* liegt die hier sichtbar werdende Polyfunktionalität intuitiv nicht unbedingt auf der Hand und wird erst bei der Sichtung von Belegen deutlich. So verbergen sich hinter der Verwendung als Partikel relativ häufig gebrauchte Konstruktionen wie *überhaupt kein* oder *überhaupt nicht*. Ähnlich verhält es sich bei der sehr häufigen Verwendung von *immer* als Partikel; hierbei handelt es sich um Verbindungen wie *immer wieder*, *immer öfter* usw. Erstaunlich ist, dass diese Verwendungsweise im Essaykorpus fast genauso häufig ist wie die des temporalen Adverbs. Dieses Beispiel zeigt deutlich die Vorteile der Quantifizierbarkeit der „ADV"-Typen: Es können nicht nur Lexeme syntaktischen Wortklassen (und umgekehrt) zugeordnet werden, sondern auch Aussagen über Häufigkeiten von bestimmten Verwendungen in unterschiedlichen Textsorten, Registern usw. gemacht werden, was vorher in dieser Weise unmöglich war. Dies ist beispielsweise für Studien in der Varietätenlinguistik oder in der Fremdsprachenlehre relevant

(z.B. für die Frage, welche Lexeme in welchen Verwendungsweisen vermittelt werden sollten).

Mithilfe der vorgestellten Subklassifikation adverbieller Wörter können nicht nur die vier syntaktisch relevanten Oberklassen (Adverb, Modalwort, Partikel, Modalpartikel) gefunden werden, sondern auch bestimmte syntaktische Konstruktionen, wie sie in Kapitel 2 angesprochen wurden.

Als weitere Beispiele für Anwendungen soll kurz illustriert werden, wie aus den Daten bestimmte syntaktische Modifikationsklassen extrahiert werden können.

Die erste Suche bezieht sich auf adverbiell komplexe Strukturen im Sinne einer Partikel, die ein Adverb attribuiert. Beide zusammen sollen eine Konstituente im Satz bilden. Mithilfe einer entsprechenden Suche nach einer Partikel, gefolgt von einem Adverb, lässt sich aus dem Testkorpus folgende Liste (für rechtsköpfige Einheiten; ebenso lässt sich Linksköpfigkeit definieren) generieren:

Match
bewegt so viele Menschen dazu **immer wieder** auf Gewalt oder derartige Mittel
zu tun , da sie **immer wieder** mit dem Gesetz in Konflikt
Abgesehen von den Strafen die **immer wieder** auf einen warten ist es
hat , ist es doch **noch lange** kein Grund Kriminäl zu werden
Lösung , weshalb die Kriminalitätsrate **immer weiter** ansteigt . Es wir nie
. Ich werde es auch **fast nie** verstehen . Geldangelegenheiten sind immer
werden würde , würde es **weniger oft** zu Diskussionen um das liebe
sicher , es ist auch **ganz oft** die Langeweile , der Nervenkitzel
Villa in der Karibik ... **Immer wieder** bekommt man in den Zeitungen
und dass das Leben denn **erst recht** von noch mehr Problemen bedrückt
die vorher schon Exestierten sind **immr noch** da , vieleicht auch weg
leidet , unmöglich , sich **ebenso lange** auf seine Arbeit zu konzentrieren

Abb. 7: Ausschnitt aus der Trefferliste komplexer rechtsköpfiger adverbieller Einheiten im Testkorpus

Es lässt sich ebenso eine Frequenzliste solcher komplexer Phrasen erstellen.

Die zweite Suche bezieht sich auf Partikeln, die eine Präpositionalphrase attribuieren. Mittels der Suche nach einer Partikel, gefolgt von einer Präposition lässt sich die folgende Belegliste erstellen:

Match
Arme Menschen leben und sich **nur durch** Kriminalität " über Wasser "
Neueinsteiger " werden sehr oft **nur mit** Geldstrafen und Haus- oder Ladenverbot
zu leisten und kommt vielleicht **auch mit** Gewalt an Sachen , die
bekommen . Die meisten haben **schon in** jungen Jahren mit der Polizei
ihr Eigentum eventuell hart und **besonders auf** legalen Wegen erarbeiten mussten .
immer ein heikles Thema . **Besonders in** der heutigen Gesellschaft , wo
sein würden . Es sollte **nur nach** der erbrachten Leistung geguckt werden
. Mädchen gehen wenn überhaupt **nur bis** sie 14 Jahre sind zur
schätzen gelernt wird , möglicherweise **erst nach** dem Ableben entsprechender Person .
legen somit oft einen Grundstein **erst für** weitere Generationen . Wie sollte
Wert , sind aber auch **nur aufgrund** der ursprünglichen Leistung des Wissenschaftlers
immer mehr von Bombenanschläge und **vorallem von** Geidelnahmen in anderen Ländern ,
... oder sogar Banküberfälle finden **auch in** der HOffnung statt um sich

Abb. 8: Ausschnitt aus der Trefferliste von Partikeln mit PP-Anbindung im Testkorpus

Auch für diese syntaktische Klasse lässt sich quasi per Knopfdruck eine Frequenzliste erstellen, die beispielsweise als Grundlage für die Behandlung von PP-fokussierenden Partikeln genutzt werden kann, die mit Listen zu NP-fokussierenden Partikeln kontrastiert werden kann, usw.

Die Beispiele demonstrieren, dass sich durch die vorgeschlagene Subklassifikation neue Möglichkeiten für die korpusbasierte Untersuchung adverbieller Wörter und adverbialer Strukturen ergeben. Es lassen sich im Korpus automatisch adverbielle Klassen suchen, die zuvor nur in syntaktisch annotierten Korpora definiert werden konnten.

Dieser Aufsatz zeigt exemplarisch, wie abhängig Wortkategorien von den zugrunde liegenden Klassifikationskriterien sind. Hier wurde auf der Grundlage des STTS-Tagsets erarbeitet, wie eine morphologisch 'vorgefilterte', syntaktisch heterogene Wortklasse nach strikt syntaktischen Kriterien subklassifiziert werden kann, so dass die resultierenden Wortklassen für syntaktische

Fragestellungen anwendbar sind. Andere morphologisch definierte Klassen des STTS können nach den hier angewendeten Kriterien in dieselben „ADV"-Subkategorien fallen. Für eine syntaktische Klassifikation aller Wörter (nicht nur der „ADV"-Kandidaten) nach den aufgestellten Kriterien muss das Schema entsprechend auf andere STTS-Klassen erweitert werden. Um bei der Bearbeitung der STTS-getaggten Korpusdaten keinen Informationsverlust (z.B. Verlust morphologischer Informationen) zu haben, wird hierbei auf eine Mehrebenen-Korpusarchitektur zurückgegriffen, bei welcher die zugrunde liegenden Annotationen beibehalten werden. Zugleich wird an der Automatisierung der Wortartenzuweisung mittels trainierbarer Tagger gearbeitet.

Danksagung

Ich danke Anke Lüdeling für ihre hilfreichen Kommentare zu dem Artikel, des Weiteren Cedric Krummes und Marc Reznicek für ihre Unterstützung bei der Korpusannotation und der Erarbeitung der Annotationsrichtlinien.

Literatur

Admoni, Wladimir (1982): Der deutsche Sprachbau. 4., überarb. u. erw. Aufl. München: Beck.

Atwell, Eric (2008): Development of tag sets for part-of-speech tagging. In: Lüdeling, Anke / Kytö, Merja (Hg.): Corpus linguistics. An international handbook. Berlin: de Gruyter.

Brants, Thorsten (2000): TnT – A statistical part-of-speech tagger. In: Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000. Seattle, WA, 224-231.

Dimroth, Christine / Klein, Wolfgang (1996): Fokuspartikeln in Lernervarietäten. Ein Analyserahmen und einige Beispiele. In: Zeitschrift für Literaturwissenschaft und Linguistik 104: 73-114.

Eisenberg, Peter (2004): Grundriß der deutschen Grammatik. Bd. 2. Der Satz. 2., überarb. u. aktual. Aufl. Stuttgart / Weimar: Metzler.

Garside, Roger / Leech, Geoffrey / McEnery, Tony (Hg.) (1997): Corpus annotation: Linguistic information for computer text corpora. New York: Addison Wesley Longman.

Helbig, Gerhard / Buscha, Joachim (2001): Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht. Berlin u.a.: Langenscheidt.

Hoffmann, Ludger (Hg.) (2007): Handbuch der deutschen Wortarten. Berlin / New York: de Gruyter.

Knobloch, Clemens / Schaeder, Burkhard (Hg.) (2005): Wortarten und Grammatikalisierung: Perspektiven in System und Erwerb. (= Linguistik – Impulse & Tendenzen 12). Berlin / New York: de Gruyter.

Möllering, Martina (2001): Teaching German modal particles: A corpus-based approach. In: Language Learning & Technology 5: 130-151.

Pasch, Renate / Brauße, Ursula / Breindl, Eva / Waßner, Ulrich Hermann (2003): Handbuch der deutschen Konnektoren. Linguistische Grundlagen der Beschreibung und syntaktische Merkmale der deutschen Satzverknüpfer (Konjunktionen, Satzadverbien und Partikeln). (= Schriften des Instituts für Deutsche Sprache 9). Berlin / New York: de Gruyter.

Pittner, Karin (1999): Adverbiale im Deutschen. Untersuchungen zu ihrer Stellung und Interpretation. (= Studien zur deutschen Grammatik 60). Tübingen: Stauffenburg.

Pittner, Karin (2009): *Wieder* als Modalpartikel. In: Zeitschrift für germanistische Linguistik 37: 296-314.

Rapp, Reinhard / Lezius, Wolfgang (2001): Statistische Wortartenannotierung für das Deutsche. In: Sprache und Datenverarbeitung 25, 2: 5-21.

Rehbock, Helmut (2009): Ein Zeitadverb als Diskursmarker. In: Zeitschrift für germanistische Linguistik 37: 236-265.

Schiller, Anne / Teufel, Simone / Stöckert, Christine / Thielen, Christine (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical Report. Stuttgart: Institut für maschinelle Sprachverarbeitung. Internet: http://www.ifi.uzh.ch/~siclemat/man/SchillerTeufel99STTS.pdf (Stand: 26.10.2009).

Schmid, Helmut (1994): Probabilistic Part-of-speech tagging using decision trees. In: International Conference on New Methods in Language Processing. Manchester, UK, 44-49.

Sütterlin, Ludwig (1923): Die deutsche Sprache der Gegenwart. Ihre Laute, Wörter, Wortformen und Sätze. 5. Aufl. Leipzig: Voigtländer.

Vyatkina, Nina A. (2007): Development of second language pragmatic competence: The data-driven teaching of German modal particles based on a learner corpus. Ph.D. diss; Pennsylvania State University.

FRANZISKA MÜNZBERG

Korpusrecherche in der Dudenredaktion: Ein Werkstattbericht

Abstract

Thema des Beitrags ist der Einsatz des Dudenkorpus in der Zusammenarbeit von Grammatikautoren und Dudenredaktion. Das annotierte Korpus und die Recherchemöglichkeiten, die es bietet, werden anhand aktueller Beispiele aus der Werkstatt einer Dudenredakteurin beschrieben. Einen Schwerpunkt bildet neben einfachen Vergleichen zwischen zwei oder drei morphologischen Varianten die komplexere Frage, ob temporales *wo* (*der Zeitpunkt, wo*; *jetzt, wo*) in der Dudengrammatik weiterhin als standardsprachlich bezeichnet werden soll. Zugleich wird versucht, die Attraktivität alternativer Konstruktionen (*der Zeitpunkt, zu dem*; *jetzt, da*) für Schreibende und Lesende zu messen. Diese 'Alternativen' verhalten sich jedoch keineswegs wie die eingangs erwähnten morphologischen Varianten zueinander – zu unterschiedlich sind semantische und syntaktische Leistungen, zu unterschiedlich die Restriktionen, die für ihre Verwendung im Satz gelten, zu unterschiedlich sind schließlich die untersuchten Texte, aus denen die mittels Hochrechnung ausgewerteten über 30 000 Sätze stammen. Zur Diskussion steht, welche Konsequenzen in einer Grammatik für ein breites Publikum zu ziehen sind. Diese Frage wird für die 'Wortgrammatik' anders beantwortet als für die 'Regelgrammatik'.

1. Anlass für die Erhebung empirischer Daten

Undenkbar wäre die Arbeit der Dudenredaktion und ihrer Autor(inn)en ohne Korpora. Aussagen über Morphologie, Syntax und Semantik in Wörterbüchern, Grammatiken und Ratgebern wurden früher anhand von Exzerptsammlungen formuliert, während sie heute mit den Ergebnissen aus der Recherche in elektronischen Korpora konform sein sollen. Das setzt voraus, dass immer wieder anhand neuer Datensätze, mit neuen Recherchemethoden und offen für unterschiedliche Formen von Normbewusstsein alte Aussagen überprüft und Aussagen über neue Themen gemacht werden. Das Ideal ist hier noch nicht die sich selbst schreibende korpusbasierte Grammatik oder die 'emergente' Grammatik, die eine reale Sprachentwicklung in einer Gesellschaft, bei einem Kind, in einer bestimmten Sprechergruppe imitieren kann, indem sie Frequenzen in Regeln umrechnet. Eine solche Grammatik würde

beispielsweise anhand der vorliegenden Korpora geschriebener Sprache voraussichtlich bei einer großen Anzahl von Verben keine 2. Person und keinen Imperativ erzeugen oder das Paradigma zwar aufstellen, aber eine Warnung ausgeben, dass diese Phänomene selten und daher vermutlich nicht standardsprachlich seien. Von der Dudenredaktion erwartet die Öffentlichkeit vielmehr Antworten auf häufig gestellte Fragen, die viele Benutzer(innen), Laien ebenso wie Sprachprofis, brennend interessieren. Besonders was in der telefonischen Sprachberatung erfragt wird, müssen die Dudenbände 4, „Die Grammatik", und 9, „Richtiges und gutes Deutsch", oder die Wörterbücher beantworten. Damit geraten Themen in den Vordergrund, die als Zweifelsfälle bewusst diskutiert werden.

2. Arbeit mit dem Dudenkorpus (Grammatik)

Das Dudenkorpus gehört mit 1,79 Milliarden Wortformen und Satzzeichen (Stand September 2009) zu den größten der deutschen Sprache. Neben Tages- und Wochenzeitungen aus Deutschland, Österreich und der Schweiz enthält es Zeitschriften, Fach- und Sachbücher sowie belletristische Werke unterschiedlicher Genres. Die Teilkorpora sind lemmatisiert, flach annotiert und einzeln oder in beliebiger Zusammenstellung mithilfe der Abfragesprache CQP (Corpus Query Processor) durchsuchbar. Zu den hinterlegten Metadaten gehören neben Quelle und Datum auch etwa die Angabe, ob es sich um eine Übersetzung handelt oder ob in alter Rechtschreibung geschrieben wurde, und eine Zuordnung zu Themen und Sachgebieten. Da das Korpus noch nicht die angestrebten ausgewogenen Anteile einzelner Korpora enthält, werden in regelmäßigen Abständen kleinere ausgewogene Teilkorpora erstellt. Die Redaktion verwendet sie, um aus dem Gesamtkorpus gewonnene Ergebnisse zu kontrollieren. Das Korpus ist nicht öffentlich verfügbar und wird ausschließlich von der Dudenredaktion und ihren Autor(inn)en genutzt.

In der Lexikografie wird das Korpus genutzt, um Einzelbedeutungen zu unterscheiden, Belege zu finden und grammatische Eigenschaften wie Genus und Flexionsklasse zu bestimmen. Zu den typischen Rechercheaufgaben bei der Redaktion einer Grammatik gehört der Vergleich zweier Varianten. Die einfachste Form einer solchen Recherche besteht z.B. in der Anfrage, wie oft das Adjektiv *gesund* im Komparativ und Superlativ umgelautet wird. Dazu muss lediglich festgestellt werden, wie häufig die Formenreihe *gesündere.**, *gesündest.** (2 125 Treffer im Gesamtkorpus) im Vergleich zu *gesundere.**, *ge-*

*sundest.** vorkommt (72 Treffer im Gesamtkorpus; theoretisch könnte sich hierunter ein das Ergebnis leicht verwässerndes *du gesundest* befinden). Ergänzt werden kann diese Anfrage durch die Suche nach der undeklinierten Form in Wortfolgen wie *ist gesünder / gesunder* und *gesünder / gesunder ist* (jeweils ohne Unterbrechung durch Satzzeichen); ein falsch positives Ergebnis erhält man dabei durch die Wortfolge *und der Appetit ein gesunder ist.* Raffinierter und unmöglich ohne Annotation ist die Suche nach [finites Verb, gefolgt von] *gesunder* und *gesunder* [, gefolgt von finitem Verb]; hier ist jedoch mit mehr falsch positiven Ergebnissen zu rechnen. Schon diese Routineanfrage zeigt, um wieviel komfortabler eine Korpusrecherche im Vergleich zu einer Internetrecherche ist. Sie veranschaulicht auch, auf welche Klippen googelnde Laien stoßen könnten. So würde eine Suche nach der Form *gesunder* im Vergleich zur Form *gesünder* die Ergebnisse verfälschen: *Birnen sind ein gesunder Teil unserer Ernährung*, weiß www.misterinfo.de.

Besonders hilfreiche Features des Dudenkorpus und seiner Benutzeroberfläche *Koala* sind, dass die Groß- und Kleinschreibung bei der Suche beachtet wird und dass nach Satzzeichen gesucht werden kann. Der Bereich von Trunkierungen (Platzhaltern, optionalen Buchstaben und Wortformen) und anderen Bedingungen ist steuerbar. Den Nutzer(inne)n bleibt es überlassen, ob sie nach Lemmata oder nach konkreten Wortformen oder Endungen suchen möchten, nach Wortarten bzw. 'Parts of Speech' oder aber nach Kategorisierungen wie etwa Numerus und Kasus. Schließlich kann die Ergebnisanzeige erweitert werden, wenn sich herausstellt, dass der gewählte Kontext von ein oder zwei Sätzen nicht ausreicht, um den Beleg zu interpretieren.

'Interpretieren' ist ein bewusst gewähltes Schlagwort. Tatsächlich wird bei dieser Form der Korpusrecherche nicht nur gezählt und gerechnet, sondern auch gelesen und ausgelegt. Nur so können geeignete Filter gefunden werden, um in einer neuen Suche oder in einer Hochrechnung unerwünschte positive Ergebnisse (im Folgenden „Fehlbelege" genannt) auszuschließen und um schließlich eine geeignete Darstellung der Ergebnisse in einem Wörterbuch oder einer Grammatik herauszuarbeiten. Um beispielsweise herauszufinden, ob und wo es regelmäßig gebildete Steigerungsformen zum Adjektiv *nah(e)* gibt, muss die Anfrageroutine im Vergleich zu *gesundest.** geändert werden. Sonst werden als Fehlbelege die Wortformen, die mit *nahestehend* beginnen, eingerechnet. Eine unter mehreren Möglichkeiten ist die Suche nach den einzelnen Superlativformen *naheste*, *nahesten*, *nahestem*, *nahester* – mit immerhin 20 erstaunlichen Treffern und ohne jegliche Entsprechung im Komparativ. Genauso gut

hätte man aber auch die optionalen Buchstaben am Ende der Wortform *naheste* auf höchstens einen einschränken können. Wegen der verschiedenen Suchmöglichkeiten, die die Ergebnisse unter Umständen beeinflussen könnten, wird meist zusammen mit den Ergebnissen auch die Anfrage gespeichert, bevor die Suchergebnisse an die Autor(inn)en weitergeleitet werden. Anschließend werden die Ergebnisse diskutiert und bewertet; nicht alles irgendwie Auffällige wird gleich nach seiner Entdeckung berücksichtigt. Die *nahesten*-Belege beispielsweise sind noch in keinem Dudenband veröffentlicht. Viel komplexer als die Suche nach morphologischen Eigenschaften wird die korpusbasierte Arbeit an einer Grammatik, wenn syntaktische und semantische Leistungen von Varianten beschrieben werden sollen.

3. Beispiel für eine Recherche: *wo* temporal

Im Abschnitt über Nebensätze (Randnummer 1659) in der 8. Auflage der Dudengrammatik (Duden 2009) beschreibt Peter Gallmann, dass Pro-Adverbien (*wie*, *wo*) anstelle einer Verbindung aus Präposition und Relativpronomen Nebensätze anschließen können:

> Die einfachen Pro-Adverbien entsprechen relativen Präpositionalphrasen:
>
> [...] Über die Art, wie (= in der) man schwermütige, tolle und rasende Menschen behandeln müsse, sollte billig ein philosophischer Arzt ein eigenes Werk schreiben. (A. Knigge) [...] – (Auch temporal:) Wenn in unserem Leben etwas Tragisches passiert, kommt der Zeitpunkt, wo (= zu / in / bei dem) wir das Dunkel ins Auge fassen müssen. (Internetbeleg) [...]
>
> Neben Nebensätzen mit temporal zu verstehendem *wo* finden sich auch Nebensätze mit *da* sowie mit *als* und *wenn*, wobei *als* und *wenn* wohl nicht als Adverbien, sondern als Subjunktionen zu bestimmen sind. Alle vier Konstruktionen sind standardsprachlich korrekt:
>
> Aber auch der Zeitpunkt, wo das Kind zum ersten Mal „Nein" sagen wird, rückt näher. – Dies war der Zeitpunkt, da der Vorstand der Paul-Martini-Stiftung anregte, das Konzept der Stiftungsarbeit zu aktualisieren. – Just zum Zeitpunkt, als das neu errichtete SOS-Kinderdorf hätte besiedelt werden sollen, brach in Liberia Bürgerkrieg aus. – Am besten stellt man Fragen zu dem Zeitpunkt, wenn sie aktuell auftreten. (Internetbelege)

Die Darstellung entspricht seit der 7. Auflage in wesentlichen Zügen der in den großen wissenschaftlichen Grammatiken, vgl. Zifonun et al. (1997, Bd. I: 42), und Eisenberg (2006, Bd. II: 277), mit berechtigter Kritik an der 6. Auflage der

Dudengrammatik (Duden 1998, Randnummer 1332) und einleuchtender Erklärung (das Lokaladverb *wo* kann den Charakter eines „universellen Relativadverbs" annehmen). Im Folgenden (Duden (2009), Randnummer 1660) zeigt Gallmann, dass man in regionalen Varietäten noch wesentlich kreativer mit *wo* umgeht. Die Vielfalt der regionalen und gesprochensprachlichen Verwendungen von *wo* beschreiben über die genannten Standardwerke hinaus eindrücklich z.B. Günthner (2002), Pittner (2004), Fleischer (2004) und Elspaß / Möller (http://www.philhist.uni-augsburg.de/lehrstuehle/germanistik/sprachwissenschaft/ada/runde_3/f12a-b/ Stand: 08 / 2010). Trotz seines guten Standings in den Grammatiken fällt das Relativadverb *wo* jedoch regelmäßig Aufsatz- und Manuskriptkorrekturen zum Opfer, sobald es andere als lokale Bezüge herstellen soll. Besonders Lehrer(innen) sind nach Davies (2007: 56f.) dafür verantwortlich, dass die Liebe der Grammatiker(innen) insbesondere zum „temporalen *wo*", wie es auch hier im Folgenden verkürzend genannt wird, bis jetzt eine eher unglückliche Liebe geblieben ist. Als dieses Auseinanderklaffen zwischen Referenzgrammatik und Grammatikalitätsurteilen nun auch noch auf der IDS-Jahrestagung 2008 offen angesprochen wurde, beschlossen Autor und Redaktion, ihre Meinung nochmals zu überdenken. Dass temporales *wo* in gesprochener Sprache, auch im elaborierten akademischen Vortrag jenseits der Nähesprache, wenigstens in bestimmten Kontexten völlig unmarkiert und sprachlicher Brillanz keineswegs abträglich zu sein scheint, zeigten einige der Tagungsbeiträge selbst. Aber von der Dudengrammatik wird ja unter anderem gefordert, dass sie den geschriebenen Standard beschreibt. Dieser geschriebene Standard wird höchstwahrscheinlich durch Aufsatzkorrekturen in der Schule beeinflusst.

Zunächst muss eingegrenzt werden, welches temporale *wo* überhaupt zum Gegenstand einer sinnvollen Diskussion über die Standardnähe werden kann. Unbestritten nähesprachlich sind freie Adverbialsätze mit *wo* als einer Art Konjunktion; gleich drei Belege finden sich im folgenden Gesprächsausschnitt:

> S2: Ja, ja, wo wir haben fort müssen, anno vierzig. [hier vielleicht noch Relativadverb: Bezug zwischen *wo* und *anno vierzig*]
>
> [...]
>
> S1: (gleichzeitig) Und wer hat hingesehen wo wir heimgekommen sind? Wo ich mal heimgekommen bin, hat ganz NAME hat gebrannt. [keinerlei relativer Anschluss]
>
> (Datenbank gesprochenes Deutsch, Zwirner-Korpus: Alemannisch, 1955; Interaktion ZW038, http://dsav-wiss.ids-mannheim.de/DSAv/KORPORA/ZW/ZW0/ZW038/ZW038TRA.HTM; zur regionalen Zuordnung vgl. auch Interaktion OS023 und den bei Pittner (2004: 365) zitierten Hamburger Beleg aus dem Pfeffer-Korpus, PF022).

Interessanter für die Frage nach der Standardkonformität von temporalem *wo* ist die Verwendung nach Adverbphrasen mit *jetzt* oder *heute* als Kern. Pittner (2004: 63) ist der Ansicht, dieser Gebrauch von *wo* müsse schon deshalb als standardsprachlich gelten, weil es in der Standardsprache „keinen gleichwertigen Ersatz" gebe. So könnten manche Sprecher(innen) *da* anstelle von *wo* in den folgenden Belegen vielleicht als „gehoben" oder „veraltend" empfinden; semantisch würde *da* vermutlich dasselbe leisten:

> Dieser Aspekt wurde in diesem Kapitel bereits erwähnt (siehe Abbildung 16.3), aber es ist wichtig, ihn jetzt, wo Sie mehr über Transkription und Translation wissen, nochmals aufzugreifen. (Übersetzung eines Biologie-Lehrbuchs von Campbell / Reece) [rein temporal oder auch kausal lesbar?]

> Wieso soll der Gaspreis jetzt, wo Öl wieder billiger wird, weiter steigen? (taz, 01.08.2005) [rein temporal oder auch konzessiv bzw. äußerungsbegründend lesbar?]

Zum Vergleich ein Originalbeleg mit *da:*

> Zwischen 1995 und 2003 hat sich der Umsatz des schwäbischen Familienunternehmens verdreifacht, und das in einem Zeitraum, da der deutsche Hemdenmarkt von 90 Millionen verkauften Stück auf knapp 62 Millionen geschrumpft ist. (SZ, 31.12.2004)

Überprüft werden sollte, ob die Konstituentenstruktur (Adverbphrase) des Bezugsausdrucks eine Rolle für die Frequenz von *wo*-Relativsätzen spielt. Schließlich sollte untersucht werden, unter welchen Bedingungen *wo* nach Bezugsausdrücken mit *Jahrhundert*, *Tag*, *Augenblick*, *Moment*, *Zeitpunkt* steht, auch um festzustellen, ob die für die Dudengrammatik ausgewählten Beispielsätze, die Nominalphrasen mit *Zeitpunkt* enthalten, typisch und damit geeignet sind.

4. Erste Sichtungen und Arbeitshypothesen

Die unterschiedliche Beurteilung des temporalen *wo* durch Grammatiker(innen) und Lehrkräfte könnte nicht nur auf Unterschieden im Normbewusstsein beruhen, sondern es könnte auch zahlenmäßig erfassbare syntaktische oder semantische Gründe dafür geben, dass *wo* und speziell temporales *wo* als relativer Anschluss einmal markiert, einmal unmarkiert erscheint. Zu Beginn der Untersuchung sah es freilich so aus, als müsste diese Hypothese verworfen werden: Ausgetestet wurden jeweils unterschiedliche Bezugsausdrücke – in verschiedenen Kasus, mit und ohne Präposition (*der Augenblick, wo* vs. *in dem Augenblick, wo*), mit Wechsel im Numerus (*der Augenblick, wo* vs. *die Augenblicke, wo*), mit Variation in der Struktur bzw. Wortstellung (*die langen Au-*

genblicke des Wartens, wo; *die Augenblicke, wo* ... [Verb] *und* [Verb]), determiniert (*der, dieser Augenblick*) oder unbestimmt (*ein Augenblick*) usw. Meistens dienen die Bezugsausdrücke als Adverbialien bzw. als Teile von Adverbialien – aber das gilt nicht speziell für *wo*-Belege. Keine der Datenreihen lieferte eine sensationell einfache Erklärung.

Ferner könnte temporales *wo* ja tatsächlich eher in Interviews und umgangssprachlich gefärbten Glossen als in Berichten und Analysen vorkommen. Wenn es wirklich klar als nähesprachlich markiert ist, dann lässt sich das nicht nur anhand von Metadaten, sondern eventuell auch anhand häufiger Personalpronomina und anhand typisch nähesprachlicher Wortwahl und Grammatik im selben Satz nachweisen. Auch hier ließen erste Stichproben allerdings noch keine Beweiskraft erwarten – der eine *wo*-Beleg gibt einen Mannheimer Fährmann möglichst im Original wieder, ein anderer einen Berliner Taxifahrer, die nächsten Belege aber stammen aus Fachbüchern über mittelalterliche Geschichte oder über Biotechnologie, aus Theaterkritiken oder Wirtschaftsprognosen. Es konnte keine Textsorte ausgemacht werden, in der temporales *wo* niemals vorkommt, und es wird zu oft auch im gehobenen und im akademischen Stil verwendet, um schlicht als umgangssprachlich abgetan zu werden.

Immerhin aber kristallisierten sich aus diesen ersten Versuchsreihen einige „*wo*-Konkurrenten" heraus, die den Schreibenden je nach Kontext als Alternativen zur Verfügung stehen. So wurden für die Vergleiche zwischen der Nebensatzeinleitung durch *wo* und alternativen Ausdrucksmöglichkeiten schließlich folgende Kombinationen ausgewählt (die Datenreihen mit Bezugsausdruck *jetzt* stehen als voraussichtlicher Sonderfall gleichsam außer Konkurrenz daneben):

Augenblick Moment Zeitpunkt Jahr (mit Jahreszahl) Jahrhundert Nacht Tag damals [Sonderfall: jetzt] tagsüber tags	direkt gefolgt von Komma und	*wo* *in* (*der* / *dem*) *an* (*der* / *dem*) *zu* (*der* / *dem*) *wenn* *als* *da* *während* (*deren* / *derer* / *dessen*)

Gefunden wurden über 30 000 Sätze, Belege mit Bezugsausdruck *jetzt* nicht eingerechnet. 1 780 davon enthielten die gesuchte Nebensatzeinleitung mit *wo*. Aus jeder Datenreihe wurden wenn nicht alle, so doch mindestens je 50 Sätze aus unterschiedlichen Quellen gelesen und ausgewertet. Der Anteil an Fehlbelegen wurde hochgerechnet, d.h., bei 10 Fehlbelegen in 100 durchgesehenen Sätzen wurden vom Gesamtergebnis 10 % Fehlbelege abgezogen. Als Fehlbelege wurden beispielsweise gewertet:

> Aber jetzt, als Rentner, zieht es ihn auch ab und zu weit fort. (Mannheimer Morgen 19.04.2000) Auch das Schnüffeln des Hundes klingt seltsam jetzt, als wäre es nicht wirklich da, nur dumpfe Erinnerungen an lautere Tage. (SZ, 05.09.1995)
>
> Sie schwieg einen Moment, während sie beschleunigte und in einer uneinsehbaren Kurve zu einem besonders waghalsigen Überholmanöver ansetzte. (P. Mayle [Übersetzung])

Erhellend erscheint die Frage, der derzeit Ulrike Stölzel, IDS Mannheim, in einer eigenen Erhebung nachgeht: ob der Bezugsausdruck von *wo* eher einen Zeitpunkt oder einen Zeitraum bezeichnet. Im Zusammenhang damit steht allgemein die Frage nach Tempus und Modus im übergeordneten Satz und im *wo*-Nebensatz, die beim genaueren Lesen der Belege aus dem Dudenkorpus aufkam. So könnte generalisierendes Präsens („immer wenn" mit Betonung nicht des Iterativen, sondern des Regelhaften) die Verwendung von *wo* begünstigen. Ein typischer Beleg wäre der folgende; referiert wird auf einen Zeitpunkt:

> Anmeldung ist neun Monate vor der Einschulung, das ist der Moment, wo in meinen Augen Schule losgehen kann. (Zeit, 13.02.2002)

Entsprechend im Präteritum (auch *augenblick, wo* bezieht sich auf ein für sich genommen punktuelles Ereignis, das einer „immer wenn"-Regel nach immer dieselbe Folge hat):

> der zuschlag erfolgte z.b. in dem augenblick, wo ein angezündetes licht erlosch oder das ins licht gesteckte geldstück beim herunterbrennen zu boden fiel (Deutsches Wörterbuch, „Versteigerung", gefunden über ein Zitat in der Computerzeitschrift c't, 08/2001)

Als mündlich, umgangssprachlich oder stärker regional markiert könnten demgegenüber Belege empfunden werden, in denen *Augenblick, wo* sich eher auf ein einzelnes erzähltes (auch hier wieder punktuelles) Ereignis bezieht. Charakteristisch ist dann möglicherweise das Auftreten der im Korpus ansonsten selteneren Pronomina *ich*, *wir*, *Sie*:

> Und drei Monate später kam der Augenblick, wo ich die Zehen ganz schwach bewegen konnte. (K. Meyer 1996)

Auf der anderen Seite im Varietätenspektrum würde man wohl einige der zahlreichen Belege ansiedeln, in denen aus der „immer wenn"-Beziehung eine regelrecht konditionale geworden ist – hier wird nun nicht erzählt, sondern vermutet, behauptet und analysiert:

> Doch in dem Augenblick, wo eine entsprechende technische Vorrichtung handlungsbestimmend wird, verändern sich die Bedingungen der Bewertung. (K. Weber 2005)
>
> In dem Moment, wo man den Raum des Rechts betritt, muss man nicht nur den anderen anerkennen, sondern sich auch selbst rechtfertigen. (Zeit, 08.12.2004)

Sucht man hingegen mit Bezugsausdrücken wie *Jahrhundert* gezielt nach der Referenz auf einen längeren Zeitraum, so stößt man auf beinahe lokale Bezüge, in denen das „Jahrhundert" wie die „Gewölbe" und „Katakomben" eher als Bühne oder Setting, mehr als virtueller Raum denn als Ereigniszeit erscheint. Anders als in anderen Datenreihen sind hier auch nichtrestriktive relative Anschlüsse typisch:

> Nach einem Ausflug in das 18. Jahrhundert, wo er in den „Gewölben des Dr. Hahnemann" die Geschichte des Urvaters der Homöopathie unterhaltsam und spannend erforschte, kehrt er jetzt zurück in das Reformationszeitalter: „Der Bader von St. Denis" heißt sein neuester Roman, in dem er auf den Spuren von Ambroise Par [sic], dem Wegbereiter der modernen Chirurgie wandelt. (Mannheimer Morgen, 02.09.2004)
>
> Oder er stürzt in die dunklen Katakomben der Weltgeschichte und landet im 15. Jahrhundert, wo er mit den blutrünstigen Hussiten für die tschechische Unabhängigkeit kämpfen soll. (SZ, 14.01.1995)

Folgende Schwierigkeiten ergeben sich für die Auswertung:

1) Die wichtigsten 'Konkurrenten' von *wo* unterliegen anderen syntaktischen Restriktionen als *wo* und leisten – abgesehen wohl von *da* – semantisch nicht dasselbe wie *wo*. So sind *als* und *wenn* im Deutschen kaum füreinander einsetzbar, und der präpositionale Anschluss ist nicht möglich nach einem Adverb (**tagsüber, an dem*).

2) Textausschnitte können aus anderen Quellen zitiert sein, vgl. den oben zitierten Artikel aus dem Grimm'schen Wörterbuch.

3) Häufig zitierte Titel von Filmen und Theaterstücken, auch Zitate etwa aus Liedtexten können Ergebnisse verfälschen.

4) Besonders in Interviews und in humoristischen Beiträgen wird manchmal der Versuch unternommen, 'unterschichtige' oder anders stigmatisierte Redeweisen nachzuahmen.

5) Aussagen über regionale Variation können aus überregional erscheinenden Blättern nur unter Vorbehalt gefolgert werden, weil die Schreibenden oder die Interviewten selbst keineswegs am Erscheinungsort des Blattes geboren oder aufgewachsen sein müssen. Auch könnte die Nachbearbeitung etwa von Interviews in einzelnen Redaktionen und sogar bei einzelnen freien Journalisten, die einer Zeitung besonders viele Beiträge liefern, rigoroser sein als bei anderen.

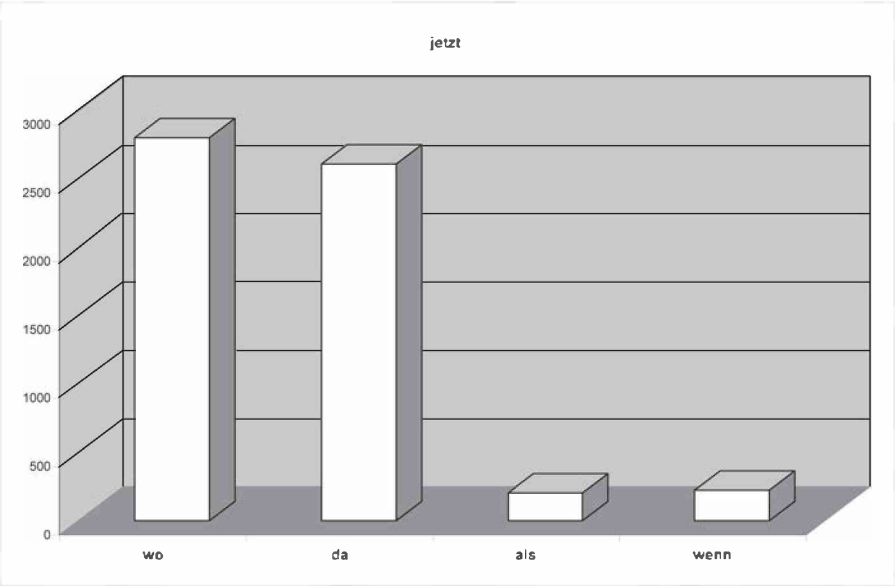
6) Einzelne Faktoren, denen ein Einfluss auf die Verwendung von *wo* zugeschrieben wird, könnten (auch unbemerkt) miteinander korrelieren. Wenn etwa zwei Zeitungen aus unterschiedlichen Regionen als Beispiele für regionale Schreibsprache gegeneinandergehalten werden, so unterscheiden sich diese beiden Zeitungen notwendigerweise noch durch andere Merkmale als den Erscheinungsort. Dieselben Zeitungen könnten z.B. unterschiedlich viele Tokens aus nähesprachlich geprägten Interviews mit biografischem Schwerpunkt und insgesamt unterschiedlich viele Tokens mit Merkmal 1. oder 2. Person, Präsens usw. enthalten.

7) Im Zusammenspiel von Bezugsausdruck und Nebensatzeinleitung können neben der temporalen Bedeutung weitere Bedeutungen mitschwingen. Es ist nicht sicher, ob alle Sprecher(innen) den Wert „temporal" für alle gefundenen Belege vergeben würden. Eine Auswertung nach dem genauen semantischen Wert, nach Aktionsarten u.Ä. („punktuell", „immer wenn"; „eher lokal", „eher konditional") trägt auch bei möglichst exakter Arbeitsanweisung subjektive Züge. Es ist auch nicht sicher, ob sie alle dieselben Sätze als 'Konkurrenten' zu Relativsätzen mitzählen würden. Als besonders schwierig erwies sich diese Entscheidung in den Datenreihen mit der Nebensatzeinleitung *wenn*; gezählt wurde hier nach dem Grundsatz „im Zweifel für den Beleg".

8) Es stehen keine Hilfskräfte zur Verfügung, um alle Daten auf diese Weise semantisch auszuwerten, alle Korrelationen auszutesten und dabei auch satzübergreifend den Kontext einzubeziehen.

5. Korpusdaten

Alle nachfolgend präsentierten Zahlen sind mit entsprechender Vorsicht zu lesen. Zunächst werden die Datenreihen mit Bezugsausdruck *jetzt* gesichtet – hier lautete die Arbeitshypothese, dass *wo* der einzig mögliche standardsprachliche Anschluss eines Relativsatzes sei –, danach die übrigen Datenreihen, in denen *wo* wohl nur unter bestimmten Umständen unmarkiert wirkt.



Die Ergebnisse zu *jetzt, wo* und seinen Konkurrenten (insgesamt 5 834 Belege) bestätigen die Annahme, dass *wo* hier standardsprachlich ist. Allerdings sprechen die Zahlenverhältnisse im Duden-Gesamtkorpus (Stand Frühjahr 2008, 1, 3 Mrd. Tokens) dafür, *da* ebenfalls als standardsprachlich zu bewerten.

Aus den *jetzt*-Belegen selbst zeichnen sich wiederum Muster ab. Ein großer Teil der *wo*- und *da*-Sätze kann nicht nur temporal, sondern zugleich auch kausal oder konzessiv verstanden werden. Referiert wird regelmäßig auf einen Sachverhalt, den die Leser(innen) voraussichtlich nicht bestreiten werden. Die unanfechtbare Aussage im *wo*- / *da*-Nebensatz soll der im Matrixsatz vorgetragenen (unter Umständen durchaus kontroversen) Meinung mehr Nachdruck verleihen. Begünstigt werden diese Lesarten durch Partikeln wie *gerade*, *ausgerechnet*, *selbst*. Wirklich ersetzbar sind *da* und *wo* aber nicht durch die Subjunktionen *weil* und *obwohl*, denn *da* und *wo* sind besser geeignet, einen relativen Bezug herzustellen:

> Gerade jetzt, da [weil] in jeder Klinik das Zauberwort „Kostensenkung" die Runde mache, seien Ärzte mit kaufmännischer Ausbildung bitter nötig. (SZ, 11.04.1996)
>
> Gerade jetzt, wo [obwohl(?)] Reklameaussagen als Eigenschaftszusagen verbindlich werden, sind die Werbetexte seltsam aussagelos. (Computerzeitschrift c't 10/2001)

Die wenigen Nicht-Fehlbelege (nur etwa die Hälfte von 442 Treffern) für *jetzt, als* entstammen eher Texten, die im Präteritum oder im historischen Präsens aus der Perspektive einer Figur erzählen. Auch hier kann der auf *jetzt* folgende Nebensatz zugleich auf Bekanntes, Unbestreitbares referieren und eine begründende oder kontrastierende Funktion haben:

> Erst jetzt, als nichts mehr getan werden konnte, merkte ich, wie meine Hände zitterten. (B. Jaumann 2002)

Mit dem ebenfalls seltenen *wenn* wird – wiederum in einer Datenreihe mit nur gut 200 validen von insgesamt über 300 gefundenen Belegen – eine Vorzeitigkeit, Bedingung oder Begründung ausgedrückt.

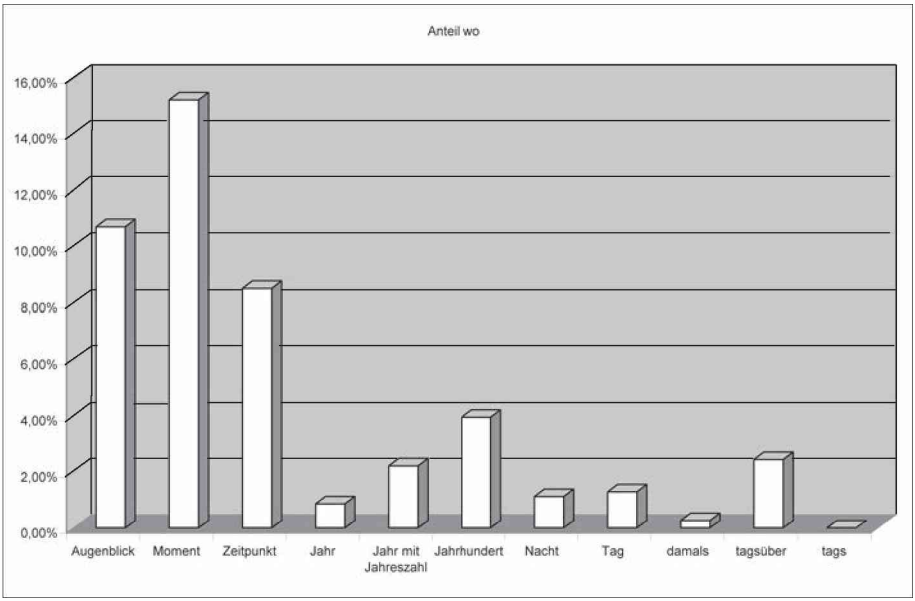
> Ein wahres Schnäppchen sei das Unternehmen doch jetzt, wenn es in Insolvenz gehe. (SZ, 29.09.2006)

Was für *jetzt* gilt, trifft aber nicht auf andere Adverbien wie *tagsüber*, *tags*, *nachts* (präferierte Nebensatzeinleitung: *wenn*) oder *damals* (präferierte Nebensatzeinleitung: *als*) zu.

> So sieht es also im Gehirn aus, nachts, wenn wir schlafen, oder tags, wenn wir träumen. (SZ, 28.09.1998)
>
> Das war damals, als noch der Guckkasten angesagt war. (taz, 17.06.2004)

Die Merkmale, die den ersten Vermutungen nach zur Verwendung von *wo* führen könnten, treten keineswegs ausschließlich in *wo*-Sätzen auf. Erst der Vergleich mit lexikalisch unterschiedlichen Bezugsausdrücken liefert die ersehnten deutlichen und relativ interpretationsunabhängigen Zahlenunterschiede, die kein Vergleich grammatischer Merkmale erbringen konnte: Nach *Augenblick* und *Moment* ist *wo* häufiger – hier gewinnt es seinen Konkurrenten etwa 15% 'Marktanteil' ab –, wenn auch lange nicht so gebräuchlich wie nach *jetzt*. Das in der Dudengrammatik gewählte Beispiel *Zeitpunkt* kann wahrscheinlich ebenfalls als akzeptabel für ein nachfolgendes *wo* bestätigt werden, auch wenn der Anteil an *wo*-Sätzen gering im Vergleich zur Datenrei-

he *Moment, wo* erscheint und der Nebensatz in 92% der Fälle anders eingeleitet wurde. In allen anderen Datenreihen ist *wo* im Vergleich zu den Alternativen *als*, *wenn*, *in der / dem* usw. ausgesprochen selten:



Greift man nun *Augenblick* und *Moment* als typische Kerne des Bezugsausdrucks zu einem *wo*-Relativsatz heraus, so müssten sich regionale Unterschiede in der Verwendung von *wo* wohl recht gut innerhalb dieser vergleichsweise *wo*-affinen, ansonsten unmarkierten Umgebung nachweisen lassen. Aber abgesehen davon, dass die Prozentsätze zwischen *Augenblick* und *Moment* überhaupt variieren (hier gibt es möglicherweise regionale oder diastratische Unterschiede), liefert dieser Vergleich wiederum keine bahnbrechenden Ergebnisse. Die erwartbare süddeutsch-österreichisch-schweizerische Vorliebe für temporales *wo* lässt sich aus den Daten (Gesamtkorpus, Stand Herbst 2009, knapp 1, 8 Mrd. Tokens, und Probe mit ausgewogenerem Teilkorpus) nicht ablesen. Eine künftige Umfrage mit leicht modifizierten Beispielen im *Atlas zur deutschen Alltagssprache* kann hier möglicherweise aussagekräftigere Ergebnisse liefern.

Schließlich kann die Verwendung von *wo* (oder einer Subjunktion) der stilistischen Variation geschuldet sein. Einzelne Schreiber, Lektorate oder Redaktionen könnten die Formulierung *in dem Moment, in dem* wegen der Wiederholung von *in dem* als ungelenk empfinden (nur 734 Vorkommen im Dudenkorpus gegenüber immerhin 541 Vorkommen von *in dem Moment, wo* – bei insgesamt

2 571 Belegen für *Moment, in* und nur 974 für *Moment, wo*). Ein weiterer Grund für stilistische Variation ist eine ähnliche Formulierung im Kontext, der für solche Aspekte der Untersuchung eigentlich über den einzelnen Satz hinaus ausgewertet werden müsste:

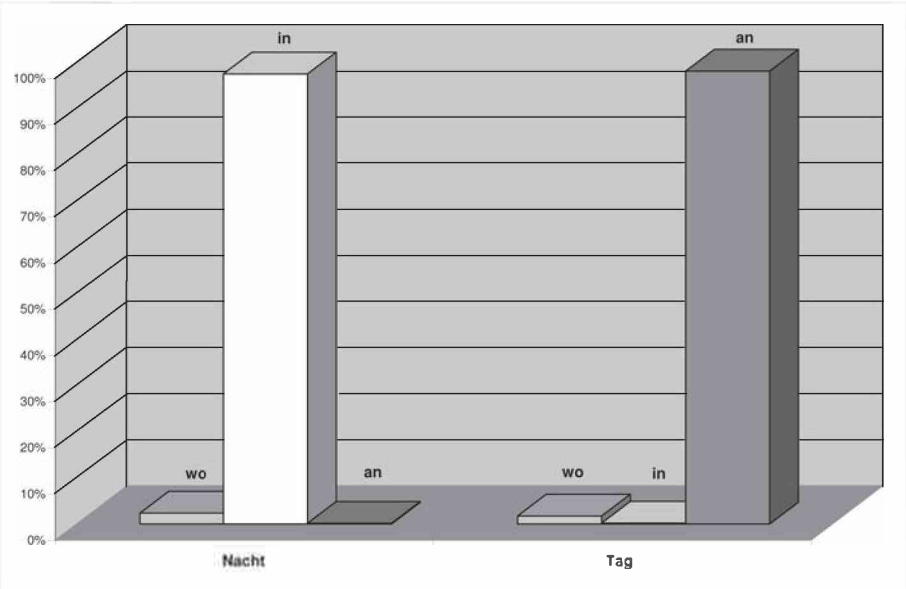
> Über das Gefühl, wenn der Boden unter den Füßen wegbricht, über den Moment, wo man sich entscheiden muss. (Zeit, 31.07.2002)

6. Diskussion

Diese Ergebnisse beweisen keine der Eingangshypothesen zur Semantik. Sie stehen aber auch nicht im Widerspruch dazu: Schließlich können Eigenschaften wie „auch konditional zu verstehen" oder „nicht auf einen längeren Zeitraum in seinem Verlauf bezogen" den Wortfolgen *jetzt / Moment / Augenblick, wo* meist leichter zugeschrieben werden als den Wortfolgen *Jahr / Jahrhundert, als / in dem.*

Insgesamt scheinen *wo* und *da* in den beschriebenen Kontexten die einzigen Nebensatzeinleitungen zu sein, die niemals ganz agrammatisch sind. Damit würde *wo* sich in beinahe überall einsetzbare Relativa in anderen Sprachen einreihen, vgl. neugriech. *pou* („wo", anders als seine Konkurrenten *to opoio / i opoia* etc. ohne Bestimmung nach Kasus, Genus und Numerus und auch nicht, wie in *apo / stin / me / gia tin opoia*, direkt mit einer Präposition kombinierbar), oder es könnte mit temporalen / konditionalen Nebensatzeinleitungen wie frz. *au moment où* verglichen werden, die ebenfalls als Ausweitungen eines ursprünglich lokalen Gebrauchs gelesen werden können. Verfasser einer reinen Wort- oder Kollokationengrammatik würden unter den Stichwörtern *Augenblick* und *Moment* die verschiedenen Varianten des relativen und / oder temporalen (konditionalen) Anschlusses beschreiben. Die Nebensatzeinleitung mit *wo* dürfte man als standardsprachlich eher selten, aber nicht als falsch bezeichnen, und man müsste die 'Arbeitsteilung' zwischen *wenn* und *als* erklären. Schließlich wäre auf die Wahl der Präposition (meist *in*, aber auch *an* ist möglich) und auf die Vermeidung von Wortwiederholungen einzugehen.

Unter den Stichwörtern *Tag* und *Nacht* wäre über die *als*-oder-*wenn*-Regel hinaus vor allem für Zweitsprachlerner der Hinweis von Nutzen, dass auch der präpositionale Anschluss nicht frei gewählt wird: Es heißt *der Tag, an dem*, aber *die Nacht, in der.* Die Nebensatzeinleitung mit *wo* hingegen ist so selten (< 2%), dass man sie sogar verschweigen könnte.



Ähnlich wären häufig nachgefragte Stichwörter wie *Jahr* und *Jahrhundert* zu behandeln. Bei *jetzt* und *heute* könnte man umgekehrt zwischen *wo* und *da* wählen lassen und die anderen Möglichkeiten als Ausnahmen darstellen oder gar vereinfachend unterschlagen. Für eine Grammatik, die über die Wortgrammatik und über beliebte Kollokationen hinaus allgemeine Regeln aufstellen soll, ist nun eine Formulierung zwischen zwei Extremen gesucht, die sich beide ausgehend von objektiven Zahlen begründen lassen:

> Extrem Nr. 1: Temporales *wo* ist nie ganz falsch, es ist in allen Datenreihen zu finden. Darüber hinaus hat es sogar den Charme, einzelsprachübergreifend mit der Bedeutungsverlagerung von lokal über temporal zu konditional, kausal und konzessiv 'im Trend zu liegen'. Deshalb soll es in der Dudengrammatik uneingeschränkt empfohlen und seinen Konkurrenten gegenüber positiv herausgehoben werden.
>
> Extrem Nr. 2: In den meisten Datensätzen ist temporales *wo* weitaus seltener als seine Konkurrenten und wirkt daher auch mehr oder weniger stark markiert. Deshalb ist die Dudenredaktion nicht bereit, es als standardsprachlich anzuerkennen.

Zwischen den beiden Extremen liegt der Status quo. Der Autor des Syntaxkapitels, Peter Gallmann, und die Dudenredaktion (Kathrin Kunkel-Razum, Franziska Münzberg) haben sich anhand der hier vorgestellten Ergebnisse dagegen entschieden, Grundlegendes an der Darstellung in der Dudengramma-

tik zu ändern. Würde man vor temporalem *wo* in bestimmten Kontexten warnen – denn auf eine Warnung laufen deskriptiv gemeinte Attribuierungen wie „selten", „besonders in der gesprochenen Sprache", „umgangssprachlich" in der Praxis der Benutzer(innen) meist hinaus –, dann käme es womöglich vermehrt zu weitaus auffälligeren Formulierungen und Hyperkorrekturen wie *das Jahrhundert, an dem* oder *in dem Moment, zu dem*. Und solche auffälligen Formulierungen wären dem individuellen Ziel der Schreibenden, in einer Grammatik nachzuschlagen und sich daraufhin sprachlich möglichst konform und standardnah verhalten zu können, entgegengesetzt. Was im Dudenband 4 noch fehlt, ist eine Bemerkung zu den außergewöhnlichen Zahlenverhältnissen bei *jetzt, da* und *jetzt, wo.* In den Kontext von Randnummer 1659 passt *jetzt* als Adverb nicht. Nun bleibt zu diskutieren, ob das Thema besser nur im Dudenband 9 (*Richtiges und gutes Deutsch*) abgehandelt wird oder ob (und wo, zulasten welches anderen Themas) es in der nächsten Auflage der Dudengrammatik seinen Platz finden soll. Der je nach Kontext und vielleicht wiederum je nach Bezugsausdruck schwankenden Beurteilung von *da* als Relativsatzeinleitung trägt Damaris Nübling in ihrem Kapitel über die Nichtflektierbaren Rechnung (in Randnummer 943 mit einem als „gehoben veraltend" markierten und einem unmarkierten Beispiel). In Randnummer 943 könnte man auch *wo* einordnen – was die Lehrkräfte aber kaum besänftigen dürfte.

Literatur

Davies, Winifred V. (2007): Die Geschichte vom „schlechten" Deutsch. In: Der Deutschunterricht 3: 52-62.

Duden (1998): Duden – Die Grammatik. (= Duden 4). 6. Aufl. Mannheim u.a.: Dudenverlag.

Duden (2009): Duden – Die Grammatik. (= Duden 4). 8. Aufl. Mannheim u.a.: Dudenverlag.

Duden (2007): Duden – Richtiges und gutes Deutsch. (= Duden 9). 6. Aufl. Mannheim u.a.: Dudenverlag.

Eisenberg, Peter (2006): Grundriss der deutschen Grammatik. 2 Bde. 3. Aufl. Stuttgart: Metzler.

Elspaß, Stephan / Möller, Robert: Atlas zur deutschen Alltagssprache (AdA). Internet: http://www.philhist.uni-augsburg.de/lehrstuehle/germanistik/sprachwissenschaft/ada/ (Stand: 09 / 2009).

Fiehler, Reinhard / Wagener, Peter: Datenbank gesprochenes Deutsch. Internet: http://dsav-wiss.ids-mannheim.de/DSAv/DSAVINFO.HTM (Stand: 01 / 2010).

Fleischer, Jürg (2004): A typology of relative clauses in German dialects. In: Bernd Kortmann (Hg.) Dialectology meets typology. Dialect grammar from a cross-linguistic perspective. Berlin / New York: de Gruyter, 211-243.

Günthner, Susanne (2002): Zum kausalen und konzessiven Gebrauch des Konnektors *wo* im gesprochenen Umgangsdeutsch. In: Zeitschrift für Germanistische Linguistik 30, 3: 310-341.

Pittner, Karin (2004): *Wo* in Relativsätzen – eine korpusbasierte Untersuchung. In: Zeitschrift für germanistische Linguistik 32, 3: 357-375.

Zifonun, Gisela / Hoffmann, Ludger / Strecker, Bruno et al. (1997): Grammatik der deutschen Sprache. 3 Bde. (= Schriften des Instituts für Deutsche Sprache 7). Berlin / New York: de Gruyter.

PER BÆRENTZEN

Einige neue Regularitäten im Gebrauch der Pronominalformen *deren* und *derer*

Abstract

Mein Beitrag basiert auf zwei früheren Publikationen (Bærentzen 1995 und 2008), die hier in Auszügen wiedergegeben werden.

Über die Verwendung der beiden Pronominalformen *deren* und *derer* herrscht im tatsächlichen Sprachgebrauch und in den grammatischen Darstellungen eine gewisse Unsicherheit. Anhand zahlreicher Korpusbelege wird im Folgenden zweierlei gezeigt: Im ersten Kapitel werden für die zum Teil komplementäre Distribution von *deren* und *derer*, vor allem in der Verwendung als Attribut eines Substantivs, gegenüber den traditionellen Beschreibungen präzisere Regeln sowie deren Begründung gegeben. Im zweiten Kapitel wird dann die ungewöhnliche, aber immer häufiger vorkommende Verwendung von *derer* als vorangestelltem Attribut eines Substantivs beschrieben und interpretiert.

1. Regeln für die komplementäre Distribution von *deren* und *derer*

Das Pronomen *der* hat in demonstrativer und relativer Funktion an gewissen Stellen im Paradigma erweiterte Formen entwickelt, die aus dem Stamm *d-* plus einem ersten und einem zweiten Flexiv bestehen: *dessen* (*d-es*(*s*)*-en*), *deren* (*d-er-en*), *derer* (*d-er-er*), *denen* (*d-en-en*). Die Formen *deren* und *derer* sind beide sowohl Genitiv Singular Femininum als auch Genitiv Plural und haben somit denselben Stellenwert. Da sie nicht als freie Varianten, sondern nach gewissen Regeln auftreten, gilt es, diese möglichst adäquat zu erfassen.

1.1 Traditionelle Regelformulierung

Die traditionelle Regelformulierung geht von der Verweisrichtung des Pronomens aus und besagt, dass bei Rückverweis (Anaphora) *deren*, bei Vorverweis (Kataphora) *derer* zu wählen sei. Da das Relativpronomen obligatorisch rückverweisend ist, folgt aus der genannten Verweisrichtungsregel, dass beim Relativum nur die Form *deren* zu erwarten ist, wie in (1) und (2), während beim Demonstrativum, das sowohl rück- wie in (3) als auch vorverweisend wie in

(4) sein kann, beide Formen möglich sind. (In den Beispielen ist das Wort, auf welches das Pronomen nach der gängigen Auffassung verweist, durch Unterstreichung kenntlich gemacht:)

(1) Dort begegnete man Dichtern, *deren* Werke man kannte [...]. (MK1/MHE.00000 Heuss 1964: 86)

(2) [...] die Gerichtsklausel, kraft *deren* der Schützling die unmittelbare Gerichtsbarkeit des Schutzherrn [...] in Anspruch nehmen darf. (LIM/LI1.00245 Kisch 1970)

(3) Der Fachschaftsvorstand ist der Fachschaftsversammlung verantwortlich und an *deren* Weisung gebunden. (LIM/LI1.00083 [Diverse Flugblätter], 1971)

(4) Märtyrer steigern die Kräfte *derer*, denen sie Vorbild wurden. (MK1/WJA.00000 Jaspers 1962: 71)

1.2 Überprüfung dieser Regel anhand der Belege

Um die Distribution von *deren* und *derer* näher zu untersuchen, habe ich in den neunziger Jahren (Bærentzen 1995) das Mannheimer Korpus und das Limas-Korpus, die mir schon damals in elektronischer Form zur Verfügung standen, ausgewertet. Die Korpora lieferten insgesamt 1379 Belege, davon 1314 *deren*-Belege gegenüber nur 65 *derer*-Belegen.

Normalerweise sind *deren* und *derer* in allen Verwendungen kasusnonkongruent, d.h. sie bekommen ihren Kasus Genitiv nicht durch eine Kongruenz mit einem anderen Wort, sondern auf andere Weise zugewiesen. Die gängigen Verwendungsmöglichkeiten von *deren* und *derer* im heutigen Deutsch lassen sich anhand von fünf Funktionstypen beschreiben, die so definiert sind, dass der Unterschied im Gebrauch der beiden Pronominalformen möglichst deutlich hervortritt:

- Funktionstyp 1

  Das Pronomen ist vorangestelltes Attribut eines Substantivs:

  (5) Dort begegnete man Dichtern, *deren* **Werke** man kannte [...]. (MK1/MHE.00000 Heuss 1964: 86)

- Funktionstyp 2

  Das Pronomen ist nachgestelltes Attribut eines Substantivs:

(6) Märtyrer steigern die **Kräfte** *derer*, denen sie Vorbild wurden. (MK1/WJA.00000 Jaspers 1962: 71)

– Funktionstyp 3

Das Pronomen wird von einer Präposition regiert:

(7) Der einzelne Wissenschaftler ist nur noch ein Spezialist, der in einem Zusammenhang ausgebildeter Verfahrensweisen steht, **innerhalb** *derer* er seine besondre Aufgabe erfüllt. (MK1/WBM.00000 Bollnow 1962: 64)

– Funktionstyp 4

Das Pronomen wird vom Satzprädikat regiert:

(8) Die meisten Güter, *deren* der Mensch **bedarf**, sind vermehrbar. (LIM/LI1.00495 Beer 1970)

– Funktionstyp 5

Das Pronomen ist einem quantitätsbezeichnenden Wort untergeordnet:

(9) Mit diesem Gedicht [...] lässt sich keines aus den anderen Gattungen, *deren* Ennius **viele** versuchte, vergleichen. (LIM/LI1.00410 Schönberger 1970)

Entscheidend für die Definition der fünf Funktionstypen ist die Art des dem Pronomen übergeordneten Wortes, welches im Folgenden als der Definitor des jeweiligen Funktionstyps bezeichnet und in den Beispielen durch Fettdruck kenntlich gemacht wird. Für die Definition der Funktionstypen 1 und 2 ist außerdem von Belang, dass das Pronomen dem Definitor vorangestellt (Funktionstyp 1) bzw. nachgestellt (Funktionstyp 2) ist. Keine Relevanz für die Definition der fünf Funktionstypen hat dagegen das unterstrichene Wort, auf das der pronominale Hinweis gerichtet ist. Aus Tabelle 1 ist zu ersehen, wie die insgesamt 1 379 Belege sich in relativer und demonstrativer Verwendung auf die fünf Funktionstypen verteilen.

Wie aus der Tabelle hervorgeht, hat *deren* die breitere Streuung, während *derer* die speziellere Variante ist. Bemerkenswert ist, dass die Form *derer* nicht nur als Demonstrativum, sondern auch als Relativum verwendet wird. Besonders auffällig ist, dass beim Relativum in Funktionstyp 3, wo das Pronomen von einer Präposition regiert wird, die Form *derer* viermal häufiger vorkommt als die Form *deren*.

Alle 18 Belege mit *derer* als Relativum, das ja obligatorisch rückverweisend ist, verstoßen gegen die gängige Regelformulierung, die von der Verweisrichtung des Pronomens ausgeht. Dieser Befund lässt vermuten, dass nicht die Verweisrichtung des Pronomens, sondern andere Prinzipien die Wahl zwischen *deren* und *derer* steuern. Diese Prinzipien sollen im Folgenden aufgedeckt werden.

	Relativum		Demonstrativum		Insgesamt	
	deren	*derer*	*deren*	*derer*	*deren*	*derer*
Funktionstyp 1	903	0	349	0	1252	0
Funktionstyp 2	0	0	0	45	0	45
Funktionstyp 3	4	16	0	2	4	18
Funktionstyp 4	26	2	0	0	26	2
Funktionstyp 5	24	0	8	0	32	0
Insgesamt	957	18	357	47	1314	65

Tab. 1: Verteilung von *deren* und *derer* auf die Funktionstypen

1.3 Neuer Vorschlag

Da die Funktionstypen 1 und 2, denen gemeinsam ist, dass ihr Definitor ein Substantiv ist, sich durch die topologische Position des Pronomens vor bzw. nach dem Definitor und durch die konsequente Wahl von *deren* bzw. *derer* unterscheiden, liegt es nahe, die topologische Position des Pronomens im Verhältnis zum Definitor mit der Wahl der jeweiligen Pronominalvariante zu korrelieren und folgende Regel zu formulieren: Wenn das Pronomen vor dem Definitor steht, wird *deren* verwendet, wenn das Pronomen nach dem Definitor steht, wird *derer* verwendet. Diese Positionsregel besitzt für die Funktionstypen 1 und 2 fast volle Gültigkeit. (Erst in jüngster Zeit wird bisweilen von ihr abgewichen, und zwar begegnen Fälle mit *derer* als vorangestelltem Attribut, also in Funktionstyp 1, wie in Kapitel 2 gezeigt wird.) Die Positionsregel scheint auch für Funktionstyp 3 zu gelten, denn in sämtlichen Belegen zu Funktionstyp 3 ist das Pronomen dem Definitor, also der Präposition, nachgestellt, und die Form *derer* hat hier eindeutig den Vorrang. Auf die Funktionstypen 4 und 5, wo die Positionsregel nicht gilt und *deren* aus anderen Gründen die Norm geworden ist, soll hier nicht eingegangen werden. Sie wurden an anderer Stelle ausführlich behandelt (Bærentzen 1995).

Funktionstypen 1 und 2

In den Funktionstypen 1 und 2 beruht die komplementäre Verwendung von *deren* und *derer* in subtiler Weise auf morphologischen Zwängen kasusnonkongruenter Attribute im Genitiv Singular Femininum und Genitiv Plural. In der Funktion als Attribut eines Substantivs unterliegen beide Pronominalformen nämlich den in (10a-c) angeführten allgemeinen und zum Teil gegensätzlichen formalen Forderungen:

(10)

a) Es muss an den Attributen selbst zum Ausdruck kommen, dass es sich um kasusnonkongruente Attribute handelt.

b) Syntagmen im Genitiv Singular Femininum und Genitiv Plural müssen an ihrem ersten kasusflektierten Wort durch das starke (pronominale) Flexiv *-er* gekennzeichnet sein. Es heißt (*die Qual*) *der Arbeit*, (*die Qual*) *viel schwerer Arbeit*, (*die Qual*) *der schweren Arbeit*, (*die Qual*) *vieler Arbeit*, nicht aber *(*die Qual*) *Arbeit*, *(*die Qual*) *viel Arbeit*, *(*die Qual*) *schweren Arbeit*.

c) Kasusnonkongruente Attribute werden dem Substantiv nachgestellt. Kasuskongruente Attribute werden dem Substantiv vorangestellt.

(Eine Ausnahme von den Forderungen (10b) und (10c) bilden die kasusnonkongruenten Einwortattribute im Genitiv Singular Femininum auf *-s*, die erstens nicht durch *-er* gekennzeichnet sind und zweitens vor dem Substantiv stehen können: *Mutters Kleid*, *Sophies Bruder*.)

Den in (10a-c) genannten Regularitäten werden *deren* und *derer* in folgender Weise gerecht:

Die in (10a) formulierte Forderung, dass der kasusnonkongruente Status von *deren* und *derer* deutlich sein muss, wird durch das bloße Vorhandensein der Doppelflexivik erfüllt. Die entsprechende kasuskongruente Form ist *der* mit nur einem Flexiv.

Die in (10b) formulierte Forderung nach formaler Kennzeichnung eines Syntagmas im Genitiv Singular Femininum und Genitiv Plural durch *-er* erfüllt das erste Flexiv beider Formen.

Der in (10c) beschriebenen Normalkorrelation zwischen Voranstellung und Kasuskongruenz bzw. zwischen Nachstellung und Kasusnonkongruenz des Attributs passen sich *deren* und *derer* eben durch die Varianz von *-en* und *-er*

im zweiten Flexiv an. Diese Varianz spiegelt das Streben nach morphologischen Normalzuständen beim vorangestellten bzw. nachgestellten Attribut wider:

Dass beim vorangestellten Pronomen (Funktionstyp 1) im zweiten Flexiv die Variante *-en* gewählt wird, lässt sich folgendermaßen erklären: Vorangestellte kasusnonkongruente Attribute sind fast ausschließlich Einwortgenitive auf *-s*. Deshalb ist ein vorangestelltes nonkongruentes Attribut, das nicht auf *-s* ausgeht, eine irreguläre Erscheinung, und es besteht die Neigung, ein solches Attribut als kasuskongruent aufzufassen. Dieser Neigung kommt das zweite Flexiv *-en* entgegen, das – wie aus dem Adjektivparadigma bekannt – keinen eindeutigen formalen Stellenwert hat und somit geeignet ist, die Nonkongruenz des Pronomens zu verschleiern und eine Kongruenz mit dem übergeordneten Substantiv vorzutäuschen. Die Annahme, dass am vorangestellten Pronomen eine Kongruenz angestrebt wird, unterstützen die gelegentlich auftretenden Fälle mit der Form *derem*, wie in Beispiel (11), wo das zweite Flexiv *-em* eine volle Kongruenz mit dem Substantiv *Kleinkind* realisiert, sowie die in Kapitel 2 angeführten Fälle (20), (21), (25), (28) und (29) mit *derer* als vorangestelltem Attribut, in denen das zweite Flexiv *-er* sich als kongruent interpretieren lässt.

(11) Grössinger nahm eine Arbeit bei einer Reinigungsfirma an und wohnte mit einer jungen Frau und *derem* **Kleinkind** zusammen. (N92/OKT.39618 Salzburger Nachrichten, 24.10.1992)

Beim nachgestellten Pronomen (Funktionstyp 2), wo eine (Pseudo-)Kongruenz fehl am Platz wäre, wird im zweiten Flexiv die Variante *-er* gewählt. Mit *-er* als finalem Element zeigt *derer* volle formale Übereinstimmung mit allen übrigen nachgestellten Attributen im Genitiv Singular Femininum und Genitiv Plural, deren erstes kasusflektiertes Wort gemäß der in (10b) formulierten Regel auf *-er* ausgehen muss.

Funktionstyp 3

Auch in Funktionstyp 3, wo das Pronomen von einer Präposition regiert wird, erfolgt die Wahl zwischen *deren* und *derer* gemäß der Positionsregel. Das beruht darauf, dass Funktionstyp 3 sich in Wirklichkeit als eine Abwandlung der Funktionstypen 1 und 2 betrachten lässt, was Beispiel (12) aus Funktionstyp 1 und Beispiel (13) aus Funktionstyp 3 deutlich machen. In (12) begegnet in Übereinstimmung mit der Positionsregel die Form *deren*, weil das Pronomen

dem Substantiv *Hilfe* vorangestellt ist. In (13) hat die Fügung *mit Hilfe* präpositionalen Charakter gewonnen, so dass das Pronomen nach dem Substantiv *Hilfe* zu stehen kommt und in Übereinstimmung mit der Positionsregel die Form *derer* aufweist:

(12) Die Bedienung hatte sich zurückgezogen und auf eine Klingel verwiesen, mit *deren* **Hilfe** sie jederzeit herbeizuzitieren sei. (MK1/TPM.00000 Pinkwart 1963: 76)

(13) Die Physiker [...] haben eine vollständige Theorie des Weltalls, **mit Hilfe** *derer* sie alle durch Experimente festgestellten Tatsachen erklären können. (MK1/WBO.00000 Bamm 1963: 86)

Wie die Fügung *mit Hilfe* lassen sich auch die übrigen genitivregierenden Präpositionen, die in den 22 Belegen zu Funktionstyp 3 vorkommen, sprachgeschichtlich auf Substantive zurückführen. Die belegten Präpositionen sind: *innerhalb* (10mal), *auf Grund / aufgrund* (4mal), *mittels* (3mal), *mit Hilfe*, *kraft*, *an Stelle*, *seitens*, *hinsichtlich* (je einmal). Somit steht in Funktionstyp 3 das Pronomen in Wirklichkeit als Attribut bei einem Substantiv präpositionalen Charakters. In allen Belegen zu Funktionstyp 3 ist das Pronomen der Präposition, d.h. dem präpositionalen Substantiv, nachgestellt, und deshalb ist *derer* die nach der Positionsregel zu erwartende Form, die denn auch 18mal begegnet und also eindeutig den Vorrang hat. Nur in 4 Belegen wird – regelwidrig – *deren* verwendet, was wohl auf den normativen Einfluss derjenigen Grammatiken zurückzuführen ist, die *derer* als Relativum ablehnen. Das sind Fälle wie (14) und (15):

(14) [...] die Gerichtsklausel, **kraft** *deren* der Schützling die unmittelbare Gerichtsbarkeit des Schutzherrn [...] in Anspruch nehmen darf. (LIM/LI1.00245 Kisch 1970)

(15) Die ganze Aktivität des Abendlandes rührt ja nicht von einer theoretischen Einsicht her, **auf Grund** *deren* unsere Vorfahren sich berechtigt gefühlt hätten zu handeln, sondern es war ganz anders. (MK1/WHN.00000 Heisenberg 1963: 45)

So weit meine ersten Beobachtungen zu *deren* und *derer* in den Funktionstypen 1 bis 3, wie ich sie in den neunziger Jahren (Bærentzen 1995) beschrieben habe. Sie waren nur möglich, weil mir damals das Mannheimer Korpus und das Limas-Korpus in elektronischer Form zur Verfügung standen.

2. Neue Beobachtung: *derer* als vorangestelltes Attribut

In letzter Zeit wurde mein Interesse für *deren* und ins Besondere *derer* erneut geweckt, da mir beim Studium jüngerer Korpora auffiel, dass nun – im Widerspruch zu den im ersten Kapitel beschriebenen Regularitäten – in einigen Fällen *derer* als vorangestelltes Attribut, also in Funktionstyp 1, begegnet, wie ich es schon früher in Bærentzen (2008) geschrieben habe. Im *Archiv der geschriebenen Sprache* des IDS finden sich unter den ersten 4000 *derer*-Belegen (alle aus neueren Zeitungen) nicht weniger als 68 Fälle mit *derer* als vorangestelltem Attribut. Die Fälle verteilen sich auf zwei Varianten, je nachdem, ob das zweite Flexiv *-er* sich allein als nonkongruent oder sowohl als nonkongruent wie auch als kongruent interpretieren lässt. Diese Interpretationen werden im Folgenden durch Fälle mit den ungewöhnlichen Pronominalformen *desses* und *desser* unterstützt, die im gesamten IDS-Archiv der geschriebenen Sprache 14mal bzw. 10mal als vorangestelltes Attribut vorkommen.

Zur ersten Variante gehören Beispiele wie (16) und (17):

(16) Das zuständige Privatisierungsministerium will dagegen nur die Entstehung einer Aktiengesellschaft genehmigen, *derer* **Tätigkeit** auf zwölf Monate befristet sein würde. (M96/606.23275 Mannheimer Morgen, 10.06.1996)

(17) Marie geht es so wie vielen Mädchen und Jungen. Wenn sie sich wehrt, weil ihre Tante sie wieder einmal herzen will, dass ihr an *derer* **Busen** fast die Luft wegbleibt, gilt sie als unartig. (R99/SEP.77142 Frankfurter Rundschau, 24.09.1999)

Im ersten Flexiv von *derer* manifestiert sich wie immer der nonkongruente Kasus Genitiv des Femininums Singular oder des Plurals. Aber auch das zweite Flexiv ist in (16) und (17) eindeutig nonkongruent. Der Stellenwert des Attributionsnomens *Tätigkeit* in (16) ist Nominativ Singular Femininum, der Stellenwert des Attributionsnomens *Busen* in (17) ist Dativ Singular Maskulinum. Diese Stellenwerte sind für das zweite Flexiv *-er* der Pronominalform *derer* ausgeschlossen. Das zweite Flexiv *-er*, das anstelle des Standardflexivs *-en* eingetreten ist, muss hier also ein nonkongruentes Flexiv mit dem Stellenwert Genitiv Singular Femininum sein und ist als eine Dublette des ersten Flexivs zu sehen und verdeutlicht eben die Nonkongruenz des Pronomens mit dem Attributionsnomen. Genitiv Singular Femininum muss es in beiden Fällen sein, weil *derer* in (16) auf das Substantiv *Aktiengesellschaft*, in (17) auf das Substantiv *Tante* verweist.

Die Analyse von *derer* in (16) und (17) wird durch Fälle wie (18) und (19) mit der ungewöhnlichen Pronominalform *desses* unterstützt. Der Stellenwert des Attributionsnomens *Identität* in (18) ist Nominativ Singular Femininum, und der Stellenwert des Attributionsnomens *Beschwerden* in (19) ist Nominativ Plural. Das zweite Flexiv *-es* des Pronomens, das anstelle des Standardflexivs *-en* eingetreten ist, kann hier kein kongruentes Flexiv sein, sondern muss ein nonkongruentes Flexiv mit dem Stellenwert Genitiv Singular Maskulinum sein, da das Pronomen auf die Maskulina *der Erkrankte* bzw. *einem Hexenschuss* verweist. Das zweite Flexiv in *desses* ist als eine Dublette des ersten Flexivs zu sehen und verdeutlicht die Nonkongruenz des Pronomens mit dem Attributionsnomen:

(18) Im Tal starb der Erkrankte, *desses* **Identität** nicht bekannt gegeben wurde. (N00/JAN.00496 Salzburger Nachrichten, 07.01.2000)

(19) Der Nationalstürmer laboriert an einem Hexenschuß, *desses* **Beschwerden** auch trotz intensiver Behandlung eines Chiropraktikers nicht beseitigt werden konnten. (M98/OKT.83429 Mannheimer Morgen, 21.10.1998)

Zur zweiten Variante der Fälle mit *derer* als vorangestelltem Attribut gehören Beispiele wie (20) und (21).

(20) Beide treten diesmal nicht als offizielle Kandidaten ihrer Parteien an, erfreuen sich aber dennoch *derer* **Unterstützung**. (R99/NOV.91571 Frankfurter Rundschau, 11.11.1999)

(21) Meine Ehefrau wollte dem Nationaltheater die Lüster wegen *derer* gewerblichen **Bestimmung** für eine so lange Zeit (zwei volle Spielzeiten!) nur gegen ein angemessenes Entgelt zur Verfügung stellen. (M01/JUL.54035 Mannheimer Morgen, 23.07.2001)

Im ersten Flexiv von *derer* manifestiert sich wie immer der nonkongruente Kasus Genitiv des Femininums Singular oder des Plurals. Das zweite Flexiv dagegen hat einen mehrdeutigen Stellenwert. In (20) und (21) mit dem Attributionsnomen *Unterstützung* bzw. *Bestimmung* im Genitiv Singular Femininum lässt sich nicht entscheiden, ob das zweite Flexiv *-er*, das anstelle des Standardflexivs *-en* getreten ist, nonkongruent oder kongruent ist. Bei der nonkongruenten Interpretation hätte es wie das erste Flexiv den Stellenwert Genitiv Plural, da das Pronomen auf das pluralische Substantiv *Parteien* bzw. *Lüster* verweist.

Bei der kongruenten Interpretation hätte es abweichend vom ersten Flexiv aber in Übereinstimmung mit dem Attributionsnomen *Unterstützung* bzw. *Bestimmung* den Stellenwert Genitiv Singular Femininum.

Dass beide Interpretationen möglich sind, zeigt ein Vergleich mit den ungewöhnlichen Pronominalformen *desses* und *desser* in den Parallelfällen (22) bis (24):

(22) Krimi (BRD, 1996) mit Dieter Landuris, Stefan Reck, Lene Beyer, Robert Voigt u.a. Die zehnjährige Lisa Timmermann wird auf der Kirmes entführt und in ein abgelegenes Kellerverlies gesperrt. Sie ist die Tochter des bekannten und exzentrischen Rockmusikers Chris Timmermann und *desses* **Ex-Frau** Liane. (O96/AUG.88683 Neue Kronen-Zeitung, 28.08.1996)

(23) Alexander III. ist [...] der einzige russische Zar, während *desser* langen **Regierungszeit** kein einziger Krieg stattfand und Rußland wirtschaftlich, sozial und moralisch geblüht hat. (R99/FEB.14770 Frankfurter Rundschau, 23.02.1999)

(24) [...] nachdem Heinz-Harald Frentzen seine Forderung auf Abwahl des Kerpeners wegen *desser* zuletzt rüder **Fahrweise** („Wer so fährt, darf nicht Sprecher unserer Vereinigung sein") zurückgezogen hatte. (R98/JUL.59237 Frankfurter Rundschau, 25.07.1998)

Das zweite Flexiv *-es* der Pronominalform *desses* in (22) ist mit Sicherheit ein nonkongruentes Flexiv, das abweichend vom Attributionsnomen *Ex-Frau* aber in Übereinstimmung mit dem ersten Flexiv *-es-* den Stellenwert Genitiv Singular Maskulinum hat, da das Pronomen auf das maskuline Substantiv *Rockmusikers* verweist. Im Unterschied hierzu ist das zweite Flexiv *-er* des Pronomens *desser* in (23) und (24) mit Sicherheit ein kongruentes Flexiv, das abweichend vom ersten Flexiv *-es-* aber in Übereinstimmung mit dem Attributionsnomen *Regierungszeit* in (23) bzw. *Fahrweise* in (24) den Stellenwert Genitiv Singular Femininum hat. Da nicht nur (22), sondern auch (23) und (24) Parallelfälle zu (20) und (21) sind, lässt sich also nicht entscheiden, ob bei *derer* in (20) und (21) das zweite Flexiv *-er* ein nonkongruentes oder ein kongruentes Flexiv ist. Die Tatsache, dass das attributive Adjektiv *gewerblichen* in (21) schwache Flexionsform aufweist, ist kein Indiz dafür, dass das zweite Flexiv von *derer* in diesem Fall als ein kongruentes Flexiv zu sehen ist, denn ein Vergleich mit (23) und (24), wo das zweite Flexiv von *desser* mit Sicherheit ein

kongruentes Flexiv ist, zeigt, dass ein Adjektiv nach dieser Pronominalform nicht nur schwach wie *langen* in (23), sondern auch stark wie *rüder* in (24) flektieren kann.

Auch in (25) vor einem Attributionsnomen (*Abkürzungsname*) im Nominativ Singular Maskulinum lässt sich nicht entscheiden, ob das zweite Flexiv *-er* von *derer* nonkongruent ist und wegen des Verweises auf das Femininum *Gartenausstellung* wie das erste Flexiv den Stellenwert Genitiv Femininum Singular hat, oder ob es kongruent ist und abweichend vom ersten Flexiv aber in Übereinstimmung mit dem Attributionsnomen den Stellenwert Nominativ Singular Maskulinum hat. Die Parallelfälle (26) und (27) zeigen, dass beide Interpretationen möglich sind. In (26) ist das zweite Flexiv von *desses* mit Sicherheit nonkongruent, in (27) ist das zweite Flexiv von *desser* mit Sicherheit kongruent:

(25) Die „Blumenstadt" Erfurt mit ihrer seit Jahrhundert angestammten Pflanzen- und Samenzucht überflügelt den Frankfurter Palmengarten mit ihrer „Gartenausstellung" an Größe und Vielfalt deutlich, auch wenn *derer* schlichter **Abkürzungsname** „EGA" das gar nicht vermuten läßt. (R97/DEZ.100345 Frankfurter Rundschau, 18.12.1997)

(26) Der Abrahamhof, *desses* **Rohbau** im Vorjahr errichtet wurde, soll fertiggestellt werden, sodaß der mächtige Lungauer Einhof 1993 eröffnet werden kann. (N92/MAR.10439 Salzburger Nachrichten, 19.03.1992)

(27) Als eine Satire auf den Umgang mit Helden und die Sensationsgier der Medien legte der britische Regisseur Stephen Frears [...] sein moralisches Märchen an, *desser* kommerzieller **Erfolg** nicht hauptsächlich auf sein Konto, sondern auf das des Hauptdarstellers ging. (R97/JUL.51602 Frankfurter Rundschau, 5.07.1997)

In (28) steht *derer* vor einem Attributionsnomen (*Umgebung*) im Dativ Singular Femininum. Das zweite Flexiv *-er* lässt sich als nonkongruent und als kongruent interpretieren. Im ersteren Fall hätte es wegen des Verweises auf das pluralische Substantiv *Wohnräumen* wie das erste Flexiv den Stellenwert Genitiv Plural. Im letzteren Fall hätte es abweichend vom ersten Flexiv aber in Übereinstimmung mit dem Attributionsnomen *Umgebung* den Stellenwert Dativ Singular Femininum.

(28) Ferner muss das Angebot am Arbeitsplatz oder in den Wohnräumen (oder in *derer* unmittelbaren **Umgebung**) des Käufers, in öffentlichen Verkehrsmitteln oder auf öffentlichen Strassen und Plätzen oder an einer Werbeveranstaltung erfolgt sein, die mit einer Ausflugsfahrt oder einem ähnlichen Anlass verbunden war. (A97/JUN.07073 St. Galler Tagblatt, 03.06.1997)

In (29) steht *derer* vor einem Attributionsnomen (*Kehlen*) im Genitiv Plural. Das zweite Flexiv *-er* lässt sich als nonkongruent und als kongruent interpretieren, hat aber in beiden Fällen den Stellenwert Genitiv Plural. Die nonkongruente Interpretationsmöglichkeit ergibt sich daraus, dass der pronominale Hinweis auf das pluralische Substantiv *Passanten* gerichtet ist. Die kongruente Interpretationsmöglichkeit ergibt sich daraus, dass das Attributionsnomen *Kehlen* im Genitiv Plural steht:

(29) Mit Popcorn versuchten die Renggli-Junioren Samuel, Seraina, Benjamin und Melanie die vielen Passanten zu locken. Das haute nicht besonders, wohl wegen *derer* trockener **Kehlen**. (E99/JUL. 19333 Züricher Tagesanzeiger, 26.07.1999)

3. Ausblick

Der im zweiten Kapitel beschriebene neue Gebrauch von *derer* als vorangestelltem Attribut bedeutet eine Normalisierung der formalen Struktur, da die neuen Interpretationsmöglichkeiten den allgemeinen Strukturregeln für vorangestellte Attribute im Deutschen folgen. In Fällen wie (16) und (17) mit eindeutigem nonkongruentem Stellenwert des zweiten Flexivs entspricht *derer* einem sächsischen Genitiv (wie in *Mutters Kleid, des Kaisers neue Kleider*). In Fällen wie (20), (21), (25), (28) und (29) mit mehrdeutigem nonkongruentem/kongruentem Stellenwert des zweiten Flexivs entspricht *derer* bei nonkongruenter Interpretation einem sächsischen Genitiv und bei kongruenter Interpretation einem kongruierenden Artikelwort. Bei kongruenter Interpretation ist ein Syntagma wie *wegen derer gewerblichen Bestimmung* als eine Parallele zum Syntagma *wegen ihrer gewerblichen Bestimmung* zu sehen. In Fällen, wo die kongruente Interpretation möglich ist, lässt sich das Pronomen als ein relatives/demonstratives Possessivum charakterisieren. Dass eine solche Entwicklung nicht auszuschließen ist, zeigt das Lateinische, wo auf der Basis der Genitivform *cujus* des Pronomens *qui* ein kongruierendes relatives und interrogatives Possessivum *cujus*, *cuja*, *cujum* entstand (Wackernagel 1928: 81f.).

Während der neue Gebrauch von *derer* als vorangestelltem Attribut als eine Normalisierung der formalen Struktur zu sehen ist, stellt der im ersten Kapitel als Funktionstyp 1 beschriebene Standardgebrauch von *deren* als vorangestelltem Attribut eine formale Anomalie dar, weil das zweite Flexiv *-en* weder eine Nonkongruenz noch eine Kongruenz deutlich zum Ausdruck bringt, sondern allenfalls eine Pseudokongruenz signalisiert, die im Formsystem des Deutschen sonst keine Entsprechung hat.

Deshalb ist zu erwarten, dass in der Verwendung als vorangestelltes Attribut *deren* allmählich durch *derer* verdrängt wird.

Literatur

Bærentzen, Per (1995): Zum Gebrauch der Pronominalformen *deren* und *derer* im heutigen Deutsch. In: Beiträge zur Geschichte der deutschen Sprache und Literatur 117: 199-217.

Bærentzen, Per (2008): Die Pronominalform *derer* als vorangestelltes Attribut. Anfänge einer grammatischen Umstrukturierung. In: Valentin, Jean-Marie (Hg.): Akten des XI. Internationalen Germanistenkongresses Paris 2005: „Germanistik im Konflikt der Kulturen". Bd. 4. (= Jahrbuch für Internationale Germanistik, Reihe A, Kongressberichte 80). Bern: Lang, 105-110.

Wackernagel, Jacob (1928): Vorlesungen über Syntax. 2. Reihe. 2. Aufl. Basel: Birkhäuser.

Belegquelle

Alle Beispielsätze stammen aus den Korpora des Instituts für Deutsche Sprache (http://www.ids-mannheim.de/service/#Korpora).

GEERT STUYCKENS

Zum Wesen der Subjektlücken in Verbzweitkoordination auf der Grundlage eines deutsch > niederländischen Übersetzungskorpus

Abstract

Dieser Beitrag vergleicht die Frequenz der SLF-Koordination (Subjektlücke in finiten / frontalen Sätzen)[1] gegenüber alternativen Koordinationskonstruktionen zwischen Deutsch und (daraus übersetztem) Niederländisch auf der Grundlage eines literarischen unidirektionalen deutsch > niederländischen Übersetzungskorpus. Dabei werden grammatische Beschreibungskategorien wie die thematischen Rollen beider involvierten Subjekte sowie die Form und der Bekanntheitsgrad des ersten Subjekts ermittelt und mit den Frequenzdaten in Beziehung gesetzt; mit dem Ziel zu überprüfen, ob und inwieweit eine Korrelation zwischen den grammatischen Kategorien und der Wahl der jeweiligen Koordinationsvariante gegeben ist und insbesondere welche Faktoren das Vorkommen einer Subjektlücke fördern bzw. hemmen. Die Ergebnisse dieser Beschreibung erlauben insofern theoretische Rückschlüsse auf die bis dato noch ungeklärte Natur der Subjektlücken, als sie darauf hindeuten, dass die Subjektlücken in den verschiedenen Koordinationsalternativen weder im Sprachvergleich noch einzelsprachlich-intern einheitlich zu analysieren sind.

1. Theoretischer Hintergrund, Methode und Ziel

Die zweite Konstituente eines deklarativen Hauptsatzes ist im Deutschen (Dt.) und im Niederländischen (Ndl.) grundsätzlich das finite Verb. Das allgemeine Muster dieser Verbzweit- (V2-)Stellung wird im Rahmen des sog. topologischen Feldermodells durch die Beispiele (1) für das Deutsche und (2) für das Niederländische illustriert (GG 77-78):[2]

	Vorfeld	linke Satzklammer	Mittelfeld	rechte Satzklammer	Nachfeld
(1)	Vielleicht	hatte	er dem verunglückten Freund	helfen wollen	–
(2)	Misschien	had	hij zijn verongelukte vriend	willen helpen	–

[1] Terminus nach Höhle (1983).

[2] Der Code verweist auf das einschlägige Korpusbeispiel: Die beiden Buchstaben sind die Initialen des Schriftstellers (hier: Günter Grass) und bezeichnen das Subkorpus, die Zahl ist die Kennnummer des Beispiels innerhalb des Subkorpus.

Das finite Verb an zweiter Stelle, wie *hatte / had* in (1-2), bildet die linke Klammer vorn im Satz. Die rechte Klammer hinten ist die feste Position für infinite Verbformen, wie die Infinitive *helfen wollen / willen helpen* in (1-2). Die beiden Klammern zusammen teilen den Satz in drei Felder: Die Konstituente vor der linken Satzklammer besetzt das Vorfeld, die Konstituenten zwischen den beiden Klammern bilden das Mittelfeld und was sich hinter der zweiten Klammer befindet, gehört zum Nachfeld. Welche Konstituente das Vorfeld füllt, hängt in V2-Aussagesätzen von der Informationsverteilung des Satzes ab. Traditionell spricht man von Inversion, wenn eine andere Konstituente als das Subjekt, wie z.B. die Adverbialbestimmung der Modalität *vielleicht / misschien* in (1-2), die Satzanfangsposition besetzt. Das Subjekt tritt dann ins Mittelfeld, sodass das finite Verb auch in diesem Fall an zweiter Stelle steht.

Deutsch (3) weist eine Koordinationskonstruktion mit Inversion im ersten und einem unsichtbaren Subjekt (einer Subjektlücke) im zweiten V2-Teilsatz auf: die sog. SLF-Koordination. Die Subjektlücke (SL) entspricht dem vorangehenden ersten Subjekt.

SLF kommt auch im Niederländischen vor (siehe Beispiel (4)), ist aber aus normativer Sicht unter Umständen weniger akzeptabel als im Deutschen (Van de Velde 1986: 509).

(3) Vielleicht hatte er dem verunglückten Freund helfen wollen und war dabei gleichfalls unter die Eisdecke geraten. (GG 77)

(4) Misschien had hij zijn verongelukte vriend willen helpen en was daarbij eveneens onder het ijs geraakt. (GG 78)

Die Art der Lücke ist in der Literatur umstritten. Nach der traditionellen Grammatik (Duden 2005: 912, 1410; E-ANS:[3] 27·5·1) entsteht die Lücke durch Zusammenziehung, d.h. durch die Einsparung des gemeinsamen Subjekts der gereihten Teilsätze. Der Zusammenziehungsansatz wird auch in einigen generativen Grammatiken unter dem Namen 'Konjunktions- oder Koordinationsreduktion', 'Koordinationsellipse' oder 'Subjekttilgung' vertreten. Gemeint ist dann, vereinfachend gesagt, dass das phonologische Material in einer Art Tiefenstruktur vorhanden ist und auf dem Wege zur Oberflächenstruktur getilgt wird (Johannessen 1998, Zwart 1991). Der Hauptvorteil dieses Ansatzes liegt darin, dass 'Zusammenziehung' bzw. 'Tilgung' sich „mit einer Erweiterungsprobe nachweisen" lässt (Duden 2005: 1033) bzw. die Umkehroperation 'Rekonstruktion' impliziert und insoweit die Beziehung zwischen SLF-Koordi-

[3] Elektronische Algemene Nederlandse Spraakkunst.

nation und konkurrierenden Konstruktionen mit wiederholtem Subjekt einschließt, als das Subjekt an der Tilgungsstelle rekonstruiert (wiederaufgenommen) werden kann. An der Frage, ob die Tilgungsstelle links (Zwart 1991) oder rechts (Höhle 1990) vom zweiten finiten Verb zu finden ist, scheiden sich jedoch die Geister. Es ist mit anderen Worten unklar, ob der zweite Teilsatz invertiert ist oder nicht. Ein verwandtes Problem liegt in der Definition von gemeinsamen Subjekten, bei der es nicht mehr allein um Wortfolge, sondern um gleiche oder ungleiche Wortfolge in den Teilsätzen (strukturelle Identität) bzw. um Identität im weiteren Sinne geht. Die Definition reicht nämlich von der strengsten Anforderung formaler, funktionaler, struktureller und semantischer Gleichheit bis hin zur einigermaßen vagen Bestimmung als Referentengleichheit. Die Anforderung strengster Identität beruht auf einer Regel, die wohl aufgrund der Bemühungen präskriptiver Grammatiker im 17. Jahrhundert besonders in der niederländischen Tradition gängig ist. Diese Regel besagt, dass nur eine parallele Wortfolge in beiden Teilsätzen (strukturelle Gleichheit) erlaubt ist, sodass in dem Beispiel (4) die SL nach dem zweiten Verb angesiedelt werden müsste. Die Anforderung bloßer Referentengleichheit würde auch die Ansiedlung der SL vor dem Verb (strukturelle Ungleichheit) zulassen und erscheint angemessener im Hinblick auf Fälle wie (5-6), in denen sich das fehlende Subjekt eher als *er / hij* denn als *ein Jäger / een jager* rekonstruieren ließe (formale Ungleichheit):

(5) Dann kam auf einmal ein Jäger an und erschoss das Häschen.

(6) Toen kwam er opeens een jager aan en schoot het haasje dood.
(Zwart 1996: 265)

Andere generative Analysen umgehen die Ellipsenprobleme, indem sie die SL als *pro* (Fanselow 1990, Hartmann 1994) oder *PRO* (Höhle 1990 als Möglichkeit neben Ellipse) betrachten. Sowohl *pro* als auch *PRO* sind basisgenerierte Pronomina ohne phonetische Eigenschaften, die anderen Anwendungsbereichen, nämlich dem Bereich der *pro*-drop-Sprachen wie Italienisch oder Spanisch bzw. dem der infiniten Verben, entstammen. Sie unterscheiden sich dadurch, dass *pro* nicht anaphorisch und *PRO* anaphorisch ist. Die Analyse der SL als eines leeren Pronomens stößt aber auch auf (theorieinterne) Schwierigkeiten, von denen eine aus Fällen wie (7-8) hervorgeht:

(7) Gestern gingen einige Jäger in den Wald und fingen einen Hasen.

(8) Gisteren gingen enkele jagers naar het bos en vingen een haas.

Ein leeres Pronomen kann nicht von einem echt quantifizierenden Antezedens wie *einige Jäger / enkele jagers* gebunden werden. Da (7-8) grammatisch sind, kann die SL also kein leeres Pronomen sein (siehe Kathol 1999: 318).

Im Rahmen der kognitiven Grammatik, deren Syntaxmodell mehr auf dem sichtbaren als auf dem unsichtbaren Material basiert, argumentiert Kathol (1999), dass das erste Subjekt zwischen beiden Teilsätzen geteilt wird. Die SL ist daher keine leere Kategorie. In der generativen Grammatik wäre ein geteiltes Subjekt nur möglich ohne Inversion im ersten Teilsatz (d.h. nicht in SLF-Koordination), wie in (9-10):

(9) [CP Der Jäger$_i$ [C' ging t_i in den Wald] und [C' fing t_i einen Hasen]].

(10) [CP De jager$_i$ [C' ging t_i naar het bos] en [C' ving t_i een haas]].

Das Subjekt wird hier aus beiden Konjunkten in den nicht koordinierten Teil des Satzes bewegt und hinterlässt eine Spur *t* in beiden C'-Konjunkten. Da das Subjekt am Satzanfang steht, kann es über beide Konjunkte distribuieren, sodass die Agreement- und Valenzbedingungen der jeweiligen Verben erfüllt sind. Im Grunde betrachtet Kathol (1999) die Koordinationskonstruktion in (9-10) und die SLF-Koordination in (11-12) als zwei Linearisierungsvarianten derselben Koordination von Prädikaten:

(11) Gestern ging der Jäger in den Wald und fing einen Hasen.

(12) Gisteren ging de jager naar het bos en ving een haas.

Dies scheint für das Deutsche zuzutreffen, wo das Subjekt von (9) sowie von (11) nur vor dem zweiten Verb wiederaufgenommen werden kann (13 und 15), nicht jedoch für das Niederländische, wo sich das Subjekt von (10) zwar nur vor dem zweiten Verb (17) wiederholen lässt, jenes von (12) sich aber sowohl davor als auch danach wiederholen ließe (19-20):

(13) Der Jäger ging gestern in den Wald und er fing einen Hasen.

(14) *Der Jäger ging gestern in den Wald und fing er einen Hasen.

(15) Gestern ging der Jäger in den Wald und er fing einen Hasen.

(16) *Gestern ging der Jäger in den Wald und fing er einen Hasen.

(17) De jager ging gisteren naar het bos en hij ving een haas.

(18) *De jager ging gisteren naar het bos en ving hij een haas.

(19) Gisteren ging de jager naar het bos en hij ving een haas.

(20) [OK]Gisteren ging de jager naar het bos en ving hij een haas.

Es dürfte nun deutlich sein, dass keiner der obigen drei Ansätze – Tilgung, leeres Pronomen und geteiltes Subjekt – die Art der SL in SLF-Koordination adäquat erfasst. Außerdem vermeiden die Vorschläge eine systematische Konfrontation mit der tatsächlichen Distribution von SLF und den konkurrierenden Konstruktionen ohne Inversion im ersten und/oder mit wiederaufgenommenem Subjekt im zweiten Teilsatz. Die Argumentation zugunsten des einen oder anderen Ansatzes ist ziemlich intratheoretisch und vernachlässigt Kontraste zwischen Deutsch und Niederländisch.

Deshalb führe ich im Folgenden auf der Grundlage eines unidirektionalen deutsch > niederländischen Übersetzungskorpus eine deskriptive und kontrastive Korpusuntersuchung zur Distribution der Koordinationsalternativen durch. Ziel der Untersuchung ist, auf einer deskriptiven Ebene herauszufinden, welche mit dem Subjekt zusammenhängenden Faktoren die Distribution beeinflussen und welche dieser Faktoren das Vorkommen einer SL begünstigen.[4] Auf einer theoretischeren Ebene soll das Wesen der SL im Deutschen und im Niederländischen beleuchtet werden.

2. Korpusuntersuchung

2.1 Klassifizierung der Koordinationstypen

Zuerst muss als Ausgangspunkt eine Klassifizierung der verschiedenen alternativen Koordinationstypen in einige wenige Haupttypen durchgeführt werden, um die Recherchen kontrollierbar zu machen. Dazu schlage ich die Klassifikation in Tabelle 1 vor, in der die SLF-Merkmale als Eigenschaften dienen, die einen positiven oder negativen Wert haben können. Unklare Wertigkeiten werden durch ein Fragezeichen gekennzeichnet. Jede grammatische Kombination der Werte ergibt einen anderen Typ. Alle (Sub-)Typen mit besetztem zweitem Vorfeld werden unter Typ 5 subsumiert:

[4] Dieses *Begünstigen* kommt dem Begriff der *Lizenzierung* aus der generativen Grammatik nahe. Diesen Begriff vermeidet die vorliegende Arbeit aber, weil sie möglichst theorieübergreifend sein will.

	1. Teilsatz invertiert	2. Teilsatz invertiert	2. Vorfeld besetzt	2. Subjekt explizit
Typ 1 – SLF	+	?	–	–
Typ 2	–	–	–	–
Typ 3	?	?	–	–
Typ 4	+	+	–	+
Typ 5				
Subtyp 5.1.	+	+	+	+
Subtyp 5.2.	+	–	+	+
Subtyp 5.3.	–	+	+	+
Subtyp 5.4.	–	–	+	+

Tab. 1: Klassifikation der Alternativen

Vier Typen sind oben schon durch Beispiele veranschaulicht worden: Typ 1, d.h. SLF-Koordination, in (3-8 und 11-12), Typ 2 in (9-10), Typ 4 in (16 und 20) und Typ 5 in (13, 15, 17 und 19). Lediglich Typ 3 gilt es hier noch zu illustrieren:

(21) {Danach zapfte er wieder Bier, [SL? legte SL? Soleier oder zum Mostrichklacks Buletten auf Teller,} goß Lage nach Lage Korn bis zum Strich ein.] (GG 209)

(22) {Daarna tapte hij weer bier, [SL? legde SL? gepekelde eieren of gehaktballen naast de klodder mosterd op borden,} schonk het ene rondje jenever na het andere tot het streepje in.] (GG 210)

Ein Fall vom Typ 3, (wie in (21-22) der Teil in eckigen Klammern), folgt prinzipiell in einer Koordinationsreihe auf einen solchen vom Typ 1, vgl. ebd. den Part zwischen geschweiften Klammern. Der erste Teilsatz des Ersteren ist dem zweiten Teilsatz des Letzteren gleich und seine Wortfolge, Inversion oder nicht, gleichermaßen fraglich. Typ 3 ist also im Grunde ambig zwischen Typ 1 und Typ 2.

Die Klassifizierung in fünf Grundtypen orientiert sich an der traditionellen Dreiteilung der generativen Grammatik zwischen SLF-Koordination (= mein Typ 1), Koordination vollständiger Teilsätze (CP-Großkonjunkte), die sich mit meinem Typ 5 deckt, und phrasaler Koordination (C'-Kleinkonjunkte). Dabei ist die Kategorie der phrasalen Koordination für den kontrastiven Zweck dieser Arbeit allerdings nicht feinkörnig genug, da phrasale Koordination mit

nicht invertierten Teilsätzen (= mein Typ 2) sowohl im Deutschen (9) als auch im Niederländischen (10) grammatisch ist, während solche mit invertierten Teilsätzen (= mein Typ 4) unter Umständen zwar im Niederländischen (20) grammatisch im Deutschen (16) aber hochmarkiert, wenn nicht gar völlig ungrammatisch, ist (vgl. Van de Velde 1986: 506-508, 510). Deswegen sind Typ 2 und 4 hier unbedingt getrennt zu halten. Überdies kommt in meinem Vorschlag zusätzlich zu den Typen 1 und 2 noch der zwischen beiden doppeldeutige Typ 3 hinzu, der für die Theoriebildung der generativen Grammatik irrelevant ist: Offene Fragen, u.a. bzgl. der SL-Position im ersten Teilsatz von Typ 3, behandeln generative Ansätze nämlich alle im Rahmen des zweiten Teilsatzes der SLF-Koordination (= Typ 1). In der vorliegenden Arbeit ist die Annahme eines dritten Typs indes für die Aufarbeitung der Korpusdaten, zumal zur Auflösung mehrfacher Koordination in binäre, unerlässlich. Mein Vorschlag ist also eine formalistisch inspirierte, auf dem generativen Grammatikmodell basierende Klassifikation mit einer kontrastiv und einer deskriptiv einzelsprachlich motivierten Modifikation.

Von einem funktionalistischen Standpunkt aus könnte man an dieser Stelle einwenden, Typ 5 fasse unangemessen und inkonsequent Typen mit und ohne Inversion zusammen. In einer solchen Perspektive würde es tatsächlich einen Unterschied machen, ob der Sprachbenutzer, um einen bestimmten Inhalt auszudrücken, eine Inversion wählt oder unterlässt, denn jeder Veränderung in der Form der Konstruktion entspräche auch eine Veränderung in deren Bedeutung. Diese Arbeit zielt aber gerade darauf, aufgrund einer auf die Korpusuntersuchung zugeschnittenen formalen Klassifikation zu erforschen, wie die verschiedenen Typen im diskursiven Kontext fungieren und wie die Funktion der Typen im Kontext (kontextuelle Faktoren) ihre Form (Auftreten oder Unterbleiben einer SL) beeinflusst, d.h. zu untersuchen, wie das Form-Funktion-Mapping im Deutschen gegenüber dem Niederländischen genau geregelt ist. Die Absicht ist folglich nicht, a priori eine funktionale Klassifikation anzusetzen.

2.2 Korpus

In Ermangelung eines elektronisch verfügbaren deutsch > niederländischen Übersetzungskorpus habe ich manuell aus vier deutschen Romanen, geschrieben von Grass, Seghers, Süskind und Walser, 853 zweiteilige Koordinations-

konstruktionen,[5] die einem der fünf Typen angehören, und ihre niederländischen Übersetzung exzerpiert und sie in *Abundantia Verborum*, ein an der Universität Leuven entwickeltes Computertool zur Durchführung linguistischer Fallstudien (Speelman 2005, Fachbereich Linguistik, Universität Leuven), eingeführt. Dabei habe ich alle Korpusfälle mit der relevanten Typnummer (1, 2, 3, 4 oder 5) und Sprache (Deutsch oder Niederländisch) versehen. Anders als zwei einsprachige Korpora ermöglicht ein Übersetzungskorpus, parallele Beobachtungen zu vergleichen und aufzudecken, durch welche Form die jeweilige Sprache denselben Inhalt ausdrückt. Ferner erlaubt es eine systematische Analyse der bilingualen Intuition des Übersetzers. Johansson (2003) zufolge ist dies die beste Basis für den Sprachvergleich. Vorteil eines literarischen Korpus ist, dass man berechtigterweise annehmen darf, dass es die Standardsprache repräsentiert.

Die Fallstudie ist quantitativ-qualitativ gestaltet (Lemnitzer / Zinsmeister 2006: 36-37). Einerseits werden die absolute und relative Häufigkeit der Alternativen berechnet, andererseits werden die Daten, ausgehend von den konkreten Fällen, qualitativ interpretiert. Insbesondere operationalisiere ich die Forschungsfrage, welche mit dem Subjekt verbundenen Faktoren die Distribution beeinflussen und welche dieser Faktoren das Vorkommen einer SL begünstigen, indem ich die folgenden drei Variablen teste:

1) die Form des ersten Subjekts:

   Ich habe die Form des ersten Subjekts in den Korpusfällen auf einer dreistufigen Skala zunehmender Komplexität als Null, Pronomen oder volle NP annotiert.

2) die thematische Rolle beider Subjekte:

   In der Annahme, dass nicht nur die formale Seite des Subjekts, sondern auch seine funktional-semantische Seite, und zwar seine thematische Rolle, Einfluss auf die Verteilung haben könnte, habe ich die Korpusfälle danach unterschieden, ob ihre beiden Subjekte identische oder unterschiedliche thematische Rollen aufweisen.

3) der Informationsstatus des ersten Subjekts:

   Ebenfalls verwandt mit der Form des ersten Subjekts ist sein Informationsstatus. Van de Velde (1986: 506) suggeriert, dass die Thema-Rhema-Struktur, d.h. die Tatsache, ob das erste Subjekt aus alter / gegebener oder neuer

[5] Mehrfache Koordinationskonstruktionen wurden dabei rekursiv in binäre aufgelöst.

Information besteht, einen Einfluss auf die Distribution nehmen könnte. Strube / Hahn (1999) haben in Anlehnung an Prince (1981) eine dreistufige Gegeben-neu-Skala entwickelt: „höreralte", d.h. „(situational) evozierte" oder „ungebrauchte" Diskurseinheiten (z.B. allgemein bekannte Eigennamen); „vermittelte", d.h. „inferierbare" oder „brandneue verankerte" Diskurseinheiten; und „hörerneue" Diskurseinheiten. Ich habe das erste Subjekt der Korpusfälle als zu einer dieser drei Stufen gehörend codiert.

Zur qualitativen Interpretation werden also Kategorien herangezogen, die nicht unmittelbar aus den Daten hervorgehen, sondern Faktoren beschreiben, die möglicherweise die Distribution der Alternativen beeinflussen. Ob tatsächlich eine signifikante Korrelation zwischen dem jeweiligen Faktor und dem Gebrauch der Alternativen besteht, wird an einem Chi-Quadrat-Test für Kontingenztafeln überprüft.

Abbildung 1 stellt den Beobachtungseditor in *Abundantia Verborum* dar und soll zeigen, dass das Tool hier einen doppelten Zweck erfüllt: Einmal dient es zur Datensammlung („Contents") und dann zur Anreicherung der gesammelten Daten mit linguistischen Annotationen („Labels").

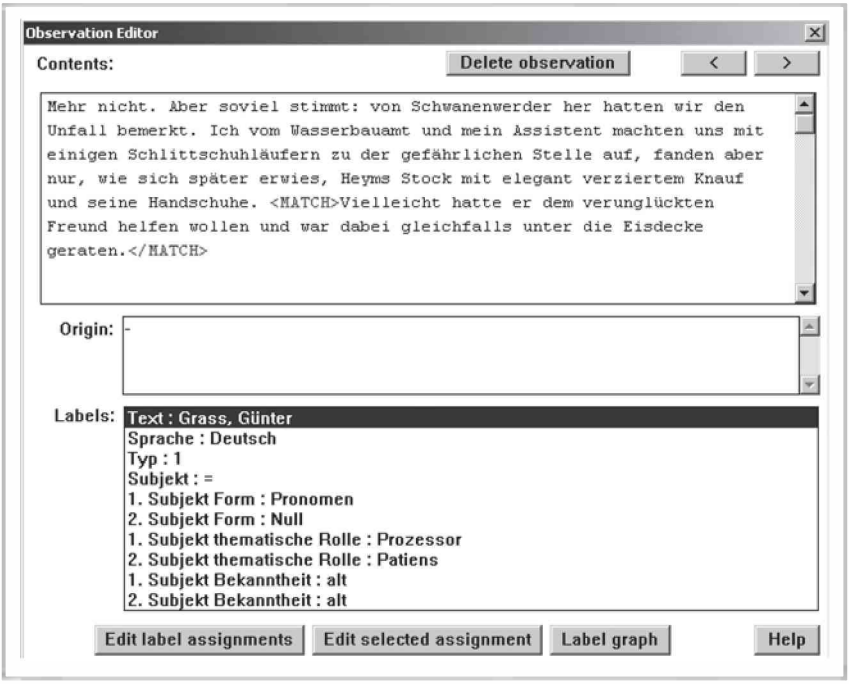


Abb. 1: Beobachtungseditor in *Abundantia Verborum*

An den angereicherten Daten werden statistische Berechnungen angestellt, die im nun folgenden Kapitel 3 wiederum in Grafiken und Kontingenztafeln zusammengefasst werden.

3. Ergebnisse

Die wichtigsten Ergebnisse gehen aus Abbildung 2 zur relativen Frequenz der Alternativen und den Tabellen 2 bis 4 hervor, die in Kontingenztafeln die Koordinationstypen und Sprachen in Kombination mit jeweils einer der drei Variablen darstellen. Die Resultate werden jeweils unter den Zahlenwerten erläutert:

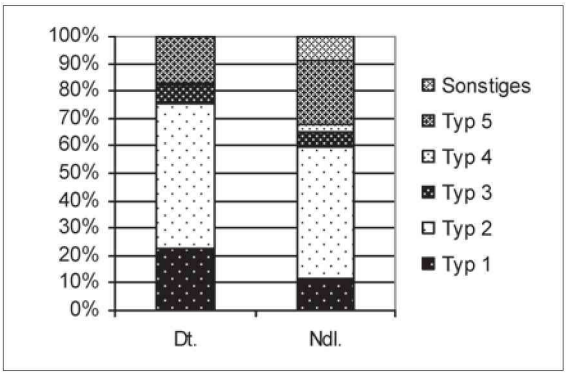


Abb. 2: Relative Häufigkeit der Alternativen

- Typ 1 ist ca. 15% häufiger im Deutschen als im Niederländischen (ca. 25% vs. 10%).
- Typ 2 ist ca. 5% häufiger im Deutschen als im Niederländischen (ca. 55% vs. 50%).
- Typ 3 ist im Deutschen und im Niederländischen ungefähr gleich häufig (ca. 5% = 5%).
- Typ 4 ist ca. 5% weniger häufig im Deutschen als im Niederländischen (ca. 0% vs. 5%).
- Typ 5 ist ebenfalls ca. 5% weniger häufig im Deutschen als im Niederländischen (ca. 20% vs. 25%).
- Im Allgemeinen verlieren Typ 1 und 2 (mit SL) und umgekehrt gewinnen Typ 4 und 5 (mit Wiederaufnahme) Treffer im Zuge der Übersetzung.

		Dt.		Ndl.	
		absolut	relativ	absolut	relativ
volle NP	Typ 1	59	21,00	26	11,26
	Typ 2	160	56,94	144	62,34
	(Typ 3	n.v.	n.v.	n.v.	n.v.)
	Typ 4	1	0,36	5	2,16
	Typ 5	61	21,71	56	24,24
Pronomen	Typ 1	134	37,43	71	21,19
	Typ 2	146	40,78	160	47,76
	(Typ 3	n.v.	n.v.	n.v.	n.v.)
	Typ 4	3	0,84	20	5,97
	Typ 5	75	20,95	84	25,07
(Null	Typ 1	n.v.	n.v.	n.v.	n.v.
	Typ 2	145	67,76	110	52,63
	Typ 3	59	27,57	43	20,57
	Typ 4	n.v.	n.v.	n.v.	n.v.
	Typ 5	10	4,67	56	26,79)

Tab. 2: Form des 1. Subjekts[6]

Dt.: $\chi^2 = 23{,}3$; df = 3; p = 0,000

Ndl.: $\chi^2 = 17{,}8$; df = 3; p = 0,000

- Typ 1 ist im Deutschen ca. 15% (ca. 20 vs. 35%) und im Niederländischen ca. 10% (ca. 10 vs. 20%) weniger häufig bei einer vollen NP als bei einem Pronomen als erstem Subjekt. Das Umgekehrte gilt für Typ 2 (ca. 55 vs. 40% im Deutschen bzw. ca. 60 vs. 50% im Niederländischen).

- Die Korrelation zwischen der Form des ersten Subjekts und der Wahl einer Alternative ist in beiden Sprachen äußerst signifikant (p = 0,000).

[6] Typ 3 weist als erstes Subjekt per definitionem eine Subjektlücke, d.h. die Nullform, auf. Demgegenüber kann die Nullform nicht als erstes Subjekt der Typen 1 und 4 auftreten. Es scheint daher vertretbar, diejenigen Teile der Tabelle, die entweder Typ 3 oder die Nullform betreffen, einzuklammern und von den Chi-Quadrat-Tests auszuschließen.

		Dt.		Ndl.	
		absolut	relativ	absolut	relativ
= tR	Typ 1	151	22,64	89	14,81
	Typ 2	365	54,72	332	55,24
	Typ 3	50	7,50	38	6,32
	Typ 4	3	0,45	16	2,66
	Typ 5	98	14,69	126	20,97
≠ tR	Typ 1	42	22,58	8	4,60
	Typ 2	86	46,24	82	47,13
	Typ 3	9	4,84	5	2,87
	Typ 4	1	0,54	9	5,17
	Typ 5	48	25,81	70	40,23

Tab.3: Semantik beider Subjekte [= bzw. ≠ thematische Rollen (tR)]

Dt.: $\chi^2 = 14{,}0$; df = 4; p = 0,007

Ndl.: $\chi^2 = 38{,}2$; df = 4; p = 0,000

- Typ 1 ist im Niederländischen ca. 10% häufiger bei identischen als bei unterschiedlichen thematischen Rollen für beide Subjekte (ca. 15% vs. 5%). Im Deutschen ist kein Einfluss der Semantik feststellbar.

- Typ 2 ist sowohl im Deutschen wie im Niederländischen ca. 10% häufiger bei identischen als bei unterschiedlichen Rollen (ca. 55% vs. 45%). Umgekehrt ist Typ 5 im Deutschen ca. 10% (ca. 15% vs. 25%) und im Niederländischen sogar ca. 20% (ca. 20% vs. 40%) seltener bei identischen als bei unterschiedlichen Rollen.

- Typ 5 gewinnt im Zuge der Übersetzung ca. 10% mehr Treffer bei unterschiedlichen als bei identischen Rollen (ca. +15% vs. +5%).

- Die Korrelation zwischen der Semantik beider Subjekte und dem Gebrauch einer oder der anderen Alternative ist in beiden Sprachen sehr signifikant.

		Dt.		Ndl.	
		absolut	relativ	absolut	relativ
alt	Typ 1	169	22,32	88	12,72
	Typ 2	400	52,84	377	54,48
	(Typ 3	59	7,79	43	6,21)
	Typ 4	4	0,53	24	3,47
	Typ 5	125	16,51	160	23,12
vermittelt	Typ 1	20	22,73	8	10,00
	Typ 2	49	55,68	36	45,00
	(Typ 3	n.v.	n.a.	n.v.	n.v.)
	Typ 4	0	0,00	1	1,25
	Typ 5	19	21,59	35	43,75
(neu	Typ 1	4	50,00	1	33,33
	Typ 2	2	25,00	1	33,33
	Typ 3	n.v.	n.v.	n.v.	n.v.
	Typ 4	0	0,00	0	0,00
	Typ 5	2	25,00	1	33,33)

Tab. 4: Informationsstruktureller Status des 1. Subjekts[7]

Dt.: $\chi^2 = 1{,}19$; df = 3; p = 0,7554

Ndl.: $\chi^2 = 13{,}79$; df = 3; p = 0,0032

- Im Niederländischen ist Typ 2 ca. 10% häufiger bei einem alten als bei einem vermittelten 1. Subjekt (ca. 55% vs. 45%), umgekehrt ist Typ 5 ca. 20% seltener (ca. 25% vs. 45%). Im Deutschen ist kein Einfluss des Informationsstatus feststellbar.
- Die Treffer des Typs 5 nehmen im Zuge der Übersetzung ca. 15% mehr zu bei einem vermittelten als bei einem alten Subjekt (ca. +15% vs. +5%).
- Typ 2 verliert im Zuge der Übersetzung ca. 10% mehr Treffer bei einem vermittelten als bei einem alten Subjekt (ca. –10% vs. –0%).

[7] Typ 3 weist eine Subjektlücke und dementsprechend inhärent eine alte Diskursentität als erstes Subjekt auf. Daneben sind nicht genügend Fälle mit neuem erstem Subjekt vorhanden. Diese Teile der Tabelle werden deshalb eingeklammert und nicht mit in die Chi-Quadrat-Tests einbezogen.

– Die Korrelation zwischen dem Informationsstatus des 1. Subjekts und der Wahl einer Alternative ist im Niederländischen (p = 0,0032) im Gegensatz zum Deutschen (p = 0,7554).

4. Schlussfolgerungen

Tabelle 5 fasst die Ergebnisse zusammen:

	Dt.		Ndl.
Typ 1	volle NP << Pron (= tR) = (≠ tR)	≈ ≠ =	volle NP < Pron (= tR) > (≠ tR) alt = vermittelt
Typ 2	volle NP >> Pron (= tR) > (≠ tR)	≈ = ≠	volle NP > Pron (= tR) > (≠ tR) alt > vermittelt
Typ 5	volle NP = Pron (= tR) < (≠ tR)	= ≈ ≠	volle NP = Pron (= tR) << (≠ tR) alt < vermittelt

Tab. 5: Übersicht

Alle drei Faktoren – Form des ersten Subjekts, Semantik beider Subjekte und Informationsstatus des ersten Subjekts – beeinflussen auf irgendeine Weise die Distribution. Das Vorkommen der SL in Typ 1 wird im Deutschen durch ein Pronomen als erstes Subjekt und im Niederländischen zusätzlich durch identische thematische Rollen für beide Subjekte erleichtert. Die Lücke in Typ 2 wird hingegen im Deutschen durch eine volle NP als erstes Subjekt sowie identische thematische Rollen für beide Subjekte und im Niederländischen darüber hinaus durch eine bekannte Entität als erstes Subjekt begünstigt.

Eine Schlussfolgerung besteht deshalb darin, dass in beiden Sprachen ein Faktor mehr Typ 2 als Typ 1 beeinflusst und der Einfluss des Formfaktors auf beide Typen genau entgegengesetzt ist: Bei Typ 1 begünstigen Pronomina die SL, während dies bei Typ 2 volle NPs tun. Typ 1 und 2 dürfen also nicht einheitlich analysiert werden (contra Kathol 1999).

Eine weitere Schlussfolgerung ist, dass im Niederländischen ein Faktor mehr als im Deutschen die SL sowohl in Typ 1 als in Typ 2 beeinflusst. Bei Typ 1 umfasst der zusätzliche Faktor eine Präferenz für identische thematische Rollen; bei Typ 2 besagt er, dass alte erste Subjekte bevorzugt werden. Subjektlü-

cken im Niederländischen ähneln daher eher einer richtigen Koordinationsellipse als Subjektlücken im Deutschen. Dementsprechend werden sie stärker als fehlendes Subjekt empfunden, das unter Umständen unbedingt wieder aufgenommen werden muss. Dies könnte erklären, warum Typ 4 und 5 in den niederländischen Übersetzungen frequenter sind als in den deutschen Originaltexten. Um sicher zu gehen, dass dieser Effekt nicht dem Übersetzungsprozess zuzuschreiben ist, werde ich die Resultate in künftiger Forschung an einem niederländisch > deutschen Übersetzungskorpus überprüfen.

Typ 5 wird schließlich im Deutschen durch unterschiedliche thematische Rollen und im Niederländischen zudem durch ein vermitteltes erstes Subjekt gefördert.

Zusammensetzung des Korpus

Grass, Günter (1999a): Mein Jahrhundert. Göttingen: Steidl.

Grass, Günter (1999b): Mijn eeuw. Übers. v. J. Gielkens. Amsterdam: Meulenhoff.

Seghers, Anna (1942): Das siebte Kreuz. Nachdr. 2007. Berlin: Aufbau.

Seghers, Anna (1984): Het zevende kruis. Übers. v. N. Rost. Amsterdam: Van Gennep.

Süskind, Patrick (1985): Das Parfum. Die Geschichte eines Mörders. Zürich: Diogenes.

Süskind, Patrick (2006): Het parfum. De geschiedenis van een moordenaar. Übers. v. R. Jonkers. Amsterdam: Prometheus.

Walser, Martin (1998): Ein springender Brunnen. Frankfurt am Main: Suhrkamp.

Walser, Martin (1999): Een springende fontein. Übers. v. R. van Hengel. Breda: De Geus.

Literatur

Duden (2005): Duden – Die Grammatik. (= Duden 4). Mannheim et al.

Elektronische Algemene Nederlandse Spraakkunst (E-ANS). Internet: http://www.ru.nl/e-ans/ (Stand: 08/2010).

Fanselow, Gisbert (1990): Minimale Syntax. (= Groninger Arbeiten zur germanistischen Linguistik 32). Groningen. [Zugl. Habil.schrift Univ. Groningen].

Hartmann, Katharina (1994): Zur Koordination von V2-Sätzen. In: Zeitschrift für Sprachwissenschaft 13, 1: 3-19.

Höhle, Tilman N. (1983): Subjektlücken in Koordinationen. Unveröff. Ms. Internet: http://www.uni-tuebingen.de/Deutsches-Seminar/hoehle/SLF-W5.1_neu.pdf (Stand: 08/2010).

Höhle, Tilman N. (1990): Assumptions about asymmetric coordination. In: Mascaró, Joan / Nespor, Marina (Hg.): Grammar in progress: Glow essays for Henk van Riemsdijk. Dordrecht: Foris, 221-235.

Johansson, Stig (2003): Contrastive linguistics and corpora. In: Granger, Sylviane / Lerot, Jacques / Petch-Tyson, Stephanie (Hg.): Corpus-based approaches to contrastive linguistics and translation studies. Amsterdam / New York: Editions Rodopi, 31-44.

Johannessen, Bondi (1998): Coordination. Oxford / New York: Oxford University Press.

Kathol, Andreas (1999): Linearization vs. phrase structure in German coordinate constructions. In: Cognitive Linguistics, 303-342.

Lemnitzer, Lothar / Zinsmeister, Heike (2006): Korpuslinguistik. Eine Einführung. Tübingen: Narr.

Prince, Ellen F. (1981): Toward a taxonomy of given-new information. In: Cole, Peter (Hg.): Radical pragmatics. New York: Academic Press, 223-255.

Speelman, Dirk (2005): Abundantia Verborum. A computer tool in support of corpus-based linguistic case studies. Internet: http://wwwling.arts.kuleuven.ac.be/genling/abundant/ (Stand: 12 / 2010).

Strube, Michael / Hahn, Udo (1999): Functional centering: Grounding referential coherence in information structure. In: Computational Linguistics 25, 3: 309-344.

Van de Velde, Marc (1986): Zum 65. gratuliere ich Ihnen und biete ?(ich) Ihnen diesen Beitrag an. In: Cox, Heinrich L. / Vanacker, Valeer F. / Verhofstadt, Edward (Hg.): Wortes anst – Verbi gratia: donum natalicum Gilbert A. R. De Smet. Leuven: Amersfoort, 503-512.

Zwart, C. Jan-Wouter (1991): Subject deletion in Dutch: A difference between subjects and topics. In: Kas, Mark / Reukland, Eric / Vet, Co (Hg.): Language and cognition. Yearbook 1991 of the Research Group for Linguistic Theory and Knowledge Representation of the University of Groningen. Bd. 1. Groningen, 333-350.

Zwart, C. Jan-Wouter (1996): Morphosyntax of verb movement: A minimalist approach to the syntax of Dutch. Dordrecht: Kluwer Academic Publishers.

ELMA KERZ

The role of low-level schemas in English academic writing

A usage-based constructionist approach

Abstract

Recently, considerable attention has been devoted to corpus-based studies on recurring multiword expressions in English academic language (e.g., Biber / Conrad / Cortes 2004, Biber 2006, Hyland 2008). However, these studies place their focus on uninterrupted sequences of lexical items, the so-called 'lexical bundles'. At the same time, several studies have indicated that the vast majority of prefabs, i.e., frequently recurrent strings of lexical items, are of flexible nature, i.e., including optional slots as well as semantically constrained slots (see, for instance, Schmitt / Carter 2004: 7).

The present study investigates the use and function of flexible sequences in the register of English academic writing. These prefabs will be referred to as 'flexible formulaic sequences' (henceforth FFSs). The focus is on FFSs in two groups of high frequency lexical items in the academic subcomponent of the BNC, viz. the 'research' verbs and nouns (e.g., *analyze*, *investigate* or *study*) and the 'coming-to-know' verbs (e.g., *find*, *suggest* or *show*). From a usage-based constructionist perspective, FFSs can be conceived of as register-specific low-level schemas: some slots of these constructions semantically (collocationally) constrain which items can fill them, while others are more schematic and can be understood as register-specific frame elements.

One of the major reasons why FFSs have been hitherto neglected has to do with their identification and extraction. The fact that the sequences involve slots whose filling is not restricted to a single lexical item or expression, but rather to a homogeneous group of lexical items, leads to difficulties, especially concerning their automatic extraction. Whereas lexical bundles, as defined by Biber, are easily extractable with the help of available software programs, the FFSs can only be appropriately distilled by a manual inspection of concordance lines.

This paper demonstrates how FFSs serve important pragmatic functions in the register of academic writing. Academic texts (regardless of which text type (articles, monographs)) take a completed or ongoing research process as their major topic and the academic discussion about this process constitutes a major and essential part of their content. This is, among other things, reflected in the use of FFSs which signal the various stages of the research process, starting from the problem-definition phase and ending with the scientific communication of the research findings. Furthermore, FFSs

are also used in the same way as lexical bundles, viz. as discourse organizers (see, for instance, Biber 2006) as well as markers of "competent participation in a given community" (Hyland 2008: 5).

To sum up, the present paper shows that the set of FFSs found within the two groups of target lexical items play a considerable role in establishing cohesive and persuasive discourse, and performing register-specific rhetorical functions, both for authors and readers of academic texts.

1. Introduction

Recent years have seen an increasing interest in studying formulaic language (see, for instance, Wray 2008, Schmitt (ed.) 2004, Corrigan et al. 2009). The existence of formulaic expressions or sequences underlines the 'Idiom Principle' put forward by Sinclair (1991: 110), according to which "a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices". Corpus-based research on formulaic language (cf. Eeg-Olofsson / Altenberg 1994, Moon 1998) has indicated a large proportion of formulaic material in natural language, stretching as high as 80 per cent. The extensive use of such prefabs has even motivated some scholars (cf. Sinclair 1991, Hoey 2005) to propose an alternative approach to traditional conceptions of grammar: "instead of seeing lexical choices as constrained by the slots which grammar make available for them, they regard lexis as systematically structured through repeated patterns of use" (Sinclair 1991: 108).

The reason for different estimates of the proportion of formulaic material in natural language has to do with

> what is being counted and the nature of the texts examined, including the proportion of written to spoken material, for although speech and writing both feature formulaic sequences, the forms and distributions are different (Butler 1998: 28, Moon 1998: 72f.)" (Wray 2000: 464).

As pointed out by Wray (ibid.),

> A full appreciation of what formulaic language is requires us to recognize that we are not dealing with a single phenomenon, but rather with a set of more and less closely related ones, across different data types [...].

Due to the lack of an unequivocal definition of recurring sequences of linguistic elements, there is a proliferation of terms referring to various types of these sequences (see Wray 2000: 465 for a list of around fifty terms used to describe various facets of formulaicity).

Wray's recent (2008) study introduces the Morpheme Equivalent Unit (MEU), which is a unit of linguistic representation, and Needs Only Analysis (NOA), which is a processing principle. MEU is a notion which describes a sequence of morphemes or lexemes with a single meaning, regardless of internal composition, and is used to capture a wide range of linguistic configurations that fall under the rubric of formulae: phrasal verbs, slot-and-frame patterns, idioms, and full constructions. However, why not use the notion of 'construction' in the Construction Grammar sense? The notion of 'construction' may supplant the various linguistic terms (e.g., formulae, holophrases, lexical phrases, idioms, chunks, prefabricated chunks, collocations, and routines) that have been utilized in the literature.

According to Wray (2008), oral traditions heavily draw on formulaic material in order for speakers to accommodate large amounts of information using limited processing capacity. She notes that written registers / genres, in contrast, do not pose the same processing constraints and thus do not make extensive use of formulaic material. Although production circumstances differ between spoken and written registers, it may in fact be the case that the latter relies on formulaic material to the same degree as the former. As we will see later in this study, written registers, such as English academic writing, may also heavily rely on formulaic material but for reasons different than those for oral registers.

Wray (2008) examines the relationship that speech and writing have with formulaic material more deeply and suggests several different lines of interpretation: individual motivations for guaranteeing novelty (as opposed to cliché) in the spoken or written registers, pragmatic constructs of shared knowledge between addresser and addressee, and processing notions pertinent to pressure created by the discourse context. In contrast to spoken registers, which are characterized by their online production, academic texts are not produced spontaneously, i.e., constructions are not assembled on the spot but "are carefully planned, edited, and revised" (Biber / Conrad / Leech 1999: 23). Furthermore, while the authors of literary texts, for instance, strive for originality and creativity, the authors of academic texts deliberately signal the register by drawing on specific rhetorical strategies, which, among other things, include the use of recurring multiword expressions of the type discussed here.

As noted by Biber / Barbieri (2007: 265),

> [...] it is not the case that there is a single pool of lexical bundles that speakers and writers draw from for these discourse functions. Rather, each register employs a distinct set of lexical bundles, associated with the typical communica-

tive purposes of that register. The following sections show that there are fundamentally different sets of lexical bundles associated with spoken university registers in contrast to written registers. Thus, an ESP perspective, considering each register on its own terms, is required to adequately describe the use of lexical bundles in the university context.

As observed by Hyland (2008: 6), our formulations are "shaped by the way we regularly encounter them in similar texts". From a usage-based constructional perspective, it can be argued that frequently recurring strings of items in specific text types or registers become entrenched for the language users of that language domain. This study contents that there is higher probability of encountering routinized patterns of language use in specific language domains. Since the language system is 'experience-driven', the permanent confrontation, especially with domain-specific linguistic input, has an enormous impact on the repository of constructions accessible and used by the language user.

Despite the fact that many scholars have recognized the role and importance of studying formulaic sequences including optional as well as semantically constrained slots (see, for instance, Schmitt / Carter 2004), the focus of research on prefabs has been on the collocations of the type *data showed*, i.e., uninterrupted sequences of several lexical items. For instance, Biber's notion of lexical bundles is based on a frequency-driven approach (Biber / Conrad / Leech 1999). It is defined as common recurrent sequences of words (e.g., 4-word sequences "I don't know if"). Lexical bundles are regarded as not complete structural units and not idiomatic in meaning. They are structurally complex, i.e., often composed of a matrix clause / phrase and the beginning of an embedded clause / phrase. Lexical bundle is a useful notion as it underlines the necessity of looking at the use of recurring multiword expressions within a given register, the view also supported by the present study. Still, as pointed out by Schmitt / Carter (2004: 7), the vast majority of formulaic sequences are flexible in nature. According to them, once software is developed which permits their automatic extraction, "we may find that flexible formulaic sequences are even more prevalent than totally fixed ones".

The present study investigates the use and function of what will be referred to as 'flexible formulaic sequences' (henceforth FFSs) in the register of English academic writing. From a usage-based constructionist perspective, FFSs can be conceived of as register-specific low-level schemas: some slots of these constructions semantically (collocationally) constrain which items can fill them, while others are more schematic and can be understood as register-specific

frame elements. The focus is on FFSs in two groups of high frequency lexical items in the academic subcomponent of the BNC, viz. the 'research' verbs and nouns (e.g., *analyze*, *investigate* or *study*) and the 'coming-to-know' verbs (e.g., *find*, *suggest* or *show*).

The majority of corpus-based studies on prefabs view frequency as the decisive definitional criterion. Although frequency plays a crucial role in identifying formulaic sequences, there are several difficulties associated with it. One of the difficulties has to do with determining a sufficient frequency threshold or cut-off for the 'formulaic status' of a sequence. Another difficulty with solely relying on frequency as a means of spotting formulaicity has to do with the infrequent use in general English (the concept of 'general English' has been questioned, see Killgarriff 2003: 54) of certain sequences which are typical of specific language domains, i.e., genres / registers. This results from the fact that high frequency items – typical candidates for entering formulaic sequences – tend to display a range of senses with specific ones being entrenched within a particular genre / register. High frequency lexical items tend to be highly polysemous and of 'phraseological nature' with an entrenched sense in specific register / genres. As we will see later, this study underlines the importance of investigating register-specific low-level schemas.

2. A usage-based constructionist approach to recurring multiword expressions

Although Construction Grammar and the Usage-Based Model are two independent theoretical frameworks, they are commonly combined in linguistic analyses. The former refers to a family of grammatical theories which regard constructions – form-function amalgamations – as the basic units of grammar (cf. Langacker 1987; Fillmore / Kay 1993; Goldberg 1995, 2006; Croft 2001), whereas the latter assumes that language use permeates linguistic knowledge. There are several reasons for pursuing a usage-based constructionist approach in the study of recurring multiword expressions. One of the main reasons has to do with the fact that the existence of such expressions proves to be perfectly compatible with one of the major tenets of usage-based constructionist approaches: that a language system is not entirely economic since it simultaneously incorporates both schematic / abstract structural configurations and prefabricated chunks of concrete expressions that occur with sufficient frequency in everyday language situations. These approaches to grammar can be treated

as an instance of what has been termed the 'exemplar model' (cf. Nosofsky 1988), which assumes that the same information is frequently stored at different levels of abstraction. As pointed out by Langacker (2000: 1), rather than being 'minimal' and 'economical', grammar is 'maximal' and 'nonreductive'.

A further reason for pursuing a usage-based constructionist approach to recurring multiword expressions has to do with the fact that this approach assumes the so-called 'syntax-lexicon continuum'. According to Wray (2008:3), the term 'formulaic language' is generally used to "refer to the large units of processing – that is, lexical units that are more than one word long". However, as pointed out by Wray (ibid.), a fundamental clear-cut boundary between small and large lexical units is only sustainable within certain theoretical frameworks, such as Construction Grammar. One of the central tenets of this theoretical framework is the existence of constructions displaying various degrees of complexity, ranging from morphemes, words via substantive (i.e., fully lexically-filled) idioms to partially lexically-filled (i.e., semi-schematic) and fully schematic or abstract linguistic configurations.

Wray (2008: 10) provides a list of various types of expressions which are considered formulaic. These expressions can be treated as different construction types. One type of these expressions is of idiomatic nature of the type *kick the bucket* or *spill the beans*, which from a constructional perspective can be described as fully lexically filled (or totally lexically specified) constructions occupying the end (lexicon) point of the syntax-lexicon continuum. Her list also includes what she labels 'partly-fixed frames': "complete phrases or clauses partly realized with specific lexical material and partly left open for interchangeable items, for example, 'the end of the N'; 'as a result of NP'; 'the way in which CLAUSE'." From a constructional perspective, these expressions can be described as partially lexically filled constructions occupying an intermediate level of the syntax-lexicon continuum. The present study puts forward a further type of formulaic sequences not included in Wray's (2008) list. These sequences occupy some intermediate level of the continuum but do not subsume any totally specified lexical material. Rather, some of the slots of these sequences are semantically (collocationally) constrained, while others are describable in terms of register-specific frame elements. Throughout this study the term 'flexible formulaic sequences' (FFSs) will be used to refer to linguistic configurations which subsume lexically constrained slots, slots being occupied by rather abstract categories (register-specific frame elements) as well as optional ones. From a constructional point of view, these sequences can be treated

as low-level schemas, occupying an intermediate level on the syntax-lexicon continuum. This intermediate level links the semantically constrained elements to more schematic elements (register-specific frame participants). This continuum is not only constituted by varying numbers of slots which may or may not be lexically filled (substantial) in a construction, but also by a continuum which exists for each slot between 'schematic' and 'substantial'. Some of the slots are more schematic / abstract (e.g., a constituent encoding the register-specific frame element), others more or less severely constrained by selection restrictions concerning their lexical filling (e.g., they are constrained to a particular lexical field or a small group of referents). The FFSs discussed here can be conceived of as low-level schemas and they illustrate the 'bottom-up' orientation of Cognitive Grammar and the observation that "low-level schemas, expressing regularities of only limited scope, may on balance be more essential to language structure rather than high-level schemas representing the broadest generalizations" (Langacker 2000: 30f.).

Not only has research on prefabs paid the most attention to semantically and syntactically opaque linguistic configurations, a similar situation is found in earlier works conducted within "Berkeley" Construction Grammar and the Goldbergian version of Construction Grammar. Goldberg's (1995) original definition of 'construction' subsumes the aspect of semantic unpredictability. Within these versions of Construction Grammar the focus has hitherto been on either rather abstract or schematic constructions, such as the resultative or the caused-motion construction, or on fully lexically filled constructions such as idioms or frozen collocations. The role of partially lexically filled constructions involving one or more flexible slots has been reduced to the study of partially substantive idioms (e.g. *jog <someone's> memory, under the auspices of* NP). Thereby, the 'constructionhood' of such configurations is explained in terms of 'semantic unpredictability'. Within constructionist approaches to grammar, the notion of 'partially lexically filled construction' has primarily been used to refer to semantically unpredictable constructions involving one or more fixed lexical items as, for instance, linguistic structures of the type of the *way*-construction (e.g., *He made his way through the crowd*) or the *into*-causative (e.g., *She tricked him into marrying her*). But what about constructions which contain several slots which are constrained concerning their lexical filling, or, to put it differently, constructions involving what may be termed as 'collocationally' constrained slots?

The usage-based model takes for granted the fact that more schematic/abstract constructions emerge through usage since the statistical properties of the input shape the language user's repertoire of linguistic configurations. Within this model, it is assumed that two usage-based properties impinge on grammatical representation in the language user's mind: the frequency of occurrence of particular grammatical forms and structures, and the meanings of the words and constructions in use. The input modifies the repository of linguistic units available to the language user. Human beings acquire language through encountering an array of arbitrary samples and using them to build systematic representations. More schematic constructions emerge through the processes of schematization and analogy of the frequently recurring patterns. The density of similar schemas leads to the emergence of more abstract/schematic constructions.

In contrast to generative grammar, a usage-based model is compatible with probability theory, i.e., statistical models of language. Generative grammar, which is viewed as a rationalist approach, assumes that it is not possible to model linguistic knowledge since essential parts of language are innate – a blueprint in the brain at birth as part of the human genetic inheritance. A usage-based model, on the other hand, is an empiricist approach, which assumes that there have to be certain cognitive abilities in the brain, since learning new structures or configurations is not possible from a completely blank slate, a tabula rasa. The central tenet of a usage-based approach is to assume that the mind is not initially equipped with a detailed array of principles and procedures specific to language systems, but that general operations such as schematization, statistical input, analogy, etc., enable a child to acquire the detailed structure of natural language. We are capable of acquiring the complicated and extensive structure of language by applying statistical pattern recognition. A usage-based type of modeling a language requires statistical, as opposed to categorical observations. Human cognition is considered to be probabilistic and language, since it is an integral part of cognition, is also probabilistic. Manning/Schütze (1999: 15) put forward that "the argument for a probabilistic approach to cognition is that we live in a world filled with uncertainty and incomplete information". And hence a felicitous interaction with the world requires our capability to deal with this type of information.

The usage-based model assumes that the emergence of flexible and creative language use hinges on the predisposition to discover regular and recurring patterns found across a range of familiar utterance types. Such patterns display

a relatively high degree of schematicity capable of sanctioning a (potentially) open-ended set of utterance tokens. Initially, such patterns will be concrete lexically filled constructions, involving simple slot and frame structures based on specific uses of specific lexical items. In later stages, these may provide the basis for further abstractions, as more schematic constructions are built on more substantial constructions to capture increasingly higher-order grammatical generalizations.

FFSs are particularly amenable to analysis combining the theoretical framework of Frame Semantics with that of Construction Grammar. As FFSs are semi-fixed low-level schemas, they encode both constructional and frame-semantic information. Some slots of this construction are semantically (collocationally) constrained as to their filling, while others are more schematic and can be conceived of as register-specific frame elements. It may be argued that the theory of Frame Semantics complements the Construction Grammar theory. It provides an explanation for the fact that for certain domains (frames) we find linguistic routines. Such frames, according to Lakoff (1987: 68) are 'idealized cognitive models' which we use to organize our knowledge of the world. The contexts of situation to which entrenched constructions contribute and their underlying frames can be identified.

A first step in a frame-based semantic characterization of lexical units is to identify the phenomena, experiences and scenarios linked to a frame in question. The next step involves an elaboration of a list of predicates which have the potential to trigger this frame. Finally, the so-called 'frame elements', i.e., the semantic roles associated with a predicate-argument structure are identified and labeled.

Research lies at the core of the scholarly world and plays a vital role in conceptualizing and understanding reality. Phenomena in the outside world are apprehended and computed via the research process; daily experience of research may be captured in terms of the following schema: a researcher consciously and deliberately engages in the process of contemplation of an entity, an event that unfolds within a certain period of time. When investigating an entity or a phenomenon, researchers frequently focus on particular aspects or facets of the phenomenon, in most cases motivated by the complexity of the phenomenon under investigation. Many phenomena occurring in the 'research world' are highly complex and hence scholars are frequently not in a position to study them in their entirety but have to focus on certain aspects with the intent to learn more about the nature of these phenomena.

The Research Frame is a register-specific frame, and the situation described by this frame is not an everyday situation. This specification is responsible for a greater differentiation of participants which figure in this frame and which can be seen as a sort of register-specific frame-elements such as 'Object Scope', 'Method', 'Purpose', 'Result', 'Parameter', or Precondition. The register-specificity of this frame may be considered to be a reason why specific frame elements, such as Parameter, should be treated as belonging to the 'core' frame elements.

3. Method

The focus here is placed on the use of two groups of high frequency lexical items denoting two key stages of the research process: verbs of studying (e.g., *study*, *examine*, or *investigate*) and 'coming-to-know' verbs (e.g., *find*, *show*, or *indicate*).

In the relevant literature on English for Specific Purposes (ESP) and English for Academic Purposes (EAP) (cf. Coxhead / Nation 2001: 252), a classification of academic vocabulary into three levels, viz. general service or basic vocabulary, technical vocabulary, and sub-technical vocabulary, is sometimes made. The latter is differently termed as 'generally useful scientific vocabulary' (Barber 1962), 'frame words' (Higgins 1966), 'subtechnical vocabulary' (Cowan 1974, Anderson 1980, Yang 1986), 'academic vocabulary' (Martin 1976, Coxhead 2000), 'specialised non-technical lexis' (Cohen et al. 1988), and 'semi-technical vocabulary' (Farrell 1990). In particular it involves lexical items that occur more frequently in academic texts than in non-academic texts and do so consistently across different disciplines and discourse registers without being academic-discipline specific. Both groups of the target items belong to this group of sub-technical vocabulary.

Coxhead (2000) overlooks the fact that especially high-frequency lexical items tend to display polysemous structure with their specific senses being entrenched within a specific language domain (i.e., register / genre). Hence, determining whether a given word should be listed on the Academic Word List[1] by comparing its frequency in academic writing to its frequency in 'general English' can be misleading. For instance, the lexical item *study* is not only frequently found in academic English but also in general English and would thus

[1] The AWL is a list of lexical items which appear with high frequency in English-language academic texts. For more details see http://www.victoria.ac.nz/lals/resources/academicwordlist/ (last visited: 10 / 2010).

not be listed on the Academic Word List. But the fact is that only its specific sense is highly entrenched in the register of academic writing, i.e., a researcher studying a specific entity / phenomenon. High frequency lexical items tend to display a polysemous structure, with entrenched senses in specific domains of language use. For this reason, the selection criteria on which the Academic Word List is based seem to be inadequate, especially for the purposes of the register / genre-specific studies.

The majority of studies on prefabs have relied on a corpus-driven approach to their identification. Prefabs are most commonly extracted with the help of n-grams (in the form of bi-grams, tri-grams, etc.) or 'phrase frames' (Fletcher 2006). Cheng / Greaves / Warren (2006) use the notion of 'phraseological profile' to refer to the identification of the meaningful word associations in a text or a corpus (linked to what Phillips (1989) refers to as the 'aboutness' of a text).

Despite the fact that corpus software is under constant improvement, there are still limits to what it can find in a large corpus. One of the main difficulties of automatical extraction of FFS has to do with the fact that corpus-based methods are generally only applicable to the constitutive slots of a given construction, which makes it difficult to accommodate optional slots. The identification and extraction of FFSs necessitates a parsed corpus. Unfortunately, the BNC is not a syntactically annotated and balanced parsed corpus allowing for extraction of such sequences. Furthermore, as noted above, FFSs may also involve optional material, making their extraction an even more difficult task. FFSs were manually extracted via inspection of randomly selected concordance lines around the target items. The present study is based on the academic subcomponent of the BNC which consists of 15 429 582 words (approx. 15.5% of the entire BNC) and incorporates academic texts from a wide range of academic disciplines, including engineering, applied and natural sciences, arts, and social sciences.

The present study was based on an interpretation of patterns emerging from manual inspection of randomly selected concordance lines providing the target items in their contextual environment (KWIC display). As noted by Johannson (1992: 334) "[...] grammars do not grow magically from corpora. They require the intelligent analytical mind of a grammarian who draws knowledge of previous studies, on his or her own intuition as well as on observations of text".

The term 'lemma', which has hitherto been commonly used in computational and corpus linguistics, does not prove to be an adequate one for the purposes of the present study. Recent research has questioned the notion of 'lemma' as a relevant unit which has hitherto been commonly used in computational and corpus linguistics (cf. Rice / Newman 2005). More attention has been devoted towards constructions associated with inflected forms of a word (cf. Bybee / Hopper (eds.) 2001; Thompson / Hopper 2001; Newman / Rice 2004, 2006). There have been several corpus-based and cognitively / functionally oriented studies reporting findings based on inflected form levels of a word. The focus of these studies was on the collocational, constructionist and grammaticalizational distributional profile of inflected forms of individual verbs.

According to inflectional island hypothesis (Rice / Newman 2005), adults make use of particular inflectional forms of individual verbs on a register-specific basis. As observed by Newman (2008):

> A corpus-based approach which explores low-level generalizations in the use of language invariably yields a very large number of observations which must be reconciled with other kinds of empirical and theoretical ideas about language if we are to make progress in linguistics [...].

Sinclair (1991: 8) contended that

> it is now possible to compare the usage patterns of, for example, all the forms of a verb, and from this to conclude that they are often very different one from another [...] There is a good case for arguing that each distinct form is potentially a unique lexical unit, and that forms should only be conflated into lemmas when their environments show a certain amount and type of similarity.

The present study adopts the inflectional island hypothesis and provides a fine-grained level of analysis since it examines each single inflectional form of the target verb with respect to its occurrence within a particular low-level schema (or FFS). Due to space limitations, the present paper will focus solely on the following forms of the target items: the *to*-infinitive forms of the research verbs, the nominalized versions of the research verbs and the past participle forms of 'coming-to-know' verbs.

4. Analysis

Table 1 provides the distributional profiles of the *to*-infinitive forms of the research verbs across various FFSs in the academic sub-corpus of the BNC.

FFS	Lexical items				Sum
	examine	*study*	*analyse*	*investigate*	
I	93	49	28	30	200
IIa	25	21	28	40	114
IIb	15	20	11	16	62
III	34	8	14	27	83
IV	25	4	6	3	38
Total	192	102	87	116	497

Table 1: The distribution of the *to*-infinitive forms of the research verbs across FFSs

Legend:

I (AP)$_{\text{Reason / Temporal}}$ *It be* AP (*for* NP)$_{\text{Scholar}}$ *to* V$_{\text{research verb}}$ NP1 / (*wh*-clause)$_{\text{Object Scope}}$ (*in terms of / with regard to* NP2)$_{\text{Parameter}}$

e.g.: *It is possible to analyse possible interactions between the syntactic cue and* [...] [B2X 101]

IIa NP1$_{\text{Method}}$ *be used / utilized / employed / applied* to V$_{\text{research verb}}$ NP2$_{\text{Object Scope}}$

e.g.: *The benefit analysis instrument was used to investigate the possible* [...] [B2M 962]

IIb NP1$_{\text{Scholar}}$ *use / make use of / employ* NP2$_{\text{Method}}$ to V$_{\text{research verb}}$ NP2$_{\text{Object Scope}}$

e.g.: *Psychologists use quasi-biological or anthropological observational procedures to investigate women* [CMR 74]

III NP1$_{\text{Aim / Goal / Objective}}$ (*of* NP2$_{\text{Study}}$) *be to* V$_{\text{research verb}}$ (AP)$_{\text{Manner}}$ NP3$_{\text{Object Scope}}$

e.g.: *The aim is to analyse the levels of social process that* ...[CGY 1075]

IV NP1$_{\text{Scholar}}$ *move on / proceed / turn to / go on / paused to* (AP)$_{\text{Manner}}$ V$_{\text{esearch verb}}$ NP2$_{\text{Object Scope}}$

e.g.: *We can now move on to analyse the middle class and* ...[EDH 1586]

The most frequent FFS around the *to*-infinitive form of the research verbs has already been discussed by Biber / Conrad / Leech (1999) under the term 'anticipatory *it*-construction'. The impersonal *it* can be used in order to take the focus off the author of a text. According to Biber / Conrad / Reppen (1998: 76), extraposed constructions with adjectival predicates

> present a proposition that cannot be attributed to any particular person, and they frame the proposition in terms of a static condition (with an adjectival predicate such as *possible* or *difficult*) rather than in terms of a dynamic action or process (such as *think* or *feel*).

Table 2 presents two most frequent FFSs around the past participle forms of the 'coming-to-know' verbs. The most frequent FFS around the 'coming-to-know' verbs is the-so called NCI construction. The meaning of this construction can be formulated in the following way: certain facts / properties are ascribed (in the *to*-infinitive complement clause) to an entity under investigation, encoded by the NP. The passive formulaic expression 'be found / shown ... to' serves to indicate the nature of the evidence for the truth of a given statement. The use of this FFS also supports information packaging in academic writing: placing the entity / phenomenon under investigation in the initial position and leaving the research unspecified.

FFS	Lexical items			Sum
	found	*shown*	*seen*	
I	71	104	92	267
II	38	68	31	137
Total	109	172	123	404

Table 2: The distributional profiles of the past participle forms of the 'coming-to-know' verbs across various FFSs

Legend:

I NP1 *be* $V_{found/shown/seen}$ *to* V-infinitive
e.g.: *The computed value of fracture initiation pressure gradient was found to be* 1.02 psi/ft. [B2J 1711]

II *It be* $V_{found/shown/seen}$ *that*-clause
e.g.: *It was found that increases in protein synthesis and breakdown could be correlated with the progression of disease severity* ... [HU2 1777]

Frequently employed FFSs around the target items also reflect conceptualization of events in English academic writing. One of the central tenets of Cognitive Linguistics is that language users are equipped with the capacity to construe one and the same situation in alternate ways and that linguistic configurations reflect language users' construal. There have been several proposals accounting for the relation holding between conceptualization of events and the selection of particular constructions. Fisher / Gleitman / Gleitman (1991) say that constructions serve as a 'zoom lens' which the writer uses to direct the reader's attention to a particular perspective on a scene. Talmy (1996) describes the use of constructions to highlight certain aspects of a scene, at the expense of other aspects, as the 'windowing of attention'. Langacker (1987) speaks of constructions forcing a certain 'construal' of a situation.

Let us briefly consider the latter approach. Langacker (1987) lists three parameters which are used to construe a specific situation: 'selection', 'perspective' and, 'abstraction', He uses the term 'focal adjustment' to refer to variation in terms of these parameters. The language user selects particular focal adjustments and thus configures a perceived scene in a particular way by means of language, and hence provides a particular construal of the situation in question.

The first parameter of 'selection' specifies which facets of a scene are being focused on. This parameter subsumes the notion of 'conceptual domain', i.e., a coherent organization of knowledge within our conceptual system against which a conceptualization is achieved. This background knowledge, which is essential for understanding the concepts evoked by lexical items, may range from "'basic' notions of time, space, color or temperature, to complex and rather specific background knowledge such as the rules of cricket" (Taylor 2002: 203). Furthermore, "a domain may consist in knowledge of typical scenarios, cultural conventions, and metalinguistic notions of dialectal and stylistic variation" (ibid.). This may figure prominently in the researcher's / author's selection of linguistic configurations in academic texts. In this connection, it may be argued that the awareness of writing in the academic register serves as the background against which the researcher's / author's selection of linguistic constructions takes place.

The parameter of 'perspective' relates to the viewing's position of a certain scene and thus "consequences for the relative prominence of its participants" (Langacker 1987: 117). When discussing the parameter of perspective, Langacker (ibid.: 120ff.) deals with the following aspects: figure / ground alignment, viewpoint, and the related problems of deixis, and subjectivity / objectivity. The latter is of special importance for the present study. The dichotomy subjectivity / objectivity is related to measuring the degree of how subjective / objective the construal of the entity in question is. The research process on the part of a researcher tends to be construed 'objectively', i.e., "it is solely an object of conceptualization, maximally differentiated from the conceptualizer (i.e., the speaker and / or hearer)" (Langacker 2006: 41). Academic texts preferably adopt what Langacker calls "the optimal viewing arrangement": The researcher / author of academic texts tends not to be involved as a conceptualizer. He observes a situation from a vantage point external to its setting. Or as Gray / Malins (2004: 22) put it,

> research *about* (into) practice has tended to be carried out by other academic researchers (historians, educationalists, sociologists, psychologists, and so on) from an external perspective. These approaches reflect more the classic scientific method, where the researchable is objectified, and the researcher remains detached.

In a similar vein, Biber (2006: 5) notes that authors of academic texts "rarely express their own attitudes or feelings explicitly, and they almost never use language that indicates their actual thought processes". The informational focus of academic texts and presentation of information in an objective way is also reflected in the use of FFSs. In many cases, the use of FFSs is definitely related to the researcher's / author's desire to adopt the optimal viewing arrangement. This static character of academic texts may also be achieved via the use of FFSs.

The third parameter is "abstraction". Each of the aforementioned three parameters refers to the language user's capacity to focus on certain facets of a perceived situation. One way in which the language user can adopt a certain perspective on a scene has to do with the level of "specificity" at which he perceives something: this parameter can also have an impact on the author's preference for nominalized forms since these forms allow one to leave out the details of the research process. Table 3 provides the distributional profiles of the nominalized forms of the research verbs across various FFSs in the academic sub-corpus of the BNC.

FFS	Lexical items				Sum
	study	*analysis*	*examination*	*investigation*	
I	140	34	9	6	189
II	116	34	18	4	172
III	62	40	10	18	130
IV	19	20	–	5	44
Total	337	128	37	33	535

Table 3: The distributional profiles of the nominalized forms of the research verbs across various FFSs

Legend:

I $NP1_{\text{research noun}}$ (*of* $NP2_{\text{Object Scope}}$) $V_{\textit{found/indicated/revealed}}$ (*that*-clause)$_{\text{Result}}$
e.g.: *Restriction analysis of the rescued plasmid revealed that it had the expected structure.* [K5Y 336]

II $NP1_{\text{research noun}}$ (*of* $NP2_{\text{Object Scope}}$) $V_{\textit{found/indicated/revealed}}$ $NP3_{\text{Result}}$
e.g.: *Our study found a rate ratio of 0.66. for acute gastroenteritis deaths* [...] [HWU 368]

III $NP1_{\text{research noun}}$ (*of* $NP2_{\text{Object Scope}}$) *be* $V_{\textit{conducted/performed/carried out/performed}}$ verb (*by* $NP3_{\text{Scholar}}$)
e.g.: *A prospective, randomized, comparative study was perform*ed [...] [HU3 5472]

IV $NP1_{\text{research noun}}$ (*of* $NP2_{\text{Object Scope}}$) $V_{\textit{led to/raised/contributed}}$ NP3
e.g.: *The studies led to the revelation of reasons why the programme was less successful.* [CMR 512]

Discourse-pragmatic functions of FFSs around the target research nouns (see Table 3) can be assigned in terms of moves. According to Swales (2004: 151), trying to understand a genre is "typically a process which starts from macro features and only later tries to align these with particular linguistic realizations, and then looks for explanatory links between the macro and the micro". This is sometimes done in terms of the so-called 'moves' (Swales 1990) that reflect the conventionalized structuring of genre determined by its communicative purpose. These moves are fairly stable functional units belonging to certain genres, however, the concrete rhetorical strategies of their realization in different texts depend, on the one hand, upon the individual preference of a writer (within the scope of possible rhetorical choices of this or that genre), and, on the other hand, upon the socio-cultural context of the genre origin and function.

The focus of EAP/ESP studies dealing with 'moves' has been on the genre of research articles, which because of their rather inflexible organization (as required by the journals) prove to be particularly amenable to the analysis in terms of moves. It goes without saying that in the case of the research article which follows a rigid required textual organization (predominantly the Introduction-Method-Results-Discussion structure, the so-called 'IMRAD'[2] structure) and has to accommodate space limitations, the identification of moves is far easier than in the case of textbooks or dissertations. But still all different types of academic texts reveal similar organizational structure due to the fact that they all share the same content, i.e., the research process with its various phases and the academic discussion about it constitute a major and necessary part of their content (cf. Meyer 1997: 74). The academic sub-component of the BNC does not exclusively contain research articles but also textbooks, dissertations, etc. Despite this diversity of text types, it is possible to deduce an ideal-

[2] IMRAD is an acronym for **I**ntroduction, **M**ethods, **R**esults **A**nd **D**iscussion.

ized model of the research process with its various stages as laid down in academic texts. By comparing various works on the research process and identifying commonalities across these different studies, I was able to deduce a model of the research process with six key phases. According to this model, the research process includes various phases, starting from the problem-definition and ending with scientific communication of the research findings: Phase 1: defining the scope and objectives of the study > Phase 2: constructing or developing a theoretical framework > Phase 3: employing a convenient method for obtaining results > Phase 4: finding results > Phase 5: drawing conclusions > phase 6: communicating the research findings. In Lakoff's (1987) terms this would be called an 'idealized cognitive model' of the research process. It is important to keep in mind that the different stages of the research process do not generally follow each other in a linear sequence, but "are rather part of a continuous iterative cycle, or helix, of experience (consistent with Kolb's 1984 'experiential learning cycle')" (Gray/Malins 2004:12). The model of the research process is reflected in the global format and content schemata for structuring information in academic texts.

The key phases of the research process may be signaled by specific moves, which in turn can be realized by the use of specific FFSs. To illustrate this, let us briefly consider FFSs around the target items given in Tables 1-3: Some of these FFSs are realizations of the following three moves which in turn highlight three essential phases of the research process:

Phase 1: identifying and formulating a viable research question; Move 1: outlining the aim and scope of the research endeavor (see FFS III, Table 1)

Phase 3: gathering and evaluating data; Move 3: addressing methodological issues (see FFSs IIa and IIb, Table 1)

Phase 4: finding and interpreting results; Move 4: presenting and discussing research findings (see FFSs I and II in Table 2 and FFSs I, II, and IV in Table 3).

According to Hyland (2008: 5), the extensive use of lexical bundles in written registers, such as academic texts, "helps to signal the text register to readers and reduce processing time by using familiar patterns to link elements of new information". This is also true of FFSs. Their semantically constrained slots fulfill this function, whereas their more flexible slots, i.e., slots containing register-specific frame element information, carry new information. In this world

of steadily increasing information load, including the academic world, the necessity of FFS use is given. This means the packaging of information by making use of constructions involving preassembled parts. According to the corpus data used in the present study, target items commonly enter constructions which can be considered to be "pre-packaged", pre-formulated, constituting an established cognitive routine that no longer requires a processing effort when used by the author as well as the reader of academic texts.

To sum up, the following set of functions is found to motivate the use of FFSs (which are by the way intimately linked): (i) packaging information in a neutral and objective way; (ii) signaling the text register to readers; (iii) highlighting the key stages of an idealized cognitive model of the research process and (iv) guiding readers through the text.

In addition, FFSs also facilitate the writing process as their more constrained parts (i.e., semantically constrained slots) are familiar elements used to link elements of new information encoded by more abstract / schematic parts.

5. Concluding remarks

The picture that emerges from this study is that FFSs deserve more attention in the treatment of recurring multiword expressions (especially in specific registers, such as that of English academic writing) rather than uninterrupted fixed strings of words. Furthermore, this work shows the importance of studying these register-specific low-level schemas and the role they play within the theoretical framework of Usage-Based Construction Grammar. From a usage-based constructional perspective, it can be argued that frequently recurring strings of items in specific text types or registers become entrenched for the language users of that language domain.

This study illustrates that a synthesis of construction grammar, corpus, and register-specific methodology may lead to new and valuable insights about the role and function of frequently recurrent strings of lexical items in the register of English academic writing. The flexible formulaic sequences presented here were found to enclose restricted sets of semantically related lexical units whose selection is conditioned by the collocational restrictions as well as pragmatic background of texts and linguistic conventions of the register.

While spoken registers primarily employ formulaic sequences for supporting online processing, academic writing makes use of these sequences, among

other reasons, for signaling a formal, objective, 'academic' style. The informational focus of academic texts and presentation of information in an objective way is also reflected in the use of FFSs. In many cases, the use of FFSs is definitely related to the researcher's / author's desire to adopt the optimal viewing arrangement. FFSs may also be realizations of moves which in turn highlight the key stages of an idealized cognitive model of the research process.

A usage-based constructionist approach to a language which assumes a syntax-lexicon continuum and strives to account for the entirety of a given language system, must also include the discussion of register-specific low-level schemas referred to here as flexible formulaic sequences.

References

Anderson, Janet I. (1980): The lexical difficulties of English medical discourse for Egyptian students. In: English for Specific Purposes 37: 3-5.

Barber, Charles L. (1962): Some measurable characteristics of modern scientific prose. In: Barber, Charles L. / Behre, Frank / Ohlander, Urban / Olsson, Yngve / Stubelius, Svante / Söderlind, Johannes / Zandvoort, Reinard Willem (eds.): Gothenburg studies in English 14. Contributions to English syntax and philology. Gothenburg: University of Gothenburg, 21-43.

Biber, Douglas (2006): University language: A corpus-based study of spoken and written registers. Amsterdam / Philadelphia: Benjamins.

Biber, Douglas / Barbieri, Federica (2007): Lexical bundles in university spoken and written registers. In: English for Specific Purposes 26: 263-286.

Biber, Douglas / Conrad, Susan / Reppen, Randi (1998): Corpus linguistics: Investigating language structure and use. Cambridge: Cambridge University Press.

Biber, Douglas / Conrad, Susan / Leech, Geoffrey (1999): Longman grammar of spoken and written English. Harlow: Longman.

Biber, Douglas / Conrad, Susan / Vivian, Cortes (2004): If you look at ...: Lexical bundles in university teaching and textbooks. In: Applied Linguistics 25: 371-405.

Butler, Christpoher S. (1998): Collocational frameworks in Spanish. In: International Journal of Corpus Linguistics 3, 1: 1-32.

Bybee, Joan / Hopper, Paul (eds.) (2001): Frequency and the emergence of linguistic structure. Amsterdam / Philadelphia: Benjamins.

Cheng, Winnie / Greaves, Chris / Warren, Martin (2006): From n-gram to skipgram to concgram. In: International Journal of Corpus Linguistics 11, 4: 411-433.

Cohen, Andrew et al. (1988): Reading English for specialized purposes: Discourse analysis and the use of student informants. In: Carrell, Patricia L. / Devine, Joanne / Eskey, David (eds.): Interactive approaches to second language reading. New York: Cambridge University Press, 152-167.

Corrigan, Roberta et al. (2009): Formulaic language 1: distribution and historical change. Amsterdam / Philadelphia: Benjamins.

Cowan, J. Ronayne (1974): Lexical and syntactic research for the design of EFL reading materials. In: TESOL Quarterly 8, 4: 389-400.

Coxhead, Averil (2000): A new academic word list. In: TESOL Quarterly 34, 2: 213-238.

Coxhead, Averil / Nation, Paul (2001): The specialised vocabulary of English for Academic Purposes. In: Flowerdew, John / Matthew, Peacock (eds.): Research perspectives on English for Academic Purposes. Cambridge: Cambridge University Press, 252-267.

Croft, William (2001): Radical construction grammar: Syntactic theory in typological perspective. Oxford: Oxford University Press.

Eeg-Olofsson, Mats / Altenberg, Bengt (1994): Discontinuous recurrent work combinations in the London-Lund corpus. In: Fries, Udo / Tottie, Gunnel / Schneider, Peter (eds.): Creating and using English language corpora. Amsterdam: Rodopi, 63-77.

Farrell, Paul (1990): Vocabulary in ESP: A lexical analysis of the English of electronics and a study of semi-technical vocabulary. (= CLCS Occasional Paper 25). Dublin: Trinity College.

Fillmore, Charles J. / Kay, Paul (1993): Construction Grammar Coursebook. Ms., Department of Linguistics. University of California at Berkeley

Fisher, Cynthia / Gleitman, Henry / Gleitman, Lila R. (1991): On the semantic content of subcategorization frames. In: Cognitive Psychology 23, 331-392.

Fletcher, William H. (2006): "Phrases in English" Home. Internet: http://pie.usna.edu (last visited 10 / 2010).

Goldberg, Adele (1995): Constructions: A construction grammar approach to argument structure. Chicago: Chicago University Press.

Goldberg, Adele (2006): Constructions at work: The nature of generalization in language. Oxford: Oxford University Press.

Gray, Carole / Malins, Julian (2004): Visualizing research. A guide to the research process in art and design. Ashgate: Aldershot.

Higgins, John J. (1966): Hard facts. In: ELT Journal 21, 1: 55-60.

Hoey, Michael (2005): Lexical priming: A new theory of words and language. London: Routledge.

Hyland, Ken (2008): As can be seen: Lexical bundles and disciplinary variation. In: English for Specific Purposes 27: 4-21.

Kilgarriff, Adam (2003): Linguistic search engine. In: Simon, Kiril (ed.): Shallow processing of large corpora: Workshop held in association with Corpus Linguistics 2003. Lancaster, March.

Johansson, Stig (1992): Comments. In: Svartvik, Jan (ed.): Directions in corpus linguistics: Proceedings of Nobel Symposium 82 Stockholm, 4-8 August 1991. Berlin / New York: de Gruyter, 332-334.

Lakoff, George (1987): Women, fire and dangerous things. Chicago: University of Chicago Press.

Langacker, Roland (1987): Foundations of cognitive grammar 1: Theoretical prerequisites. Stanford, CA: Stanford University Press.

Langacker, Roland (2000): A dynamic usage-based model. In: Barlow, Michael / Kemmer, Suzanne (eds.): Usage-based models of language. Stanford: CSLI Publications, 1-63.

Langacker, Roland (2006): Cognitive grammar. In: Geeraerts, Dirk (ed.): Cognitive linguistics: Basic readings. Berlin / New York: de Gruyter, 29-69.

Manning, Chris / Schütze, Hinrich (1999): Foundations of statistical natural language processing. Cambridge, MA: MIT Press.

Martin, Anne V. (1976): Teaching academic vocabulary to foreign graduate students. In: TESOL Quarterly 10, 1: 91-97.

Meyer, Paul Georg (1997): Coming to know: Studies in the lexical semantics and pragmatics of academic English. (= Forum für Fachsprachen-Forschung 35). Tübingen: Narr.

Moon, Rosamund (1998): Fixed expressions and idioms in English. Oxford: Clarendon Press.

Newman, John (2008): Aiming low in linguistics: Low-level generalization in corpus-based research. Paper presented at the 11th International Symposium on Chinese Languages and Linguistics, National Chiao Tung University, Hsinchu, Taiwan. Internet: http://www.ualberta.ca/~johnnewm/ (last visited: 10 / 2010).

Newman, John / Rice, Sally (2004): Patterns of usage for English SIT, STAND, and LIE: A cognitively-inspired exploration in corpus linguistics. In: Cognitive Linguistics 15: 351-396.

Newman, John / Rice, Sally (2006): Transitivity schemas of English EAT and DRINK in the BNC. In: Gries, Stefan / Stefanowitsch, Anatol (eds.): Corpora iiguistics 2. The syntax-lexis interface. Amsterdam: Benjamins, 225-260.

Nosofsky, Robert M. (1988): Exemplar-based accounts of relations between classification, recognition, and typicality. In: Journal of Experimental Psychology: Learning, Memory and Cognition 14, 700-708.

Phillips, Martin (1989): Lexical structure of text. (= Discourse Analysis Monographs 12). Birmingham: University of Birmingham.

Rice, Sally / Newman, John (2005): Inflectional islands. Presentation at the 9th International Cognitive Linguistics Conference, Yonsei University, Seoul, Korea. Internet: http://www.ualberta.ca/~johnnewm/ (last visited: 10 / 2010).

Schmitt, Norbert (ed.) (2004): Formulaic sequences: acquisition, processing and use. Amsterdam / Philadelphia: Benjamins.

Schmitt, Norbert / Carter, Ronald (2004): Formulaic sequences in action: An introduction. In: Schmitt (ed.), 1-22.

Sinclair, John (1991): Corpus, concordance, collocation. Oxford: Oxford University Press.

Swales, John M. (1990): Genre analysis: English in academic and research settings. Cambridge: Cambridge University Press.

Swales, John M. (2004): Research genres: exploration and applications. Cambridge: Cambridge University Press.

Talmy, Leonard (1996): The windowing of attention in language. In: Schibatani, Masayoshi / Thompson, Sandra A. (eds.): Grammatical constructions: Their form and meanings. Oxford: Oxford University Press, 235-287.

Taylor, John (2002): Cognitive grammar. Oxford: Oxford University Press.

Thompson, Sandra / Hopper, Paul (2001): Transitivity, clause structure and argument structure: evidence from conversation. In: Bybee / Hopper (eds.), 27-60.

Wray, Alison (2000): Formulaic sequences in second language teaching: Principle and practice. In: Applied Linguistics 21, 4: 463-489.

Wray, Alison (2008): Formulaic language: Pushing the boundaries. Oxford: Oxford University Press.

Yang, Huizhong (1986): A new technique for identifying scientific / technical terms and describing science texts. In: Literary and Linguistic Computing 1, 2: 93-103.

MARKÉTA MALÁ

Copular clauses in English and in Czech – a comparative corpus-based approach

Abstract

Copular clauses, i.e. clauses with a verbo-nominal predicate comprising a copular verb and a subject complement, are used in both English and Czech to ascribe a quality, property or value to the subject. While both languages make use of copular verbs *be* and *become* (*být*, *stát se*, respectively, in Czech), the repertoire of copular verbs is much broader in English, making it possible to distinguish between various types of attribution (e.g., verbs of 'seeming', attribution based on perception, verbs of 'remaining' etc.). The question then arises of what means are employed in Czech to express such 'modified attribution' and, on the other hand, what the constructions used in Czech can suggest of the meaning of the respective copular verbs in English.

The paper is based on the material drawn from a parallel translation corpus of Czech and English fiction texts. We hope it will therefore also illustrate some ways in which multilingual corpora can be employed in contrastive research.

1. Introduction

The present article sets out to explore two areas. First, given the differences between the system of copular verbs in English and Czech, the contrastive approach may reveal the means used to render the meaning of English copular clauses in Czech. At the same time, the paradigms of Czech correspondences can suggest something about the meaning and classification of English copulas. The second goal is a more methodological one: using a bidirectional parallel corpus of Czech and English, we would like to test some of the possibilities translation corpora offer for the study of comparable patterns of usage in different languages.

2. The material and method

The study is based on the material drawn from a parallel Czech – English corpus being put together as a part of a larger project of multilingual corpora:

InterCorp.[1] A pilot bidirectional balanced subcorpus of aligned Czech and English translations (c. 800 000 tokens) was used for the present study (Figure 1, on the methodology cf. also Johansson 2007, Dušková 2004, 2005).

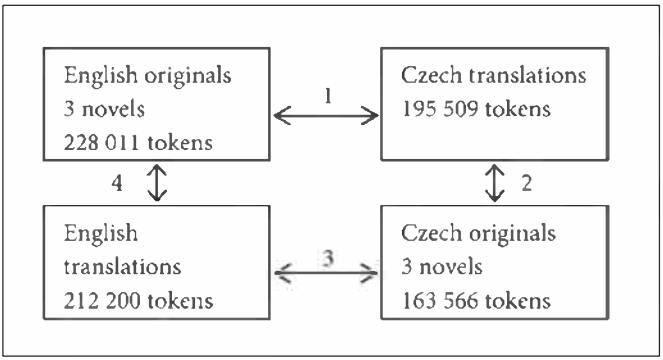


Figure 1: The corpus used in the present study (a sub-corpus of the InterCorp multilingual corpus)

The arrows in Figure 1 show the comparisons that were made: first, we proceeded from the English source texts, looking for the Czech translation counterparts of copular verbs; a pattern of classes of copular verbs became quite prominent after having classified and grouped together the correspondences. While the first step of the analysis focussed on formal correspondences, the use of a bidirectional corpus in the next one made it possible for us to proceed from function to its formal realization. Epistemic adverbials were identified as a frequent means of rendering the modification achieved by the copular verbs *seem*, *appear* and *look* in Czech. The Czech epistemic adverbials were used as query terms in Czech original texts; their English correspondences (step 3), though formally varied, may be considered functionally equivalent due to sharing the same translation counterpart. They all fall within the semantic domain of epistemic modification. The patterns of choice among the constructions available within this domain in English originals were found to be different from those in English translations (step 4).

[1] The study is a part of the research project *Czech National Corpus and Corpora of Other Languages* MSM 0021620823, InterCorp, http://www.korpus.cz/intercorp. The concordancer used was ParaConc (http://www.athel.com/para.html).

3. The scope of the study

Both in English and in Czech the repertoire of copular verbs includes the verb *be / být*, which "does not add any semantic content to the predicate phrase it is contained in" (Pustet 2003: 5). Since we are concerned with the types of meaning modification achieved by copular verbs, the verb *be* will be excluded from the present study. Instead, we shall focus on the other copular verbs ('semi-copulas', 'quasi-copulas' or 'complex-intransitives'), which display the same syntactic behaviour as the copula *be* but "add meaning to the predicate phrases in which they are contained. This semantic function, while not directly affecting the inner core of the predicate phrase, that is, its lexical nucleus, by altering the intrinsic semantic content of the latter, consist in 'importing' ... meaning components into the predicate phrase." (ibid.: 5-6). In Czech, the only 'semi-copula' is *stát se*, a resultative verb equivalent to *become (*cf. Grepl / Karlík 1998: 212). In English, the range of copulas is much broader. For the present study we only selected the basic prototypical members of the two groups of these verbs as listed in Huddleston / Pullum (2002: 263-264): a) verbs with depictive predicative complements (current copulas): *feel*, *continue*, *appear*, *look*, *keep*, *seem*, *smell*, *remain*, *sound*, *stay*, *prove*, *taste*; b) verbs with resultative predicative complements (resulting copulas): *become*, *grow*, *come*, *turn*, *fall*, *get*, *go*. The number of copular verbs in the corpus is given in Table 1.

	word-count (tokens)	number of copular verbs	
		absolute	per 1000 tokens
English originals	228 011	1 054	4.6
English translations	212 200	760	3.6

Table 1: The number of copular verbs in the corpus used

4. The correspondences between *become* and *stát se*

The system of copular verbs in English and Czech seems to overlap in the resultative verb *become* and *stát se*. However, a closer look at the correspondences of *become* reveals that while there is functional similarity, the two languages differ in the syntactic realization (cf. Teich 2003: 51). *Become* corresponds to *stát se* in 14.7 percent of examples only (Example 1). Although generally the same structural choices are available in the two languages, the patterns of choice differ. Czech appears to prefer focussing on the resultant state, indicating the change by

temporal adverbials (Example 2). There are also two types of correspondences, both more frequent than the copular verb *stát se*, which may be accounted for by the typological differences between the two languages. Czech, being a synthetic language, displays a preference for expressing aspectual modification (including resultativeness) by affixation (cf. prefixes *z-*, *vy-* in Examples 3 and 4). The English resultative copular predication will then be rendered in Czech as a lexical verb whose prefix indicates a change. The lexical verb may be derivationally related to an adjective corresponding to the English subject complement (Example 3, *wise* = *moudrá*) or morphologically unrelated to it (Example 4).

(1) A small bolt from a cockpit *became* jewellery. (MOE)
Matice z pilotní kabiny *se stala* šperkem.

(2) The mountains around the school *became* icy gray ... (JRH)
Hory kolem školy *byly teď* ledově šedé ...
"The mountains around the school *were now* icy gray ..."

(3) You that demon for pleasure who *became so wise*. (MOE)
Ty, která sis tak potrpěla na zábavu a která jsi tolik *zmoudřela*.

(4) We do, after all, wish him to *become* someone we can be proud of, don't we? (KIA)
Chceme přece, aby *vyrostl* v člověka, na nějž budeme moci být hrdí, ne?
"we ... wish him to grow up into someone ..."

The English copular predication was also found to correspond to a Czech catenative construction *začít* (*začínat*) / *přestat být* (i.e. *start* / *cease to be*) + complement (Example 5). Four correspondences were described as zero counterparts: here the overall semantic equivalence of the sentences is maintained, yet an explicit counterpart of the copular predicate cannot be identified in the translation, e.g. due to a shift in semantic roles and / or clause element functions (Example 6).[2]

(5) In jail he *became* serene and devious. (MOE)
Ve vězení *začal být* vážný a nevyzpytatelný.

(6) "Bulstrode, Millicent" then *became* a Slytherin. (JRH)
"Bulstrodeovou, Millicent" zařadil klobouk do Zmijozelu.
"«Bulstrode, Millicent»-object assigned the hat-subject to Slytherin"

[2] Instances where the counterpart of the English copular clause is missing in the Czech translation were excluded from the description.

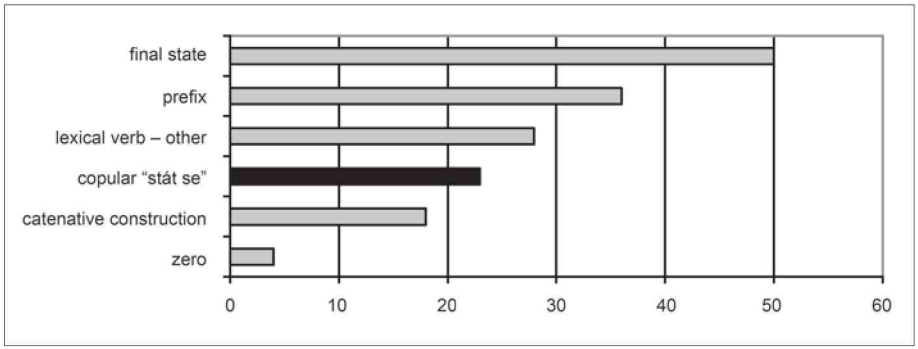


Figure 2: The correspondences of *become*

5. Czech translation counterparts of English copular verbs

Following the same procedure as in the case of *become*, we can identify a range of translation correspondences available for the other English copular verbs. Again, there may be zero[3] or overt correspondences. The overt counterparts may be classified as shown in Table 2 below:

Type of correspondence		Examples
verbal	lexical verb (resultative, perception, etc.)	For in all other respects, she *looks* a self-pitying sort. (KIA) Ve všech ostatních ohledech přitom *působí* sebelítostivě.
	copular verb	It *gets* very scary. (KIA) *Bude* to hrozně strašidelné.
	catenative construction	He would *get* restless ... (MOE) *Začínal být* neklidný ...
	verbal prefix (resultative)	They crossed the Bulaq Bridge and the traffic *got worse*. (MOE) Přejeli přes most Búláq a provoz *se zhoršil*.
verbo-nominal		He *felt* capable only of reconnaissance ... (MOE) *Měl pocit*, že je schopen jenom rekognoskace ...

[3] Apart from lexical-syntactic transpositions illustrated in Example 6, the category of 'zero correspondence' also includes instances where a direct counterpart of the copular verb (and the modification achieved by it) is missing in the translation, the rest of the sentence being, however, identifiable, e.g. "... there seemed to be a lot of strangely dressed people about" (JRH) – "...na ulici je spousta podivně oblečených lidí" (i.e. "... in the street there are a lot of strangely dressed people").

Type of correspondence		Examples
adver-bial	(epistemic, temporal, etc.)	It *seems* unimportant now, with the war and such things. (MOE) Teď za války a vůbec to *zřejmě* není důležité.
clausal	'comment clause'	Noriko, however, *seems* very proud of her apartment, ... (KIA) Noriko je však, *jak se mi zdá*, na svůj byt velice hrdá ...

Table 2: Types of Czech translation counterparts of English copular verbs

However, it is not primarily the paradigm of translation choices itself that we aim at. Since "one of the most fascinating aspects of multi-lingual corpora is that they can make meanings visible through translation patterns" (Johansson 2007: 28), we shall try to use the correspondence patterns to highlight the meaning modification achieved by copular verbs (cf. Table 3).

	total	zero	verbal							vb-nom.	adverbial		clau-sal
			lexical verb			copular verb		cate-nat.	pre-fix res.		epist	temp man-ner	
			re-sult.	per-cept.	other	*stát se*	*být*						
become	156	4	30	3	10	23	23	18	36	1	0	8	0
turn	16	0	2	0	0	1	2	2	9	0	0	0	0
go	35	0	8	0	0	0	2	0	25	0	0	0	0
fall	30	0	4	0	0	1	0	0	23	0	0	2	0
grow	31	0	10	0	3	0	3	3	12	0	0	0	0
get	48	2	12	3	1	0	9	6	12	2	0	1	0
come	2	0	1	0	0	0	0	0	1	0	0	0	0
prove	8	0	4	1	1	0	0	0	0	2	0	0	0
appear	38	7	0	17	5	0	0	0	0	0	8	0	1
look	169	4	6	141	1	0	7	0	2	0	7	1	0
seem	317	43	0	135	5	0	5	0	1	7	108	3	10
sound	26	1	0	14	5	0	2	0	0	0	1	3	0
feel	79	0	0	37	8	0	15	0	1	16	2	0	0
taste	2	0	0	2	0	0	0	0	0	0	0	0	0
remain	39	6	4	1	22	1	1	0	0	1	0	3	0
continue	49	17	0	0	1	0	0	0	1	0	0	30	0
stay	7	0	1	0	4	0	0	0	0	0	0	2	0
keep	2	0	0	0	1	0	1	0	0	0	0	0	0
total	**1054**	**84**	**82**	**354**	**67**	**26**	**70**	**29**	**123**	**29**	**126**	**53**	**11**

Table 3: Translation counterparts and the meaning of English copular verbs (the individual types of correspondence are exemplified in Table 2 above)

The correspondence pattern may be approached as a type of multi-dimensional variation (cf. Biber 1995: 18-20), i.e., a 'translation correspondence variation'. In a way similar to Biber's methodology, the approach uses text-corpora and computational tools to identify the corresponding structures in the parallel texts – the 'dimensions' of translation correspondence. Since the present study relies on a relatively small pilot corpus, multivariate statistical techniques were not used to analyze the co-occurrence relations among translation counterparts. We only recorded the absolute numbers of the individual types of correspondences for each copular verb in Table 3. The most frequent type of counterpart of each copula is marked in dark grey, the second most frequent in a lighter shade of grey. Even with this degree of simplification, a distinct pattern of correspondence starts to emerge.

First, there is a group of copular verbs which are predominantly translated by a lexical resulative verb. The resultative meaning of the verb is typically expressed by the prefix. This group comprises the copular verbs *become*, *turn*, *go*, *fall*, *get*, *grow*, *come*, *prove*. These verbs, indeed, coincide with those classified as verbs with resultative predicative complements in Huddleston / Pullum (2002: 264). The second group of copular verbs share a preference for lexical verbs related to perception as their counterparts: *appear*, *look*, *seem*, *sound*, *feel*, *taste*. Within this class, three verbs, *appear*, *look*, *seem*, are also frequently translated using epistemic adverbial modification in Czech. We shall return to this sub-group in Section 6 below. The third group of copular verbs – *remain*, *continue*, *stay*, *keep* – typically invites three types of counterparts: a lexical verb which comprises the meaning of the copula and its complement, adverbial modification, and a zero counterpart. The aspectual modification (durative) achieved by the copular verb in English is manifested in the morphological aspect marking in Czech. This applies both to the translation by a lexical verb and to the zero correspondence, where a direct counterpart of the copula is missing in the translation and the Czech verb corresponds directly to the infinitival complement of the copula (Example 7). The duration may be lexically reinforced by an adverbial of time (Example 8).

(7) Neither of us spoke for a few moments, while I *continued to light* lanterns. (KIA)
Chvíli jsme oba mlčeli, a já *rozsvěcel* lucerny.

(8) Mori-san *remained absorbed* by his pictures. (KIA)
Mori-san *si dál* zkoumavě *prohlížel* obrázky.

6. Copular verbs as means of epistemic modification

While the copular verbs in the first and third group may be considered means of aspectual modification, often corresponding to morphological marking of aspect in Czech, the second group – *appear*, *look*, *seem*, *sound*, *feel*, *taste* – seems to prefer lexical verbs related to perception or specific verb-based constructions closely tied to the individual copulas. For instance, 46.8 percent of the counterparts of the copula *feel* are lexical verbs. The range of these verbs is quite limited though: in 89.2 percent of instances the verb chosen is (*po*)*cítit* (*se*) or *připadat* (*si*). The second and third most prominent types of counterparts of the copula *feel* can be characterized syntactically either as copular predicates (with the copula *být*) or verbo-nominal constructions. Both types, however, comprise highly fixed expressions, often semantically non-compositional, e.g. *feel sorry – být* (komu) *líto*, *feel wide awake – být vzhůru*, *feel strange – mít podivný pocit*, *feel hungry – mít hlad*.[4]

Three copular verbs in this group – *appear*, *look*, and *seem* – share a preference for an epistemic adverbial modifier as their second most frequent counterpart. The adverbial typically takes either the form *jakoby / jako by* ("it was as if") (Example 9) or that of a modal adverb (Example 10).

(9) A soft rustling and clinking *seemed* to be coming from up ahead. (JRH)
Zepředu *jako by* k nim doléhalo tiché šustění a cinkání.

(10) ... he seems capable in that category. (MOE)
... je *zřejmě* v tomhle směru schopný.

The latter type of adverbial, being one of the paradigmatic choices of expressing epistemic modification available in both languages, poses the question of whether different ways of conveying this meaning are systematically preferred in the source and in the target language. Since copular verbs and modal adverbs are members of a larger set of means of expressing epistemic modification in the two languages, the bidirectional corpus can be used to identify the functionally equivalent constructions. Proceeding from the Czech translations to the English source texts, the Czech modal adverbs *zřejmě*, *očividně*, *zjevně*, *zdánlivě*, *nejspíš*,

[4] The correspondence between the copular predicate in English and the verbo-nominal construction with the verb *mít* ("have") in Czech suggests the closeness of *have* to copulas. *Have* in verbo-nominal constructions (e.g. *have the feeling that* ...) can indeed be considered a "copular verb with an object-like complementation" (Dušková et al. 2006: 417).

asi, *možná*, *nepochybně*, which serve as counterparts of the copular verbs *seem*, *appear* and *look*, were also found to correspond to English modal adverbs, adjectives and verbs, or comment clauses.[5] All these English forms can be considered functionally equivalent.[6] This approach thus makes it possible to highlight functional patterns in English by grouping together English constructions which, although formally varied, share the same Czech counterpart (Figure 3).

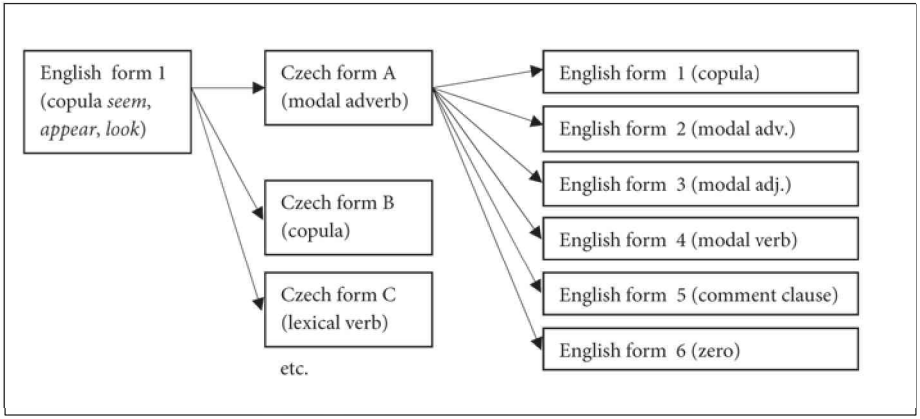


Figure 3: Looking for patterns of meaning in the parallel corpus[7]

The same modal adverbs were then used as query terms in Czech original texts. While the range of correspondences remains the same, the preferences are different (Table 4). In both directions of translation epistemic modification tends to be expressed by modal adverbs in English. However, the distribution of the other means seems to reveal translation effects, i.e. "differences between choices in original and translated texts in the same language" (Johansson 2007: 32, see also Baker 1993). Copular verbs constitute the third most frequent means of expressing the modification in the original texts (14.8 percent); when English is the target language, their representation drops to a mere

[5] 8.2 percent of examples had zero counterparts.

[6] While English constructions whose Czech translation counterparts share the same markers of discourse function can generally be considered functionally equivalent, manual checking is always an inevitable step in the analysis.

[7] The Czech counterparts (A, B, C, etc.) of the English copular verbs represent formally different means of rendering the meaning of the copulas in Czech. Proceeding from each of these Czech forms to its English translations, we can group together various English constructions (1, 2, 3, etc.) which can be assumed to be functionally equivalent (e.g. expressing epistemic modification) due to the fact that they share the same Czech form as their counterpart.

3 percent. More research is needed to answer the question of whether the differences in the preference patterns in the two languages are linked to the narrow repertoire of copular verbs in Czech and the tendency to prefer adverbial clausal modification to analytic modification by a modal verb within the verb phrase. Parallel corpora appear to be useful tools in investigating the area of such language-specific preferences.

	Czech translations > English originals		Czech originals > English translations	
	Σ	(%)	Σ	(%)
modal adverb	210	40.3	111	46.8
copular verb	77	14.8	7	3.0
modal verb	110	21.1	34	14.3
comment clause	60	11.5	27	11.4
modal adjective	18	3.5	5	2.2
zero	46	8.2	53	22.4
total	**521**	**100**	**237**	**100**

Table 4: The translation correspondences of modal adverbs *zřejmě očividně, zjevně, zdánlivě, nejspíš, asi, možná, nepochybně*

7. The dative with copular verbs

The epistemic modification conveyed by the copular verbs of perception and 'seeming' may be explicitly related to the observer or experiencer: "… the sense verbs and verbs of seeming license a *to*-phrase where the oblique NP expresses the experiencer." (Huddleston / Pullum 2002: 263) Johansson points out that the meaning of the copula *seem* can therefore be defined as: "somebody or something gives the experiencer the impression of being something or doing something." (Johansson 2007: 118)

In the Czech counterparts of this class of copular verbs the epistemic evaluation may be explicitly ascribed to an experiencer using a noun phrase in the dative case (Example 11).[8]

(11) *To one as young as you*, I'm sure it seems incredible … (JRH)
Někomu tak mladému jako ty to jistě zní neuvěřitelně …

[8] Similar to the English *to*-prepositional phrase, the Czech dative is typically the case form referring to the recipient.

The overt expression of the experiencer by the Czech dative was 4.6 times more frequent than overt reference to the experiencer by the *to*-prepositional phrase in the original English texts (Table 5). The reasons for the difference may be sought in the syntactic structure of the Czech counterparts. The dative is an obligatory complement of some of the verbs corresponding to the copular verbs of perception and 'seeming' (Example 12). However, even when optional, the dative tends to be expressed overtly in Czech where in English it remains implicit (Example 13).

(12) It felt as though he was sitting on some sort of plant. (JRH)
Připadalo *mu*, že snad sedí na nějaké rostlině.

(13) After what seemed an age, she turned and left. (JRH)
Zdálo se *jim*, že to trvá celou věčnost, pak se však paní Norrisová otočila a vyšla ven.

Moreover, our data have shown that the optional dative can also occur in the correspondences of English copular verbs other than verbs of perception and 'seeming' (Table 5). The semantic role of the participant referred to by the dative noun phrase, however, is not that of an experiencer or observer in these cases. The free dative denotes a participant affected by the change in situation or the resultant state (cf. Poldauf 1964). The change may have a negative effect on the participant ('dativus incommodi' – Example 14) or affect the participant in a positive way ('dativus commodi' – Example 15). However insufficient the size of the corpus, the Czech dative counterparts may suggest that even copular verbs other than those of perception and 'seeming' can convey an implicit 'relativization' of the situation with respect to a particular participant.

(14) The cut had turned a nasty shade of green. (JRH)
Rána *mu* ošklivě zezelenala.
"The cut had turned a nasty shade of green to / on him"

(15) Wood was now looking as though all his dreams had come true at once. (JRH)
Wood se teď tvářil, jako by se *mu* naráz splnily všecky jeho sny.
"... as though all his dreams had come true at once to him"

	English *to*-PP	Czech dative
become	0	1
turn	0	1
go	0	4
fall	0	0
get	0	0
grow	0	0
come	0	1
prove	0	0
appear	1	9
look	2	7
seem	17	43
sound	0	6
feel	1	25
taste	0	0
remain	0	0
continue	0	0
stay	0	1
keep	0	0
total	21	97

Table 5: The English *to*-prepositional phrase with copular verbs and the corresponding Czech dative construction

8. Conclusion

As pointed out by Tognini-Bonelli (1993: 209), "corpus studies have brought about a major shift in the relationship between data and theory. [...] Patterns of usage [...] are now surfacing and becoming noticeable in the light of the available evidence." This paper set out to examine the opportunities for a systematic study of comparable patterns of usage in different languages provided by parallel corpora. Based on the analysis of English copular verbs and their Czech correspondences, we hope to have shown several ways in which a bidirectional translation corpus can be employed to reveal such patterns in both languages and their correspondences as well as to highlight the effects of translation from one language to the other.

First, parallel corpora can be used to "make meanings visible through translation patterns" (Johansson 2007: 28). English copular verbs were grouped together according to their preferences with respect to the translation counter-

parts. The three classes of copulas established in this way correspond to the classification given by reference grammars – resulting copular verbs, current copulas and copular verbs of perception and 'seeming'. Secondly, parallel corpora can serve as a tool which makes it possible to proceed from a particular discourse function to its realization forms. Assuming that the various forms which share the same translation counterpart are functionally equivalent, various realizations of the discourse function can be grouped together. While parallel paradigms of means of expressing the particular function in the two languages may exist, the actual patterns of choice may be language-specific. Moreover, the pattern of preferences in the source language can leave its mark on the translation through overuse or underuse of particular constructions. This may be illustrated by the higher proportion of copular verbs as means of epistemic modification in English source texts as compared with English translations from Czech. Last but not least, parallel corpora may be the source of small surprises, such as the emergence of the affected participant in the translation counterparts of resultative copular clauses.

Parallel corpora not only appear to be useful sources of empirical data for comparative corpus-based research but they also make it possible to approach the data in new useful ways.

References

Baker, Mona (1993): Corpus linguistics and translation studies: Implications and applications. In: Baker / Francis / Tognini-Bonelli (eds.), 233-250.

Baker, Mona / Francis, Gill / Tognini-Bonelli, Elena (eds.) (1993): Text and technology. In Honour of John Sinclair. Amsterdam / Philadelphia: Benjamins.

Biber, Douglas (1995): Dimensions of register variation. A cross-linguistic comparison. Cambridge: Cambridge University Press.

Dušková, Libuše (2004): Syntactic constancy of the subject complement 1: A comparison between Czech and English. In: Linguistica Pragensia XIV, 2: 57-71.

Dušková, Libuše (2005): Syntactic constancy of the subject complement 2: A comparison between English and Czech. In: Linguistica Pragensia XV, 1: 1-17.

Dušková, Libuše et al. (2006 [1988]): Mluvnice současné angličtiny na pozadí češtiny. [The Grammar of Contemporary English against the Background of Czech]. Praha: Academia.

Grepl, Miroslav / Karlík, Petr (1998): Skladba češtiny. [The Syntax of Czech]. Praha: Votobia.

Huddleston, Rodney / Pullum, Geoffrey K. (2002): The Cambridge grammar of the English language. Cambridge: Cambridge University Press.

Johansson, Stig (2007): Seeing through multilingual corpora. On the use of corpora in contrastive studies. Amsterdam / Philadelphia: Benjamins.

Poldauf, Ivan (1964): The third syntactical plan. In: Travaux linguistiques de Prague 1, 241-255.

Pustet, Regina (2003): Copulas. Universals in the categorization of the lexicon. Oxford: Oxford University Press.

Teich, Elke (2003): Cross-linguistic variation in system and text. A methodology for the investigation of translations and comparable texts. Berlin / New York: de Gruyter.

Tognini-Bonelli, Elena (1993): Interpretative nodes in discourse – *actual* and *actually*. In: Baker / Francis / Tognini-Bonelli (eds.), 193-212.

Sources

InterCorp: http://www.korpus.cz/intercorp.

English – Czech direction:

JRH: Rowling, Joanne K. (1997): Harry Potter and the Philosophers' Stone. London: Bloomsbury. / (2000): Harry Potter a Kámen mudrců. Translated by Medek, Vladimír. Praha: Albatros.

KIA: Ishiguro, Kazuo: An artist of the Floating World. London: Faber. / (1999): Malíř pomíjivého světa. Translated by Hanuš, Jiří. Praha: Argo.

MOE: Ondaatje, Michael: The English patient. London: Bloomsbury. / (1997): Anglický pacient. Translated by Masnerová, Eva. Praha: Argo.

Czech – English direction:

LFM: Fuks, Ladislav (1963): Pan Theodor Mundstock. Praha: Československý spisovatel. / (1969): Mr Theodore Mundstock. Translated by Urwin, Iris. London: Cape.

MKL: Kundera, Milan (1985): Nesnesitelná lehkost bytí. Toronto: Sixty-Eight Publishers. / (1984): The unbearable lightness of being. Translated by Heim, Michael Henry. London: Faber.

MVV: Viewegh, Michal (1984): Výchova dívek v Čechách. Praha: Český spisovatel. / (1997): Bringing up girls in Bohemia. Translated by Brain, A.G. London: Readers International.

SVETLANA GOROKHOVA

The role of frequency effects in the selection of inflected word forms: A corpus study of Russian speech errors

Abstract

A comparison of Russian speech error data with the corpus data suggests that probabilistic information about the frequency of inflected word forms is available in the speaker's production lexicon, and higher-frequency inflected forms are more likely to be selected during language production.

1. Introduction

Background: The role of frequency effects in the generation and perception of syntactic structures has been a point of debate for a number of years. The two opposing approaches to mental grammar, known as structural / generative grammars and usage-based grammars, propose different accounts of how probabilistic information about inflected word forms is used in language production and processing.

Usage-based grammars suggest that every time a word form or construction is used, it activates a node or pattern of nodes in the mental lexicon, and that the storage of a word form or grammatical construction is affected by the frequency of its activation (see Croft / Cruse 2004).

Generative grammars, while accepting the view that frequency effects play a role in the storage and production of irregular forms, posit that regularly inflected forms are insensitive to frequency effects. According to this view, which is known as the dual-route or dual-processing theory, regularly inflected forms are produced online via certain grammatical rules, and it is only low-frequency irregular forms that are stored as whole units (Caramazza et al. 1985; Pinker / Prince 1988, 1994; Pinker 1991, 1997; Clahsen et al. 1992; Prasada / Pinker 1993; Marcus et al. 1995; Ullman 1999). However, dual-route models have met some criticism from the proponents of usage-based models, who claim that alongside irregular forms, speakers store many regular high-frequency inflected forms (see Dąbrowska 2008). Form frequency effects for regular inflected

words have been observed in English (Sereno / Jongman 1997, New et al. 2004), Finnish (Lehtonen / Laine 2003), French (New et al. 2004), and Dutch (Baayen / Dijkstra / Schreuder 1997, Baayen et al. 2002, Tabak / Schreuder / Baayen 2005 for comprehension; Bien / Levelt / Baayen 2005 for production). Similar results have been obtained for anomic speech (Bi / Han / Shu 2007). These experimental studies provided supportive evidence for Stemberger / McWhinney's (1986) hypothesis that high frequency inflected forms are stored as separate entries in the lexicon.

To solve the controversy, some authors have hypothesized that during language comprehension, both morphemes and whole-word representations could be accessed in parallel (Caramazza / Laudanna / Romani 1988, Frauenfelder / Schreuder 1992, Schreuder / Baayen 1995). According to this view, morphologically complex words can be identified via two routes: a direct route that makes use of whole-word representations, and a decompositional route that goes through morphemic units. Which of the two routes "wins the race" is determined by the words' linguistic and distributional properties such as frequency, formal and semantic transparency, morpheme productivity and lexicality (Baayen / Dijkstra / Schreuder 1997; Bertram / Laine / Karvinen 1999; Caramazza / Laudanna /Romani 1988; Laine / Vainio / Hyönä 1999; Schreuder / Baayen 1995, 1997).

There is considerable linguistic evidence that "use overrides structure" in determining lexical representation (Bybee 2006: 189). The fact that the frequency of use of grammatical constructions at different levels of schematicity is an important determinant of linguistic structure and language use is recognized by all usage-based models of grammar (see Croft / Cruse 2004, Bybee 2006, Diessel 2007). Demuth (2007) suggests that the frequency of different grammatical structures is part of the speaker's linguistic competence. Bybee (2006) argues that token frequency partly determines the strength of representation of individual lexical items in the mental lexicon while the strength of schemas (such as the formation of past tense forms of regular and irregular verbs in English) is reflected in type frequency. Diessel (2007: 124) regards the frequency of occurrence of linguistic structures as a driving force of several psychological mechanisms involved in language use. However, in his work reviewing the state of the art in probabilistic modeling of language use, Jurafsky notes that so far "the role of probabilities in non-lexical syntactic structure, while assumed in most probabilistic models, rests on very little psychological evidence" (Jurafsky 2003: 84).

Recent experimental studies have focused on the use of probabilistic information about grammatical word forms in language comprehension. Kostić/Mirković (2002) investigated the processing of inflected forms of Serbian feminine nouns. Their results seem to indicate that the average frequency of a given noun form plays a role in the processing of inflected nouns. In a study of the recognition of Serbian nouns, Milin / Filipović Đurđević / Moscoso del Prado (2008) found that the distribution of inflectional variants within a particular inflectional paradigm influences word recognition. The authors conclude that inflectional paradigms in the traditional linguistic sense influence the processing and representation of words in the mental lexicon (cf. the study by Clahsen et al. (2001) suggesting the psychological reality of inflectional paradigms).

Some evidence supporting the view that lexical processing is affected by the frequency of a grammatical word form comes from the studies of language deficit in agrammatism. Stemberger (1984, 1985) used an associative network model of sentence production, claiming that since more frequent word forms have lower activation thresholds, they tend to replace less frequent word forms in agrammatic speech (cf. Bybee 1995). Faroqi-Shah / Thompson (2004), who analyzed verb inflection errors observed in English-speaking agrammatic individuals, concluded that the errors are likely to be a consequence of a pre-phonological diacritical deficit. They further proposed that, in case of a diacritical failure, word form frequency affects sentence production (but see Janssen / Penke (2002) for contradicting evidence from the speech of agrammatic aphasics). In a number of experiments, Baayen et al. (2003) show that the (relative) frequencies of plural forms influence response latencies. This finding can be interpreted as strong evidence in favor of independent representation of plural forms for nouns and verbs, even for forms that are completely regular. Recent statistical studies provide growing evidence for the availability of probabilistic information about individual inflectional variants of a word in lexical memory (see Baayen 2007, Smolka / Zwitserlood / Rösler 2007).

Recent studies of first (Tomasello 2003, Lieven / Tomasello 2008) and second language acquisition (Ellis 2008) have revealed the importance of frequency effects in the acquisition of linguistic constructions. In particular, frequency effects have been shown to play a role in the development of children's syntax (see Demuth 2007).

Diessel (2007) argues that frequency of use reinforces the representation of linguistic expressions in memory, which in turn influences their activation and interpretation in language use. He further claims that the influence of frequency on linguistic structure challenges the rigid division between grammar and language use, suggesting a dynamical model of grammar in which linguistic structure is grounded in language use (Diessel 2007: 123-124).

Data: This paper focuses on the role of frequency effects in the production of syntactic structures. It approaches the issue of whether the selection of a word's grammatical forms is probabilistic in nature through an analysis of Russian speech error data. An advantage of speech error evidence is that it reveals the processes that occur in natural language production and are not dependent on any preset experimental conditions.

Since Russian is a highly inflected language, a failure to retrieve a target grammatical feature such as gender, number, case, person, tense, and aspect is likely to surface as an inappropriate grammatical form of a given word. Unlike picture-word naming tasks used to investigate the mechanism of grammatical feature selection, which focus on a limited number of categories such as gender and number (e.g. Schriefers 1993; Caramazza et al. 2001; Schiller / Caramazza 2002, 2003; Costa et al. 2003; Bordag / Pechmann 2008), speech errors involve a variety of different grammatical features, thus making it possible to add a wider perspective to the study of the storage of inflected word forms and the mechanism of their retrieval.

The paper reports a study of 198 naturally produced Russian speech errors (slips of the tongue) which result in the selection of a wrong inflected form of a noun, pronoun, verb, or adjective. Such errors (henceforth context-free substitutions of a grammatical feature) are seemingly unaffected by the grammatical forms of the other words in the current utterance. The available examples comprise substitutions of different grammatical features such as case, number, gender, person, tense, and aspect.

The 198 examples relevant to the present study were selected from the corpus of approximately 6000 Russian speech errors which were collected by tape-recording and digitally recording everyday conversations, telephone conversations, and live TV and radio programs such as talk shows and interviews.

2. An Overview of the Russian Inflectional System

The following is a general description of the features of the Russian inflectional system that are relevant to the present study (for a detailed description, see e.g. Shvedova et al. 1982, Jacobson 1984).

Russian is a morphologically complex language in which nouns do not require determiners expressing number, gender, and case information. Instead, grammatical features such as case, gender, number, person, tense, and aspect almost always surface as bound morphemes.

Nouns, personal pronouns and nominal adjectives are overtly marked for gender, number, and case, while verbs and predicate adjectives are marked for gender and number information. There are three grammatical genders (masculine, feminine, and neuter) and two kinds of number specification, singular and plural.

Russian has six grammatical cases: nominative, accusative, genitive, dative, instrumental, and locative, which are marked differently for singular and plural in three declension paradigms. For singular nouns, the exact form of the case marking depends on the gender of the noun. Most singular masculine and neuter nouns follow the first declension, the feminine nouns that end in *-a* or its allomorph in the nominative case and the masculine nouns that end in *-a* follow the second declension, and the feminine nouns that end in a palatalized consonant in the nominative case follow the third declension.

Plural nouns (regardless of gender) typically follow one of the three different declension paradigms depending on whether their stems end in a hard consonant, a soft consonant or sibilant, or in [*-j*]. The case markers for nominative, accusative, and genitive case forms differ depending on the declension type whereas the case markers for dative, instrumental and locative case forms are similar in all plural nouns regardless of their stem ending.

Russian case marking varies with respect to animacy, which affects the marking of the accusative case in masculine and plural nouns. In addition, allomorphy is fairly common, e.g. the phonological shape of the noun determines the specific variant of the preposition and case marker.

The declension of personal pronouns is to a large extent idiosyncratic, using full or partial suppletion in most declension paradigms.

Similarly to nouns and adjectives, Russian verbs are inflected for gender and number. Besides, verbs express tense, aspect, person, voice, and mood information. There are three tenses: past, present, and future, and two aspects (perfective and imperfective).

3. Examples

This section contains some representative examples of context-free substitutions of the different grammatical features.

In (1), the genitive case form of the noun 'interest', *procent-ov*, is selected instead of the target locative case form, *procent-ax*:

(1) Case feature substitution: LOC → GEN

Očen'	xorošo	sekonomlju	na **procent-ax**	→	... na	**procent-ov**
very	well	save:FUT	on interest-PL.**LOC**		... on	interest-PL.**GEN**

I'll save a lot on interest [*payments*]

Likewise, dative is commonly replaced by accusative / genitive as in (2), where the accusative / genitive form of the pronoun 'she', *ee*, is substituted for the target dative, *ej* (the exact substitute case feature cannot be identified due to the homonymy of the accusative and genitive case forms):

(2) Case feature substitution: DAT → ACC / GEN

Ty	**ej**	pozvonila?	→	Ty	**ee**	pozvonila?
you	3SG.F.**DAT**	call:F.PST		you	3SG.F.**ACC/GEN**	call:F.PST

Have you called her?

In (3), the genitive case form of the noun 'health', *zdorovj-a*, is substituted for the target instrumental case form, *zdorovj-em*:

(3) Case feature substitution: INS → GEN

Pod	normoj	ili **zdorovj-em**	ponimajut	takuju	formu	žiznedejatel'nosti...	→ ...	**zdorovj-a**
Under	norm-SG.INS	or health-SG.**INS**	interpret	such	form	vital activity		health-SG.**GEN**

Norm or health is interpreted as such a form of vital activity…

In (4), the target genitive case form *belka* of the word *belok* 'protein' is replaced by the nominative / accusative case feature (again, the exact substitute case feature cannot be identified because of the homonymy of the nominative and accusative case forms):

(4) Case feature substitution: GEN → NOM / ACC

– Ja tut pročla v gazete, čto zefir...
I've read in a paper that marshmallow...

– Bez	**belk-a**?	→	Bez	**belok**?
without	protein-SG.**GEN**		without	protein:SG.**NOM / ACC**

Has no protein?

Apart from the case feature, other grammatical features also appear to be involved in context-free substitutions although their examples are not as numerous in the speech error corpus as those of case feature substitutions.

In (5), the target plural number feature of the noun *akcii* 'shares' and its premodifier *eti* 'these' is replaced by the singular *akciju* 'share' and *etu* 'this', respectively:

(5) Number feature substitution: PL → SG

U každogo brokera vsegda est'na rukax opredeljonnoje količestvo akcij.
Each broker always holds a certain number of shares.

Vot	možno	vzjat'	u	nego	vzajmy	et-i	**akci-i**	→ ...et-u	**akci-ju**
So	can	take	from	him	on loan	this-**PL**.NOM/ACC	share-**PL**.ACC	this-**SG**.F.ACC	share-**SG**.ACC

So you can borrow these shares from him

In (6), the speaker declines the neuter gender noun *bljudce* 'saucer' as a feminine gender noun; moreover, in all its modifiers, the neuter gender feature is replaced by the feminine gender feature:

6) Gender feature substitution: N → F

Ja voz'mu	vot	**et-o**	**bljudc-e:**	**on-o**	**sam-oe**	**malen'k-oe**
I will take	here	this-SG.**N**.ACC	saucer-SG.**N**.ACC	3SG-**N**.NOM	most-SG.**N**.NOM	small-SG.**N**.NOM
→ ...	vot	**et-u**	**bljudc-u:**	**on-a**	**sam-aja**	**malen'k-aja**
		this-SG.F.ACC	saucer-SG.F.ACC	3SG-F.NOM	most-SG.F.NOM	small-F.NOM

I'll take this saucer: it's the smallest

In (7), the target 2 person form of the verb 'be', *budeš'*, is replaced by the 3 person form, *budet*:

(7) Person feature substitution: 2 → 3

Poslezavtra	**bud-eš'**	otdoxnuvšij	→	**...bud-et...**
day after tomorrow	be-**2**SG.FUT	well-rested		be-**3**SG.FUT

You'll feel well-rested tomorrow

(8) Tense feature substitution: FUT → PST

Ja dumaju,	ja vynuždena	**budu**	vyslušat'	plamennuju	tiradu	→	...**byla**...
I think	I have to	be:3SG.**FUT**	listen to	fiery	tirade		be:3.SG.**PST**

I think I'll have to listen to a fiery tirade

The past tense form of the verb 'be', *byla*, in (8) is substituted for the target future tense form, *budu*. Characteristically, past and present tense forms are opposed to future tense forms in that they are predominant in spoken Russian: the past and the present forms each have a frequency of occurrence of about 45 per cent whereas the frequency of the future forms is only about 10 per cent (Sandzhi-Garjaeva 2003).

There are, however, contrary examples such as (9), where the target past tense form of the verb 'come', *prišel*, is replaced by the future tense form, *pridet*:

(9) Tense feature substitution: PST → FUT

V načale 80-x godov	Uran	**priš-el**	na eto mesto	→	...**prid-et**
in early 80s	Uranus come-**PST**.SG.M		to this place		come-FUT.SG

In the early 80s, Uranus moved to this point

In (10), *vzjat'* and *brat'* are two suppletive forms (perfective and imperfective respectively) of 'get': the imperfective form is substituted for the target perfective:

(10) Aspect feature substitution: PFV → IPFV

Mogla	prekrasnuju	putevku	**vzjat'**	→	...**brat'**
Could	gorgeous	voucher	get:INF.**PFV**		get:INF.**IPFV**

I could've got a gorgeous voucher

At the same time, (11) is a reverse example, where the target past imperfective form of the verb *dopuskat'* 'make' is replaced by the past perfective:

(11) Aspect feature substitution: IPFV → PFV

My pomnim,	kak nervničala	Alena i	**dopusk-ala**	ošibki	→	...**dopust-ila**
We remember	how was nervous	Alena and	make-PST.**IPFV**	mistakes		make-PST.**PFV**

We remember that Alena was nervous and was making mistakes

However, errors involving tense and aspect feature substitutions are sometimes difficult to interpret because the two features closely interact and as a result, the error verb form may differ from the target form both in tense and in aspect as in (12), where the target future perfective form of the verb 'succeed', *udastsja*, is replaced by the present imperfective form, *udaetsja*:

(12) Tense + Aspect feature substitution:

Možet byt'	vse-taki	emu	**uda-stsja**	dovesti	svoju ideju	do konca
maybe	still	3SG.M.DAT	succeed-3SG.FUT.PFV	bring	one's idea	to end
		→ ...	**uda-etsja** ...			
			succeed-3SG.PRS.IMPV			

Still, he might succeed in carrying his idea through

4. Comparison with the corpus data

Errors like (1) - (12) suggest that within the inflectional paradigm of a given word, some forms may be more likely to be selected.

To test whether the frequencies of inflected word forms correlate with the patterns of context-free grammatical feature substitutions, the raw and relative frequencies of each substitute word form in the *Russian National Corpus* (http://ruscorpora.ru) were compared to those of the target word. Since the speech errors under study occurred in spoken Russian, the search was run in the spoken part of the corpus. The subcorpus of spoken Russian is a 7.8 million word spoken language corpus, containing formal and informal monologs and dialogs. As only part of the corpus is annotated for grammar, the search results had to be corrected manually to remove grammatical homonymy in some cases.

The comparison between the raw frequencies of the target and error word forms indicates that speakers tend to substitute higher frequency forms for low frequency forms ($t(198) = 2.35$, $p < .05$).

For noun / pronoun case form substitutions, which make up 52.5 per cent of the total number of context-free grammatical feature substitutions analyzed, the number of available examples (104) made it possible to estimate the statistical significance of the difference between the relative frequencies of the target case forms within the given word's declension paradigm and the relative frequencies of the corresponding substitute case forms.

Sample results for case feature substitutions are presented in Table 1. The table shows the raw frequencies of the target and substitute case forms and their relative frequencies within the word's declension paradigm as per the spoken part of the *Russian National Corpus*. The general tendency, as illustrated by Table 1, is that in most cases, a case form that has a higher frequency within the word's declension paradigm is substituted for a lower frequency case form. The tendency seems to apply both to regular (nouns) and irregular (pronouns) inflected forms. Paired-samples *t*-tests performed to compare the relative frequencies of the target case forms to those of the substitute case forms reveal that the tendency is statistically significant for all case feature substitutions ($t(104) = 3.39$, $p < .001$).

Noun or personal pronoun	**Target form**	**Raw freq.**	**Relative freq. (%)**	**Error form**	**Raw freq.**	**Relative freq. (%)**
belok *protein*	GEN	7	41.2	NOM / ACC	9	52.9
točki *points*	GEN	56	21.7	NOM / ACC	161	62.4
on *he*	GEN	20 294	21.63	NOM	48 856	52.08
turisty *tourists*	DAT	9	5.7	ACC / GEN	28	17.72
rodstven-niki *relatives*	DAT	122	11.11	ACC / GEN	158	14.39

Noun or personal pronoun	Target form	Raw freq.	Relative freq. (%)	Error form	Raw freq.	Relative freq. (%)
ona *she*	DAT	3 241	7.93	ACC / GEN	9 396	22.98
my *we*	DAT	9 506	12.21	ACC / GEN	24 419	31.36
oni *they*	DAT	4 511	7.6	ACC / GEN	11 009	41.13
vy *you*	DAT.PL	14 833	16.27	ACC / GEN	18 305	20.08
papa *dad*	DAT	105	4.2	GEN	133	5.3
proekty *projects*	DAT	8	3.61	LOC	15	6.78
etap *stage*	INS	9	1.51	NOM / ACC	89	14.93
zdorov'e *health*	INS	39	5.4	GEN	181	25
holodil'nik *fridge*	LOC	48	15.1	NOM / ACC	82	25.78
procenty *interest*	LOC	26	1.05	GEN	1 087	44.77
literatura *literature*	LOC	146	21.95	GEN	211	31.73
kollektiv *staff*	GEN	62	19.13	LOC	54	16.66
stol *table*	LOC	330	13.68	DAT	212	8.79

Table 1: Sample frequencies of target and error word forms in context-free case feature substitutions

Because of the homonymy of some case forms, it is not always possible to identify the exact case feature of the substitute, e.g. in (2), the substitute form *ee* 'her' is either ACC or GEN, whereas in (4), the substitute word form *belok* 'protein' is either NOM or ACC.

The general frequency distribution of different case forms in spoken Russian is shown in Figure 1 (the frequency values were taken from Martynenko 2003).

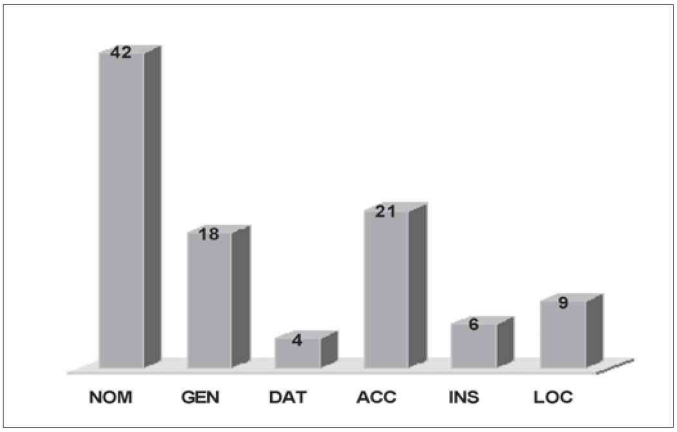


Figure 1: Frequency distribution of different case forms in spoken Russian (per cent)

A comparison of the results presented in Table 1 and the data from Figure 1 indicates that the case forms reported to occur most frequently in spoken Russian (nominative, genitive, and accusative) tend to replace the less frequent oblique case forms (dative, instrumental, and locative). At the same time, the more frequent nominative and accusative forms are substituted for the genitive. Paired t-tests show that the frequency differences appear to be significant for GEN → NOM / ACC substitutions (t (37) = 3.58, $p < .001$) and for DAT → ACC / GEN substitutions (t (39) = 3.08, $p < .01$).

Because the accusative case forms of Russian nouns and pronouns are often homonymous with either their nominative (for singular masculine inanimate and plural inanimate nouns; masculine inanimate, neuter, and plural inanimate demonstrative pronouns, etc.) or their genitive (for singular masculine animate and plural animate nouns; singular and plural 1, 2 and 3 person pronouns; masculine animate and plural animate demonstrative pronouns, etc.) case forms, the frequency of a substitute inflected form is in fact the sum frequency of NOM / ACC or ACC / GEN forms respectively, so the 'strength' of a substitute inflected form in such cases may be 'doubled', increasing the likelihood of a substitution error.

The numbers of the examples illustrating the substitutions of other grammatical features are still insufficient to estimate the statistical significance of the relative frequency differences. In addition, the available examples are quite variable, e.g. the examples of substitutions of grammatical features such as gender and number involve different parts of speech, so it is often problematic to determine the exact boundaries of a word's inflectional paradigm for a particular feature. Besides, the interaction of grammatical features such as tense and aspect (see example 12 above) in some of the substitution errors complicate their analysis. Thus far, there is no statistically significant data on the relative frequency differences for any of the above-mentioned grammatical features other than the case feature.

5. Discussion

The goal of this paper was to use speech error data to explore the mechanisms of grammatical feature selection in the speaker's production lexicon in a highly inflected language like Russian and to investigate whether speakers use any kind of probabilistic information about inflected word forms during natural language production.

Firstly, speech error data indicate that the selection of grammatical features is generally a competitive process. This conclusion runs counter to the results of most experimental studies of lexical retrieval that used picture-word interference paradigms to study the mechanism of grammatical feature selection. Their authors claim that although the selection of lexical nodes may be competitive, the selection of their grammatical properties is an automatic consequence of lexical selection (Caramazza et al. 2001; Schiller / Caramazza 2002, 2003; Costa et al. 2003; Bordag / Pechmann 2008). Contrary to this view and in line with Schriefers (1993), evidence from Russian speech errors points to the competitive nature of grammatical feature selection. However, unlike the picture-word interference studies which focus on the competition resulting from the interference of the gender or number feature of a distractor word, examples (1-12) of context-free substitutions of a grammatical feature indicate that grammatical features can compete for selection within a word's grammatical paradigm even when the selection process is unaffected by any distracting context. Moreover, context-free grammatical feature substitutions reveal that the list of grammatical features involved in the competition is not confined to gender and number. It appears that other features such as case, tense, aspect,

and person (examples 1-4, 7-12) can also be selected by competition. Thus, speech error evidence indicates that grammatical features do compete for selection during language production.

The comparison of raw frequencies of the target and substitute word forms in the corpus indicates that the general tendency is for a higher-frequency inflected form to substitute for a lower-frequency form. This finding reveals the role of token frequency effect in language production.

This is consistent with the finding that lexical processing is affected by the frequency of a grammatical word form (e.g. Kostić / Mirković 2002, Milin / Filipović Đurđević / Moscoso del Prado 2008) and the data on first and second language acquisition (Tomasello 2003, Lieven / Tomasello 2008, Ellis 2008) and agrammatic speech (Stemberger 1984, 1985; Faroqi-Shah / Thompson 2004).

We have seen that token frequency effect plays a part in language production. Assuming that context-free substitutions of a grammatical feature result from competition among the grammatical forms of a target word during lexical retrieval, it seems reasonable to suggest that some inflected forms might dominate the word's inflectional paradigm, i.e. some types of forms are more likely to be selected. Is information about type frequency available in the production lexicon?

The comparison between the relative frequencies of the target and error inflected forms of a word in the spoken part of the *Russian National Corpus* shows that, at least for the case feature, the relative frequency of the substitute case form within the word's declension paradigm is generally higher than that of the target case form.

The results suggest that during the selection of inflected word forms, some forms may have a priority within a word's inflectional paradigm. Thus, it appears that the noun and pronoun case forms that occur most frequently in spoken Russian (nominative, genitive, and accusative) tend to substitute for the less frequent oblique case forms such as dative; at the same time, the higher-frequency nominative and accusative forms tend to replace the genitive. This tendency seems to reflect type frequency effect. The fact that some noun and pronoun case forms are homonymous probably adds to the strength of the substitute case forms, e.g. the homonymy of the nominative and the accusative forms (the most robust case forms) may make a GEN → NOM / ACC

substitution more likely to occur whereas the homonymy of the accusative and the genitive forms may increase the likelihood of a DAT → ACC / GEN substitution.

A question that naturally poses itself is whether it is a certain case form that has a priority over some other case forms within the word's declension paradigm, or whether it is possible to speak of a priority that a certain grammatical feature, e.g., the genitive case feature, has over some other grammatical features, e.g., dative and instrumental case features.

It seems that it would be premature to speak of a hierarchy of grammatical features rather than inflected word forms because many of the available speech error examples involve homonymy of case forms, e.g. a case form that might be either genitive or accusative has a higher frequency compared to the dative form. Such examples leave one in doubt as to whether it is the genitive or the accusative case feature that dominates over the dative, or whether the effect is due to the combined frequency of the genitive and the accusative case forms of a given word. So far, there is not enough evidence for a hierarchy of grammatical features rather than grammatical forms. It therefore seems reasonable to suggest that it is certain inflected word forms that are more likely to be selected than other forms during language production.

A way to account for this finding is to hypothesize that the different inflected forms of a word are coded for frequency of occurrence in the production lexicon, and the more robust higher-frequency forms are more readily accessible as potential substitutes of the weaker low-frequency forms. A higher-frequency form can be used as a default form in case the target lower-frequency form is currently inaccessible. Within the framework of spreading activation models of language production (Dell 1986, Dell et al. 1997), such frequencies stored in lexical memory can be encoded as resting activation levels (see Jurafsky 2003).

The results provide supportive psycholinguistic evidence for dynamical, usage-based models of mental grammar and are in line with the view that both token frequency and type frequency play a role in the organization of the mental lexicon (cf. Bybee 2006) and that the frequency of occurrence of linguistic constructions is part of the speakers' linguistic competence.

6. Conclusions

Speech error evidence indicates that inflected word forms compete for selection during natural language production. It further suggests that probabilistic information about individual inflected forms of a word is available in the speaker's production lexicon. The selection of an inflected word form seems to be affected by the frequency of its occurrence, and the more frequently occurring forms are more likely to be selected. These findings provide psychological evidence in favor of usage-based models of grammar.

References

Baayen, R. Harald (2007): Storage and computation in the mental lexicon. In: Jarema, Gonia / Libben, Gary (eds.): The mental lexicon: Core perspectives. Amsterdam: Elsevier, 81-104.

Baayen, R. Harald / Dijkstra, Ton / Schreuder, Robert (1997): Singulars and plurals in Dutch: Evidence for a parallel dual route model. In: Journal of Memory and Language 36: 94-117.

Baayen, R. Harald / McQueen, James / Dijkstra, Ton / Schreuder, Robert (2003): Frequency effects in regular inflectional morphology: Revisiting Dutch plurals'. In: Baayen, R. Harald / Schreuder, Robert (eds.): Morphological Structure in Language Processing. Berlin: de Gruyter, 355-390.

Baayen, R. Harald / Schreuder, Robert / De Jong, Nivja H. / Krott, Andrea (2002): Dutch inflection: the rules that prove the exception. In: Nooteboom, Sieb / Weerman, Fred / Wijnen, Frank (eds.): Storage and computation in the language faculty. Dordrecht: Kluwer Academic Publishers, 61-92.

Bertram, Raymond / Laine, Matti / Karvinen, Katja (1999): The interplay of word formation type, affixal homonymy, and productivity in lexical processing: Evidence from a morphologically rich language. In: Journal of Psycholinguistic Research 28, 3: 213-226.

Bi, Yanchao / Han, Zaizhu / Shu, Hua (2007): Compound frequency effect in word production: Evidence from anomia. In: Brain and Language 103: 8-249.

Bien, Heidrun / Levelt, Willem / Baayen, R. Harald (2005): Frequency effects in compound production. In: Proceedings of the National Academy of Sciences of the USA 102: 17876-17881.

Bordag, Denisa / Pechmann, Thomas (2008): Grammatical gender in speech production: Evidence from Czech. In: Journal of Psycholinguistic Research 37: 69-85.

Bybee, Joan L. (1995): Regular morphology and the lexicon. In: Language and Cognitive Processes 10: 425-455.

Bybee, Joan L. (2006): Frequency of use and the organization of language. Oxford: Oxford University Press.

Caramazza, Alfonso / Miceli, Gabriele / Silveri, M. Caterina / Laudanna, Alessandro (1985): Reading mechanisms and the organisation of the lexicon: Evidence from acquired dyslexia. In: Cognitive Neuropsychology 2: 81-114.

Caramazza, Alfonso / Laudanna, Alessandro / Romani, Christina (1988): Lexical access and inflectional morphology. In: Cognition 28, 3: 297-332.

Caramazza, Alfonso / Miozzo, Michele / Costa, Albert / Schiller, Niels / Alario, F. Xavier (2001): A crosslinguistic investigation of determiner production. In: Dupoux, Emmanuel (ed.): Language, brain, and cognitive development: Essays in honor of Jacques Mehler. Cambridge, MA: MIT Press, 209-226.

Clahsen, Harald / Hadler, Meike / Eisenbeiss, Sonja / Sonnenstuhl-Henning, Ingrid (2001): Morphological paradigms in language processing and language disorders. In: Transcations of the Philological Society 99, 2: 247-277.

Clahsen, Harald / Rothweiler, Monika / Woest, Andreas / Marcus, Gary F. (1992): Regular and irregular inflection in the acquisition of German noun plurals. In: Cognition 45: 225-255.

Costa, Albert / Kovacic, Damir / Fedorenko, Evelina / Caramazza, Alfonso (2003): The gender congruency effect and the selection of freestanding and bound morphemes: Evidence from Croatian. In: Journal of Experimental Psychology: Learning, Memory, and Cognition 29: 1270-1282.

Croft, William / Cruse, D. Alan (2004): Cognitive linguistics. Cambridge: Cambridge University Press.

Dąbrowska, Ewa (2008): The effects of frequency and neighbourhood density on adult speakers' productivity with Polish case inflections: An empirical test of usage-based approaches to morphology. In: Journal of Memory and Language 58: 931-951.

Dell, Gary S. (1986): A spreading-activation theory of retrieval in sentence production. In: Psychological Review 93, 3: 283-321.

Dell, Gary S. / Schwartz, Myrna F. / Martin, Nadine / Saffran, Eleanor M. / Gagnon, Deborah A. (1997): Lexical access in aphasic and nonaphasic speakers. In: Psychological Review 104: 801-838.

Demuth, Katherine (2007): The role of frequency in language acquisition. In: Gülzow, Insa / Gagarina, Natalia (eds.): Frequency effects in language acquisition. (= Studies on Language Acquisition (SOLA) series). Berlin: de Gruyter, 383-388.

Diessel, Holger (2007): Frequency effects in language acquisition, language use, and diachronic change. In: New Ideas in Psychology 25: 108-127.

Ellis, Nick C. (2008): The dynamics of second language emergence: Cycles of language use, language change, and language acquisition. In: Modern Language Journal 92: 232-239.

Faroqi-Shah, Yasmeen / Thompson, Cynthia K. (2004): Semantic, lexical, and phonological influences on the production of verb inflections in agrammatic aphasia. In: Brain and Language 89: 484-498.

Frauenfelder, Uli H. / Schreuder, Rob (1992): Constraining psycholinguistic models of morphological processing and representation: the role of productivity. In: Booij, Geert E. / Van Marle, Jaap (eds.): Yearbook of morphology 1991. Dordrecht: Kluwer Academic, 165-183.

Jacobson, Roman (1984): Russian and Slavic grammar: Studies. Edit. by Linda R. Waugh and Morris Halle. (= Janua linguarum. Series maior 106). Berlin / New York: de Gruyter.

Janssen, Ulrike / Penke, Martina (2002): How are inflectional affixes organized in the mental lexicon? Evidence from the investigation of agreement errors in agrammatic aphasics. In: Brain and Language 81: 180-191.

Jurafsky, Dan (2003): Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In: Bod, Rens / Hay, Jennifer / Jannedy, Stefanie (eds.): Probabilistic Linguistics. Cambridge, MA: MIT Press, 39-95.

Kostić, Aleksandar / Mirković, Jelena (2002): Processing of inflected nouns and levels of cognitive sensitivity. In: Psihologija 35: 287-297.

Laine, Matti / Vainio, Seppo / Hyönä, Jukku (1999): Lexical access routes to nouns in a morphologically rich language. In: Journal of Memory and Language 40: 109-135.

Lehtonen, Minna / Laine, Matti (2003): How word frequency affects morphological processing in monolinguals and bilinguals. In: Bilingualism: Language and Cognition 6: 213-225.

Lieven, Elena / Tomasello, Michael (2008): Children's first language learning from a usage-based perspective. In: Robinson, Peter / Ellis, Nick C. (eds.): Handbook of Cognitive Linguistics and Second Language Acquisition. New York: Routledge, 168-196.

Marcus, Gary F. / Brinkmann, Ursula / Clahsen, Harald / Wiese, Richard / Woest, Andreas / Pinker, Steven (1995): German inflection: The exception that proves the rule. In: Cognitive Psychology 29: 189-256.

Martynenko N.G. (2003): Suschestvitel'noe: Kategoria padezhda. In: Sirotinina (ed.), 47-64.

Milin, Petar / Filipović Đurđević, Dusica / Moscoso del Prado Martín, Fermin (2007): The psychological reality of inflectional paradigms. Internet: http://cogprints.org/6188 (last visited: 01 / 2010).

New, Boris / Brysbaert, Marc / Segui, Juan / Ferrand, Ludovic / Rastle, Kathleen (2004): The processing of singular and plural nouns in French and English. In: Journal of Memory and Language 51: 568-585.

Pinker, Steven (1991): Rules of language. In: Science 253: 530-535.

Pinker, Steven (1997): Words and rules in the human brain. In: Nature 387: 547-548.

Pinker, Steven / Prince, Alan (1988): On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. In: Cognition 28: 73-193.

Pinker, Steven / Prince, Alan (1994): Regular and irregular morphology and the psychological status of rules of grammar. In: Lima, Susan D. / Corrigan, Roberta L. / Iverson, Gregory K. (eds.): The reality of linguistic rules. Amsterdam: Benjamins, 321-351.

Prasada, Sandeep / Pinker, Steven (1993): Generalizations of regular and irregular morphology. In: Language and Cognitive Processes 8: 1-56.

Russian National Corpus. Internet: http://ruscorpora.ru (last visited: 10 / 2010).

Sandzhi-Garjaeva, Z.S. (2003): Glagol: Kategorija vremeni. In: Sirotinina (ed.), 110-124.

Schiller, Niels O. / Caramazza, Alfonso (2002): The selection of grammatical features in word production: The case of plural nouns in German. In: Brain and Language 81: 342–357.

Schiller, Niels O. / Caramazza, Alfonso (2003): Grammatical feature selection in noun phrase production: Evidence from German and Dutch. In: Journal of Memory and Language 48: 169-194.

Sirotinina Olga B. (ed.) (2003): Razgovornaja retsch' v cisteme funkzional'nych stilej sovremennogo russkogo literaturnogo jazyka. Grammatika. Moskva: URSS.

Schreuder, Robert / Baayen, R. Harald (1995): Modeling morphological processing. In: Feldman, Laurie Beth (ed.): Morphological aspects of language processing. Hillsdale, NJ: Erlbaum, 131-154.

Schreuder, Robert / Baayen, R. Harald (1997): How complex simplex words can be. In: Journal of Memory and Language 37: 118-139.

Schriefers, Herbert (1993): Syntactic processes in the production of noun phrases. In: Journal of Experimental Psychology: Learning, Memory, and Cognition 19: 841-850.

Sereno, Joan A. / Jongman, Allard (1997): Processing of English inflectional morphology. In: Memory & Cognition 25: 425-437. Internet: http://www2.ku.edu/~kuppl/abstracts/poeim.html (last visited: 10 / 2010).

Shvedova Natalia Ju. et al. (1982): Russkaja Grammatika. T. 1. Moskva: Nauka.

Smolka, Eva / Zwitserlood, Pienie / Rösler, Frank (2007): Stem access in regular and irregular inflection: Evidence from German participles. In: Journal of Memory and Language 57: 325-347.

Stemberger, Joseph Paul (1984): Structural errors in normal and agrammatic speech. In: Cognitive Neuropsychology 1, 4: 281-313.

Stemberger, Joseph Paul (1985): Bound morpheme loss errors in normal and agrammatic speech: One mechanism or two? In: Brain and Language 50: 225-239.

Stemberger, Joseph Paul / McWhinney, Brian (1986): Form-oriented inflectional errors in language processing. In: Cognitive Psychology 18: 329-54.

Tabak, Wieke / Schreuder, Robert / Baayen, R. Harald (2005): Lexical statistics and lexical processing: semantic density, information complexity, sex, and irregularity in Dutch. In: Reis, Marga / Kepser, Stephan (eds.): Linguistic evidence. Berlin: de Gruyter, 529-555.

Tomasello, Michael (2003): Constructing a language: A usage-based account of language acquisition. Cambridge, MA: Harvard University Press.

Ullman, Michael T. (1999): Acceptability ratings of regular and irregular past-tense forms: Evidence for a dual-system model of language from word frequency and phonological neighborhood effects. In: Language and Cognitive Processes 14: 47-67.

FRANCESCA STRIK LIEVERS

Constructing Judgments

The Interaction between Adjectives and Clausal Complements in Italian[1]

Abstract

This paper puts forward a corpus-based analysis of two constructions in Italian: "V + ADJ + infinitive clause" and "V + ADJ + *che* ('that') clause". Some verbs, such as *essere* ('be') and *sembrare* ('seem') occur in both constructions; in addition, many adjectives can appear in the predicative complement position of both constructions. The *che*-clause and the infinitive construction are apparently semantically similar, both expressing a judgment or an attitude (codified by the adjective) toward the content of the subject clause. However, upon examining corpus data with a distinctive-collexeme analysis (Gries / Stefanowitsch 2004), subtle differences clearly emerge: the *che*-clause construction tends to be associated with epistemic / evidential adjectives while the infinitive construction shows a stronger association with evaluative adjectives. As a result of the analysis, it can be stated that constructions appear to be characterized by a meaning (see Goldberg 2006), which can be influenced by lexical choices (adjectives, verbs, clitics).

1. Introduction

This paper examines two constructions which can appear with a limited number of verbs in Italian. The constructions under analysis are the following:

- V + ADJ + infinitive clause
- V + ADJ + *che* ('that') clause,[2]

where both the infinitive and the *che*-clause are subject complement clauses. The verbs which can appear in the V position (in the third person singular) are: *sembrare*, *apparire*, *parere* (all of them roughly meaning 'seem'), *essere* ('be'), *diventare* and *divenire* ('become'), *risultare* ('prove', 'seem').

[1] This work was originally conceived and presented at GaC 2009 in collaboration with Irene Russo and then further developed by myself. I am deeply indebted to her, especially for the execution of the distinctive-collexeme analysis. I am extremely grateful to Caterina Guardamagna, for kindly revising the manuscript. All errors and shortcomings are, of course, my sole responsibility.

[2] The *che*-clause is a finite clause with the verb (in most cases) in the subjunctive mood (cf. Wandruszka 1991).

Examples[3] of the two constructions are given in (1) and (2):

(1) Infinitive clause:

Sembra improbabile convincere tutti gli italiani d'un colpo ad adeguarsi.

'It seems unlikely to persuade all of the Italians to adapt suddenly.'

(2) *Che*-clause:

Sembra improbabile che i Cobas possano decidere di sospendere lo sciopero.

'It seems unlikely that the Cobas may decide to call off the strike.'

It has to be noticed that with verbs of 'seeming' (i.e., *sembrare*, *apparire*, *parere*, *risultare*) both constructions can also occur with an overtly expressed Experiencer. In (3) and (4) the Experiencer is expressed by a clitic pronoun:

(3) Infinitive clause:

Mi sembra giusto dire la verità.

'It seems right to me to say the truth.'

(4) *Che*-clause:

Non mi pare giusto che paghi l'innocente per il peccatore.

'It doesn't seem right to me that the innocent pays for the sinner.'

Reading the examples in (1) - (4), both with and without the Experiencer, it seems that the *che*-clause and the infinitive are semantically similar. This intuition is supported by the observation that the adjectives occurring in the predicative complement position are also similar in the two constructions, expressing some kind of attitude toward the content of the clause (see Section 2 for a semantic description of these adjectives). In addition, a quick scan through the list of adjectives seems to suggest that those adjectives which can occur in one of the two constructions can also occur in the other.[4] Finally, in a number of cases the two constructions can be (at least in part) mutually paraphrasable:

[3] In this paper all the examples in Italian are extracted from the *La Repubblica* corpus (see Section 4), except for example (5), which is mine. The English translations try to follow the structure of the Italian examples, therefore they may sometimes not sound completely natural.

[4] As will be shown later in the paper, this not exactly true, though it is true for the majority of the adjectives.

(5a) *E' importante dare il buon esempio.*

(b) *E' importante che si dia il buon esempio.*

'It is important to set a good example.'

The two constructions are thus similar, but are they completely equivalent? To answer this question, a corpus-based analysis was carried out with the aim of investigating possible differences between them. As will be demonstrated, an analysis which goes beyond raw frequency data shows that the infinitive and the *che*-clause construction clearly tend to be associated with different types of adjectives, and thus with different meanings. Thanks to a distinctive-collexeme analysis (Gries / Stefanowitsch 2004; Section 4), it has been possible to identify statistically significant tendencies in the association of adjectives and constructions.

2. Adjectives in predicative complement position

Before examining the data, a preliminary step needs to be taken. A more detailed description of the adjectives found in the predicative complement position of both the infinitive and the *che*-clause construction is considered to be crucial for detecting the meaning(s) of the constructions themselves (see Stefanowitsch / Gries 2008: 131).

As already noticed, in both constructions the adjectives describe an attitude toward the content of the subject clause. We can talk, as in Biber et al. (1999), of 'stance', intended as the expression of "personal feelings, attitudes, value judgments, or assessments". In order to describe the adjectives in a more detailed way, the semantic classification adopted here is the one proposed for English adjectives in Dixon (2005: 84-85). Out of the eleven classes identified by Dixon (DIMENSION, PHYSICAL PROPERTY, SPEED, AGE, COLOUR, VALUE, DIFFICULTY, VOLITION, QUALIFICATION, HUMAN PROPERTY, SIMILARITY), only three include the adjectives which may appear in the constructions analyzed here. The relevant classes are: VALUE, DIFFICULTY and QUALIFICATION, whose members, in Biber's terms, are all adjectives expressing stance relative to the proposition or state of affairs to which they refer. Some examples of Italian adjectives are given for each class in the following table:

Semantic Type	Adjectives
QUALIFICATION	*probabile* ('probable'), *vero* ('true'), *ovvio* ('obvious'), *possibile* ('possible'), *impossibile* ('impossible'), *certo* ('certain'), ...
VALUE	*bello* ('nice'), *strano* ('strange'), *curioso* ('curious'), *necessario* ('necessary'), *cruciale* ('crucial'), *importante* ('important'), ...
DIFFICULTY	*facile* ('easy'), *difficile* ('difficult'), *semplice* ('simple'), ...

Table 1: Semantic types of ADJs (following Dixon 2005)

Within the group of QUALIFICATION adjectives it is possible to identify at least two main semantic subclasses which are particularly relevant for the present analysis: epistemic and evidential adjectives.

Epistemic modality can be broadly defined as a speaker's evaluation of the likelihood of a state of affairs (Nuyts 2001: XV), and it can be expressed in language by grammatical or lexical means.[5] Some of the QUALIFICATION adjectives, such as the highly frequent *possibile* ('possible'), *impossibile* ('impossible'), *probabile* ('likely'), may be considered as a lexical expression of epistemic modality. These adjectives may be situated on a scale ('epistemic scale') ranging from certainty that a state of affairs applies to certainty that it does not apply (Nuyts 2001: 22).

(6) *Sembra probabile che un comparto dell'acciaieria rimanga in attività.*

'It seems likely that one section of the steelworks remains active.'

(7) *A prima vista sembra possibile trovare un accordo.*

'At first sight it seems to be possible to reach agreement.'

Evidentiality can be defined as the coding of the source of information (see e.g., Anderson 1986: 274, De Haan 1999: 83). Although it is generally considered a grammatical category[6] (cf. Aikhenvald 2004) in typological studies, here

[5] For Italian it is possible to mention some verbal lexemes (such as the modals *dovere* 'must' and *potere* 'may', and verbs such as *pensare* 'think' and *ritenere* 'consider'), the use of verbal tenses and moods with a modal value (such as the epistemic future), and lexical means of many kinds (such as the adverbials *forse* 'maybe', *verosimilmente* 'probably'). See Pietrandrea (2005) for an overview.

[6] In this approach, only languages where it is obligatory to mark grammatically where the information comes from or how it is acquired ("leaving this out results in a grammatically awkward 'incomplete' sentence", Aikhenvald 2004: 6) are considered to show the (grammatical) category of evidentiality.

evidentiality is considered a conceptual category, which, as epistemic modality, can also be expressed by lexical means such as, for example, a subset of QUALIFICATION adjectives. Evidential adjectives may be described as lexemes which "convey certainty based on some kind of evidence which is there for everyone to see" (Simon-Vandenbergen / Aijmer 2007: 147).[7] This class includes lexemes such as *chiaro* ('clear'), *evidente* ('evident'):

(8) *Io preferisco questa canzone e dunque mi sembra evidente che debba vincere la gara.*

'I prefer this song, therefore it seems evident to me that it has to win the competition.'

Considering both epistemicity and evidentiality as basic functional categories, the two may be conceptually differentiated quite easily. It has, however, to be observed, as in Squartini (2004: 873f.), that

> despite the fact that a distinction between marking the evidential source of the information and signaling the epistemic degree of certainty seems to be conceptually self-evident and also empirically solid [...], in most cases interpreting empirical data turns out to be not straightforward.

This is why the distinction between epistemic and evidential adjectives will only be used here when it is sufficiently clear due to a different behavior in the constructions analyzed (see Section 4). Otherwise reference will be made to the more general class of QUALIFICATION adjectives.

Turning now to adjectives expressing VALUE, or 'evaluative' adjectives. They form a broad class which includes adjectives expressing a value judgment, such as *bello* ('nice'), *necessario* ('necessary'):

(9) *Mi sembra bello che certi film affrontino la spiritualità e la fede religiosa.*

'It seems nice to me that some movies deal with spirituality and religious faith.'

(10) *Non mi sembrava bello entrare nel Palacio de los matrimonios con le gambe nude.*

'It did not seem nice to me to enter the Palacio de los matrimonios with naked legs.'

[7] The authors are actually referring to evidential adverbs, but this description may well also apply to adjectives.

It has been observed that in English, both with *that* and with *to* complement clauses, evaluative adjectives suggesting necessity or importance (*important*, *essential*, *critical*, *crucial*, etc.) can acquire a deontic sense, indicating a course of action to be followed rather than simply evaluating a state of affairs (cf. Biber et al. 1999, Van Linden / Davidse 2009). This also applies to Italian (cf. Wandruszka 1991: 476), as shown in the following example:

(11) *Il paziente ha bisogno di cure ed è preferibile che sia tenuto sotto osservazione.*

'The patient needs care, and it is preferable that he is kept under observation.'

The last class, DIFFICULTY adjectives, is quite small, the most frequent lexemes being *facile* ('easy') and *difficile* ('difficult'). It is, however, useful to individuate DIFFICULTY adjectives as an autonomous class since, as will be illustrated in Section 4, they show a polysemy pattern that is particularly relevant for the present analysis.

3. The role of constructions (Goldberg 2006)

As already mentioned, both grammatical and lexical means of expression are considered in literature regarding modality and evidentiality. Here it is claimed that also constructions (in interaction with specific lexical items) can play a significant role in the codification of epistemicity / evidentiality, and more generally, of stance.

Syntactic alternations and lexical meanings interact and Construction Grammar provides a fruitful framework for a combined treatment of lexical data and constructions. Here the notion of construction devised by Goldberg's approach is adopted (Goldberg 2006) which allows the investigation of the distributional patterns of lexical items with respect to syntactic forms. Constructions are defined as follows:

> Any linguistic pattern is recognized as a construction as long as some aspect of its form or function is not strictly predictable from its component parts or from other constructions recognized to exist. In addition, patterns are stored as constructions even if they are fully predictable as long as they occur with sufficient frequency. (Goldberg 2006: 5).

This definition highlights the role of frequency in the identification of constructions. Corpus data analyzed with statistical methods may thus offer useful tools for analyzing the interaction between syntactic patterns and lexical items.

4. Analysis of the data

The analysis is based on the data extracted from the *La Repubblica* corpus (Baroni et al. 2004) which consists of 380 million tokens of contemporary Italian newspaper texts. The verbs listed in Section 1 appear in the two constructions with the frequencies reported in Table 2.

	Number of occurrences		
V	V + ADJ + *che*	V + ADJ + infinitive	total
essere	90 138	70 486	160 624
sembrare	2 491	1 850	4 341
parere	831	588	1 419
apparire	1 134	255	1 389
diventare	185	709	894
risultare	255	293	548
divenire	52	52	104

Table 2: Frequency of the two constructions in *La Repubblica*

In the following the quantitative analysis will be limited to *sembrare* and *essere* since they are by far the most frequent. Moreover, they are a good representation of the contrast between a semantically "neutral" verb, i.e., *essere*, and a verb with modal (evidential / epistemic) meaning, i.e., *sembrare*.

4.1 Raw frequency counts are not sufficient

As a first step, the frequency of the adjectives appearing in the predicative complement position of the *che*-clause and of the infinitive construction is considered. The goal of such analysis is to point out whether two constructions prefer different types of adjectives.

However, looking at raw frequencies proves not to be sufficient, as shown in Tables 3 and 4 which report the 10 most frequent adjectives for each construction. In these tables (and in the following ones) QUALIFICATION adjectives are in bold, VALUE adjectives are underlined, and DIFFICULTY adjectives are in italics.[8]

[8] The grouping of Italian adjectives into semantic classes has been accomplished manually. Therefore, although the classification tries to follow existing ones in the literature, it still remains subjective and, of course, contestable. The problem of this type of qualitative interpretation of quantitative results is recognized in Gries / Stefanowitsch (2010), where a possible solution is individuated in grouping semantically similar lexemes through a cluster analysis. An attempt to apply this method to the case analyzed here is left to future work.

essere + *che*			*essere* + infinitive	
Adjective	Frequency		Adjective	Frequency
vero	23 946		**possibile**	14 330
chiaro	8 659		*difficile*	12 142
possibile	5 657		necessario	7 709
probabile	5 072		*facile*	6 770
evidente	4 175		**impossibile**	3 005
giusto	3 154		giusto	2 984
necessario	1 826		inutile	2 315
ovvio	1 651		importante	1 866
sicuro	1 636		opportuno	1 519
difficile	1 403		sufficiente	1 245

Table 3: *essere* + *che* vs. *essere* + infinitive: raw frequencies of ADJs

sembrare + *che*			*sembrare* + infinitive	
Adjective	Frequency		Adjective	Frequency
difficile	211		giusto	313
strano	194		*difficile*	214
chiaro	194		opportuno	111
impossibile	192		**impossibile**	110
giusto	148		**possibile**	61
evidente	143		*facile*	57
incredibile	105		doveroso	54
improbabile	104		assurdo	45
probabile	76		necessario	42
possibile	62		utile	41

Table 4: *sembrare* + *che* vs. *sembrare* + infinitive: raw frequencies of ADJs

No clear differences emerge between the *che*-clause and the infinitive construction as regards the most frequent adjectives: the semantic types of adjecti-

ves appear to be almost equally distributed between the two constructions. So, even if frequency could seem highly plausible for the characterization of constructions, these results suggest that the analysis should be more complete and statistically refined.

For example, the most frequent adjectives occurring in just one type of construction can be considered for each verb. In this way functional tendencies begin to emerge, as the following tables show:

essere + *che*			*essere* + infinitive	
Adjective	Frequency		Adjective	Frequency
evidente	4 178		indispensabile	964
sicuro	1 638		prematuro	313
certo	1 181		esagerato	161
probabile	77		agevole	77

Table 5: *essere* + *che* vs. *essere* + infinitive: most frequent ADJs occurring in just one construction

sembrare + *che*			*sembrare* + infinitive	
Adjective	Frequency		Adjective	Frequency
chiaro	140		*facile*	52
evidente	82		**lecito**	17
probabile	68		azzardato	9
scontato	38		eccessivo	8

Table 6: *sembrare* + *che* vs. *sembrare* + infinitive: most frequent ADJs occurring in just one construction

Two main tendencies emerge from these tables. First, the *che*-clause construction shows a strong preference for QUALIFICATION adjectives, both evidential (e.g., *chiaro* 'clear') and epistemic (e.g., *probabile*, 'probable'). Secondly, the infinitive construction tends to be associated mainly with VALUE adjectives (e.g., *indispensabile* 'indispensable', *eccessivo* 'excessive'). These tendencies regard both *sembrare* and *essere*, indicating that they depend on the construction itself and not on the semantics of the verb of the main clause.

It also has to be noticed that no evidential adjective appears with the infinitive construction. The adjectives reported in the tables are, of course, just the most frequent ones. Evidential adjectives are however also absent from the complete frequency list of the adjectives occurring in the infinitive construction. Moreover, trying to create phrases composed by an evidential adjective followed by an infinitive does not seem to give acceptable results:

(12) **Sembra evidente vincere la gara.*
'It seems evident to win the competition.'

As mentioned in Section 1, 'seem' verbs may occur with or without an Experiencer, in both constructions. Is there a difference between the presence vs. absence of an Experiencer as regards the adjectives occurring in the predicative complement position? The following tables compare the most frequent adjectives occurring just with and just without an Experiencer (expressed by a clitic pronoun, CLI, as in examples (3) and (4)).

sembrare + infinitive			CLI + *sembrare* + infinitive	
Adjective	Frequency		Adjective	Frequency
conveniente	4		serio	8
illogico	3		carino	8
improbabile	2		ingiusto	8
irrealistico	2		curioso	4

Table 7: ***sembrare* + infinitive vs. CLI + *sembrare* + infinitive: most frequent ADJs occurring in just one construction**

sembrare + *che*			CLI + *sembrare* + *che*	
Adjective	Frequency		Adjective	Frequency
indicativo	4		bello	11
sicuro	3		scandaloso	8
indubitabile	3		eccessivo	5
stupefacente	2		significativo	5

Table 8: ***sembrare* + *che* vs. CLI + *sembrare* + che: most frequent ADJs occurring in just one construction**

Clearly, any observation can be made on the basis of these data since the frequency of single adjectives is not significant enough.

4.2 Distinctive-collexeme analysis

Subtle differences could however potentially emerge by examining the data through the distinctive-collexeme analysis elaborated by Stefan Gries and Anatol Stefanowitsch (Gries / Stefanowitsch 2004). The difference between the constructions under analysis becomes even clearer through this method which, as the authors claim, "is specifically geared to investigating pairs of semantically similar grammatical constructions and the lexemes that occur in them" (ibid.: 97). The distinctive-collexeme analysis makes it possible to identify the lexemes that distinguish between semantically or functionally near-equivalent constructions, such as the *che*-clause and the infinitive clause construction.

Lexemes that exhibit a strong preference for one construction over the other can be identified on the basis of a statistical measure which determines the association strength between words in contexts. In order to calculate the distinctiveness of an adjective, the following data are needed:

a) its frequency in construction A;
b) its frequency in construction B;
c) the frequencies of constructions A and B with adjectives other than the one considered (cf. Gries / Stefanowitsch 2004: 102).

Thanks to a distinctive-collexeme analysis, it will be possible not only to compare the two constructions under analysis, but also to evaluate whether functional differences are associated with: a) the presence vs. absence of an Experiencer for 'seem' verbs, and b) the verbal lexeme occurring in the main clause ('seem' verbs vs. *essere*).

What emerges clearly through the distinctive-collexeme analysis is that the two constructions tend to prefer different types of adjectives. Tables 9 and 10 list the most distinctive adjectives for each construction. The numbers reported in the right column represent the so-called collostructional strength, a statistical measure indicating how strongly the lexical item is attracted by the construction.

essere + *che*			*essere* + infinitive	
Adjective	Coll. strength		Adjective	Coll. strength
vero	30931.0600		*difficile*	13702.3322
chiaro	10357.0016		*facile*	9727.9695
probabile	5908.2225		**possibile**	7239.3295
evidente	4906.4222		necessario	5841.7658
escluso	3274.8445		**impossibile**	3310.7335
sicuro	1902.6949		inutile	2021.5703
ovvio	1554.7188		indispensabile	1597.771
certo	1369.1834		interessante	1280.2905
indubbio	811.0485		sufficiente	937.8447
improbabile	756.0833		opportuno	673.2636

Table 9: *essere* + *che* vs. *essere* + infinitive: distinctive-collexeme analysis

sembrare + *che*			*sembrare* + infinitive	
Adjective	Coll. strength		Adjective	Coll. strength
chiaro	143.642		*facile*	101.0769
evidente	82.4319		giusto	63.1941
improbabile	68.1455		opportuno	59.3545
probabile	68.0292		**lecito**	32.5405
incredibile	46.8211		ragionevole	23.7036
strano	40.9639		necessario	18.516
scontato	37.6304		**possibile**	17.3044
escluso	23.6551		azzardato	17.1679
credibile	12.7663		eccessivo	15.2538
verosimile	12.7663		esagerato	15.2538

Table 10: *sembrare* + *che* vs. *sembrare* + infinitive: distinctive-collexeme analysis

Again, it turns out even more clearly that the *che*-clause construction tends to prefer QUALIFICATION adjectives, thus exhibiting an epistemic/evidential function while the infinitive construction is mainly associated with evaluative meanings.

What about the comparison between constructions with and without clitics? The distinctive-collexeme analysis reveals that without clitics, QUALIFICATION adjectives are preferred while with clitics, there is a stronger association with VALUE adjectives, as shown in the following tables concerning *sembrare*.

sembrare + *che*			CLI + *sembrare* + *che*	
Adjective	Coll. strength		Adjective	Coll. strength
probabile	51.8354		giusto	145.7533
improbabile	33.8951		importante	18.8778
chiaro	32.5095		bello	16.674
difficile	29.9536		normale	15.1917
scontato	28.3134		**logico**	14.5944
escluso	23.7889		opportuno	12.7459
indicativo	5.0844		scandaloso	12.1145
certo	4.3822		assurdo	9.4139
indubitabile	3.8121		positivo	8.3936
sicuro	3.8121		eccessivo	7.564

Table 11: CLI vs. no CLI in *che*-clause construction: distinctive-collexeme analysis

sembrare + infinitive			CLI + *sembrare* + infinitive	
Adjective	Coll. strength		Adjective	Coll. strength
facile	66.269		giusto	80.5895
impossibile	51.5458		doveroso	31.9749
difficile	31.7562		utile	13.6686
possibile	31.0129		corretto	12.6768
lecito	10.4389		importante	11.4452
conveniente	7.1655		carino	8.4382
logico	6.1547		ingiusto	8.4382
illogico	5.3717		serio	8.4382
ragionevole	4.537		normale	7.9117
azzardato	4.2835		inutile	7.4754

Table 12: CLI vs. no CLI in infinitive construction: distinctive-collexeme analysis

An attempt to interpret these tendencies can be made on the basis of the distinction proposed by Nuyts (2001) between subjectivity and intersubjectivity. A judgment is subjective when "the speaker suggests that (s)he alone knows the evidence and draws a conclusion from it", it is intersubjective when "(s)he indicate[s] that the evidence is known to (or accessible by) a larger group of people who share the conclusion based on it" (ibid.: 34). The contrast is between personal and shared responsibility for the statement that the speaker makes. Impersonal constructions of the type *it is probable that* suggest intersubjectivity (ibid.: 66), while it is plausible to claim that the presence of a clitic produces a shift toward subjectivity, delimiting the responsibility for the statement to the Experiencer, codified by the clitic.

Epistemic / evidential adjectives typically consist in judgments given on the basis of (shared) evidence: evidentiality describes the type of evidence while epistemic judgment states the likelihood of a state of affairs on the basis of some kind of evidence. It is therefore reasonable that these adjectives tend to be preferred in an intersubjective construction, such as that without clitics. In contrast, evaluative adjectives describe judgments not necessarily based on shared evidence, therefore they fit well in a subjective construction, such as that with clitics.

One more parameter should be taken into consideration, that is, the verb occurring in the main clause. Does the presence of *essere* vs. 'seem' verbs entail a difference as concerns the adjectives that are more attracted by the construction? The following tables compare *essere* and *sembrare* both for the *che*-clause and the infinitive construction.

essere			*sembrare*	
Adjective	Coll. strength		Adjective	Coll. strength
vero	637.8466		**impossibile**	472.2369
possibile	47.5238		*difficile*	334.6671
giusto	34.9586		strano	329.6906
necessario	33.3499		incredibile	242.8814
sicuro	28.548		**improbabile**	188.467
importante	22.0985		**scontato**	100.788
pensabile	16.9435		paradossale	524.312
significativo	11.7742		**credibile**	301.527
bello	8.778		**inverosimile**	262.791
presumibile	8.778		assurdo	214.302

Table 13: *essere* vs. *sembrare* in *che*-clause construction: distinctive-collexeme analysis

essere			*sembrare*	
Adjective	Coll. strength		Adjective	Coll. strength
possibile	121.9308		**lecito**	126.0648
necessario	83.3546		azzardato	65.9433
sufficiente	19.6362		**logico**	61.8234
inutile	15.8417		**impossibile**	54.1615
bello	14.4682		ragionevole	49.3332
importante	13.767		strano	39.9532
interessante	12.5994		opportuno	34.5279
solito	8.7206		corretto	31.3941
facile	7.1054		**vero**	26.8799
obbligatorio	6.0155		paradossale	25.4993

Table 14: *essere* vs. *sembrare* in infinitive construction: distinctive-collexeme analysis

While for the *che*-clause construction no significant differences emerge, a slightly stronger preference for VALUE adjectives is recorded for *essere* in the infinitive construction. How can this difference be explained? In Table 10 it was shown that the infinitive construction tends to be associated mainly with VALUE adjectives also with *sembrare*. However, from Table 14 it emerges that "*sembrare* + infinitive" shows a stronger association with QUALIFICATION adjectives if compared to "*essere* + infinitive". It can thus be hypothesized that, although the infinitive construction does not have an epistemic/evidential meaning itself, if the lexeme filling the verb slot has an epistemic/evidential meaning, this lexeme might slightly influence the choice of adjectives in that direction. This seems to be the case for *sembrare*, in contrast with the epistemically 'neutral' *essere*.

To conclude the presentation of the data, it is useful to mention the class of DIFFICULTY adjectives. These adjectives exhibit an interesting behavior, as the following examples show:

(13) *Sembra facile fabbricare un mazzo di buone carte da gioco.*
'It seems easy to produce a good pack of cards.'

(14) *Mi sembra facile che i repubblicani possano abbandonare gli alleati.*
'It does seem easy to me that the republicans may abandon their allies.'

The same adjective, *facile*, is interpreted differently in the two phrases. In (13), an infinitive construction, *facile* becomes an evaluative interpretation, giving a value judgment about the action described in the complement clause. In (14), a *che*-clause construction, *facile* becomes an epistemic interpretation, having a meaning similar to *probabile*. This seems to happen regularly with all DIFFICULTY adjectives: they are interpreted as VALUE adjectives in the infinitive construction, and as QUALIFICATION (epistemic) adjectives in the *che*-clause construction. This is particularly significant for the present analysis since it is another confirmation of the fact that the infinitive construction shows a strong association with VALUE adjectives, and the *che*-clause with QUALIFICATION adjectives: each construction selects the most suitable sense of the polysemous adjective, i.e., the one which is consistent with the meaning of the construction.

5. Conclusions

The *che*-clause and the infinitive constructions are, at first sight, semantically similar. The analysis carried out in this paper, however, showed that they are not equivalent: each one is characterized by its own meaning, and can therefore be considered a construction in its own right.

The distinctive-collexeme analysis proved to be very useful for detecting the most distinctive connotations of the constructions since it allows the comparison of the relative importance of each significant element. To summarize, no matter which verb is considered, the *che*-clause construction shows a stronger association with QUALIFICATION adjectives (e.g., *probabile*, *chiaro*) while the infinitive construction shows a stronger association with VALUE adjectives (e.g., *strano*, *importante*). The two constructions also differ with regard to the semantic restrictions they place on the adjectives that can occur in them: evidential adjectives (e.g. *chiaro*, *evidente*) are common in the *che*-clause construction while they do not normally occur in the infinitive construction. The meaning of the construction appears to depend both on the semantics of the adjectives and on the complementation patterns. Slight shifts in meaning can depend also on other elements, such as the semantics of the verb (*essere* vs. *sembrare*) and the presence of an Experiencer.

Finally, support for the claim that constructions themselves have meaning comes from polysemous adjectives of the DIFFICULTY class: they occur in both constructions, and it is the construction that selects the relevant meaning.

References

Aikhenvald, Alexandra Y. (2004): Evidentiality. Oxford: Oxford University Press.

Anderson, Lloyd B. (1986): Evidentials, paths of change, and mental maps: Typologically regular asymmetries. In: Chafe, Wallace / Johanna Nichols (eds.) (1986): Evidentiality: The linguistic coding of epistemology. Norwood, NJ: Ablex, 273-312.

Baroni, Marco et al. (2004): Introducing the 'La Repubblica' corpus. A large, annotated, TEI(XML)-compliant corpus of newspaper Italian. In: Proceedings of LREC 2004. Internet: sslmitdev-online.sslmit.unibo.it/corpora/downloads/rep_lrec_2004.pdf (last visited: 10 / 2010).

Biber, Douglas et al. (1999): The Longman grammar of spoken and written English. London: Longman.

De Haan, Ferdinand (1999): Evidentiality and epistemic modality: Setting boundaries. In: Southwest Journal of Linguistics 18: 83-10.

Dixon, Robert M. W. (2005): A semantic approach to English grammar. Oxford: Oxford University Press.

Goldberg, Adele (2006): Constructions at work. Oxford: Oxford University Press.

Gries, Stefan Th. (2007): Coll.analysis 3.2. A program for R for Windows 2.x.

Gries, Stefan Th. / Stefanowitsch, Anatol (2010): Cluster analysis and the identification of collexeme classes. In: Newman, John / Rice, Sally (eds.): Empirical and experimental methods in cognitive / functional research. Stanford, CA: CSLI, 59-72.

Gries, Stefan Th. / Stefanowitsch, Anatol (2004): Extending collostructional analysis: A corpus-based perspective on alternations. In: International Journal of Corpus Linguistics 9: 97-129.

La Repubblica Corpus: Internet: http://sslmit.unibo.it/repubblica (last visited: 10 / 2010).

Nuyts, Jan (2001): Epistemic modality, language and conceptualization. A cognitive-pragmatic perspective. Amsterdam / Philadelphia: Benjamins.

Pietrandrea, Paola (2005): Epistemic modality. Functional properties and the Italian System. Amsterdam / Philadelphia: Benjamins.

Simon-Vandenbergen, Anne-Marie / Aijmer, Karin (2007): The semantic field of modal certainty. Berlin / New York: de Gruyter.

Squartini, Mario (2004): Disentangling evidentiality and epistemic modality in Romance. In: Lingua 114: 873-895.

Stefanowitsch / Gries (2008): Channel and constructional meaning: A collostructional case study. In: Kristiansen, Gitte / Dirven, René (eds.) (2008): Cognitive sociolinguistics. Language variation, cultural models, social systems. Berlin / New York: de Gruyter, 129-152.

Van Linden, An / Davidse, Kristin (2009): The clausal complementation of deontic-evaluative adjectives in extraposition constructions: A synchronic-diachronic approach. In: Folia Linguistica 43, 1: 171–211.

Wandruszka, Ulrich (1991): Frasi subordinate al congiuntivo. In: Renzi, Lorenzo et al. (eds.) (1991): Grande grammatica italiana di consultazione. Vol. II. Bologna: Il Mulino, 415-481.

LISA BRUNETTI / STEFAN BOTT / JOAN COSTA / ENRIC VALLDUVÍ

A multilingual annotated corpus for the study of Information Structure[1]

Abstract

This paper presents a corpus of spoken narrative texts in Catalan, Italian, Spanish, English, and German. The aim of this corpus compilation is to create an empirical resource for a comparative study of Information Structure. 68 speakers were asked to tell a story in an acoustically isolated room by looking at the pictures of three textless books. A total of 222 narrations resulted in about 16 hours of speech. The recordings have been transcribed and an original annotation of non-canonical constructions for the Romance subgroup has been proposed, namely of morphosyntactically / prosodically marked constructions that relate informational categories such as topic, focus, and contrast. Transcriptions and annotations of some selected high quality recordings have been aligned to the acoustic signal stream. The corpus is available in audio and text format.

1. Introduction

In this paper we present a corpus that has been developed within the NOCANDO project, 'Non-canonical constructions in oral discourse: a cross-linguistic perspective' at the University of Pompeu Fabra in Barcelona (Spain). The main interest of the overall project was to study the cross-linguistic variation in the overt marking of Information Structure (from now on, IS) in general, and more specifically in the spoken, narrative register.

Researchers largely agree on the fact that languages use syntactically, morphologically, and / or prosodically marked constructions to represent infor-

[1] We wish to thank Estela Puig Waldmüller for collaborating in the recording, Teresa Suñol for her help with Catalan and Spanish transcriptions, Josep Maria Fontana, Louise McNally, Gemma Boleda, and Alex Alsina for their advice at different stages and on different aspects of the preparation of the corpus. We also thank the participants of the *Corpus Linguistics Conference* in Liverpool (July 20-23, 2009) and the *Corpus and Grammar 3* conference in Mannheim (Sept. 22-24, 2009) for their comments and questions. This research has been partially funded by the Spanish Ministry of Education and Science project OpenMT (TIN2006 15307-C03-02). The NOCANDO project was funded by the *Spanish Secretaria de Estado de Universidades e Investigación* of the *Ministerio de Educación y Ciencia* (n. I+D HUM2004-04463).

mational categories such as 'topic', 'focus', 'contrast', 'background', etc. (cf. Vallduví 1992, Vallduví / Engdahl 1996, Lambrecht 1994, Erteschik-Shir 1997, Steedman 2000, among many others). A large part of the research on IS, in particular within the generative framework, has mostly or exclusively relied on introspective judgements on sentences in isolation (see e.g. the works by Belletti, Rizzi, Zubizarreta, among many others). Nevertheless, explicit marking of IS through non-canonical constructions is much more frequent in spontaneous speech than in written or controlled discourse. In addition, a written text does not represent intonation, which is extremely important for the marking of IS.[2] Furthermore, IS can only be truly understood if sentences are considered within their linguistic context. Sentences in isolation, such as those that are constructed for introspective judgements, are therefore suboptimal to understand the properties and function of informational categories.

A better source of data for the study of IS is therefore constituted by spontaneous speech corpora. Although speech corpora are available in literature, multilingual corpora that provide comparable data across languages are rather limited in number. Furthermore, access to speech corpora, in particular to the recordings, is often very restricted. These considerations led us to compile a corpus of spontaneous spoken narrative texts in five different languages: Catalan, Italian, Spanish, German, and English. The audio recordings are freely accessible for consultation. A taxonomy of non-canonical constructions (from now on, NOCANs) was also established and an annotation of the relevant subset of the taxonomy was added to the Romance subset of the corpus. The annotation is meant to facilitate the search for IS markings in the text. The corpus is available in audio and text format. Some selected recordings were also aligned to the transcription and annotation using the PRAAT software for acoustic analysis (Boersma / Weenink 2009).[3] Our corpus is publicly available under a Creative Commons license which only excludes commercial use. Use for research is free as long as the work is properly cited and all derivates of the corpus are shared under the same conditions. A more detailed description of the corpus and its annotation is given in the following two sections.

---

2 In fact, it is the most important resource for marking IS in certain languages, such as English.

3 PRAAT is available at http://www.praat.org/.

2. The corpus

A total of 68 speakers were asked to tell a story by looking at the pictures of three textless picture story books. The result is 222 narrations of about 2-9 minutes each (a total of about 16 hours of speech). The quantitative information for each language is given in the table below.

	Catalan	Italian	Spanish	German	English
Speakers	19	16	13	9	11
Recording time	4:02:43 h	4:04:32 h	2:35:20 h	2:09:13 h	2:32:20 h
Word count	37555 w	27392 w	25077 w	15944 w	21970 w (estimated)
Segment count	5856 seg	4306 seg	3801 seg	2154 seg	3140 seg (estimated)

Table 1: Quantitative information on each language represented in the NOCANDO corpus

2.1 Speakers

Participants were mostly university students. The Catalan and Spanish speakers were mostly undergraduate or graduate students with an average age of 22 for Catalan (ranging from 18 to 30) and 20 for Spanish (ranging from 17 to 29). The Catalan speakers were from Catalonia (except one speaker from Valencia). The Spanish speakers were also from Catalonia (except one from Castilla y León), but they spoke Spanish as their first language. The Italian and English speakers had recently arrived in Barcelona. The average age of the Italian speakers was 29 (ranging from 20 to 56). They spoke geographically different varieties of Italian. The English speakers' average age was 27, ranging from 20 to 41. They came from the United States and Great Britain. A large number of the German speakers were also short-term residents in Barcelona, but a smaller number exclusively resided in Germany. The German speakers' average age was 34, ranging from 22 to 67. They came from different parts of Germany.

2.2 Methodology

Speakers were asked to narrate three stories to an experimenter while looking at the pictures of three textless books by Mercer Meyer. The books were given to each speaker in random order. Speakers were allowed to browse through the

book before they started the narration. Most speakers were recorded in an acoustically isolated room while some were recorded with a portable recording device in a silent room. In both situations speakers were sitting in front of the experimenter.

The books are entitled *Frog goes to dinner*, *A frog on his own*, and *One frog too many* and are about the adventures of a boy and his pet frog. Mercer Meyer's books have already been used in literature for the study of narration strategies in monolingual or bilingual children, adults, and second language learners (cf. Berman / Slobin (eds.) 1994, Strömqvist / Verhoven (eds.) 2004, and references quoted therein). The book used for those studies is *Frog where are you?* We made a first set of pilot recordings with this book as well as with four other books. The three aforementioned books gave better results in terms of variety of NOCANs to be used by the speakers.[4] The story entitled *Frog goes to dinner* is about the disastrous effects of the presence of the frog in a very elegant restaurant. The advantage of this story is that it includes many different characters interacting with the frog in different ways so we could expect many topic changes. The story entitled *A frog on his own* tells of the adventures of a frog taking a walk by himself in the park. The frog interacts with other characters, but unlike in the dinner story, where the other characters may temporarily have a prominent role, in this story the frog always remains the central character in the narration. The story entitled *One frog too many* tells of the frog's jealousy of a younger frog who has become the boy's new pet. This story was chosen because it presents situations in which the speaker has to distinguish between the two frogs. These situations are interesting because they induce the speaker to adopt constructions that explicitly mark the informational category of *contrast*.

2.3 Transcription and segmentation

The recordings have been transcribed according to the guidelines for the transcription of the LIP corpus (*Lessico di Frequenza dell'Italiano Parlato* 'Frequency lexicon of spoken Italian', De Mauro et al. 1993). We represented the phe-

[4] It must also be noted that *Frog where are you?* has been preferred in past literature because it did not presuppose a specific socio-cultural background so it could be used with speakers of different origins. Since our purpose was not to study the sociolinguistic aspects of narration, we did not consider the socio-cultural implications of a story as a relevant factor for choosing it or not.

nomena that are typical of spoken text: pauses, false starts, truncated words, laughs, hesitations, and vowel lengthening, among others. The transcription also followed orthographic standards.

Since the notion of 'sentence' is often not clear in spoken text, the segmentation was carried out by separating clauses: each segment generally contains one main verb, except for modal, temporal, and aspectual periphrases and verbless clauses. This criterion is very similar to that used in the transcription of CHILDES data (MacWhinney 2000). In order to recognize verb periphrases, we adopted the criteria proposed by Gavarró / Laca (2002). Segments were separated as different XML-marked units; each unit was given a unique id and NOCANs were treated as attributes of segments.

Some selected recordings were aligned to the transcription and annotation using PRAAT, as exemplified in Figure 1. The alignment between text and speech signal makes it possible to quickly identify the relevant segments and locate them in the audio recording. Subsequently, all options for acoustic analysis offered by PRAAT may be exploited.

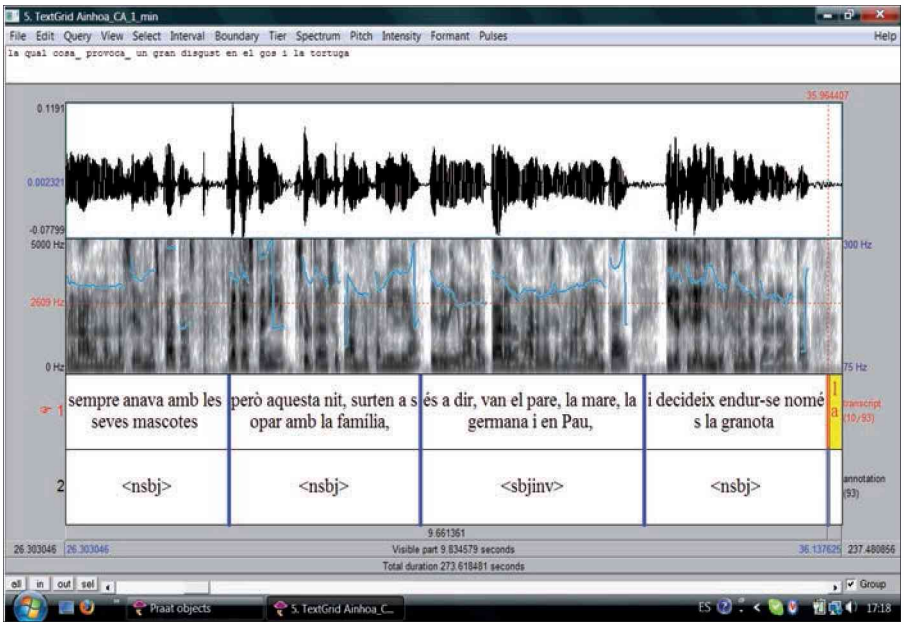


Figure 1: Audio-transcription alignment and phonetic analysis with PRAAT

3. The annotation

An annotation of non-canonical constructions (NOCANs) was carried out for the three Romance languages of the recording: Catalan, Spanish, and Italian. The reason for choosing these languages is that they are typologically similar. For instance, they all have a relatively free word order with SVO being the canonical one; they allow for null subjects, and they display left and right dislocations, including clitic dislocations. Their similar linguistic properties are parallel to their similar strategies for expressing informational notions. All these languages largely use syntax for this purpose, as opposed to languages like English, which nearly exclusively use prosody (see Vallduví / Engdahl 1996). The NOCANs found in these languages are indeed very similar or identical cross-linguistically. Despite these similarities, however, it has been shown in literature (cf. Villalba 2007, Leonetti 2008) that the use of individual NOCANs within Romance languages can vary considerably in frequency and function. A comparison between these languages is therefore important in order to show what such (quantitative or qualitative) differences are.

We must stress that an annotation of NOCANs is extremely rare in existing corpora, even those created with the explicit purpose of studying IS. The only example we are aware of is the MULI corpus ('MUltiLingual Information structure') of read German and English newspaper texts (Baumann 2006). Its annotation, which only concerns a small part of the corpus (only 250 sentences for the German part), includes NOCANs such as clefts, pseudo-clefts, extraposition, fronting, and passive sentences. Informational categories themselves are also annotated. Apart from its reduced extension, the main limit of this corpus is that it is not made of spontaneous speech, which is a crucial aspect for the study of IS, as argued above. Note further that *read* intonation is known to be different from intonation in spontaneous speech, in particular concerning the marking of informational categories (Hirschberg 1995). We therefore think that our corpus provides an empirical resource which had been missing up to now and which can enhance the research on IS and on its interaction with the other parts of grammar.

Finally, it is important to clarify the difference between NOCANs as objectively observable events in language production, and IS proper. Studying IS can be compared to hunting for ghosts in that both phenomena are not directly and objectively observable. While ghosts manifest themselves through knocks and falling objects, IS units such as topics and foci manifest themselves through

linguistic marking (i.e. NOCANs). In both cases we have to deduce the existence and presence of the object we are primarily interested in from objectively observed events. That explains why in the creation and annotation of the corpus we decided not to annotate directly IS units but the NOCANs which mark them. Although annotating IS categories is possible, in principle, the resulting annotation is less reliable because extra-sentential factors such as the linguistic context or even extra-*linguistic* factors such as the speaker's intentions have to be taken into account, and that often induces the annotator to make subjective choices. Furthermore, the annotation of IS categories is often hardly free from the bias of a specific theory on IS. In fact, a considerable amount of disagreement is still found among different schools concerning the identification and the definition of IS categories. The annotation of NOCANs is, thus, more neutral and objective.

4. Non-canonical constructions (NOCANs)

As we said above, NOCANs are marked constructions from a syntactic, morphological and / or prosodic point of view. Given the freedom of syntactic possibilities existing in the three Romance languages, the NOCANs we annotated where mostly morpho-syntactically marked constructions. The only exceptions are clear cases of deaccenting and focus fronting. In the latter construction, syntactic fronting is accompanied by prosodic fronting of the focal accent, and when the fronted focus is a preverbal element, fronting is only revealed by prosodic marking. It must be made clear that our preference for syntactic NOCANs does not mean that prosody plays no role in the representation of IS in the three Romance languages under study. Nevertheless, a fine-grained phonological annotation (e.g. of different types of pitch accents) is a very complex task that goes beyond the scope and the expertise of our team. It must be pointed out, however, that although phonological properties are not annotated, the relevant information is stored in the audio recordings and a subset of the recordings is aligned with the transcription and thus prepared for in-depth phonetic and phonological analysis. Our corpus is therefore potentially ready to be enriched with this kind of annotation in the future.

As we said above, our interest in NOCANs is in the fact that these constructions reveal the IS properties of a sentence. More precisely, NOCANs usually single out a specific information structural unit, such as a *topic* or a *focus*. Some constructions are used for instance to distinguish a topic from its *com-*

ment. A topic expression indicates the referent, about which the sentence conveys some information, which is represented by the comment (Strawson 1964, Reinhart 1981, Vallduví 1992). The construction called 'clitic left dislocation' (Cinque 1990) is used in many Romance languages precisely to make such a distinction. An example is given by (1), from the Spanish sub-corpus:

(1) *Al hombre se le cae el café*
to-the man RFL to-him falls the coffee
Ind. object obj.cl. Verb Subject
"The man drops the coffee" (Spanish)

The indirect object *al hombre* is left dislocated; the canonical position would be the post-verbal one. The left dislocation leaves a remnant clitic pronoun (*le*) in verb adjacent position and the subject occurs post-verbally. By occupying a sentence initial, pre-verbal position, the indirect object is easily identified as the sentence topic, while the rest of the sentence (verb + subject) constitutes the corresponding comment.

Another informational partition that can be marked by specific NOCANs is the *focus-background* partition. The focus is the informationally most relevant part of a sentence in a particular context. In the construction exemplified in (2), the focused direct object (capital letters indicate focal accent) occupies a non-canonical preverbal position (*without* clitic remnant, in this case).

(2) *Pure la LINGUACCIA, gli fa, la rana.*
even the tongue to-him$_{cl}$ he-puts-out the frog
Dir. Object Verb Subject
"Even the tongue did the frog put out to him" (Italian)

The construction hence explicitly and unambiguously separates the focus element from the rest of the sentence (the background). A *focus-background* construction can also be represented by a cleft sentence, as in (3):

(3) *i és ELLA que està a punt de prendre's el biberó*
and is her who is about to take for-herself the baby-bottle
"and it's him who is going to drink from the bottle"
(Catalan)

A cleft is made of two clauses: one introduced by the verb *to be*, and the other introduced by a complementizer. The two clauses allow for a clear separation

of the focus from the background: the copular clause is occupied by the sentence focus (the subject *ella* in 3), and the other clause represents the remaining background.

5. Taxonomy of NOCANs

A full list of NOCANS that are represented in our taxonomy is given in Table 2.

Label	Description	Label	Description
sbjinv	Subject inversion	*cldbl*	Clitic doubling
sbjinv_deacc	Subject inversion with deaccenting	*obj-sep*	Object separation
nsbj	Null subject	*narg*	Null argument
nsbj_c	Null subject in a coordinate clause	*focfr*	Focus fronting
arbnsbj	Arbitrary subject	*deacc*	De-accenting
sbj-sep	Subject separation	*pres*	Presentational sentence
clld	Clitic left dislocation	*pass*	Passive construction
ld	Left dislocation	*impers*	Impersonal construction
ht	Hanging topic	*cleft*	Cleft sentence
clrd	Clitic right dislocation	*pscleft*	Pseudo-cleft sentence
rd	Right dislocation	*inv-pscleft*	Inverted pseudo-cleft sentence

Table 2: The taxonomy of NOCANs for the Romance languages in the NOCANDO corpus

Whenever a certain construction represents a particular case of a more general construction, we assigned the former a label that contains the label of the latter. For instance, the label *arbnsbj* for arbitrary subjects, which are a particular case of null subjects, contains the label of null subjects: *nsbj*. We subdivided our labels into three groups: NOCANs that are specific to the subject, NOCANs that concern all arguments, and sentential NOCANs. We will describe them in details in the following subsections.

5.1 Labels specific to subjects

A subset of labels that we have proposed is specific to subjects. The three Romance languages have a default SVO order, so all constructions in which the subject did not occupy a pre-verbal position are marked as NOCANs. The label for post-verbal subjects is *sbjinv* ('subject inversion'). A Catalan example is given below:

(4) *Els va acompanyar el taxista*
them$_{cl}$ PAST take the taxi-driver
"The taxi-driver drove them" (Catalan)

The subject occurs post-verbally, while the direct object is moved to a pre-verbal position, thus rendering an OVS order. Inversion usually leaves the subject in focus or part of the focus. Note however that the informational role of the subject does not only depend on its postverbal position. If the postverbal subject is deaccented, it will be part of the background. For this reason, a deaccented post-verbal subject is marked with an additional label: *sbjinv_deacc*. In (5), *aquest nen* (orthographically separated from the verb by a comma) is the deaccented subject.

(5) *...que està disfressat, aquest nen*
for is dressed-up this child
"...for this child is dressed up" (Catalan)

Since all of the Romance languages we examined may avoid expressing the subject overtly, we frequently find null subjects (*nsbj*). We only annotated *nsbj*s in finite clauses, as infinitive ones are canonically subject-less. Null subjects generally refer to an entity that is salient in the context. They function largely in the same way as unaccented pronouns in languages like English. They can neither be focal nor topical. In (6), the *nsbj* in the *perché*-clause refers to the boy, namely an entity that is already salient in the context (it is the topic of the preceding clause).

(6) *invece il bambino è molto contento, perché ha salvato la sua rana*
instead the boy is very happy because has saved the his frog
"The boy on the contrary is very happy, because he saved his frog"
(Italian)

A special case of *nsbj*s are those which occur in coordinated clauses. We annotated them apart with the label *nsbj_c*. The reason is that these constructions can be interpreted as VP coordination, in which case subject omission in the second conjunct is expected.

(7) *Entonces la tortuga lo ve y hm se lo dice al niño*
Then the turtle it sees and uh him it tells to-the boy
"Then the turtle sees it and she tells the boy"
(Spanish)

A null subject may not refer to a definable entity but receive an arbitrary interpretation. While in English a plural pronoun is used in these cases (e.g. *they killed Kenny*), in Romance these subjects are necessarily non-overt. We assigned them the label *arbnsbj* for 'arbitrary (null) subjects'.

(8) *Y un día a este niño le regalaron pues una caja muy grande*
and one day to this boy to-him$_{cl}$ they-gave well a box very big
"And one day this boy received a large box"
(Spanish)

The importance of marking *arbnsbj*s is that they seem to play a role in sentence topic selection: the fact that the subject is arbitrary in reference makes the object a potentially better topic (cf. Brunetti 2009a).

An example of omitted argument that cannot without controversy be called subject is the argument of a copula sentence. In the subordinate clause of (9), the copula verb occurs in initial position and there is only one argument in the sentence (*la seva granota*). Since copula constructions always need two arguments, evidently one is missing here. However, it is not entirely clear that the missing argument is really the subject (Alsina 2004). In order to keep such cases apart from uncontroversial cases of *nsbj*s, we label them *narg* ('null arguments').

(9) *i llavors en Jaume es va adonar que que, home, era la seva granota*
and then the Jaume RFL PAST realizes that that well was the his frog
"and therefore Jaume realizes that that, well, it was his frog"
(Catalan)

Finally, constructions are annotated where the subject is separated from the verb by intervening material. If the separating material is another argument, then the subject is dislocated even though no subject clitics exist in these languages to mark the dislocation explicitly. We will discuss this case below when we introduce dislocation constructions. When the subject is separated from the verb by adjunct material, its syntactic position and its informational status are less clear. That is why in the annotation we tagged these constructions with specific label: *sbj-sep* ('subject separation').

5.2 Labels for all arguments

We have marked different kinds of argument detachments. Since object clitics exist in these languages, object dislocations are marked with a clitic remnant adjacent to the verb. Clitic left dislocations have been annotated with the label *clld*. As we said in Section 2, Example (1), the dislocated element is generally recognized as the sentence topic (cf. Vallduví 1992; Benincà 1988 [2001]; Zubizarreta 1998, 1999, among many others).

There is a certain variation among Romance languages with respect to the use of clitics in object dislocations. When the dislocated element is not accompanied by a clitic remnant, the construction is labelled *ld* ('left dislocation'). In Italian, the remnant clitic of an indirect object is not obligatory (cf. (10)). Its presence is associated with register, namely it is more common in colloquial speech.

(10) *A un bambino un giorno arriva un regalo*
to a boy one day arrives a present
"One day a boy receives a present" (Italian)

Given that there are no subject clitics in the three Romance languages, a preverbal subject can be demonstrated to be left dislocated only if it is separated from the verb by another argument. In that case, the subject will be assigned the same label *ld*.

(11) *Esta a mí no me quiere nada bien*
this-one to me not me_{cl} loves no good
"this one doesn't love me at all" (Spanish)

Following Vallduví (1992), among others, we assume that the informational function of an *ld* is not different from that of a *clld* (it marks the sentence topic).

Another left dislocation construction existing in these languages is the 'hanging topic' (cf. Cinque 1977, 1990). We assigned it the label *ht*.

(12) *La rana grande, la situación no le gustaba mucho*
the frog big the situation not to-her$_{cl}$ pleased much
"As for the big frog, she didn't like the situation at all"
(Spanish)

An *ht* is a detached element that has no marking of grammatical function (it is never a Prepositional Phrase, always a Noun Phrase) and is obligatorily resumed by a pronoun expressing its grammatical function. Unlike a regular *clld*, an *ht* can also be resumed by a strong pronoun or a demonstrative. If the *ht* is a subject or an object, the construction can only be distinguished from a *clld* by the presence of a strong pronoun or a demonstrative.[5] As the name itself makes clear, this construction marks topic material as well.

Clitic right dislocations (*clrd*) are dislocations of an argument to the right, with resumption of a clitic inside the clause. Unlike *clld*s, the clitic is always optional. Prosodically, a *clrd* is deaccented or has a reduced accent. Indeed, *clrd*s must always be old and salient in the discourse context (Vallduví 1992, Bott 2007, Brunetti 2009c), and prosodic weakness is precisely a marking of this constraint.[6] In (13), for instance, the frog has been mentioned in the immediately preceding discourse context.

(13) *el gat ja l'ha vist, a la granota.*
the cat already it$_{cl}$ has seen to the frog
"The cat already SAW the frog" (Catalan)

When the right dislocated argument has no resumptive clitic, the construction is simply called 'right dislocation' (*rd*). It mostly concerns cases in which the dislocated element is a subject.

A further related construction is 'clitic doubling' (*cldbl*). In terms of word order, *cldbl* is similar to *clrd*. The difference between the two lies in their intonation: while a right dislocated argument is deaccented, with *cldbl* the verb and the argument are in the same intonational unit, the nuclear accent falls on the doubled argument, and either the whole VP or the argument are in focus.

[5] In fact, prosody also contributes to distinguishing between the two constructions.

[6] Within Vallduví's (1992) model of IS, *clrd* and *clld* identify different informational units, called *tail* and *link* respectively.

(14) *Entonces la tortuga lo ve y se lo dice al NIÑO.*
so the turtle it_{cl} sees and to-him_{cl} it_{cl} says to-the boy
"So the turtle sees what happened and tells the boy everything"
(Spanish)

Sometimes the object is separated from the verb by non-argument material, although it still occurs postverbally. We label this construction *obj-sep* ('object separation'). Prosodically, the object is *not* deaccented, which means that it represents the focus or part of the focus, together with the verb.

Dislocations to the left single out topic material. An exception is focus fronting (*focfr*), which we already mentioned in Section 4, Example (2). The distinctive feature of this NOCAN is the prosodic marking of the affected phrase, which is assigned focal accent. In these languages, the focal accent canonically falls at the end of the clause, while in this construction it is in sentence initial position. Prosodic marking may not be accompanied by syntactic fronting. That happens for instance when the expression is the subject, as in (15).

(15) *A questo punto anche LARA è dispiaciuta*
at this point even Lara is sorry
"At this point even Lara is sorry about that"
(Italian)

Finally, we marked deaccented material *deacc*, namely the absence of a pitch accent on a word that would otherwise be expected to be accented (Swerts / Krahmer / Aversani 2002). Romance languages do not usually recur to *deacc*, although this is very common in other European languages (e.g. English). Nevertheless, we can sometimes find deaccented material in Romance, as in (16). The main accent would be expected to be on *rana*. On the contrary, the accent falls on *simpatica*, and the adjunct is deaccented. *Deacc* cannot be focal material, and it is not usually topic either.

(16) *ma Lara non è molto SIMPATICA, con questa rana*
but Lara not is very nice with this frog
"But Lara is NOT very nice, towards this frog"
(Italian)

Focfr and *deacc* (including *sbjinv_deacc*, see (5)) are the only NOCANs that explicitly (and in the case of *deacc*, exclusively) make reference to *prosodic*, rather than syntactic, non-canonicity.

5.3 Labels marking non-canonical types of sentences

The NOCANs described so far all affect isolated parts of a sentence. However, there is also a series of NOCANs that affect the entire sentence. We will present them in this section.

Presentational sentences (*pres*) are used to introduce new referents and states of affairs. They usually only contain new information. In Italian their most common form is *Locative clitic + verb 'to be' + NP* (see 17); in Catalan the corresponding form is *Locative clitic + verb 'to have' + NP*, and in Spanish they are typically introduced by the impersonal form of *haber* 'to have' (*hay*, *había*, etc.).

(17) *C'era una volta un bambino*
there was one time a boy
"Once upon a time there was a boy" (Italian)

Passive constructions (*pass*) also have a relation to IS. On one hand, the direct object of the active form becomes the subject of the corresponding passive form and therefore occupies a (canonical) preverbal position. Since such position is typically occupied by the sentence topic, passives may favour a topic interpretation of the direct object. Furthermore, in passives the subject of the active form corresponds to an adjunct *by*-phrase, which can be omitted. The main function of passives is in fact to omit or hide the agent of the event, which in the active form is always the subject. The agent may be hidden for various reasons, one being for instance that the speaker ignores its referent. In this sense, passives carry out a similar function as arbitrary subjects (see 9), in that they favour the presence of an indirect object (when given) as sentence topic (Brunetti 2009a).

Impersonal constructions (*impers*) were also annotated. Not even these verbs select an agent, so they are supposed to have similar effects to *arbnsbj*s and passives with respect to topic selection, as argued by Brunetti (2009a).

(18) *e lui continua hm a indicare non si sa dove*
and he keeps hum to point not IMP knows where
"And he keeps pointing who knows where"
(Italian)

Finally, we annotated three constructions where the sentence is divided into two separated clauses, and each clause typically represents a particular informational unit: cleft sentences (*cleft*), pseudo-cleft sentences (*pscleft*), and in-

verted pseudo-clefts (*inv-pscleft*). A cleft sentence has the following syntactic form: *Copula verb* + *XP* + *Comp* + *S missing XP*. As already seen in Section 4, Example (3), the XP is the focus (typically a contrastive one), and the remaining clause represents the background. *Psclefts* are related to cleft sentences. They have the form: *NP* + *relative clause* + *Copula verb* + *NP* or *S*, but unlike ordinary clefts, they do not mark a focus-background structure: it is the second part of the construction that is in focus instead (cf. *que el barquito se hunde* in (19)).

(19) *y lo que pasa es que el barquito se hunde*
and it that happens is that the little-ship IMPERS sinks
"and what happens is that the boat sinks" (Spanish)

Finally, *inv-psclefts* are *psclefts* that occur in a reversed order, namely the NP follows the copula verb.

6. Corpus exploitation

The corpus and the annotation of NOCANs provide a valuable source of *qualitative* data to be used as examples in theoretical studies on IS (see for instance Brunetti 2009a, Bott 2007). The corpus can obviously also be exploited for *quantitative* analyses of specific informational phenomena (see Mayol 2009, Mayol / Clark in press, Brunetti 2009b). In the following section we present a general overview of the annotation results on the three Romance sub-corpora, and we will propose some possible lines of research that may stem from them.

6.1 An overview of the annotation results

In Table 3 we report the most interesting results concerning the differences in frequency of NOCANs in the three Romance languages. The relative frequencies are given with respect to the total number of finite clauses.[7]

	Catalan		Italian		Spanish	
overt sbj	1561	**35.7%**	1262	**38.9%**	1027	**35.5%**
nsbj	1665	**38.1%**	1173	**36.1%**	1084	**37.5%**
arbnsbj	22	**0.5%**	7	**0.2%**	32	**1.1%**

[7] More precisely, we counted all main clauses and adjunct clauses, and we excluded all non-finite clauses (infinitives, gerunds) and relative clauses.

	Catalan		Italian		Spanish	
sbjinv	332	**7.6%**	215	**6.6%**	265	**9.1%**
clld+ld	62	**1.4%**	44	**1.35%**	39	**1.35 %**
clrd+rd	22	**0.5%**	21	**0.64%**	11	**0.38%**
ht	10	**0.2%**	2	**0.06%**	9	**0.3%**
cldbl	92	**2.1%**	7	**0.2%**	61	**2.1%**
cleft	3		4		2	
pscleft + inv-pscleft	40	**0.9%**	10	**0.3%**	37	**1.28%**
pass	5	**0.1%**	67	**2%**	7	**0.24%**

Table 3: Absolute and relative frequencies of some NOCANs with respect to segments

A first general observation to be made is that the overall number of NOCANs is relatively low. This is indeed expected as NOCANs are *marked* constructions with respect to the linguistic properties of a language. Although spontaneous speech is assumed to have more NOCANs than controlled speech or written text, this does not mean that the number of NOCANs in the former is very high.[8] An exception to low frequency is *nsbj*. Catalan, Italian and Spanish all allow for subject omission in finite clauses. In the corpus, however, *nsbj*s are in practice nearly as frequently as overt ones. Therefore, an empirical support for assuming that these languages *canonically* have overt subjects is rather weak. Although the corpus annotates *nsbj*s as NOCANs, we can actually conclude that they are at least as canonical as overt subjects. The percentages of all NOCANs including *nsbj* is 72.5% (uniformily distributed in the three languages: 72.2% in Catalan, 74.4% in Spanish, 71.1% in Italian). Without *nsbj*, the percentage drops consistently, but not dramatically, to 24.1% (23.57% in Catalan, 26.3% in Spanish, 22.8% in Italian).

Another NOCAN that has a higher frequency than the others, but low enough for the phenomenon to be still undoubtedly considered as non-canonical, is *sbjinv*. Among the factors that determine *sbjinv*, our data show that an important role is played by the type of subject. We found that a high percentage of

8 See, for example Carter-Thomas / Rowley-Jolivet (2001) on English written and spoken scientific discourse. These scholars have shown that the frequency of certain NOCANs in a written article is much lower than in the corresponding spoken presentation. What can be deduced from their data, however, is that there is a rather low number of these constructions in both kinds of discourse.

inverted subjects correspond to the pronoun *tots* (Cat.) / *todos* (Sp.) / *tutti* (It.) 'all'. We also found a strong correlation between certain situations in the storyline and inversion. There are two points in the narrations where nearly all speakers use *sbjinv*. In both cases the character referred to by the subject had been absent from the story for some time and makes a sudden re-appearance. Indeed, in these contexts the subject is in focus and we know that a focused subject tends to occupy a postverbal position whenever possible. The data also confirm the well known relation between *sbjinv* and type of verb. In general, unaccusative verbs (e.g. *come in*, *fall down*, *appear*, etc.) accompany inversion in our data.

Another interesting observation to be made on the annotation results is the rather even distribution of NOCANs among the three languages. This is in itself an interesting fact. The variation among individual speakers is higher than the variation across languages. For example, we observed that three out of the total of five passives in the Spanish subcorpus were produced by only one speaker. Another good example is impersonals. Their overall frequency in Romance is 2.1%, but we find speakers who do not use impersonals at all, while a small set use them with a high frequency of approximately 5% and one speaker even used it with a frequency as high as 10.3%. We conclude from this that NOCANs are also subject to personal style and variation.

Dislocations also behave in a very similar way in the three languages. This result is rather unexpected, as it has been claimed in literature that these three languages vary with respect to the frequency and use of dislocations. For instance, it has been argued that *clrd*s and *rd*s are much more common in Catalan than they are in Spanish (see e.g. Villalba 2007, who studied right dislocations in a Catalan theatre play and its Spanish translation). The frequencies we obtained from the corpus seem to contradict these conclusions. Upon closer inspection, however, we found that five out of the seven *clrd*s and all four *rd*s in the whole corpus were produced by only one speaker. In addition to that, the Spanish native speakers we consulted confirmed that all these instances are grammatical but highly marked. These cases in fact look like literal translations from Catalan. So we considered as plausible that the speaker in question, who was born and raised in Catalonia, showed a strong interference from Catalan. If we exclude that speaker, we find that the total of *clrd*s and *rd*s in Spanish only have a frequency of 0.2% (which would confirm Villalba's findings).

Finally, clitic doubling, although syntactically similar to dislocations, constitutes a case apart. We observe a clear difference between Catalan and Spanish on one hand and Italian on the other: the frequency of *cldbl* in the former languages is much higher than in the latter. This result is not surprising, as *cldbl* in Italian has morpho-syntactic restrictions that are totally absent in the other two languages. For instance, in Italian the clitic is only fully accepted if followed by another clitic (see Benincà 1988 [2001]: 151).

7. Conclusion and future work

The corpus we have presented in this article is an important resource for the study of information structure and potentially for the study of all phenomena found in spoken narration. The availability of high quality recordings allows the study of phonetic and phonological phenomena. The annotation of NOCANs identifies IS-related constructions within their context of appearance, and allows a quantitative analysis of them.

The developments and extensions of the corpus that we foresee in the near future take several directions. With respect to corpus compilation, we are interested in extending the corpus to other types of discourse, in particular spoken *dialogue*. Such an extension would allow a quantitative and cross-linguistic study of the differences between monologue and dialogue with respect to the use of NOCANs and the organization of information. The cross-linguistic side of the corpus can also be improved by collecting recordings of other languages, both within the Romance family (Portuguese, Romanian) and different language families. With respect to the annotation, an obvious extension concerns the annotation of NOCANs in the Germanic sub-corpora already available (English and German). Further annotations that may contribute to a better understanding of IS-related phenomena are conceivable. A phonological annotation of prosodic grouping and types of accents would offer a more detailed description of prosody-related IS phenomena. Semantic annotations of various kinds – thematic roles, animacy, degree of saliency of a referent in the discourse, etc. – would also contribute to a better understanding of certain phenomena.

As a last remark, it is important to stress that the NOCANDO corpus is publicly available and third parties are allowed (and encouraged) to enlarge and enrich the corpus both in terms of further annotations and of further languages.

References

Alsina, Alex (2004): La inversió copulativa en català. In: Anuari de Filologia 26: 9-44.

Baumann, Stefan (2006): Information structure and prosody: Linguistic categories for spoken language annotation. In: Sudhoff, Stefan et al. (eds.): Methods in empirical prosody research. (= Language, Context and Cognition 3). Berlin: de Gruyter, 153-180.

Belletti, Adriana (2001): Inversion as focalization. In: Hulk, Aafke / Pollock, Jean-Yves (eds.): Subject Inversion in Romance and the Theory of Universal Grammar. Oxford: Oxford University Press, 60-90.

Belletti, Adriana (2004): Aspects of the low IP area. In: Rizzi, Luigi (ed.): The cartography of syntactic structures. Vol. 2: The structure of IP and CP. Oxford: Oxford University Press, 16-51.

Benincà, Paula (1988 [2001]): L'ordine delle parole e le costruzioni marcate. In: Renzi, Lorenzo / Salvi, Giampaolo / Cardinaletti, Anna (eds.): Grande grammatica italiana di consultazione. Vol. 1. Bologna: Il Mulino, 129-239.

Berman, Ruth A. / Slobin, Dan I. (eds.) (1994): Relating events in narrative: A cross-linguistic developmental study. Hillsdale, NJ: Lawrence Erlbaum Associates.

Boersma, Paul / Weenink, David (2009): Praat: doing phonetics by computer (Version 5.0.47) [Computer program]. Internet: http://www.praat.org/.

Bott, Stefan (2007): Information structure and discourse modelling. PhD Diss., Univ. Pompeu Fabra, Barcelona.

Brunetti, Lisa (2009a): On the semantic and contextual factors that determine topic selection in Italian and Spanish. In: van Bergen, Geertje / de Hoop, Helen (eds.): Special issue on topics cross-linguistically. (= The Linguistic Review 26, 2 / 3), 261-289.

Brunetti, Lisa (2009b): Discourse functions of fronted foci in Italian and Spanish. In: Dufter, Andreas / Jacob, Daniel (eds.): Focus and background in romance languages. (= Studies in Language Companoion Series 112). Amsterdam: Benjamins, 43-81.

Brunetti, Lisa (2009c): On links and tails in Italian. In: Lingua 119, 5: 756-781.

Carter-Thomas, Shirley / Rowley-Jolivet, Elizabeth (2001): Syntactic differences in oral and written scientific discourse: the role of information structure. In: ASp, la revue du Geras [Groupe d'Etude et de Recherche en Anglais de Spécialité, Paris, France] 31: 19-37.

Cinque, Guglielmo (1977): The movement nature of left dislocation. In: Linguistic Inquiry 8: 397-411.

Cinque, Guglielmo (1990): Types of A'-Dependencies. Cambridge, MA: MIT Press.

De Mauro, Tullio / Mancini, Federico / Vedovelli, Massimo / Voghera, Miriam (1993): Lessico di frequenza dell'italiano parlato. Milano: Etas.

Erteschik-Shir, Nomi (1997): The dynamics of focus structure. Cambridge: Cambridge University Press.

Gavarró, Anna / Laca, Brenda (2002): Les perífrasis temporals, aspectuals i modals. In: Solà, Joan et al. (eds.): Gramàtica del català contemporani. Vol. 3. Barcelona: Empúries, 2665-2774.

Hirschberg, Julia (1995): Prosodic and other acoustic cues to speaking style in spontaneous and read speech. In: Proceedings of the International Congress on Phonetic Sciences (Stockholm, August 13-19 1995) 2: 36-43.

Lambrecht, Knud (1994): Information structure and sentence form: Topic focus, and the mental representations of discourse referents. New York: Cambridge University Press.

Leonetti, Manuel (2008): Alcune differenze tra spagnolo e italiano relative alla struttura informativa. Oral presentation at XVIII Congresso A.I.P.I. Associazione Internazionale Professori di Italiano, Universidad de Oviedo (September 3-6 2008, Oviedo).

MacWhinney, Brian (2000): The CHILDES Project: Tools for Analyzing Talk 1-2. Mahwah, NJ: Lawrence Erlbaum Associates.

Mayol, Laia (2009): Pronouns in Catalan: information, discourse and strategy. Ph.D. Dissertation, Univ. of Pennsylvania, Philadelphia, PA.

Mayol, Laia / Clark, Robin (in press): Pronouns in Catalan: Games of partial information and the use of linguistic resources. In: Journal of Pragmatics 42: 781-799.

Reinhart, Tanja (1981): Pragmatics and linguistics: An analysis of sentence topics. In: Philosophica 27: 53-94.

Rizzi, Luigi (1997): The fine structure of the left periphery. In: Haegeman, Liliane (ed.): Elements of grammar: Handbook in Generative Syntax. Dordrecht: Kluwer, 281-337.

Rizzi, Luigi (2005): On some properties of subject and topics. In: Bruge', Laura / Giusti, Giuliana / Murano, Nicola / Schweikert, Walter / Turano, Giuseppina (eds.): Proceedings of the XXX Incontro di Grammatica Generativa (February 26-28 2004, Cafoscarina, Venice, Italy).

Steedman, Mark (2000): Information structure and the syntax-phonology interface. In: Linguistic Inquiry 34: 649–689.

Strawson, Peter (1964): Identifying reference and truth-value. In: Theoria 30: 96-118. [Reprinted in: Strawson, Peter (1971): Logico-linguistic papers. London: Methuen, 75-95].

Strömqvist, Sven / Verhoven, Ludo T. (eds.) (2004): Relating events in narrative: Typological and contextual perspectives. Mahwah, NJ: Lawrence Erlbaum Associate.

Swerts, Marc / Krahmer, Emiel / Avesani, Cinzia (2002): Prosodic marking of information status in Dutch and Italian: A comparative analysis. In: Journal of Phonetics 30, 4: 629-654.

Vallduví, Enric (1992): The informational component. New York: Garland.

Vallduví, Enric / Engdahl, Elisabet (1996): The linguistic realization of information packaging. In: Linguistics 34: 459-519.

Villalba, Xavier (2007): La dislocació a la dreta en català i castellà, microvariació en la interfície sintaxi / pragmàtica. In: Caplletra: Revista Internacional de Filología 42: 273-302.

Zubizarreta, Maria L. (1998): Topic, focus and word order. Cambridge, MA: MIT Press.

Zubizarreta, Maria L. (1999): Las funciones informativas: tema y foco. In: Bosque, Ignacio / Demonte, Violeta (eds.): Gramática descriptiva de la lengua española. Vol. 3. Madrid: Espasa Calpe, 4215-4244.

III. Methodologie korpuslinguistischer Grammatikforschung / Methodologies of corpus-linguistic Grammar Research

HOLGER KEIBEL / CYRIL BELICA / MARC KUPIETZ / RAINER PERKUHN

Approaching grammar: Detecting, conceptualizing and generalizing paradigmatic variation

Abstract

This paper presents ongoing research which is embedded in an empirical-linguistic research program, set out to devise viable research strategies for developing an explanatory theory of grammar as a psychological and social phenomenon. As this phenomenon cannot be studied directly, the program attempts to approach it indirectly through its correlates in language corpora, which is justified by referring to the core tenets of Emergent Grammar. The guiding principle for identifying such corpus correlates of grammatical regularities is to imitate the psychological processes underlying the emergent nature of these regularities.

While previous work in this program focused on syntagmatic structures, the current paper goes one step further by investigating schematic structures that involve paradigmatic variation. It introduces and explores a general strategy by which corpus correlates of such structures may be uncovered, and it further outlines how these correlates may be used to study the nature of the psychologically real schematic structures.

1. Introduction

Much of what we linguists call grammar concerns schematic structures, i.e., structures that display some kind of paradigmatic variation. While, as an informal concept, the notion of schematic structures is fairly straightforward and intuitive, the precise nature of such structures as an integral part of language is not at all clear. This paper addresses the following two research questions:

1) On what empirical basis is it justified to infer schematic structures?
2) How can these schemas be captured conceptually?

These questions are pursued here as part of an overarching research program which was outlined by Kupietz and Keibel (Keibel / Kupietz 2009, Kupietz / Keibel 2009b) and is summarized in the next section. It is followed by a brief review of some previous empirical work that the current paper is based on (Section 3). The central Section 4 presents methodological and empirical explorations towards question (1) before a tentative approach to question (2) is outlined in the final section.

2. An empirical linguistic research program

The primary goal of the research program described by Kupietz and Keibel (Keibel / Kupietz 2009, Kupietz / Keibel 2009b) is to devise viable research strategies for developing an explanatory theory of grammar.[11] As the only fundamental assumption, this program adopts the general framework of Emergent Grammar (Hopper 1987, 1998), according to which any grammatical regularities are emergent by nature, being constantly influenced and reshaped by language use. These regularities are ascribed a psychological reality in the form of individual speakers' *language routines*, and these routines arise as a continuous result of each speaker's aggregating language experience. Likewise, the grammatical regularities are attributed a social reality which takes the form of *language conventions* in a language community, and these conventions may be characterized informally as the overlap between the individual grammars (i.e., language routines) of most speakers.

As one immediate consequence of their dual reality, grammatical regularities in turn necessarily shape language use: obviously, speakers routinely use their individual language routines, and in order to ensure successful communication, they are likely to use the conventions of the respective language community. If these general assumptions are valid, one would expect to find correlates of any grammatical regularity in an appropriate corpus of authentic language productions, provided that the corpus is sufficiently large and stratified.

The program proposes to adopt a strictly empirical research strategy which is founded on this prediction. Language routines of individual speakers and language conventions in a community cannot be accessed directly, but one may attempt to access and study them through their putative corpus correlates. As authentic corpus data are lexically specific, the best option to do this is by a bottom-up, inductive approach: to start from individual lexical items and to proceed by incrementally deriving increasingly complex and abstract structures around these items. Given the reciprocal dynamics of an emergent grammar that were described above, many – though not all – abstract regularities may have become psychologically real for most speakers along very similar inductive paths. In other words, the bottom-up strategy that we advocate constitutes an attempt to mimic the inductive psychological processes underlying the emergence of grammatical regularities.

[11] Any progress that we make towards this goal is published in a series of talks and papers with the same running title "Approaching grammar".

Some additional tenets of the program are:

a) to start from minimal theoretical assumptions about language;
b) to defer any linguistic classifications as long as possible;
c) to proceed conservatively, in small inductive steps;
d) to motivate each step deductively by psychological premises which in turn need to be tested independently;
e) to keep the descriptive categories and generalizations simple.

The purpose of (a) and (b) is to ensure that one studies language as a factual phenomenon, and not a theoretical construct, while (c) through (e) are intended to minimize the risk of a blind induction that any empirically driven approach is exposed to.

Despite this rather brief description (for a more detailed formulation see Kupietz / Keibel 2009b, Keibel/Kupietz 2009), the general research program is an ambitious enterprise and certainly nothing that can be completed within a short period of time. In fact, as a first skeptical response, one may doubt whether it can be accomplished at all, for how should it be possible to arrive at *any* theory of grammar if only unstructured corpus data are analyzed and even the most basic linguistic categories are excluded from the assumptions? There are two answers to this rhetorical question: first, as we argued in the work cited above, there is no viable alternative when explanation is the goal, and second, induction alone will certainly not suffice; abduction and falsification are also needed to bridge the gap between observable data and the theoretical level.

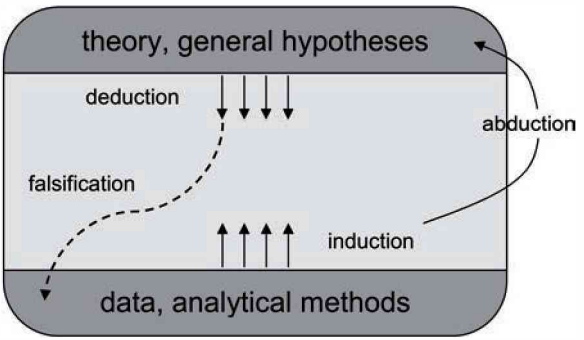


Figure 1: Towards an explanatory theory of grammar

In a nutshell, observations at the data level are incrementally generalized to more abstract structures, by means of inductive methods that are motivated by psychological facts and premises. It then has to be demonstrated that the resulting generalizations are, in general, corpus correlates of structures that are psychologically and socially real – a priori, they are merely candidates for such correlates. Exploring a large number of such corpus correlates at any level of abstraction prompts researchers to abductively formulate new hypotheses at the theoretical level. Each of these new hypotheses has to be tested empirically, in terms of deduction and falsification.

This complex, iterative strategy is not targeted at studying a specific phenomenon in a quick, direct fashion. Instead, it is an attempt to approach the very notion of *grammatical structure*, and as its focus is on explanation, centrality is given to empirical facts in at least two ways:

a) induction is driven by empirical observation;

b) any hypotheses derived from the resulting generalizations are tested against empirical data.

3. Previous work

The empirical research reported here crucially builds on previous work (an overview may be found in Keibel / Kupietz / Belica 2008), the primary goal of which was to study the emergent nature of syntagmatic structures that are psychologically and socially real. These structures may be referred to as *psychological collocations* (cf. Hoey 2005). As there is no way to access them directly, we studied them through their putative corpus correlates which take the form of *statistical collocations*. The specific notion of statistical collocation that we used to this end is that of *higher-order collocations* which are detected by processes that are meant to imitate those underlying the psychological collocations. It is a notion that is more flexible than that of n-grams, as the collocates in a higher-order collocation may be non-contiguous, and their order and distances may vary. This concept dates back to 1995 (Belica 1995, Keibel / Belica 2007) and it was rediscovered recently as the very similar, albeit not identical, concept of *concgrams* (Cheng / Greaves / Warren 2006). Due to their positional flexibility, higher-order collocations are sometimes hard to relate to the intuition of competent speakers. Therefore, each higher-order collocation is typically listed together with a *syntagmatic pattern* which summarizes the collocation's most typical word order.

To illustrate these concepts, Figure 2 shows some of the most cohesive higher-order collocations for "*why*" which were derived from a fairly small web-based corpus composed of written English (2.5 million words). Each line corresponds to one higher-order collocation, the collocates are listed in the central column while the right-most column gives the predominant syntagmatic pattern.

-1	-1	1805	**reason** one main	1	100%	reason ... main one why
			reason one	64	96%	is one reason [...] why the ...
			reason main	21	90%	The|the main reason why the ...
			reason One	24	100%	One reason [...] why the ...
			reason	181	100%	is one reason [...] why the ...
-1	-1	1282	**explain** helps	23	100%	This helps [to] explain [...] why ... the
			explain may help	3	100%	may help [to] explain why
			explain may	28	96%	This may [...] explain why the ...
			explain help	13	100%	may|might help [to] explain why
			explain	113	100%	helps may|to explain [...] why
-1	-1	575	**is** That	78	83%	That [...] is [...] why the ...
			is easy It	13	92%	It is easy to see why
			is easy	19	89%	It|it is [...] easy to see why
			is It	21	90%	It is easy to see why the
			is	393	71%	That is [... reason] why the ...
-1	-1	543	**explains** This partly	3	100%	This [...] partly explains why
			explains This	12	100%	This [partly] explains why
			explains partly	7	100%	This partly explains why the ...
			explains	49	100%	This explains [...] why the ...
-1	-1	528	**reasons** There are several	7	85%	There are several reasons why
			reasons There are	23	78%	There are several|two reasons why
			reasons There	24	100%	There are several|many reasons why
			reasons are several	9	88%	There are several reasons why
			reasons are	34	82%	There|there are [several|two] reasons why the ...
			reasons several	11	100%	There|there are several reasons why
			reasons	57	100%	are ... reasons [...] why the ...

Figure 2: Collocation profile of "*why*" (only top portion shown)

The *collocation profile* of a given node word is defined as the full spectrum of higher-order collocations around this node word, together with the dominant syntagmatic patterns and some related characteristics. Again, as an illustration, Figure 2 is the top portion of the collocation profile of "*why*". For many

higher-level research questions it is useful to have fast access to large numbers of such collocation profiles, and this is also the case for the present study. To this end, we take advantage of the collocation database CCDB (Belica 2001-2007, Keibel / Belica 2007) which currently provides collocation profiles for more than 220 000 node words. These profiles are based on the virtual corpus CCDB2007 with approximately 2.2 billion text words (Institut für Deutsche Sprache 2007b) which was composed as a subset of the German Reference Corpus DeReKo (Institut für Deutsche Sprache 2007a, Kupietz / Keibel 2009a, Kupietz et al. 2010).

It should be pointed out that node words in these collocation profiles are lemmas, whereas their collocates are word forms. We believe that this different treatment of node words and collocates imposes minimal assumptions at the psychological level. On the one hand, the fuzzy denotational and connotational structure associated with the entire paradigm of a word – irrespective of its inflectional properties – seems to be, in general, ontologically more primitive than any of its specific grammatical forms (cf. Belica et al. 2010). The units under investigation should therefore be lemmas. On the other hand, it is only the grammatically expressed word forms that can be, in general, directly observed in language use. Thus, the collocates – as observed properties of the node words – should be assessed at the level of word forms.

4. Detecting schematic structures

For the next inductive step our primary goal was to study the emergent nature of schematic structures that are psychologically and socially real. These structures may be characterized as syntagmatic structures involving paradigmatic variation. The research interest thus is on schemas as emergent psychological phenomena, but just as for psychological collocations in the previous section, there is no way to access them directly. We therefore attempted to study them through their putative corpus correlates. However, unlike the case of collocations, it is not at all clear what these correlates are and how they may be uncovered.

We are not the first ones interested in inducing syntagmatic-paradigmatic structures from corpora and there is a growing body of useful concepts and approaches in the corpus-linguistic literature, including *collocational frameworks* (Renouf / Sinclair 1991), *phrase frames* (Stubbs 2004, Fletcher 2003), *Pattern Grammar* (Hunston / Francis 2000), *local grammar patterns* (Mason

2004), and *formulaic frames* (Biber 2009). We propose a different approach that is motivated by the following rationale. As schemas are syntagmatic-paradigmatic structures, they can be thought of as being instantiated by syntagmatic structures, viz. collocations. The general idea therefore is that schematic structures may in turn be uncovered as abstractions across collocations. This idea is, again, an attempt to imitate the psychological processes underlying the emergence of schematic structures because it is likely that many schematic structures have become psychologically real for most speakers as abstractions across psychological collocations, and that the way they are constantly reshaped in speakers' minds is also driven by the same influence.

Before further outlining the idea, we first need to refine our terminology. When we talk of *schematic structures* or simply of *schemas* without further explication, we refer to structures that are real in a psychological or social sense. As stated before, what can be found from corpora – or in this case, from corpus-induced statistical collocations – are not schemas but, at best, only correlates of schemas. In the following we refer to these correlates as *schema corpus correlates* (short: *SCCs*). However, the psychological and social status of the structures that a given approach induces from collocations is not known a priori. Therefore, until this status has been tested, at least in principle, we call these induced structures *SCC candidates*. It is justified to talk of SCCs only when the induced SCC candidates are in general psychologically real. For the remainder of this paper, it is thus important to strictly distinguish between schemas, SCCs, and SCC candidates.

With these concepts we can formulate a general approach towards finding schemas which consists of three stages. The subject of the first stage is to manually explore collocations for traces of paradigmatic variation in order to obtain some inspiration as to where and how to look for schemas. In the second stage, these observations are used to devise a specific strategy for automatically inducing SCC candidates from collocations. Once such a strategy has been formulated, the general psychological reality of the SCC candidates it induces has to be tested, and these tests constitute the third stage. It is unlikely that the inductive strategy formulated upon the first attempt will be sophisticated enough to detect genuine SCCs, so these tests will prompt one to go back to stage 2 and to revise this strategy until one arrives at a setup that is believed to generally detect genuine SCCs. In the remainder of this section we explore and discuss this three-stage approach for written German, based on the CCDB and the same virtual corpus CCDB2007 as before.

4.1 Stage 1: Exploring collocations for traces of paradigmatic variation

To address the first stage of this strategy, a large number of collocation profiles have to be inspected. For instance, consider the profile of the German adjective and past participle *vergangen* (English: *last*, *past*, *elapsed*), the top portion of which is shown in Figure 3.[2]

1 1	205448	**Jahr** Umsatz Milliarden	15	73%	der Der Umsatz der im vergangenen Jahr ... Mil	
		Jahr Umsatz erwirtschaftete	12	75%	erwirtschaftete ... im vergangenen Jahr einen Ur	
		Jahr Umsatz	425	63%	Im|im vergangenen Jahr [einen der] Umsatz vor	
		Jahr Milliarden erwirtschaftete	2	100%	erwirtschaftete im vergangenen Jahr ... Milliarde	
		Jahr Milliarden	436	69%	im vergangenen Jahr ... Milliarden Mark	
		Jahr erwirtschaftete	51	64%	erwirtschaftete ... im vergangenen Jahr mit|einer	
		Jahr	22925	91%	im vergangenen [...] Jahr	
1 1	95827	**Woche** Erst Mitte	2	50%	Erst Mitte ... vergangenen Woche	
		Woche Erst angekündigt	1	100%	Erst vergangene Woche ... angekündigt	
		Woche Erst	155	52%	Erst in der vergangenen Woche hatte ...	
		Woche Mitte	103	73%	Mitte vergangener Woche	
		Woche angekündigt	84	58%	in der vergangenen Woche [...] angekündigt	
		Woche	9910	54%	in der vergangenen [...] Woche	
1 2	89453	**Jahren** zehn kontinuierlich	2	100%	vergangenen zehn Jahren [...] kontinuierlich	
		Jahren zehn zwanzig	4	75%	in den vergangenen zehn [bis] zwanzig Jahren	
		Jahren zehn	832	93%	in den vergangenen [...] zehn [...] Jahren	
		Jahren kontinuierlich zwanzig	1	100%	vergangenen zwanzig Jahren kontinuierlich	
		Jahren kontinuierlich	78	100%	in den vergangenen [...] Jahren [...] kontinuierlic	
		Jahren zwanzig	97	90%	in den vergangenen [...] zwanzig [...] Jahren	
		Jahren	14461	97%	in den vergangenen [...] Jahren	
1 1	66604	**Jahres** Ende Mai	5	100%	Ende Mai [...] vergangenen Jahres	
		Jahres Ende Juli	10	100%	Ende Juli [...] vergangenen Jahres	
		Jahres Ende	1067	98%	Ende [des] vergangenen [...] Jahres	
		Jahres Mai	349	99%	im Mai [...] vergangenen Jahres	
		Jahres Juli	289	99%	im Juli [...] vergangenen Jahres	

Figure 3: Collocation profile of *vergangen* (only top portion shown)

By scanning this collocation profile, one may uncover traces of paradigmatic variation around this word *vergangen*. However, doing so for a profile that is represented as a simple list is not very efficient, and, more importantly, a lot of paradigmatic variation may be missed in this way. The evidence for some paradigmatic structure is often scattered across the profile which in turn is gener-

[2] The full profile may be inspected at: http://corpora.ids-mannheim.de/ccdb/.

ally rather large. A more systematic approach is needed for this exploratory stage, and ideally this involves the possibility of progressively recording any evidence for paradigmatic variation as it is encountered. To this end, we took advantage of the collocation explorer VICOMTE (Perkuhn 2007).

Figure 4 shows the same collocation profile of *vergangen* in VICOMTE's default visualization. The node word is displayed in the center and its primary collocates are given on the inner-most circle around the node word, sorted by decreasing cohesion (*Jahr*, *Woche*, *Jahren*, ...). In order to keep the visualization simple, only five primary collocates are displayed at full size while the others appear miniaturized. VICOMTE offers several interactive ways of inspecting all regions of the collocation profile at normal size (e.g., mousing over the respective boxes or rotating the entire tree diagram). Secondary collocates appear attached to the corresponding primary collocates, and ternary collocates are in turn attached to their corresponding secondary collocate, and so forth. Like this, any higher-order collocation is represented by a unique radial path connecting the node with collocates on the various circles.

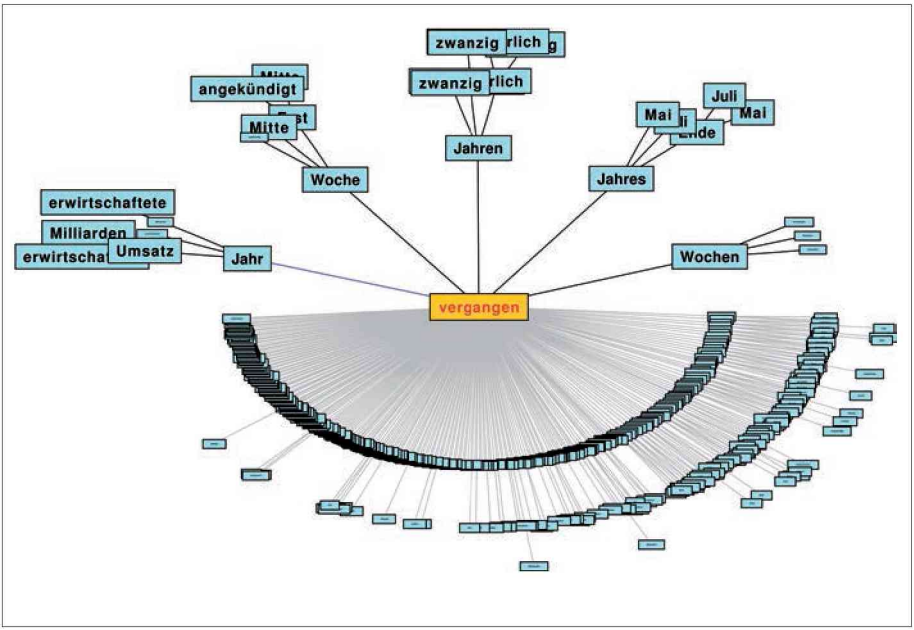


Figure 4: Collocation profile of *vergangen* (VICOMTE visualization)

To keep the explorations simple, we only look at primary collocates (in the visualization on the inner circle). Scanning through this single profile, we

encounter a lot of evidence for schematic structures around this node word. For example, many of the primary collocates of *vergangen* refer to larger units of time such as *Jahr*, *Jahrzehnt*, *Jahrhundert*, *Monat*, *Woche*, *Periode*, etc. (English: *year*, *decade*, *century*, *month*, *week*, *period*, etc.). All words of this group found in the profile are highlighted in Figure 5.

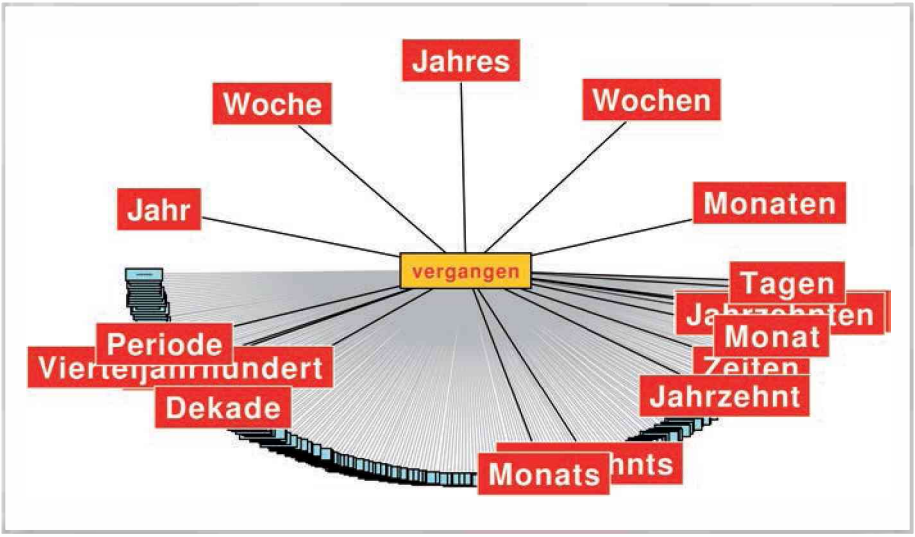


Figure 5: Profile of *vergangen*: primary collocates referring to units of time are highlighted

Based on our competence as speakers of German, we believe that this group of collocates indeed relates to a schematic structure in speakers' minds which might be summarized as follows:

(1) *vergangen*$_{+inflection}$ *<Zeiteinheit>*
last / past <unit of time>

The collocates in this description are generalized to a *paradigmatic class* (represented here by the placeholder variable *<Zeiteinheit>*). Importantly, this description is merely an intuitive label for the putative schema (and its paradigmatic class), but not necessarily the schema itself. An adequate representation of the full schema is likely to be more complex, e.g., involving relations between its components and so forth. Given the goals of this research, it seems advisable not to make any a-priori assumptions about the representation of schematic structures. Therefore, a label as in (1) is to be understood only as an intuitive shorthand for a schema, SCC or SCC candidate.

It should also be stressed that the collocations of *vergangen* with the group of collocates highlighted in Figure 5 constitute good evidence for a possible schema not because there is a nice descriptive label for it but because there appears to be some more general structure underlying these collocations in speakers' minds. In other words, there is a competence-based response in speakers who are subsequently exposed to collocations such as *vergangene Woche* and *vergangenes Jahr* which triggers schematic entities in their implicit language knowledge. It does not matter whether or not speakers are able to make this knowledge explicit – what matters is the response itself.

In many cases, therefore, it is difficult to capture the essence of a paradigmatic class of collocates by a concise label such as <*Zeiteinheit*>. Therefore, to be able to describe all putative schemas, SCCs and SCC candidates in the same universal way, we use non-interpretative labels such as (2).

(2) *vergangen* $_{+inflection}$ {*Jahr, Jahrzehnt, Jahrhundert, Monat, Woche, Periode, ...*}

last / past {year, decade, century, month, week, period, ...}

To record this particular evidence, we restructure and annotate the VICOMTE representation such that the collocates referring to units of time are grouped together (Figure 6). In this fashion, we continue to scan the remaining profile and to record any indications of paradigmatic variation by grouping together the respective collocates to a putative paradigmatic class.

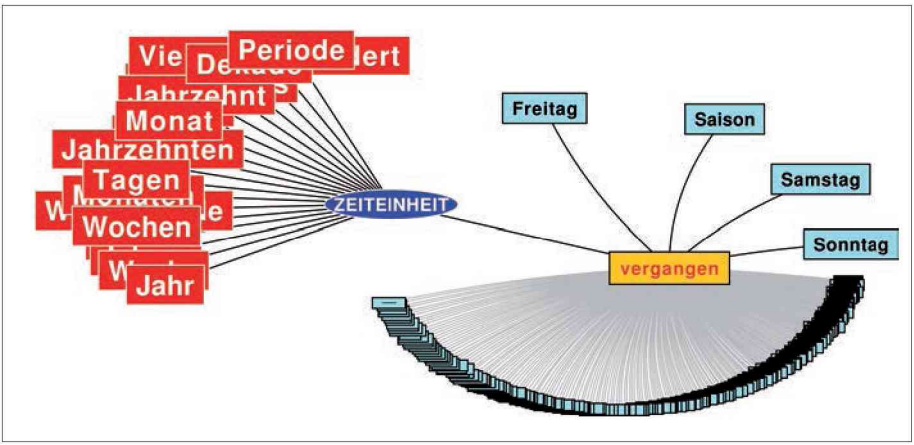


Figure 6: Annotated profile of *vergangen*

In our explorations for the specific profile of *vergangen* we identified a large number of likely paradigmatic classes of collocates. Some examples are the groups listed in (3).

(3a) *Montag, Dienstag, ..., Samstag, Sonnabend, Sonntag*
Monday, Tuesday, ..., Saturday, Sunday

(b) *zwei, fünf, sieben, zwölf, fünfzehn, anderthalb, ...*
two, five, seven, twelve, fifteen, one and a half, ...

(c) *Gewinn, Verlust, Umsatz, ...*
profit, loss, revenue, ...

(d) *gestiegen, gewachsen, zugelegt, zurückgegangen, gesunken, ...*
increased, gained, decreased, dropped, ...

(e) *deutlich, erheblich, drastisch, kräftig, stark, ...*
considerably, substantially, drastically, strongly, ...

Some of these putative paradigmatic classes in conjunction with the node word *vergangen* do not readily relate to any intuitive structures. However, inspecting the range of underlying concordances of the individual collocations does prompt respective structures in competent speakers, albeit this response may be weaker than for the earlier example (2). For instance, collocations underlying the paradigmatic classes (3d) and (3e) are often instantiated in sentence fragments such as

(4) *... in den vergangenen Jahren deutlich gestiegen ...*
... over the past few years considerably increased ...
... increased considerably over the past few years ...

and likewise for any collocate of group (3d) in the position of *gestiegen*, and any collocate of group (3e) instead of *deutlich*.

We explored a large number of collocation profiles in the same fashion, and these explorations overall lead to the following general observations. First, any evidence for paradigmatic variation that we observed for a fixed node word involved a group of collocates that are semantically fairly similar, where similarity is assessed in terms of intuitive speaker judgments. This refers to a non-categorial but rather associative psychological notion of semantic similarity

(cf. Belica et al. 2010). Second, the collocates in each such paradigmatic class tended to belong to the same lexical category. Third, where applicable, the collocates in each such paradigmatic class were often observed to share morphosyntactic features. For instance, in example (3d), all collocates grouped together were past participles in their basic form (i.e., not inflected as adjectives). Fourth, the collocations underlying each such paradigmatic class display very similar positional preferences. That is, the different collocates tend to occur in (nearly) the same position relative to the node word.

To sum up, the explorations so far suggest that simple two-word collocations for a fixed node word may be good candidates for relating to an underlying schematic structure if the respective collocates are similar in terms of their associative semantics and their positional preferences relative to the node word (observations 1 and 4). In particular, they suggest that the paradigmatic classes underlying the schemas of a node word are no language-general word classes, but rather specific to this node word, if not specific to the individual schemas (very much as in Construction Grammar, especially in the approach by Croft 2001). Note, however, that, due to our methodological framework, the collocates' agreement with respect to lexical categories and morphosyntactic features (i.e., observations 2 and 3) does not qualify as additional criterion for detecting schemas. Instead, these two observations may be construed as epiphenomena of the other two observations (cf. 4.2.1).

4.2 Stage 2: Inducing SCC candidates

With respect to automatically inducing SCC candidates, the results of the exploratory stage 1 so far are instructive in several ways. First of all, they suggest that the task may be simplified by splitting it into two subtasks (cf. Figure 7). The first subtask is to identify for a given node word – on the basis of its collocation profile – the paradigmatic classes that appear to be relevant for its schemas. Once this is done, the resulting paradigmatic classes may be used to derive from the same corpus a range of SCC candidates for this node word which constitutes the second subtask.

In the following, we briefly describe a possible way of accomplishing these two tasks. The specific operationalizations given below are only of secondary importance because our epistemic goal is to imitate and model the inductive processes underlying the emergent nature of schematic structures in language. We approach this goal here by mimicking the competence-based procedures in 4.1.

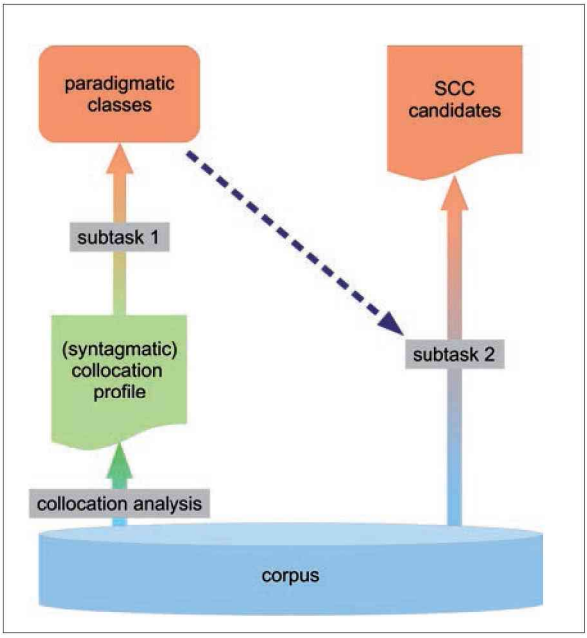


Figure 7: Two subtasks for inducing SCC candidates

4.2.1 Subtask 1: Inducing paradigmatic classes for a node word

The basic idea for accomplishing subtask 1 is to group together all primary collocates of a given node word that are sufficiently "similar", and to use a notion of similarity that imitates the four intuitive similarity criteria observed in Subsection 4.1. However, it turns out that the second and third of these observations cannot be exploited here for at least two reasons. First, they involve preexisting linguistic categories (lexical categories and morphosyntactic features) and therefore unnecessary theoretical assumptions which should be avoided given the explanatory objectives of this research program (cf. Section 2). Second, even if linguistic categories were admitted in this program, available taggers and parsers are usually not reliable enough, especially for the less frequent phenomena. This imperfect reliability would be much less problematic if it constituted evenly distributed statistical noise in the classifications – but as it involves systematic errors, the scientifically sound use of taggers and parsers generally requires time-consuming manual intervention (cf. Belica et al. in this volume) which would be too costly for the kind of research pursued here.

By contrast, the two remaining observations of Subsection 4.1 – associative-semantic and positional similarity – are fully valid concepts in this program. Both are meaningful constitutive criteria constraining the systematic search for realistic paradigmatic classes. While this is immediately obvious for the criterion of associative-semantic similarity, the following example is intended to demonstrate it also for positional similarity. Among the primary collocates of the node word *Haar* (English: hair), there are many color adjectives (in various inflected forms), including the German counterparts of *red*, *white*, *snow white*, *black*, *blonde*, *salt-and-pepper*, *brunette*, etc. Surprisingly, however, *Haar* also collocates with different forms of *blau* (English: blue), which is an unlikely hair color – at least not likely enough for the word to be traceable as a significant collocate of *Haar*. Closer inspection of the underlying concordances reveals that this collocation is mainly due to instances of the phrase (5a), or some variant of it, where *blau* is not used to refer to an attribute of *Haar*. An inductive reasoning guided by associative-semantic similarity alone would probably face difficulties to distinguish *blau* (in its various forms) from those color adjectives that are in fact used significantly as attributes of *Haar*, as in example (5b). Incorporating information about the typical word position of the collocate (relative to the node word) constrains associative-semantic induction and helps to induce appropriate paradigmatic classes.

(5a) *blonde Haare und blaue Augen*
blonde hair and blue eyes

(5b) *mit roten Haaren*
with red hair

(5c) *ihr Haar ist rot* [*gefärbt*]
her hair is [dyed] red

Note that this information also helps to distinguish predicative uses (5c) from attributive uses (5b) of the same color adjective, provided that they are different word forms (which is generally the case in German). This points to the more general observation that the positional preferences of a collocate, relative to its node word, often correlate with lexical categories and morphosyntactic features. In other words, although lexical classes and morphosyntactic features were excluded from our theoretical assumptions, the underlying type of information seems to remain available in this approach.

In the appendix we briefly describe how we operationalized the two notions of associative-semantic and positional similarity between collocates, and how both may be integrated into a single similarity measure which can be thought of as quantifying the overall *paradigmatic similarity* between any two collocates.[3]

In order to complete subtask 1, one additionally needs a way to group paradigmatically similar collocates into paradigmatic classes. Of course, it is easy to find for each individual collocate *x* the set of other collocates that are paradigmatically most similar to *x*, but it is a nontrivial optimization problem to partition the set of collocates into classes such that all collocates simultaneously are sufficiently *happy* with the other words that they are grouped with. The human mind is extremely proficient at this kind of optimization problem and performs it all the time, but it is not at all clear how this human skill could best be imitated. *Self-organizing methods* such as *hierarchical cluster analysis* are probably a good starting point. When applied to the present situation, such methods produce tree diagrams (so-called *dendrograms*) which represent the global similarity structure of all collocates. For instance, clustering the primary collocates of the node word *vergangen* produces the following dendrogram (the specific cluster analysis method is irrelevant here).

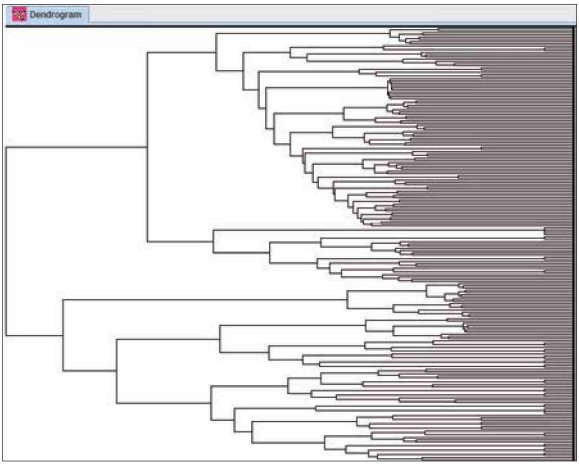


Figure 8: Dendrogram for the collocates of *vergangen* (collocates not shown)

[3] For the purposes of this study, nothing crucial hinges on these particular operationalizations – in fact, there may be more appropriate ones – they are merely tentative proposals and only serve to provide a proof of concept for the general research strategy.

Each line to the right represents one collocate (the collocates themselves are not shown to avoid overcrowding the figure), and the tree structure (from right to left) visualizes how similar collocates are successively merged to increasingly large clusters. The more to the right they are merged, the more similar they are.

One straightforward way of deriving (hypotheses on) paradigmatic classes from such a dendrogram would be to cut the tree at a given similarity level and delete the structure to the left of this level. The remaining clusters of collocates would then be interpreted as the relevant paradigmatic classes: the higher (i.e., further to the left) the cut-off level, the greater and more general the derived classes. For the above dendrogram, an intermediate cut-off level yielded, among others, the following putative paradigmatic classes for *vergangen* which contain names of weekdays (6a), months (6b), cardinal numerals (6c), past participles expressing decrease or increase (6d), and intensifying degree adverbs (6e) respectively (square brackets give each collocate's positional preference relative to *vergangen*, cf. Appendix). These examples correspond closely to the competence-based paradigmatic classes (cf. 4.1). Note that the collocates in these examples are listed in the order in which they appear in the dendrogram such that the class-internal similarity structure is still partly reflected in these lists.

(6a) *Freitagabend, Samstagabend, Freitag, Dienstag, Montag, Donnerstag, Mittwoch, Sonntag, Samstag, Sonnabend* (all: [1;1])

(b) *Mai, Oktober, Juli, März, November, September, August, Juni, April, Februar* (all: [-1;1])

(c) *zwei, drei, vier, fünf, sechs, neun, sieben, zehn, zwölf, fünfzig, fünfzehn, zwanzig, halben, zweieinhalb, anderthalb, eineinhalb* (all: [1;1])

(d) *gesunken, gestiegen, angestiegen, zurückgegangen, zugenommen, gewachsen, zugelegt* (all: [2;5])

(e) *kontinuierlich, stetig, dramatisch, drastisch, deutlich, erheblich, kräftig, stark* (all: [2;3])

We derived paradigmatic classes for a range of other node words in the same fashion and found the results to be highly plausible in almost all cases. In short, although the particular operationalizations we chose for these explorations are in part rather provisional, the explorations indicate that the pro-

posed strategy for subtask 1 is a good starting point for identifying potential paradigmatic classes that may be relevant in the schemas of a given node word.

An open issue is how to determine the *optimal* – i.e., psychologically most adequate – cut-off level which has direct consequences on the size and degree of abstraction of the resulting paradigmatic classes. Moreover, the optimal cut-off level is most likely not constant for all branches of the tree.

Importantly, the general strategy does not presuppose any predefined language-general classes but empirically derives classes that are potentially node-specific, and this is in line with the general observation in Subsection 4.1. In other words, this strategy supports the possibility that, for different node words, the same collocate word may belong to very different paradigmatic classes. However, as is illustrated in 4.2.2, the realistic classes are not only node-specific but sometimes even schema-specific – the same collocate may generalize to a different paradigmatic class in different SCC candidates around the same node word. To extend the strategy for subtask 1 to also capture this possibility, one would have to allow for a collocate to be a member of multiple classes.

4.2.2 Subtask 2: Deriving SCC candidates for a node word

In this subsection we propose a way by which the paradigmatic classes identified for a given node word by the first subtask may be used to derive SCC candidates for the same node word. The underlying corpus should be the same as the one from which paradigmatic classes were derived – in our explorations, this was again the virtual corpus CCDB2007.

Fortunately, this second subtask may be addressed in a fairly straightforward way, by exploiting an existing methodology that is already well-established within this research program. The basic idea is that SCCs are much like statistical collocations, except that at least one of their collocates is not a specific word but a whole paradigmatic class of words. Therefore, SCC candidates may be detected by re-using the same collocation algorithm as in Section 3, but this time, the paradigmatic classes (identified in subtask 1) are treated as potential collocates, as if the members of each such class were the same word. Like this, the node word may potentially be found to collocate with the classes and any other words that do not belong to these classes. The result is a new type of collocation profile which may be called *syntagmatic-paradigmatic collocation profile*. Some colloca-

tions in this profile involve one or multiple paradigmatic classes – these collocations are the SCC candidates proper – while others are entirely lexically specific and are simply higher-order collocations that were found before.

We explored this general strategy again for a range of node words. Figure 9 shows one of the SCC candidates we found for the node word *vergangen*, together with a fraction of the underlying concordances. This SCC candidate is a syntagmatic-paradigmatic collocation consisting of the node word itself, a lexically specific collocate *Jahren* (English: years), a second collocate which is a paradigmatic class {erheblich, enorm} (English: substantially, enormously), and as a third collocate the larger class {*angestiegen*, *gestiegen*, *gesteigert*, *verbessert*, *gesunken*} (English: increased, raised, improved, dropped).

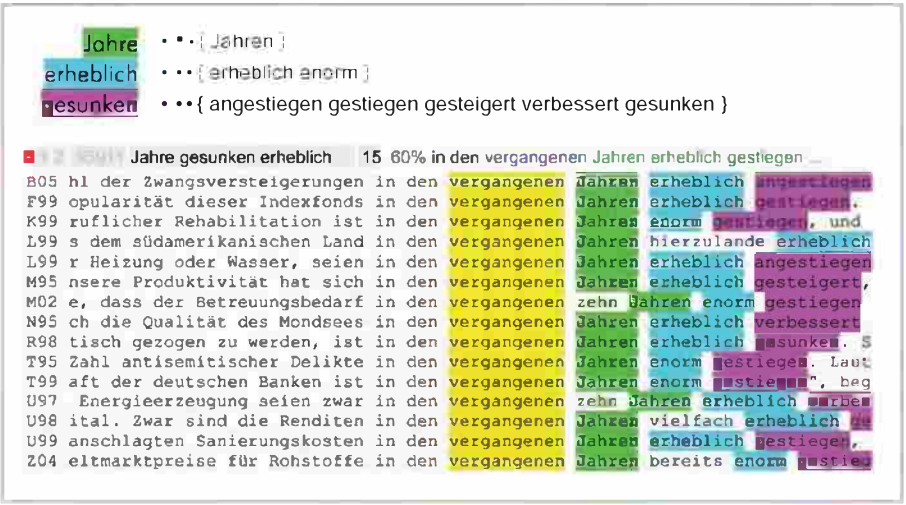


Figure 9: One SCC candidate for the lemma *vergangen* and some underlying concordances

As can be seen in the figure, the word order of the collocates is highly predictable in this example. In other cases, we observed a greater positional variability, as in the following example (Figure 10).

These figures are no full-scale representations of the identified SCC candidates. They only list the relevant paradigmatic classes and some underlying concordances which makes it easier for analyzers to refer to and talk about the SCC candidates, and to relate them to their language competence, but this kind of description does not fully capture the essence of the SCC candidate (cf. 4.1). Developing an adequate representation will be an important direction for future research (cf. Section 5).

Frühjahr • • { Herbst Frühjahr }
hatte • { hatte }
Bereits • • { bereits Bereits }

Frühjahr hatte Bereits 18 33% hatte bereits im vergangenen Herbst

```
B02 mensauflösung betroffen. Bereits im vergangenen Herbst hatte Comdirect angek
B05                  Bereits im Herbst vergangenen Jahres hatte eine vom Arktis
B06 Dieser Plan hatte jedoch bereits im vergangenen Herbst zu einen heftigen Koa
D06            Wie MacDonald bereits im vergangenen Herbst angekündigt hatte, so
E99 s Fortnum & Mason, hatte bereits im vergangenen Herbst mit dem begrenzten Ve
F95 ammlung hatte bereits im Herbst des vergangenen Jahres Papst und Bischöfe au
F01  Fachhochschule bereits seit Herbst vergangenen Jahres kommissarisch geleite
F05 gelegt werden. MMC hatte bereits im vergangenen Frühjahr von der Mitsubishi-
N91 Der 56jährige Bozo hatte bereits im vergangenen Herbst vor dem Hintergrund e
P99 interne Geschichten" an. Bereits im vergangenen Herbst hatte Windisch-Spoerk
R98 nd "kontrollieren" kann. Bereits im vergangenen Herbst hatte das Bundesfamil
R99  Doch sein ursprünglich bereits zum vergangenen Frühjahr geplanter Besuch ha
T95 arbiologe hatte bereits im Frühjahr vergangenen Jahres für Schlagzeilen geso
T01 g, Bärbel Grygier, hatte bereits im vergangen Herbst gefordert, dass Ausländ
T01 allerdings nicht. Bereits im Herbst vergangenen Jahres hatte das Fernsehmaga
U98 der Intendant ist. Er hatte sich im vergangenen Herbst bereits Hoffnungen au
G06 nische Unternehmen hatte bereits im vergangenen Herbst 22,5 Mrd. Euro für de
Z03 s Drehteam noch einmal anrücken. Im vergangenen Herbst hatte man bereits gef
```

Figure 10: Another SCC candidate for the lemma *vergangen* and some underlying concordances

Only space limitations prevent us from presenting more examples which together would demonstrate that the general strategy for subtask 2 indeed addresses a broad range of SCC candidates that vary in manifold respects, e.g., concerning their complexity (i.e., their number of collocates), their degree of abstraction (i.e., the size of their paradigmatic classes), or the distance between the collocates. All of these aspects suggest that SCC candidates may potentially delve deeply into what is commonly perceived as grammatical (rather than lexical) structure.

In interpreting these results, it is important to keep in mind that SCC candidates are derived from information that is entirely contained in the corpus. No external information is involved in the strategies for subtasks 1 and 2, especially no language competence (except for exploration and evaluation purposes). The relevant information is implicitly present in the corpus, distributed across many usage events, and the approach that we propose here is meant to uncover this hidden information by exclusively employing techniques that plausibly imitate the psychological processes underlying the acquisition and continuous emergence of schemas in language.

As a final remark on subtask 2, it is worthwhile pointing out that re-using collocation analysis for deriving SCC candidates is not only convenient but

also a way of avoiding over-generalization. For instance, suppose the result of subtask 1 included the following two paradigmatic classes for *vergangen*:

(7a) *Montag, Dienstag, ...*
Monday, Tuesday, ...

(b) *Woche, Monat(s), Jahr(es), ...*
week, month, year, ...

Given these classes, a naive solution to subtask 2 might infer from higher-order collocations (or more precisely: from syntagmatic patterns) of the form (8a) that a label such as (8b) constitutes a good SCC candidate.

(8a) *Montag vergangener Woche*
Monday last week

(b) {*Montag, Dienstag, ...*} *vergangene(r|n)* {*Woche, Monats, Jahres, ...*}
{Monday, Tuesday, ...} last {week, month, year, ...}

(c) {*Montag, Dienstag, ...*} *vergangener Woche*
{Monday, Tuesday, ...} last week

However, such an SCC would obviously be too general to correspond to anything in speakers' minds for it would be very surprising for someone to talk about "Tuesday last month" etc. A more realistic SCC candidate would be (8c). The collocate *Woche* most likely does generalize to a paradigmatic class like (7b) in some collocations of *vergangen*, but at the same time, it may have idiosyncratic properties in other collocations that are not shared by other members of (7b). More generally, in different collocations around the same node word, a collocate may instantiate different classes – including the primitive class consisting just of the collocate itself.

This observation emphasizes the necessity of subtask 2. The paradigmatic classes obtained from subtask 1 alone do not reveal much about the SCC candidates around a node word – each of these classes is likely to play a role in *some* SCC candidate, but one still needs to determine the specific SCC candidates in which they actually do, and in particular, the SCC candidates in which different classes combine.

Our provisional implementation of subtask 2 does not avoid over-generalized SCC candidates directly. However, over-generalizations are likely to receive a

low cohesion score (i.e., a low statistical significance) – and this score tends to be lower for a greater degree of over-generalization. Better treatment of the danger of over-generalization would be to work with an extended notion of paradigmatic classes that may overlap (i.e., partly include the same collocates; cf. 4.2.1) and to determine for any conflicting SCC candidates which of them would be an optimal generalization.

4.3 Stage 3: Evaluating the psychological reality of SCC candidates

In the previous subsection we described a general approach for automatically inducing SCC candidates from corpora, which involved an abstract strategy and a sequence of technical modeling decisions (e.g., choice of a corpus, similarity measures, a particular clustering algorithm, cut-off levels). Our competence-based evaluations sufficed to provide a general proof of concept, but given the ultimate goals of this line of research (cf. Section 2), a simple "looks good to me" evaluation is certainly not enough. Before SCC candidates can be used to indirectly study the real schemas that are entrenched in individual speakers and in a language community, their status as true SCCs must be established in a conclusive way.

What is needed is a systematic and rigorous evaluation of the derived SCC candidates in terms of appropriate psychological studies. SCC candidates that do not correlate to anything that is psychologically real may point either to systematic shortcomings of the general strategy, or to the deficiencies of its specific technical implementation, which in both cases might prove vital for critical revision and further improvements.

5. Future prospects

A corpus-based detection of genuine SCCs would enable us linguists to study through these SCCs the emergent schemas – i.e., syntagmatic-paradigmatic structures which are socially and psychologically real. Of particular interest might be questions like the following:

1) How do schemas operate in language processing?

2) What would an appropriate cognitive conceptualization look like?

While ultimately, such questions will necessarily involve, again, psychological investigations, it is possible to use SCCs to generate hypotheses about the nature of these structures. To this end, a good strategy would involve the following three steps. First, closely inspect a large number of corpus-derived SCCs and attempt to characterize them individually. Second, by abstracting across many SCCs, try to identify more general characteristics of this type of structure. Third, formulate these meta-descriptions as specific hypotheses (at the theoretical level) about the real schemas in language. This third step constitutes *abductive reasoning*: the inductive steps described in Sections 3 and 4 generalize from specific observations in the corpus data to more abstract structures, but all these structures still only have the status of descriptions, none of them reaches the theoretical level. To do that, abduction – in the sense of an "inference to the best explanation" (Harman 1965) – is required (cf. Figure 1).

To provide some guidelines with respect to the first step, a good starting point for characterizing a given SCC would be to inspect each of its classes relative to the SCC – and not just relative to its node word. This inspection could initially proceed along paradigmatic and syntagmatic lines. For the paradigmatic inspection, one could first attempt to describe commonalities between the class members observed in instances of this SCC. Based on these insights, one could then choose a name for the class which facilitates metadiscourse. Crucially, however, this name only constitutes a convention and is not to be confused with the class itself. An additional approach would be to attempt to generalize the class beyond observation and test its predictive power for unseen events. This may lead to insights into the dynamic nature and productivity of the SCC. After many SCCs have been inspected in this fashion, one further exciting question would be whether similar node words tend to have similar paradigmatic classes.

For the syntagmatic inspection of an SCC, one could start by attempting to characterize the relation between the SCC and each of its classes. Initially, this process should not involve pre-existing categories such as colligation, semantic preference (Sinclair 1998), or more traditional categories such as phrasal categories, subcategorization frames, thematic roles, etc. Eventually, after having inspected a broad range of SCCs in this way, the process might lead to the confirmation or modification of existing relational categories, or to the introduction of new categories, if inevitable.

Appendix

For our explorations, in order to operationalize the notions of associative-semantic and positional similarity between collocates, we took advantage of information already provided by the CCDB (Belica 2001-2007, Keibel/Belica 2007). With respect to the collocates' positional similarity, we used autofocus information that is available in the CCDB profiles for each primary collocate: the *positional focus* of a given collocate is the context window around the node word in which this collocation is most *cohesive* (i.e., statistically most significant, but not necessarily most frequent). In other words, it is a measure for the surface positions (relative to the node word) that the collocate occurs in most typically. For example, a collocate with the positional focus [-1;3] is likely to occur anywhere between one word to the left of the node word and three words to its right. Given this information, we defined the *positional similarity* between any two collocates of the same node word as the similarity between their positional foci.

The measure we used for quantifying the similarity between positional foci guarantees that two (nearly) identical positional foci are deemed the more similar the smaller they are because a smaller focus is more specific and thus conveys more information. For instance, a collocate that was assigned the largest possible focus – in the current online version of the CCDB this is the context window [-5;5] – essentially exhibits no positional preferences at all.

The operationalization of *associative-semantic similarity* between any two collocates *x* and *y* of a fixed node word is slightly more complex. To this end, we used the collocation profiles of the collocates, thus treating *x* and *y* themselves as node words.[4] The similarity between these two profiles then quantifies the degree to which *x* and *y* are used in similar ways. Formally, this similarity was assessed in terms of a measure that has proven to implement a plausible notion of similarity which is most sensitive to semantic and pragmatic factors, but also to other aspects of usage similarity between words (e.g., Belica et al. 2010).

Finally, to obtain a measure of the overall *paradigmatic similarity* between any two collocates, we combined the measures of their semantic and positional similarities (e.g., by means of multiplication). Future research should seek to match the operationalizations described in this appendix with available psychological evidence, and revise them if necessary.

[4] As node words are lemmas and collocates are word forms, we first had to lemmatize the collocates.

References

Belica, Cyril (1995): Statistische Kollokationsanalyse und Clustering. Korpuslinguistische Analysemethode. Mannheim: Institut für Deutsche Sprache. Internet: http://corpora.ids-mannheim.de (last visited: 11 / 2010).

Belica, Cyril (2001-2007): Kookkurrenzdatenbank CCDB. Eine korpuslinguistische Denk- und Experimentierplattform für die Erforschung und theoretische Begründung von systemisch-strukturellen Eigenschaften von Kohäsionsrelationen zwischen den Konstituenten des Sprachgebrauchs. Mannheim: Institut für Deutsche Sprache. Internet: http://corpora.ids-mannheim.de/ccdb/ (last visited: 11 / 2010).

Belica, Cyril / Keibel, Holger / Kupietz, Marc / Perkuhn, Rainer (2010): An empiricist's view of the ontology of lexical-semantic relations. In: Storjohann, Petra (ed.): Lexical-semantic relations: Theoretical and practical perspectives. Amsterdam et al.: Benjamins, 115-144.

Biber, Douglas (2009): A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. In: International Journal of Corpus Linguistics, 14, 3: 275-311.

Cheng, Winnie / Greaves, Chris / Warren, Martin (2006): From n-gram to skipgram to concgram. In: International Journal of Corpus Linguistics 11, 4: 411-433.

Croft, William (2001): Radical construction grammar: Syntactic theory in typological perspective. Oxford: Oxford University Press.

Fletcher, William H. (2003): Phrases in English (PIE). Internet: http://pie.usna.edu (last visited: 11 / 2010).

Harman, Gilbert (1965): The inference to the best explanation. In: The Philosophical Review 74, 1: 88-95.

Hoey, Michael (2005): Lexical priming: A new theory of words and language. London: Routledge.

Hopper, Paul J. (1987): Emergent grammar. In: Berkeley Linguistics Society 13: 139-157.

Hopper, Paul J. (1998): Emergent grammar. In: Tomasello, Michael (ed.): The New Psychology of Language: Cognitive and functional approaches to language structure. Mahwah, NJ: Erlbaum, 155-175.

Hunston, Susan / Francis, Gill (2000): Pattern Grammar: A corpus-driven approach to the lexical grammar of English. (= Studies in corpus linguistics 4). Amsterdam et al.: Benjamins.

Institut für Deutsche Sprache (2007a): Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2007-I (Release vom 31.01.2007). Mannheim: Institut für Deutsche Sprache. Internet: www.ids-mannheim.de/kl/projekte/korpora/archiv.html (last visited: 11 / 2010).

Institut für Deutsche Sprache (2007b): Virtual corpus "CCDB2007" composed from the German Reference Corpus (Institut für Deutsche Sprache 2007a).

Keibel, Holger / Belica, Cyril (2007): CCDB: A corpus-linguistic research and development workbench. Proceedings of the 4th Corpus Linguistics Conference, Birmingham. Internet: http://corpus.bham.ac.uk/corplingproceedings07/paper/134_Paper.pdf (last visited: 11 / 2010).

Keibel, Holger / Kupietz, Marc (2009): Approaching grammar: Towards an empirical linguistic research programme. In: Minegishi / Kawaguchi (eds.), 61-76. Internet: http://cblle.tufs.ac.jp/assets/files/publications/working_papers_03/section/061-076.pdf (last visited: 11 / 2010).

Keibel, Holger / Kupietz, Marc / Belica, Cyril (2008): Approaching grammar: Inferring operational constituents of language use from large corpora. In: Šticha, František / Fried, Mirjam (eds.): Grammar & Corpora 2007: Selected contributions from the conference Grammar and Corpora, Sept. 25-27, 2007, Liblice, Czech Republic. Prague: Academia, 235-242.

Kupietz, Marc / Belica, Cyril / Keibel, Holger / Witt, Andreas (2010): The German Reference Corpus DeReKo: A primordial sample for linguistic research. In: Calzolari, Nicoletta et al. (eds.): Proceedings of the seventh conference on International Language Resources and Evaluation (LREC'10), 1848-1854. Internet: www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf (last visited: 11 / 2010).

Kupietz, Marc / Keibel, Holger (2009a): The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In: Minegishi / Kawaguchi (eds.), 53-59. Internet: http://cblle.tufs.ac.jp/assets/files/publications/working_papers_03/section/053-059.pdf (last visited: 11 / 2010).

Kupietz, Marc / Keibel, Holger (2009b): Gebrauchsbasierte Grammatik: Statistische Regelhaftigkeit. In: Konopka, Marek / Strecker, Bruno (eds.): Deutsche Grammatik – Regeln, Normen, Sprachgebrauch. Jahrbuch des Instituts für Deutsche Sprache 2008. Berlin/New York: de Gruyter, 33-50.

Mason, Oliver (2004): Automatic processing of local grammar patterns. Proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics, January 6-7, 2004, University of Birmingham. Birmingham: University of Birmingham Press, 166-171.

Minegishi, Makoto / Kawaguchi, Yuji (eds.) (2009): Working Papers in Corpus-based Linguistics and Language Education, Vol. 3. Tokyo: Tokyo University of Foreign Studies (TUFS). Internet: http://cblle.tufs.ac.jp/assets/files/publications/working_papers_03/index.pdf (last visited: 11 / 2010).

Perkuhn, Rainer (2007): Systematic exploration of collocation profiles. Proceedings of the 4th Corpus Linguistics Conference, Birmingham. Internet: http://corpus.bham.ac.uk/corplingproceedings07/paper/132_Paper.pdf (last visited: 11 / 2010).

Renouf, Antoinette / Sinclair, John M. (1991): Collocational frameworks in English. In: Aijmer, Karin / Altenberg, Bengt (eds.): English corpus linguistics: Studies in honour of Jan Svartvik. London: Longman, 128-143.

Sinclair, John (1998): The lexical item. In: Weigand, Edda (ed.): Contrastive lexical semantics. Amsterdam et al.: Benjamins, 1-24.

Stubbs, Michael (2004): On very frequent phrases in English: Distributions, functions and structures. Plenary given at ICAME 25 (25th anniversary meeting of the International Computer Archive for Modern and Medieval English), Verona, Italy, May 19-23, 2004.

OLIVER MASON

Reconciling Phraseology and Grammar

Abstract

Grammar is concerned with describing the structure of sentences; phraseology is concerned with sequences of words that tend to occur together. Both can be seen to approach the description of linguistic structures from different directions, top-down in the case of grammar, bottom-up in the case of phraseology. In this paper we will argue that they have met in the middle – that grammatical descriptions can be achieved through using phraseological units, and that phraseology can be used for the description of structurally complete clauses and sentences. We will demonstrate the application of a method for analysing sentence structure with a number of examples, and will discuss the reasons why a) we believe the approach to be successful and b) what the possible explanations are for those situations in which it does not work.

1. Introduction

The primary mechanism for describing syntactic structures is based on rules – be it the rules of an evolving traditional grammar or the formal rules popularised in the middle of the previous century by Chomsky (1957). Over the past few decades, the basic phrase-structure approach has been refined and extended, and dependency has been used as an alternative driving principle besides constituency. One of the main problems of grammatical description concerns the interaction between the grammar and the lexicon: the final step in which actual words are supposed to be slotted in their respective places in the structural description.

While the initial assumption was that any noun can occupy a slot that calls for a noun, this was quickly realised as a drastic over-generalisation, and additional mechanisms were introduced to implement selection restrictions. Essentially the problem was to accommodate 'usage profiles' in abstract lexical categories. At the other end of the spectrum it was observed that words co-occurred in patterns whose regularity was not captured by systemic grammatical generalisations. Initially these regularities were described through collocation (based on Firth 1957) and the related phenomenon of discourse prosody (Louw 1993). Work by Sinclair on the environment of the word pair *naked eye*

(Sinclair 1996) extends this further, identifying a number of abstract categories that tend to co-occur with certain phrases (such as 'negative modality', or 'verbs of perception'). Current research in phraseology is mainly investigating word sequences of variable length, and also positional variation, which is a major complicating factor when working with semi-fixed phrases. Often, recurring sequences are relatively fixed, but can contain additional words or changes in the sequencing.

Unlike grammatical descriptions that look at sentence structures, phraseological units cannot easily be assigned category labels. Instead, they form an inventory of items that are used by speakers of a language, and more research is needed to investigate their usage regularities. That, however, does not mean that they are irrelevant to a grammatical description. It will be a different kind of description, which will tell us more about how 'typical' a clause or sentence is, rather than what its underlying constituent structure is.

In this paper we will first look at three approaches to the description of language structure which move further towards phraseology, and then introduce a new way of looking at grammar based solely on multi-word units extracted from corpus data.

2. Alternative approaches to language description

This section will look at three linear approaches to grammatical description. They are David Brazil's *Grammar of Speech* (1995), Susan Hunston and Gill Francis' *Pattern Grammar* (2000), and John Sinclair and Anna Mauranen's *Linear Unit Grammar* (2006). They all present more or less radical departures from traditional grammatical descriptions, which have hierarchical tree structures at their core. None of the three descriptions discussed here contains such a notion of hierarchy.

2.1 Brazil's *Grammar of Speech*

Brazil questions the notion of hierarchy in spoken utterances. In his *Grammar of Speech* (1995), he instead favours a linear approach, since speech is created incrementally in time. There are, he states, "obvious and well-recognized difficulties in reconciling this increment-by-increment presentation of speech with a hierarchical constituent-within-constituent account of how language is organized" (Brazil 1995: 4). Looking at 'telling exchanges', he establishes what

is required to fulfil the speaker's *purpose*, communicating information. This would typically be a nominal element followed by a verbal one. The sequence 'N V' is in his terminology a two-element chain; the first element alone would not be sufficient, leaving the utterance in an intermediate state. It is only the addition of the 'V' element which completes the utterance.

More information requires more elements in the chains used, and Brazil extends the description by taking into account longer chains. This he then represents in a network of sequencing rules which specify which elements can be added to a sequence at any given point to reach a 'target' state. Brazil acknowledges that this model is only suitable to describe simple utterances and subsequently elaborates the model further to include a wide range of grammatical phenomena. The key points of the grammar described are the use of increments rather than constituents, and the rejection of hierarchy in favour of linearity. Brazil still uses grammatical categories to label the various elements used.

2.2 Hunston and Francis' *Pattern Grammar*

In their *Pattern Grammar* (2000), Hunston and Francis synthesise previous approaches to the description of grammatical structure which are rooted in phraseology and avoid the traditional distinction between lexis and grammar. The resulting formalism consists of patterns describing typical environments of words, using a mixture of grammatical category labels and lexical items. For example, according to the Cobuild dictionary, *provide* has the patterns **V n** (*I'll be glad to provide a copy of this ...*), where the verb is followed by a noun group acting as a direct object, and **V n *with* n** (*The government was not in a position to provide them with food.*), where both the receiver and the object are included. A third pattern exists for a different sense, **V that**, as in *The act provides that only the parents of a child have the responsibility for that child's financial support*, illustrating the correlation between form (possible syntactic environments) and meaning (word sense). In a typical sentence, a number of patterns are involved. They can either overlap, or be adjacent to each other. The second example sentence in the previous paragraph could be analysed as follows:

	V	prep / adv						
			in	N				
				N	to-inf			
					V	n	*with*	n
The government	*was*	*not*	*in*	*a position*	*to provide*	*them*	*with*	*food*

The patterns associated with a number of words (*was*, *position*, *provide*) overlap, creating what Hunston / Francis call 'pattern flow' (2000: 215). This flow works on the principle that one element of the clause prospects what follows, or sets up an expectation: the elements fulfilling the expectation in turn set up their own expectations, which need to be fulfilled before the clause is complete.

The mixture of lexical items with grammatical categories allows for a more adequate level of granularity: the verb *be* can be followed by a number of different prepositional groups, but *position* is usually preceded by *in*, and *provide* (in this sense) is only used with *with*; there could in principle also be a pattern **V n prep n**, but in practice the choice here is too limited, and only a number of instances of this pattern with actual lexical items exists. An example of a more general pattern would be **V n prep / adv**, as in *Andrew chained the boat to the bridge.*

2.3 Sinclair and Mauranen's *Linear Unit Grammar*

An even more radical departure from traditional grammar is linear unit grammar (LUG, Sinclair / Mauranen 2006). Here all grammatical categories are abandoned, and the basic unit of analysis is the 'chunk'. A chunk is a sequence of words which form a coherent unit. This is not defined further, and thus remains a subjective element of the description. Two kinds of chunks are distinguished, those that organise the message, and those that carry information. Sinclair and Mauranen further specify a number of sub-types of those two main categories, to allow for discontinuities in the stream. While LUG is initially geared towards spoken discourse, it can also be applied to written texts, as in the following example taken from the *Independent* newspaper:

1. *Mr Kennedy* (M–)
2. *now* (OT)
3. *declares* (+M–)
4. *that* (OT)
5. *it must be* (M–)
6. *bold* (+M)
7. *in its thinking* (MS)
8. *and* (OT)
9. *ready to plan* (MS)
10. *long-term.* (MS)

Each item in the above list represents a separate chunk, with the following chunk labels:

- M– incomplete message unit
- OT text-oriented organisational element
- –M+ partial completion of message unit
- +M completion of message unit
- MS supplement to message unit

The example is very heavy on message-oriented chunks, as would be expected from a newspaper text. Other kinds of texts might contain more organisational elements.

LUG is a refreshingly different approach to grammar which offers a new perspective on the description of linguistic structures. Unfortunately, Sinclair and Mauranen are not ambitious enough when discussing its applications. They seem to envisage it being used mainly to 'clean up' messy spoken discourse as a pre-processing step before the application of traditional grammars. Essentially, all type-O chunks would be removed, and the type-M chunks put together to provide a tidy sentence more amenable to grammars developed with written language in mind.

3. The multi-word unit in a phraseological analysis of clauses and sentences

Taking up and developing ideas from the three models described above, the approach to phraseologically based analysis advocated in this paper is non-hierarchical, assumes units that can overlap ('flow' into each other), and does not make use of abstract categories. Traditional grammatical units and categories are dispensed with in favour of phraseological units, which are derived empirically from corpus data and would not normally correspond to pre-defined grammatical units. Obviously, the procedure is not without risk, as it poses a serious issue of evaluation. Since in this bottom-up approach we do not know beforehand what to expect (and not even what exactly we are looking for), we cannot easily tell whether our search is successful or not.

At the core of the phraseological description of language is the multi-word unit (MWU), akin to what Sinclair (1996) and Danielsson (2001) refer to as a *unit of meaning*. This is based on the observation that single words in isolation mostly have no unambiguous meaning. Instead they have a meaning potential, realised in a particular context. This context would typically be provided by the other elements that make up the units of meaning. An example from Danielsson is *stroke*, which among others forms the meaning units *on the stroke of half-time*, referring to a particular point in time, and *a stroke of genius*, describing a particular action or idea. Given those two units, it does not make sense to ask what the meaning of *stroke* on its own is.

There are several possible ways of constructing multi-word units. A common method is to use sequences of words (*n*-grams) of either fixed or variable length. Danielsson (2001) argues for using collocations to expand a node word into a larger unit. In this section we will describe two approaches to the identification of multi-word units, originally inspired by the chunking of linear unit grammar. Multi-word units appeared as a good starting point for an objective specification on chunks, though it remains uncertain if intersubjectivity is high on the agenda here. The two approaches are meant to complement each other, producing a number of candidate multi-word units from a text (or a set of concordance lines). Units can then be weighted according to their length and frequency to create a list of the most relevant ones, but this is not necessary for using it to describe the grammatical structure of a sentence. It might be useful for identifying which analyses are more likely than others, though this has yet not been investigated.

3.1 Frames

The first algorithm is related to the concept of collocational frameworks (Sinclair / Renouf 1991). Starting with a frame such as *the __ of*, or *as __ as*, they investigate which other words can occur in the gap. When using this procedure to identify MWUs, we simply turn the sequence around: we start with the word in the blank slot, and see what other word we can attach to it. The criterion for attaching a word is its frequency: if the next word to the left (or right) of the starting word is equally or more frequent, then we attach it to our larger unit; if it is less frequent, we stop. The MWU thus grows in two directions, both to the left and to the right.

The underlying idea is that there are content words, which are lower in frequency, and connecting grammatical words, which have a higher frequency. The function of the grammatical words is to act as 'glue' between content words. It is a bit like breaking up a brick wall: typically one ends up with individual bricks that still have some parts of mortar attached to them. The content words here are like the bricks. Instead of having an a priori specified set of grammatical 'mortar' words, we simply use each word's frequency, as it does not involve any bias. In order to account for random variation in observed frequencies, we convert actual frequency values into frequency bands (Quasthoff 1998). This makes the whole procedure more robust against individual variation.

Consider the (randomly selected) sentence fragment *Schools are invited to register for a free online education resource pack* [...]. Looking at each word in turn, we would get the MWUs shown in the table on the following page (frequency values based on the written part of the BNC).

The least frequent word obviously selects the whole sentence as a MWU, which would at a later stage be filtered out when considering the overall frequency of occurrence of each MWU candidate. Alternatively, the algorithm could be changed to require a strictly monotonous change in frequency only, in which case we would get the MWU *a free* ***on-line*** *education*, rather than the full sentence. High-frequency words (such as *are*, *to*, and *a*) do not form MWUs as core words, as their neighbours are too infrequent. Interestingly, this also applies to *education* in this example, which, however, is included in two different MWUs triggered by its surrounding words.

word form	raw frequency	frequency band	MWU (core word in bold)
the	*5 143 707*	*0*	(*for calculation only*)
schools	12 801	9	**schools** are
are	412 256	4	–
invited	4 046	10	are **invited** to
to	2 403 172	1	–
register	2 562	11	schools are invited to **register** for a free
for	783 824	3	**for** a
a	1 898 737	1	–
free	20 558	8	for a **free**
online	463	13	schools are invited to register for a free **online** education resource pack
education	18 386	8	–
resource	2 191	11	education **resource** pack
pack	2 680	11	education resource **pack**

3.2 Chains

While Frames are based on the individual frequency (band) of their component elements, Chains are simply *n*-grams. They are calculated for all values of n between 2 and 8; looking at the example from the previous section we would get the following MWUs for the core word *register*:

(1) *to register, register for*

(2) *invited to register, to register for, register for a*

(3) *are invited to register, invited to register for, to register for a, register for a free*

(4) *schools are invited to register, are invited to register for, invited to register for a, to register for a free, register for a free online*

(5) *schools are invited to register for, are invited to register for a, invited to register for a free, to register for a free online, register for a free online education*

(6) *schools are invited to register for a, are invited to register for a free, invited to register for a free online, to register for a free online education, register for a free online education resource*

(7) *schools are invited to register for a free, are invited to register for a free online, invited to register for a free online education, to register for a free online education resource pack*

There is an obvious trade-off between length and frequency of MWUs: shorter ones are much more versatile and can be re-used in different contexts more often as they contain less information; longer ones, on the other hand, are more specific, and hence can be less easily used, which results in a lower overall frequency.

3.3 Interim summary

Both ways of identifying multi-word units described in the previous section are used to obtain a set of candidates. The original way in which they were used in Mason (2006) was as part of a lexical item's profile. For this, each MWU candidate would be assigned a weighting based on its length (in component words) and frequency (in a reference corpus). This list would then be sorted according to the MWUs' weightings, and all MWUs with a weighting of less than 10% of the highest ranking weight would be discarded. For a grammatical description of an utterance this filtering step is not necessary, so we simply keep all MWU candidates. We will discuss the application of MWUs to grammar in the following section.

Danielsson (2001) uses collocation as the driving principle behind selecting constituent parts of MWUs. While this leads to comparable results, we would argue that the causal relationship is reversed: collocations are a side-effect of phraseology, rather than an explanation for it. Words that form larger units with each other will co-occur out of necessity, and thus statistical methods for extracting collocations will pick them up. However, if we take MWUs which have been selected purely on grounds of frequency differential between their elements, or recurrence of variable-length *n*-grams, then we can observe that a mere frequency list of the elements of those units will provide us with a list of words that seems close to what we would call collocations. There are of course differences, as there are a number of different procedures to compute colloca-

tions (with no obvious preference of one over the other), but here we have a single method leading to such a list without the need for complex statistical significance measures.

One weakness of the approach to MWU extraction described here is that it is restricted to those sequences which are repeated verbatim, without any variation. This produces many MWUs which vary slightly from each other, even though they should probably be seen as related variants of the same underlying phraseological unit. The same applies to discontinuous structures. Such "linguistic skeletons of full-bodied phrases" (Sinclair 2008: 408) cannot be discovered with the current procedures. However, one possible way to advance the study of multi-word units would be to look at the inventory of phrases that have been found in order to identify similarities between them, and to find phrases which are variants of each other.

Another issue is that we cannot be sure that the MWUs we extracted have any 'real' existence if judged by external standards such as semantic interpretation or psycholinguistic processing. We are, in a sense, in a dark room, trying to grope for something vague and fuzzy, and even if we can identify something, we cannot be sure that it is the item we were looking for. All we can do is to apply general principles (based on, for example, the length / frequency trade-off, or perhaps measures from information theory or quantitative linguistics) and see how far we get with what we have found.

4. Stringing together a sentence

The initial idea of using multi-word units in a grammatical analysis was to model the chunking used in Sinclair / Mauranen (2006). The chunks posited in linear unit grammar are subjective, and not specified in a way that can be algorithmically derived from text data. However, multi-word units of the kind described in the previous section might be a first approximation to their identification. The phraseological structure of a sentence is identified by matching pre-existing multi-word units against the word sequences in the sentence. Matching units can then be tabulated as in the example below (from Mason 2008: 238). Note that some stretches of the text are covered by several multi-word units. The core word of each MWU is given in bold.

at	the	top	of	the	stairs	was	...
at	the	top	of				
at	the	top	of	the			
at	the	top					
at	the	top	of	the	stairs		
	the	top	of	the			
	the	top	of	**the**			
	the	top	of				
	the	top	of	**the**	stairs		
	the	top	of	the	stairs		
		top	of	the			
		top	of	the	stairs		
at	the	top	of	the	**stairs**	was	

A more concise representation of the degree of matching can be achieved by simply showing continuous stretches, with non-overlapping boundaries marked through a vertical bar ('|') and words which are not part of a matching MWU enclosed in parentheses:

```
The Laird of Raasay | perceiving the ship in the harbour
(went aboard) to buy (some) wine and other commodities
```

The sentence (an example from Mason 2006) was selected randomly (based on the occurrence of *ship*) from the internet. As far as I am aware, it did not form part of the corpus that was used to generate the multi-word units. Its analysis illustrates several points:

- The first break (between *Raasay* and *perceiving*) coincides with a phrase boundary in a traditional description.
- The second segment (up to the gap in the coverage after *harbour*) is equivalent to a non-finite clause.
- The remaining covered stretches are an infinitive and a partial object.
- The last gap in the coverage concerns an optional element (*some*).

While these similarities to traditional ways of describing syntax are encouraging, this is not a necessary requirement or even desired outcome of the phraseological approach, as the latter constitutes an independent way of analysis.

5. Gaps in the description

There are, of course, gaps in the description of most sentences. Application of the phraseological algorithm to a set of authentic sentences (see Mason 2008) has shown that the method works well with more formulaic data, less so with more creative writing, and breaks down completely with an artificially constructed example that Hoey (2005) used to demonstrate what a sentence would look like if the effects of lexical priming were negated. This outcome, that most 'normal' or authentic language has a reasonable coverage using the phraseological analysis, but that artificially reformulated sentences cannot be analysed, might hold an answer to the question posed by Stubbs (2001: 59): "[w]hy is it that some language sounds natural, whereas other language, which is fully grammatical, 'doesn't quite sound right?'": Natural-sounding language is constructed from multi-word units, whereas not-quite-right language is constructed by combining individual words taken out of their usual contexts. While this explains the complete breakdown of the analysis for Hoey's artificial sentence, it cannot give a reason for the gaps occurring in the coverage of the authentic examples (see *went aboard* and *some* in the sentence described above). However, since the phraseological approach did mostly succeed, it remains very promising despite those gaps, since, after all, as Sinclair (1987: 158) aptly put it, "grammar is not easily applied to text". The gaps in coverage are usually due to two main reasons.

5.1 Incomplete data

If we assume that an individual's language experience shapes the way they acquire a language, then the (hypothetical) corpus of all the language they encountered in their lifetime should be the best choice as the data set from which to generate MWUs. Even if we were able to construct such a corpus, however, it would still be incomplete, as other speakers would have a different set of utterances that they base their usage on. Despite those issues we are still able to understand each other. Unknown words can usually be decoded from context (unless they are obscure terms in non-specific contexts), and so we should also be

able to deal with unknown phraseology in a similar way. Certain phrases might not be recognised as idiomatic until they have been encountered several times.

Using currently available corpora for the extraction of multi-word units will therefore always leave gaps. Not only are modern corpora still comparatively limited in size compared to a speaker's lifetime exposure to language; they are also typically not broad enough in scope. Leaving aside the issue of representativeness (see Rieger 1979 for a discussion of why corpora can never be representative), modern corpora tend to focus on written language, whereas most speakers would experience far more spoken language. Unless we are at some point in time able to record all linguistic input of a human being, we will have to live with the fact that corpora can only approximate any individual's language experience in a very limited way.

5.2 Idiom Principle vs. Open Choice

The second possible explanation for gaps in the coverage of phraseological syntactic analysis is that only parts of grammar work in the pre-supposed way: joining together pre-fabricated pieces of language to form a whole. This so-called idiom principle, the tendency of words to influence which other words occur near them in a syntagmatic sequence, is in opposition to the open-choice principle, which states that words in a paradigmatic relationship can be substituted for each other. While there are occasions where words can indeed replace each other, the scope for doing so seems to be much more restricted than originally thought by, for example, the proponents of phrase structure models of grammar. In practice, most words seem to interact with their environment in such a way that they cannot simply be taken out and be replaced with other words (see Hoey 2005 for an example of a sentence constructed in such a way). Notable exceptions here are proper nouns, as everybody with two or more children will know: it is very easy to choose the wrong name when addressing a child, as names will often have no distinct lexical environment that aids in the selection process. Other words are confused much less frequently, however.

In the example sentence described above, the word *some* has not been attached to any MWU. This could be an instance of the open-choice principle, as it is a point of variation: a set of other words can be substituted here, without necessarily affecting the idiomatic nature of the sentence. In the context of this sentence, *some* seems quite appropriate, but other quantifiers would also be pos-

sible, including not using a quantifier at all. A more complete description of grammar would have to be hybrid, taking into account both phraseology (the idiom principle) and possible slots for paradigmatic selection (the open choice principle). From this perspective it is not surprising that a purely phraseological model is not able to provide complete coverage.

6. Concluding thoughts on grammar and phraseology

Stubbs (2001: 120) argues that "a theory of language must find a balance between creative and routine language use". A hybrid model as outlined in the previous paragraph seeks to provide such a balance, whereby the routine aspects are covered through multi-word units, with the additional possibility of having scope for open choices which diverge from the routine and instead be more creative.

The lack of more abstract grammatical categories can be seen as a weak point of the description presented in this paper; however, one can also argue that they are not strictly necessary. If lexis and phraseology are the driving forces behind the creation of utterances, then one could view syntax as a mere by-product of linearising thought. Just as collocations are an epiphenomenon of phraseology, syntax could be an epiphenomenon of phraseological patterning, in that MWUs would typically be arranged in certain common ways to support the hearer's expectations and thus ease understanding. Traditional syntactic descriptions would then pick up on regularities in that patterning, but rather than providing a true explanation for the observed structural arrangements, the posited grammatical patterns and constructions would be based on a post-hoc fallacy.

Sinclair and Mauranen have demonstrated that one can get quite far with a grammatical analysis using only a minimal set of categories, and by shifting more towards a lexicalised approach, we could probably do without most traditional categories. One difficulty is assessing the explanatory power of such an approach, as we will typically not be able to assign any structural or even functional categories to the MWUs we can identify in a sentence. This then means we would essentially downgrade syntax to merely the mechanical arrangement of larger units which would have no contribution to make to the structure of an utterance. The latter would simply be a side-effect of the (preferred) sequencing of the MWUs.

The division of utterances into phraseological units reminds us of Harris' procedure for identifying morphemes from a stream of phonemes (Harris 1955). He postulates that morpheme boundaries are typically found at those places in an (unsegmented) stream of phonemes where there is a local maximum regarding the possible options for the following phoneme. This procedure does not assign categories to the identified morphemes, and thus would not be different in principle to locating phraseological boundaries in an utterance.

Even if we are only at the very beginning of a phraseological approach to grammar, the initial results are quite promising. Given the lack of any major preconceptions and theoretical assumptions about language, it is surprising how well the mechanism described in this paper can indicate the degree of 'naturalness' of a given sentence. Using this as a starting point for further study, especially extending the basic MWUs by including positional variations and insertions/deletions, we can begin to see how traditional grammar could possibly be reconciled with a corpus-based phraseological analysis of 'grammar in use'.

References

Brazil, David (1995): A grammar of speech. Oxford: Oxford University Press.

Chomsky, Noam (1957): Syntactic structures. The Hague: Mouton.

Danielsson, Pernilla (2001): The automatic identification of meaningful units in language. Språkdata, Department of Swedish, Göteborg University.

Firth, John Rupert (1957): Papers in linguistics. London: Oxford University Press.

Harris, Zellig S. (1955): From phoneme to morpheme. In: Language 31, 2: 190-222.

Hoey, Michael (2005): Lexical priming: A new theory of words and language. London: Routledge.

Hunston, Susan / Francis, Gill (2000): Pattern grammar: A corpus-driven approach to the lexical grammar of English. Amsterdam: Benjamins.

Louw, Bill (1993): Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In: Baker, Mona / Francis, Gill / Tognini-Bonelli, Elena (eds.): Text and technology. Philadelphia / Amsterdam: Benjamins, 157-176.

Mason, Oliver (2006): The automatic extraction of linguistic information from text corpora. PhD diss., Univ. of Birmingham.

Mason, Oliver (2008): Stringing together a sentence: Linearity and the lexis-syntax interface. In: Gerbig, Andrea / Mason, Oliver (eds.): Language, people, numbers: Corpus linguistics and society. Amsterdam: Rodopi, 231-248.

Quasthoff, Uwe (1998): Deutscher Wortschatz im Internet. In: LDV-Forum 15, 2: 4-23.

Rieger, Burghard (1979): Repräsentativität: Von der Unangemessenheit eines Begriffs zur Kennzeichnung eines Problems linguistischer Korpusbildung. In: Bergenholtz, Henning / Schaeder, Burkhard (eds): Textcorpora: Materialien für eine empirische Textwissenschaft. Kronberg: Scriptor, 52-70.

Sinclair, John (1987): The nature of the evidence. In: Sinclair, John (ed.): Looking up: An account of the COBUILD project in lexical computing. London: Collins, 150-159.

Sinclair, John (1996): The search for units of meaning. In: Textus 9, 1: 75-106.

Sinclair, John (2008): The phrase, the whole phrase, and nothing but the phrase. In: Granger, Sylviane / Meunier, Fanny (eds.): Phraseology: An interdisciplinary perspective. Amsterdam: Benjamins, 407-410.

Sinclair, John / Mauranen, Anna (2006): Linear Unit Grammar. Amsterdam: Benjamins.

Sinclair, John / Renouf, Antoinette (1991): Collocational frameworks of English. In: Aijmer, Karin Altenberg, Bengt (eds.): English corpus linguistics: Studies in honour of Jan Svartvik. London: Longman, 128-144

Stubbs, Michael (2001): Words and phrases: Corpus studies of lexical semantics. Oxford: Blackwell.

MILENA HEBAL-JEZIERSKA / NEIL BERMEL

Frequency and oppositions in corpus-based research into morphological variation

Abstract

Research into variation using corpora is changing our methods of describing variation itself. The concept of *opposition* has proved useful as a starting point. Through comparing the relationship between two or more morphological variants, we can reach a fuller picture of the linguistic situation. Of special interest are the mutual frequency relationship and the distribution of elements investigated.

In our contribution we will attempt to classify variants from the point of view of their role in an opposition and will exemplify this on Czech morphological variation. Oppositions can be constructed on the basis of the interplay of several different factors, for example: the relationship of frequency and acceptability of data; the relationship between frequency and usage in a certain text type; the distribution of forms in phrases and co-texts, etc.

We focus on categories determined by the frequency of the items in question. Some researchers have used frequency labels such as *dominant*, *majority*, *equifrequential*, *minority*, *sporadic* (see Šimandl 2008, Hebal-Jezierska 2007), or *central* and *peripheral* (see Tušková 2006). We attempt to ascertain what these labels tell us and how they relate to the judgments of native speakers and to usage patterns (see Bermel in press). Here we believe morphology offers a particularly useful tool for analysis, as morphological variation (as opposed to syntactic variation) most often occurs in a closed set of two or perhaps three variants, and as such, data from one variant can provide useful information about the status of the other(s).

Our conclusions are specifically directed at the situation in Czech, but we believe that these suggestions may be more widely applicable.

For a number of years we have been focusing on the use of corpus data to give a perspective on the grammaticality of variation in language and specifically on variation between grammatical forms in Czech. We approach this as *non-native speakers* of the language we study. This perspective means we have limitations on how we approach our data. Most importantly, we do not have any *native intuition* in Czech against which we can measure our results. Moreover, we have become convinced that relying on intuition is a dangerous way to use corpus data. If the results of corpus searches are only used to confirm a native speaker's intuition of what is 'good' or 'odd', then results that do not conform to intuition are in danger of being discarded, regardless of their potential value.

In this article we will suggest some objective measures that can be used to evaluate corpus data. These measures come from studies of nominal morphology in Czech. We do not claim that they are universally valid and admit that their validity could in fact be quite limited. We will discuss them here so that they can be taken and tested elsewhere.

It will be useful to begin with a few words about the language we are studying. Czech is a morphologically rich inflected language. In its nominal morphology, it has seven cases and three genders. Each gender has several major declension patterns. There are numerous points in these declension patterns where two, or occasionally more, grammatical endings are found in a single slot. In describing which forms are considered 'correct' for which slot, grammars oscillate between categorical pronouncements and vagueness.

Our goal has been to find a way of using corpus data to arrive at descriptions that more accurately reflect usage. In doing so, we have found it useful to rely on the overall picture of usage of forms within a single environment. In other words, in developing our analysis, we pay attention to the *competition and coexistence* of forms, and also look at the prosperity of individual forms. We will call this 'competition and coexistence' an *opposition*.

We are interested in *morphology* and thus have an advantage over syntacticians in that morphological oppositions are often more limited than syntactic ones. In any specific morphological context there are usually no more than two variants commonly possible.[1] In rare instances there might be three and unless strongly marked varieties of the language are considered, we almost never see four possibilities. We therefore always discuss one variant against the background of the other variants (we could call this a *proportionality principle*) and try, where possible, to use information about one variant to deduce information about another (we could call this a *complementarity principle*).

Looking through the literature on corpus data and morphology, we found two basic approaches. The first is to use raw data only, in other words, simply recording the numbers of competing forms returned by a corpus search. In doing so, scholars also set limits on what constitutes a sufficient number of occurrences to describe the feature studied.

[1] A *possible variant* is one actually attested within the declension pattern. That means that e.g. for masc. anim. nouns *-i*, *-ě*, *-ově* are possible variants, but *-y*, *-e* are not, although they form the nom. pl. elsewhere. For many words, we might only consider one or two of these as the others are never found (**turistově*, **Iré* / **Iři*).

An example of this is Tušková's (2006) study of variation in the endings of feminine nouns. According to her, there must be at least 20 attestations and these must fulfil certain conditions of representativity (in other words, they must not be stylistically limited, found in a single source, and so forth.). She uses the words 'central' and 'peripheral' to compare two competing endings. However, she does not define them in any way.

A second way to use the data is to classify them into frequency bands. This approach is, by its nature, *proportional* and relies implicitly on the existence of an opposition.

For Czech, one such method is presented by Šimandl (2010). He expresses the frequency of variants in percentages, describing the grammatical features with the following scale (where *A* and *B* are variants and *f* is frequency).

1) A has f ≤ 5 % and is labelled the **marginal** (*marginální*) variant,
   B has f > 95.0 % and is labelled the **monopoly** (*monopolní*) variant
2) A has f <5.1-39.9> % and is labelled the **minority** (*minoritní*) variant,
   B has <94.5-60.1> % and is labelled the **majority** (*majoritní*) variant.
3) A and B have f <60-40> % and are labelled the **equity** variants
4) A has f <60.1-94.5> % and is labelled the **majority** variant,
   B has <39.9-5.1> % and is labelled the **minority** one
5) A has f > 95.0 % and is labelled the **monopoly** variant,
   B has f ≤ 5 % and is labelled the **marginal** one

Table 1: Corpus frequency (Šimandl 2010)

Šimandl does not use this scale to describe variation inside the frequency band in any way. He does, however, label forms according to the scale. For example: in analysing the genitive singular of the word *ječmen*, the possible ending *-e* is described as the *monopoly* ending. The scale is useful as a shorthand for classifying corpus frequency.

However, there is a common issue surrounding all scales based primarily on frequency. It is not clear whether they give us any information about the behaviour or character of the variant in question within the corpus, or whether they describe any external reality. Halliday (1991a, 1991b, 1992) attempted to marry corpus frequency and external reality by comparing data about syntac-

tic constructions in Chinese to principles of information load. He found that in cases of simple oppositions, where there are no interactions with other features, the ratio of 9 : 1 was very common. Halliday attributed this to findings in psychology showing that at a ratio of 9 : 1, people easily label one item as 'unmarked' and the other as 'marked'. Where ratios were more evenly balanced, there was always an interaction with other features. It might be, for example, that a particular expression or context favoured the use of one construction or the other.

In Halliday's model, then, the opposition plays a significant role. We do not look exclusively at the raw frequency of individual features; instead, they are always considered against the frequency of a complementary feature. If Halliday is correct, then a ratio of 9 : 1 is effectively the tipping point between 'features in competition' and 'marked vs. unmarked features'.

Another scale, proposed in Hebal-Jezierska (2007), looks explicitly for such information about behaviour in defining its bands. According to her, the dominant endings oscillate between 85 and 100%. The variant endings are those which do not occur sporadically or dominantly (her analysis always starts with the band of the infrequent variant):

- 85-100%: dominant
- 1-84%: variant
- sporadic: not defined numerically but rather by non-representative occurrence, for example: unusual / specific style, etc.

One disadvantage of this scale is the range of the middle band. However, as we will see, the *complementarity principle*[2] can increase the usefulness of this banding.

The final method we will look at here is the relating of corpus findings to some external reality, in this instance studies of acceptability conducted on native speakers. Here we can cite studies of Czech morphology such as Bermel (2009), which draw on earlier work in syntax such as Divjak (2008) and Kempen / Harbusch (2005). On the basis of this, Bermel proposes the following scale:

- sporadic (under 2%, predicts low acceptability)
- minority (2 to 49%, cannot predict acceptability)
- majority (over 50%, predicts high acceptability)

[2] The *complementarity principle* or its elements are also used in Šimandl (2010), Tušková (2006).

Now we will return to look at the two scales proposed and see how viewing them through the prism of opposition allows us to show a common view of the relationship of two variants.

1. Commonalities in the behaviour of forms within bands

The following examples, drawn from Hebal-Jezierska (2007), concern the considerable variation in form that is found throughout the declension patterns of masculine animate nouns. In Czech, this group is treated as a subgender, with a separate group of distinct endings and agreement patterns.

In the dative and locative singular, for example, the so-called *soudce* paradigm, which includes masculine animate nouns ending in *-e* in the nominative singular, has two potential endings for the dative and locative singular: *-i* and *-ovi*. These endings have the following frequency in the corpus:

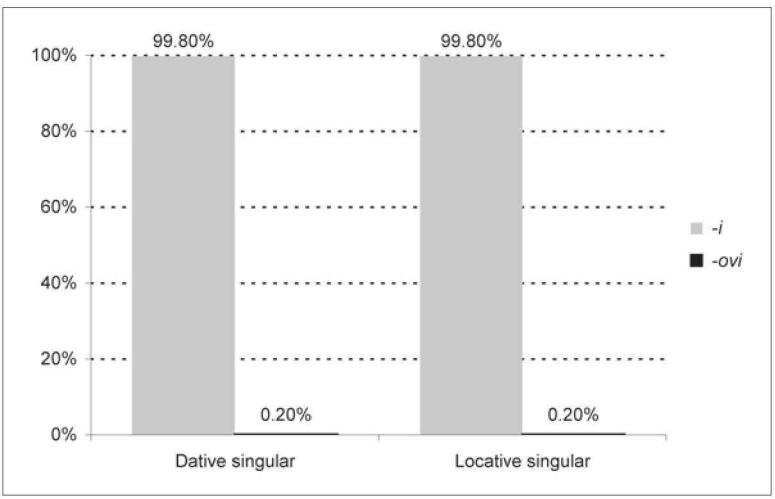


Table 2: Total frequency of the occurrence of inflectional endings *-ovi* / *-i* in the dative and locative singular (paradigm *soudce*)

The imbalance (one variant close to 0%, the second one close to 100%) in our findings very often indicates the exceptional occurrence of the infrequent variant or its acceptability within only a small group of words. A simple sorting of the examples will show whether we are dealing with the second type of result. For the first type of result, there is often a stylistic factor that has an influence on the choice of variant. For example, it may concern spoken language (in

written corpora) or a kind of stylization (for example, archaic style etc.). Ascertaining this requires some more detailed analysis by text type or by date, or may require examination of individual examples.

Our second example concerns the nominative plural of these animate masculine nouns, in particular the so-called hard declension pattern exemplified by the noun *pán*.

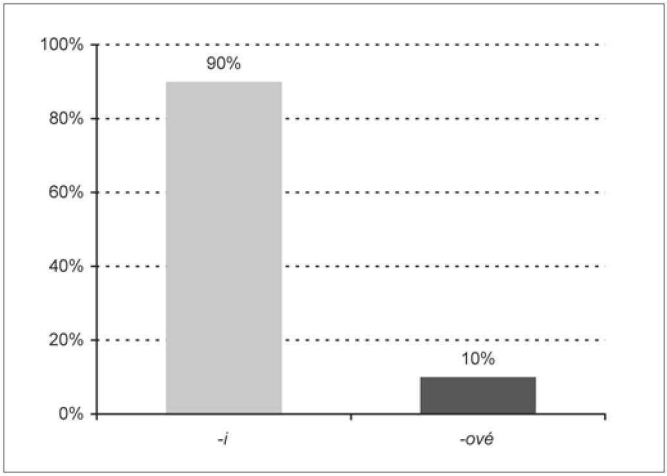


Table 3: Total frequency of the occurrence of inflectional endings *-ové* / *-i* in the nominative plural (paradigm *pán*)

We can see a second type of imbalance in Table 3 (one variant close to 10-15%, the second one 85-90%). This often indicates a *system variant*[3] or the existence of a small group of words having the infrequent variant.

By a system variant we mean a variant occurring in a word where the alternate variant is stylistically marked or sporadically attested in the corpus.

The dominant ending *-ové* appears in the corpus with the following nouns: *živočich* (99 %), *mustang* (100 %), *tvor* (99 %), *kur* (100 %), *šiml* (100 %), *makak* (100 %), *mýval* (99 %), *racek* (83 %). The reason for the predominance of the morph *-ové* in these expressions lies in their structure, euphony of phones, the presence in the language of homonymous and quasi-homonymous forms, but primarily of course in usage. Opaque morphological structure (here mostly borrowed or unmotivated words) conditions the use of the ending *-ové*,

[3] The system variant is the only possible inflectional ending (according to the system).

which gives the word a more apprehensible format.[4] There is a tendency to avoid consonant alternations.[5] (Quasi-)homonymous forms are found with the words *šiml*, *mýval*, *racek*.

These corpus results sound all the more interesting when compared with the current codification. For all the nouns mentioned above aside from *mýval* and *racek*, the morph *-ové* is the only codified ending.

PDP se vzepjaly jako <mustangové> a divoce se roztočily. (***marcin)***
'The diving transport devices reared up like <mustangs> and spun wildly round'

<Živočichové> už to takhle dělají dlouho, velice dlouho. (***matskol)***
'<Mammals> have been doing it that way a long time, a very long time.'

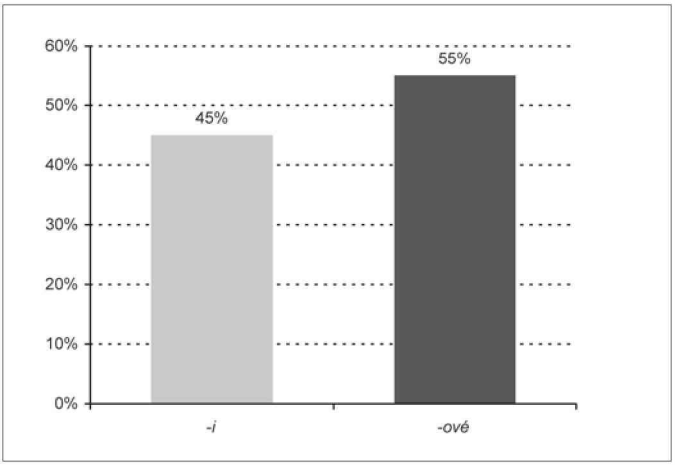


Table 4: Total frequency of the occurence of inflectional endings *-ové* / *-i* in the nominative plural (paradigm *despota*)

The third graph shows variation between two forms of nearly equal frequency. In our research the total frequency often shows that the variants are not determined by any factor or indicates the occurrence of a large group of words (used in the language) having only one of two variants.

Our final example here concerns differentiation between the standard and non-standard language layer. This can be seen in the nom. pl. ending of masculine animate nouns ending in *-asta*, *-ista*, *-ita*.

[4] This idea is explained in Rusínová (1987: 73).

[5] Rusínová (1987: 74) goes into this factor in more detail.

If we examine corpora frequency in Common Czech and in Standard Czech, we see that sometimes quite extreme differences can be explained by the fact that one variant is restricted to a particular register or variety.

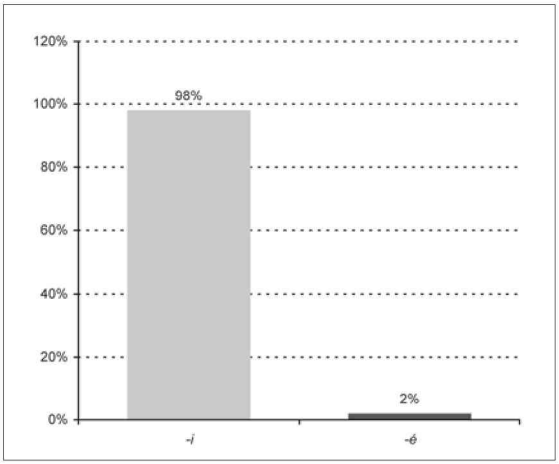


Table 5: Total frequency of the occurrence of inflectional endings *-é* / *-i* in the nominative plural in words ending in *-asta*, *-ista*, *-ita* in Common Czech (in the corpus: ČNK – ORAL2006)

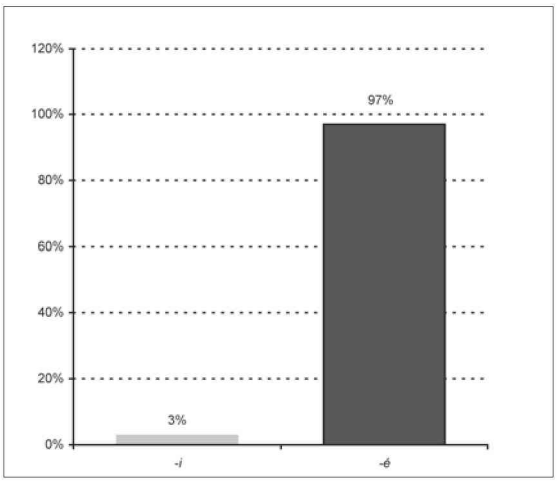


Table 6: Total frequency of the occurrence of inflectional endings *-é* / *-i* in the nominative plural in words ending in *-asta*, *-ista*, *-ita* in Standard Czech (In the corpus: ČNK – SYN2000)

Soundings taken on spoken corpora show only one occurrence of the ending *-é*. The speaker is from central Bohemia. The word that is used with the ending

-é is used with the ending *-i* in the same sentence. The meaning of the word is 'Adventist'. This is not a usual word and this factor may determine the use of the ending.

ano, ano. vířiví adventisti. jo. jo, vířiví <adventisté> já nevim kolikátýho dne. (ČNK – ORAL2006)

'*yes, yes, whirling Adventists. yeah, yeah, whirling whichever-day Adventists I'm not sure which.*'

Other occurrences only contain the ending *-i*.

a voni taky ty <fotbalisti> pořád čoveče sou jako panenky. (ČNK – ORAL2006)

'*and those football players man they're still just like dolls.*'

A different situation is seen when we analyse results from corpora of Standard Czech. Apart from the fact that the ending *-i* occurs in a quote from speech, the ending occurs in some words in Standard Czech. It allows us to construct another opposition based on the semantics of the word. There is a tendency to connect the ending *-i* with words with negative semantics. The following group of words belong to those with the frequent ending *-i*.[6]

- Nouns indicating the follower of some point of view, movement, social or political institution (*fašista* 'fascist', *imperialista* 'imperialist', *komunista* 'communist', *marxista* 'Marxist', *neonacista* 'neo-Nazi', *rasista* 'racist', *sionista* 'Zionist', *socialista* 'socialist', *šovinista* 'chauvinist', *terorista* 'terrorist');
- Nouns indicating the follower's philosophical beliefs and life attitude (*altruista* 'altruist', *egoista* 'egoist', *existencialista* 'existentialist', *idealista* 'idealist', *individualista* 'individualist', *kariérista* 'careerist', *konformista* 'conformist', *masochista* 'masochist', *moralista* 'moralist', *nihilista* 'nihilist', *oportunista* 'opportunist', *optimista* 'optimist', *pesimista* 'pessimist');
- Other... (*bandita* 'bandit', *jezuita* 'Jesuit', *husita* 'Hussite', *penzista* 'pensioner', *turista* 'tourist').

After the analysis of occurrences, there are other oppositions: a contextual opposition (the ending *-é* occurs in the historical context whereas the ending *-i* does not) and an opposition constructed on the basis of a semantic relationship (the opposition concerns a situation where the word loses its basic meaning and comes to serve as an invective).

Mussolini a ***fašisté*** *dostali 62 procent hlasů.* **(pearc)**

'*Mussolini and the <Fascists> got 62 percent of the votes.*'

[6] This division into groups follows that proposed by Dokulil / Daneš / Kuchař (1967: 422f.).

Musíme přijmout i tyto životy, napadlo mě, viděl jsem vás, vy ***fašisti*** *a komunisti, tebe, Micko, jako Rváče v roztržené kožešině s krví kapající z rozseklé ruky...* **(top3)**

'*We have to accept those lives as well, it occurred to me: I have seen you, you Fascists and communists, and you, Micka, like a thug in a ragged fur with blood dripping from your mangled hands...*'

2. Commonalities between frequencies and acceptability

The second type of banding we would like to consider is one based on a relationship with external observed facts: in this instance, bandings that reflect the acceptability of forms to native speakers.

This sort of analysis is useful in deciding whether corpus frequency allows us to draw any conclusions about the linguistic system. We can rewrite this as a question of *parole* and *langue*: if the corpus represents a very large and well-balanced sample of *parole*, what conclusions can we draw from it about *langue*, the linguistic system that supposedly exists in the minds of native speakers? Earlier in this article, we looked at this exclusively with reference to corpus data. Now we will attempt to do the same thing by comparing corpus data with data from studies on linguistic informants. To do so, we will need to make some assumptions.

To begin with, we assume, based on experience, that the main subcorpora of the Czech National Corpus (ČNK), SYN2000 and SYN2005, are large enough to give adequate data on most morphological oppositions, although not for some peripheral declension classes and conjugation types. We will probably not have enough data to study individual lower-frequency words in the lexicon. In other words, at 100 million tokens, we can study common inflectional patterns and high-frequency words in standard written Czech, although we will encounter problems at lower frequencies.

Let us also assume that the representative balancing in these two subcorpora gives a reasonably accurate picture of the written environment surrounding the average Czech. The composition of these two corpora does differ, but again, let us assume that both these corpora are based on different interpretations of valid sociological research (see e.g. Králík 2001, Šulc 2001 for information on corpus structure). This means we will be limited to using these representative corpora. We will not combine them or add other non-representative corpora to them, as is done in some studies.

Therefore, when we find, for example, that the corpus SYN2005 contains 118 examples of the genitive singular form *sýru* and 1185 of the competing form *sýra*, we can suppose that this is close to the proportions in which native speakers are used to seeing these forms in print. We can further ask whether this might influence native speakers' perception of the acceptability of these forms, or whether it might influence their tendency to produce one or another form.

These questions can be tested in two different types of study:

Acceptability judgments ask native speakers to rate the degree to which they find a form acceptable. These test reception and should thus be closer to corpus data than studies of production. They are common in studies of syntax. However, questions do arise over whether native speakers can accurately gauge their intuitions on an exact scale.

Cloze tests ask native speakers to fill in the correct form in a gap. These have traditionally been very common in studies of morphology because they reduce our uncertainty over informants' ability to rate forms. However, they are somewhat less informative because they result in an 'either / or' answer and of course they test production, which is not as closely linked to corpus data, rather than reception.

We undertook a pilot study of the Czech genitive singular that tested the link between corpus frequency and acceptability, testing lexemes of medium to high frequency (over 150 tokens in the corpus SYN2005 for the given case form).

	form in -u	*N =*	*percent*	*gloss*	*form in -a*	*N =*	*percent*
1	les-u	0	0.00%	'forest'	les-a	4316	100.00%
2	sýr-u	118	9.10%	'cheese'	sýr-a	1185	90.90%
3	rybník-u	125	11.90%	'pond'	rybník-a	929	88.10%
4	týl-u	115	34.60%	'rear, nape'	týl-a	217	65.40%
5	dvork-u	150	74.60%	'little courtyard'	dvork-a	51	25.40%
6	pokojík-u	199	93.00%	'little room'	pokojík-a	15	7.00%
7	průchod-u	194	100.00%	'passage'	průchod-a	0	0.00%

Table 7: Forms selected with their absolute and relative frequency in SYN2005

Twenty native speakers of Czech were asked to fill in a questionnaire that had them rate the acceptability of sentences that included these forms.[7] The results of the questionnaire were then correlated with the corpus data and the correlations were found to be highly significant (0.88 for forms ending in *-a*, 0.81 for forms ending in *-u*).

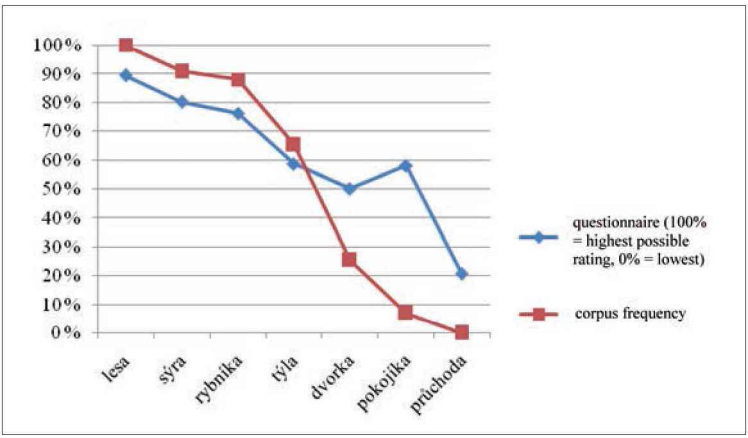


Table 8: Comparison of acceptability and corpus frequency for the formant *-a*[8]

[7] The study is described in greater detail in Bermel (2010), and draws on some of the general procedures for questionnaire set-up and empirical study design described in Cowart (1997) and Schütze (1996). The questionnaire asked for opinions on 70 sentences, rated on a scale of 1-7. Respondents were asked to answer according to their first instinct and not to revisit sentences later, and were told that this was not a test of their knowledge of standard Czech. The scale had descriptors at the anchor points (1: "only possible form / I would definitely use this form"; 7: "unacceptable usage / in this context I would definitely use another form; I don't regard this as normal Czech"). There were no descriptors for points 2-6, allowing respondents to subdivide the scale as they saw fit.

The questionnaire came in four versions, differing in the ordering of examples and in the particular sentences used for individual word forms. The relevant word form was highlighted and thus identified (as compared to syntax, this is not as significant a problem for morphological studies and reduces the number of contextual effects). Due to the length of the questionnaire and the fact that we were not particularly hiding the feature investigated, distracter sentences were not used.

Twenty native speakers of Czech completed the questionnaire in person in the presence of the investigator. Most finished in around 15 minutes, with the longest taking c. 20 minutes. ANOVA and t-tests were performed on a combined score (acceptability of *-a* form minus acceptability of *-u* form) to confirm their statistical significance.

[8] So that both runs of data can appear on the same scale, the acceptability data in both charts have been assigned percentage values, such that 1.0 (best possible score) = 100% and 7.0 (worst possible score) = 0%. For example, a score of 25% for questionnaire data means that on average, the form was regarded as '25% of normal', i.e. very low acceptability.

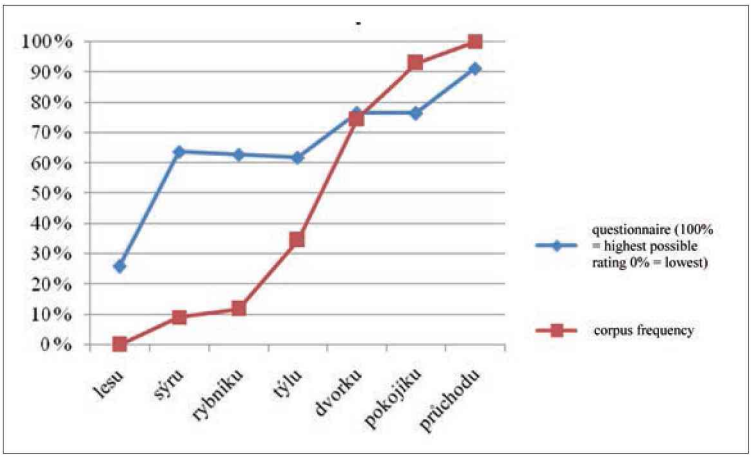


Table 9: Comparison of acceptability and corpus frequency for the formant *-u*[8]

The statistical tests performed on these data show the following:

1) Word choice, rather than context, is the most significant factor in the acceptability of particular forms.

2) When words were compared against each other, some pairs adjacent on our corpus frequency scale and some non-adjacent pairs showed significant differences in their acceptability to native speakers.

3) When corpus data are compared to acceptability data, there is a strong correlation between the relative frequency of a form in the corpus and its acceptability to native speakers. Data from representative corpora thus do have a connection to the acceptability rating of forms.

However, these very clear results do not show that corpus frequency and acceptability data have a linear relationship. In other words, we are not able to say that as relative frequency in the corpus increases, acceptability increases by a similar amount.

Instead, our results show that there is an asymmetric relationship between corpus data and judgment data. Very low proportions in the corpus – essentially sporadic occurrences for a non-sporadic feature – correlate with low acceptability for native speakers. However, from c. 2% on up to 49% of occurrences, there is no clear correlation between frequency and acceptability. Many items in this range of frequency are highly acceptable to native speakers, even if they

may not be the native speaker's first choice for usage. Only at proportions over 50% in the corpus can we reliably state that this correlates with high acceptability for native speakers.

This sort of study yields a three-part division into bands: sporadic frequency (under 2%, predicts low acceptability), minority frequency (2 to 49%, cannot predict acceptability) and majority frequency (over 50%, predicts high acceptability). It follows research on syntax by e.g. Divjak (2008) and Kempen / Harbusch (2005) that suggested similar asymmetric patterns for syntax, making it likely that this is a general principle.

Notice that we say 'sporadic occurrences *for non-sporadic features*', which in itself implies the existence of a contrast between variants in the same morphological context. In other words, because we are using nouns with relatively high frequency, we regard sporadic occurrence of a form as significant. For nouns of low frequency, sporadic occurrence in a particular case form would tell us nothing.

Here the complementarity principle described elsewhere can help us to fill in the gap. The Czech masculine genitive singular has two competing endings for its so-called hard declension pattern: these are *-a* and *-u*. A form such as Gsg *mosta* has two occurrences in the corpus SYN2005, while Gsg *mostu* has 3014. Given information about one, we can predict the acceptability of the other. Looking at the relatively low (11.86%) percentage of examples of Gsg *rybníku*, we can predict that Gsg *rybníka* will be highly rated by informants. However, the opposite would not be true: starting from the highly frequent *rybníka*, we would not necessarily deduce the correct acceptability of Gsg *rybníku*.

3. Conclusions

We can only make more general use of schemes and bands like the ones discussed in this talk if we believe *a priori* in *extrapolation*. In other words, we must believe that all morphological competition operates according to the same principles, and therefore, studying one morphological feature gives us enough material to determine frequency banding principles for all morphological features.

We are not yet at the point where we can confidently assert that we have found proof that our principles are more generally applicable. Still, further studies that look at oppositions, rather than the frequency of individual features, will

make a contribution to this effort. Most importantly, we hope that these studies attempt to anchor corpus data in detailed examination of its composition and in its relationship to linguistic descriptions from outside the corpus.

References

Bermel, Neil (2010): A pilot study on the relationship between corpus data and acceptability judgments of competing forms. In: Krčmová, Marie (ed.): Languages in the integrating world. Munich: Lincom Europa, 233-245.

Cowart, Wayne (1997): Experimental syntax: Applying objective methods to sentence judgments. Thousand Oaks, CA: Sage Publications.

ČNK: Český národní korpus – SYN2000, SYN2005, ORAL2006. Ústav českého národního korpusu FF UK. Internet: http://www.korpus.cz (Last visited: 10/2010).

Daneš, František / Dokulil, Miloš / Kuchař, Jaroslav (eds.) (1967): Tvoření slov v češtině 2. Prague: Academia.

Divjak, Dagmar (2008): On (in)frequency and (un)acceptability. In: Lewandowska-Tomaszczyk, Barbara (ed.): Corpus linguistics, computer tools and applications – state of the art. Frankfurt a.M.: Peter Lang, 213-233.

Halliday, Michael A.K. (1991a): Corpus studies and probabilistic grammar. In: Aijmer, Karin / Altenberg, Bengt (eds.): English corpus linguistics. New York / London: Longman, 30-43.

Halliday, Michael A. K. (1991b): Towards probabilistic interpretations. In: Ventola, Eija (ed.): Functional and systemic linguistics: Approaches and uses. Berlin / New York: de Gruyter, 39-61.

Halliday, Michael A. K. (1992): Language as system and language as instance. In: Svartvig, Jan (ed.): The corpus as a theoretical construct: Directions in corpus linguistics. Berlin / New York: de Gruyter, 61-77.

Hebal-Jezierska, Milena (2007): Wariantywność końcówek fleksyjnych rzeczowników męskich żywotnych w języku czeskim. Warsaw: Department of West and South Slavic Studies, University of Warsaw.

Kempen, Gerard / Harbusch, Karin (2005): The relationship between grammaticality ratings and corpus frequencies: A case study into word order variability in the midfield of German clauses. In: Kepser, Stephan / Reis, Marga (eds.): Linguistic evidence: Empirical, theoretical and computational perspectives. Berlin / New York: de Gruyter, 329-349.

Králík, Jan (2001): Vyvážení zdrojů Synchronního korpusu češtiny SYN2000. In: Slovo a slovesnost 62: 38-53.

MČ = Petr, Jan (ed.) (1986): Mluvnice češtiny. Praha: Academia.

Rusínová, Zdeňka (1987): K distribuci alomorfů -i/-ové v Nom. Pl. maskulin životných. In: Pačesová, Jaroslava (ed.): Sborník prací Filozofické fakulty Brněnské univerzity. Brno: Masarykova univerzita, 72-78.

Schütze, Carson T. (1996): The empirical base of linguistics: Grammaticality judgments and linguistic methodology. Chicago: University of Chicago Press.

Šimandl, Josef (2010): Dnešní stav skloňování substantiv takzvaných typů kámen a břímě. Prague: Nakladatelství Lidové noviny.

Šulc, Michal (2001): Tematická reprezentativnost korpusů. In: Slovo a slovesnost 62: 53-61.

Tušková, Jana Marie (2006): Variantní a dubletní tvary v současné deklinaci apelativních feminin, Brno: Pedagogická fakulta Masarykové univerzity.

Sources

marcin: Marcinko, Richard (1997): Profesionální válečník II. Ivo Železný.

matskol: Fulghum, Robert (1996): Všechno, co opravdu potřebuji znát ... Knižní klub.

pearc: Pearce, Edward (1994): Machiavelliho děti. Nakladatelství Lidové noviny.

top3: Topol, Jáchym (1994): Sestra. Atlantis.

STELLA NEUMANN

Contrasting frequency variation of grammatical features

Abstract

This paper presents a quantitative approach to contrasting grammatical structures in English and German as structured by different registers in the two languages. It approaches the contrastive comparison by organising it into registers, assuming that the options provided by the language systems are not equally distributed and are used in different situational contexts. The functional categories of register analysis also serve as a basis of comparison. The study uses a corpus of English-German original texts in eight different registers and reference corpora in both languages. All subcorpora are annotated with a wide range of linguistic information. The inclusion of the reference corpora as a basis of comparison allows overcoming the methodological issues of contrastive comparisons by introducing the concept of relative register values and only comparing these contrastively. The features discussed in this paper are mood, tense, and voice. On the basis of quantitative findings for these features, the paper will give an outlook on integrating a probabilistic representation of grammatical features in the contrastive description.

1. Introduction

The use of corpora in linguistic research is a very widespread method nowadays. The field of grammar has been one of much interest, especially for monolingual research (e.g. Biber et al. 1999 and contributions in this volume).

While the use of corpora in contrastive linguistics has long been advocated and advanced – in particular by Stig Johansson (e.g., 1998, 2007) –, there is still a need for large-scale descriptions of language contrasts that go beyond the focus on exemplary linguistic features. The present paper proposes a methodology for a quantitative contrastive comparison within the framework of register analysis (e.g., Halliday / McIntosh / Strevens 1964, Halliday / Hasan 1989). This register-based approach has the advantage that it employs a well structured grid for the analysis of individual grammatical features. Moreover, it supplements the contrastive description of language systems with an additional layer that organises the features of the system into preferred and dispreferred options, thus adding a usage-based dimension to the comparison.

The paper also proposes a method of avoiding the methodological pitfalls of direct feature comparisons – in the case of quantitative studies, in the form of direct frequency comparisons – which is based on the comparison of register-specific and register-neutral reference frequencies.

The remainder of this paper is organised as follows. The discussion of contrastive linguistics and quantitative approaches to language contrasts in Section 2 motivates the research presented. This discussion also includes register linguistics as the theoretical framework of this study. Section 3 presents the methodology including the corpus design used in the study as well as the procedure used to obtain relative register values that are subsequently contrasted. Section 4 then discusses exemplary findings yielded by this methodology. This section also provides an outlook on a probabilistic representation of grammatical features. The paper is rounded off by some concluding remarks as well as a reference to future work (Section 5).

2. State of the art

2.1 Contrastive linguistics

Contrastive linguistics is the scientific discipline concerned with the systematic comparison of two or more languages (Johansson 2007). A more traditional approach to the comparison of a given language pair is to analyse and contrast options afforded by the two language systems and compare them (e.g. Hellinger 1977, Hawkins 1986, Mair 1995, Legenhausen / Rohdenburg 1995, etc.). König / Gast (2007) provide the most recent description of the comparison of English and German, giving a detailed account of the main areas of differences between the two language systems from the phoneme inventory to a wide range of aspects of sentence grammar. Besides giving some selective usage-based examples from the British National Corpus (BNC) and other corpora to underpin the system-based description, their study does not attempt to describe frequency comparisons or rather preferences in usage as reflected by (differing) frequencies.

If corpora are used, as championed by Johansson (e.g., 1998, 2007), and thus a usage-based view is taken up, the quantitative studies typically concentrate on individual features, consequently not giving a comprehensive overview like the classical works do. A more comprehensive picture of the range of quantita-

tive variation in a given language pair would then only arise from consolidating individual studies, which is only possible if the studies are comparable as to their theoretical foundation and methodology.

2.2 Quantitative approaches to contrastive comparisons

A notable exception to the concentration on contrasting options in the language system is Elke Teich's study (Teich 2003). She does not only give a comprehensive and comparative overview of the main grammatical systems in English and German based on systemic functional linguistics (Halliday / Matthiessen 2004) but also quantifies the differences and commonalities for the register of popular scientific writings. The fact that she draws on a functional theory allows her to use the function of the respective grammatical system as the basis of comparison for the contrastive analysis. Since the author also includes translations in the corpus analysis, the findings of the study extend beyond contrasting the two languages, making statements on how the contrastive differences and commonalities impact the language in translation.

By combining a clear theoretical framework in the form of systemic functional grammar elaborated for both English and German with the analysis of a balanced corpus of textual, usage-based instances, Teich overcomes the typical limitations of the contrastive approach. Furthermore, its combination of deduction from systemic-functional grammar and induction from a bilingual corpus also goes beyond the limitations of general statements obtainable in a strictly inductive approach like Biber's (see below).

Teich's work is mainly concerned with the influence of contrasting language systems on translations. She is not chiefly interested in the more general variation of registers in a given language pair.

This is one of the major contributions of Douglas Biber's work (e.g., 1988, 1995) to the field. His 1988 study, a corpus-based description of register variation across speech and writing, still concentrated on English. The study proposed a genuinely new inductive approach to the investigation of variation based on the use of the statistical technique of multidimensional analysis. The data for English was later integrated in a cross-linguistic study (Biber 1995), again using the multidimensional approach.

The 1995 study compares the register variation in four different languages: English, Nukulaelae Tuvaluan, Korean, and Somali. After a discussion of the

dimensions identified independently for each language, Biber compares the dimensions across the four languages. Since the corpora for each register consist of different genres – a plausible reflection of different situational contexts in the four languages –, it is only natural that the language-specific dimensions are not directly comparable across the four languages. In order to overcome this cross-linguistic incompatibility, Biber introduces so-called communicative functions – some of them language-specific – again derived on the basis of inductive analyses. Biber then compares the monolingual dimensions, their relevant features, and characteristic registers in the four languages along these functions. Additionally, he compares what he labels 'equivalent registers' (1995: 237) across the languages on the basis of their respective position on the dimension. A final set of intra-lingual interpretations is concerned with the internal variation of the registers.

Biber offers a far-reaching discussion of different aspects of intra- and interlingual register variation. His approach is innovative in that it does not only entail quantitative, bottom-up analyses and detailed interpretations but also introduces a new systematic use of inferential statistical techniques to linguistic interpretation. His work, however, also raises some questions: the linguistic features Biber uses to carry out the general analysis of English registers were originally compiled for the 1988 study of the spoken–written continuum. This results in a bias of the whole study towards findings along this dimension. Even though many of the features under investigation are not restricted to the spoken–written continuum or – in systemic terms – to medium of discourse (see below) and therefore shed light on other areas of linguistic variation, certain important features serving as indicators of other aspects of register variation are not included.

One might also object to the use of features deduced from other theory-based studies in an inductive study, a procedure highlighting the fact that there is no such thing as an exclusively inductive study. Furthermore, the communicative functions required for cross-linguistic comparisons of the corpora are inferred inductively from the findings supplied by his monolingual inductive analyses. As a result, their scope is limited to generalisations that merely apply to the features originally investigated. A combined deductive and inductive methodology – particularly one that is functional[1] – would

[1] See Neumann (2003) for a discussion of the advantages of founding contrastive comparisons on the systemic functional approach.

ensure a systematic and comprehensive contrastive comparison while still taking into account relevant insights from the study of empirical evidence.

A theoretical framework well-suited to complement and motivate bottom-up analyses is the systemic functional one, certain aspects of which are related to Biber's work, for instance, the functional view of language and the role of contextual information. Teich's (2003) analysis exemplifies this for one register. As will be shown in this paper, this approach can be extended to the study of multiple registers in quantitative studies.

2.3 The register approach

The study of texts in 'classical' text linguistics focuses on individual texts, often more specifically on individual features in texts or groups of texts. These features can be linguistic in nature, like the focus on the topic of a text (e.g., Brinker 2005), or social, like the "social and political thought relevant to discourse and language" in Critical Discourse Analysis (Fairclough 1992). Register theory in its systemic functional flavour, by contrast, situates texts in a general framework viewing registers as a concretion of the language potential that incorporates the situational context in which language in use is embedded and that is realised by means of lexico-grammar (Matthiessen 1993).

'Register' has emerged as the linguistic framework for describing the context of situation since the 1960s. Three parameters were introduced for the description of registers (e.g. Halliday / McIntosh / Strevens 1964, Gregory / Carroll 1978, Halliday / Hasan 1989, Ghadessy (ed.) 1993): *field of discourse*, specifying the topic of the linguistic exchange in the given situation, its referential meaning, *tenor of discourse*, characterising the relationship between the participants in the situation, i.e. its pragmatic meaning, and, finally, *mode of discourse*, describing the way in which the exchange is transmitted, the textual meaning. These three parameters correspond to the three metafunctions of language assumed to cover human experience in terms of the ideational metafunction, enacting personal and social relationships as the interpersonal metafunction, and organising the discursive flow as the textual metafunction (Halliday / Matthiessen 2004: 29f.).

Just as situations tend to recur and form types, registers represent recurring ways of using language in a recurrent situation. The language system can even be grouped into typical co-occurrences and non-occurrences according to the respective situation. Registers can thus be described as subsystems of the lan-

guage system or, in a different perspective, as types of instantiated texts reflecting a similar situation (cf. Matthiessen 1993). The concept of types (of situations or of instantiated texts) implies a certain frequency of recurrence of features or patterns. In a methodological sense this means that, strictly speaking, a description of these types requires a quantification of their characteristic features. The analysis of groups of instantiations can thus be regarded as a requirement of the theory. Otherwise, we can only describe a given specimen of the assumed type, which does not permit any statements on the type itself.

Systemic functional register analysis has the advantage that it provides a comprehensive set of functional parameters that reflect referential, pragmatic, and textual meaning. As will be seen in the following section, it integrates abstract concepts and concrete operationalisations in a way that allows plausible reasoning. Its attachment to a functional theory of language makes register analysis particularly well-suited for contrastive comparisons where the functional concepts can be used as a *tertium comparationis* as shown by Teich (2003). An orthodox systemic functional analysis may, however, be extremely costly and certainly requires a substantial amount of potentially subjective interpretation. Furthermore, the integration of concepts and their operationalisation in terms of linguistic features may not always be pursued to its full potential. The present study tries to avoid these pitfalls by using a somewhat more rigorous – but by the same token slightly less theory-driven – analysis building on computational tools as far as possible to push the interpretation to a later stage in the research design (cf. Hansen-Schirra / Neumann / Steiner 2007). Connecting the register parameters to observable indicators via intermediate categories that help to narrow down the interpretative space is seen as a pivotal step to ensure that the features under investigation do actually allow statements on the abstracts concepts without too many confounding factors. Establishing this connection also includes a statement on which feature value points to which concept. This derivation process is discussed in the following section as well as other aspects concerning the methodology of the study.

3. Methodology

3.1 Derivation of indicators in register linguistics

Register theory organises the analysis in the three general parameters named above. Since these are highly general, they have been specified into a range of categories that are frequently called subdimensions. For instance, field of dis-

course can be specified in terms of 'experiential domain'. Experiential domain is the category describing the subject area covered by a register. As this is still a latent variable that cannot be observed directly in texts, observable indicators are required that reflect the subject matter. Among the indicators or operationalisations pointing to subject matter are vocabulary, lexical chains, paragraphing, etc. (cf. Steiner 2004, Neumann 2008). Some subdimensions require an additional level of subcategorisation. Social role relationship as a subdimension of tenor of discourse is a case in point. It is concerned with the linguistic reflexes of diverging social roles taken up by the interactants. Social roles are determined by the level of authority and expertise of the interactants but also by factors such as gender, ethnicity, religion, etc. (cf. e.g., Poynton 1985; cf. Dreitzel 1980 from the point of view of sociology). Only these factors can be related to observable indicators in texts. A higher level of authority on the part of the speaker, for instance, can be determined by assessing the frequency of imperatives in a given register. These various levels of categorisation highlight the importance of comprehensive analyses in order to interpret the abstract parameters in a meaningful way.

As can be gathered from the example of medium, a subdimension under mode of discourse, the categories mediating between the subdimensions and the observable indicators also facilitate the interpretation of quantitative findings. Medium is concerned with the assignment to spoken or written mode. Written mode, for instance, can be characterised by a high lexical density (a high proportion of content words per all words, Ure 1971), less variation in thematic structure, i.e., the grammatical function in sentence-initial position, etc. A combination of features diverging from these values would then point to the other end of the cline, to the spoken mode. A more complete derivation of indicators for the three parameters is described in Neumann (2008). The indicators thus derived can then be queried in the corpus.

3.2 Methodology for quantitative contrastive comparisons

The present study uses the following procedure to yield results that are interpreted subsequently. In the first step, the indicators are queried and quantified in the corpus. The top-down feature derivation is hence complemented by a bottom-up feature analysis. Then, the quantitative results are processed statistically yielding descriptive statistics. The final step, significance testing, is not discussed here due to the reasons given below.

The contrastive perspective of this study highlights a methodological problem relevant for any cross-linguistic investigation. It is well known that direct comparisons of lexico-grammatical features across languages can be misleading. Johansson (2007: 3) addresses the insufficiency of formal categories in contrastive comparisons citing the example of modal auxiliaries which, according to him, could be expressed in quite different ways. To some degree, this is certainly the case in the language pair English-German, where German relies to some extent on modal adverbs, particles, etc. As Teich (2003) shows, the options of two language systems become increasingly divergent the more fine-grained the comparison is. Therefore, some kind of basis of comparison is required. König and Gast, for instance, employ general semantic notions such as 'temporal relation', 'inalienable possession', 'co-reference', and 'understood subjects' and argue that nearly all of the formal and semantic distinctions used by them are also employed in typological studies (König/Gast 2007: 6). In a similar vein, Teich (2003) argues in favour of comparing underlying functions. Johansson (2007: 3) suggests the use of translations as "a source of perceived similarities across languages."

These problems on the level of language systems equally apply to quantitative studies. Comparing the frequency of occurrence of a given feature directly may be a questionable procedure since it may fail to recognise the different usage preferences associated with the feature in the two language systems. A feature may be comparable in terms of the language system but may be more or less common in the two languages as expressed by frequencies of occurrence.

One way of circumventing this problem is the comparison of relative values only. The present study proposes to first compare the frequencies of a given feature in one language to a reference value in the same language and to only use the resulting relative values in the contrastive comparison. The reference value can be obtained from a reference corpus consisting of a mixture of registers that represent some kind of average of the respective language (see below). With this method, only the magnitude of difference of a given frequency in relation to the reference corpus is compared contrastively. This method factors out systematic frequency differences in the language pair and concentrates on the register-specific deviations from the respective baseline.

This approach has the additional advantage of taking into account register-specific preferences for given linguistic features. This reflects the view of registers as filtering the options provided by the language system (Matthiessen 1993). While the description of the options of the language system is un-

weighted, registers can be seen as boosting or blocking certain features in terms of frequency of occurrence. We can then describe features as relatively frequent or absent in a given register, allowing a representation of linguistic options in terms of probability (Halliday 2005, Nesbitt / Plum 1988; see Section 4).

3.3 Corpus design and enrichment

The corpus analysed on the basis of these theoretical considerations has the following characteristics. It is a sample of the CroCo Corpus[2] (Neumann / Hansen-Schirra 2005, Neumann 2008) of texts drawn from the following eight registers: political essays (ESSAY), literary texts (FICTION), instruction manuals (INSTR), popular scientific writings (POPSCI), letters to the shareholder (SHARE), political speeches (SPEECH), tourism brochures (TOU), and websites (WEB). Each register consists of at least ten texts totalling approximately 31 250 words per register, thus meeting the thresholds for register descriptions discussed by Biber (1990, 1993; cf. Hansen-Schirra et al. (forthcoming) for a detailed description of the corpus). The CroCo Corpus was designed for the study of linguistic properties of translations distinctive from original, non-translated texts. This explains the selection of registers: only those registers are included in the corpus which are translated in both directions. In addition to this prerequisite, each register is assumed to represent a different contextual configuration. The corpus used in the present study is a subset of the CroCo Corpus excluding translations and totalling 574 649 tokens.

In order to compute relative register values as introduced in the previous section, the corpus has to be supplemented by reference corpora in both languages. The reference corpora used here were built specifically for the purpose of register analysis (Neumann 2003) and were enlarged in the CroCo project. The English and the German component consist of 2 000 word samples each in 17 different registers respectively totalling approximately 42 000 tokens each. They were built on the FLOB Corpus (Hundt / Sand / Siemund 1998). However, in order to reduce the bias towards fictional registers, three fictional registers were replaced by other registers (calls for tender, cooking recipes, prepared speeches). Furthermore, two registers – travel guide books and court decisions – were added.

The complete CroCo Corpus is annotated with several layers of linguistic information. It contains meta-information on each text including a coarse regis-

[2] http://fr46.uni-saarland.de/croco (last visited: 08 / 2010).

ter analysis to provide filter options. On word level, parts of speech and morphology are annotated, the latter being of particular importance to the German part of the corpus. On phrase level, phrase structure and grammatical functions are manually annotated. Finally, clauses and sentences are segmented. For the purpose of investigating properties of translations, the corpus is also aligned on word, phrase, clause, and sentence level. The representation in various stand-off formats allows queries into any combination of features on all annotation layers (cf. Hansen-Schirra / Neumann / Steiner forthcoming).

4. Findings and Discussion

In the following, we will exemplify the methodology introduced in the previous sections for the three features voice, tense, and mood. The discussion of potential contrasts is embedded in the comparison of registers under the above-mentioned assumption that registers filter the options provided by language systems – and thus motivating the quantitative approach to contrastive comparisons. Moreover, the features are not simply discussed as such but related to their place in the classificatory framework of register analysis, i.e., to the abstract concepts. The presentation of findings will start with voice, an indicator of subdimensions of field of discourse.

4.1 Voice

One of the main functions of voice is to shift the focus of an utterance to the actor (in active voice) or away from this role (in passive voice). In this function, voice is one of the indicators used for determining the goal pursued by speakers in a given register, i.e., for the category of 'goal orientation'. A frequent use of the passive voice points to an orientation towards presenting facts rather than the agents responsible for these facts (Götze / Hess-Lüttich 2002: 107f.). A high frequency of passive voice can then be interpreted as pointing to an expository goal. It can be expected to co-occur with features supporting a nominal style like a high lexical density, frequent nominalisations, complex phrases, etc. Central passives can be queried by searching for forms of *to be / werden* and past participles on the basis of the part of speech annotation (allowing for a certain number of intervening words depending on the language). In the German language, an additional query considering verb-last word order is included as well.

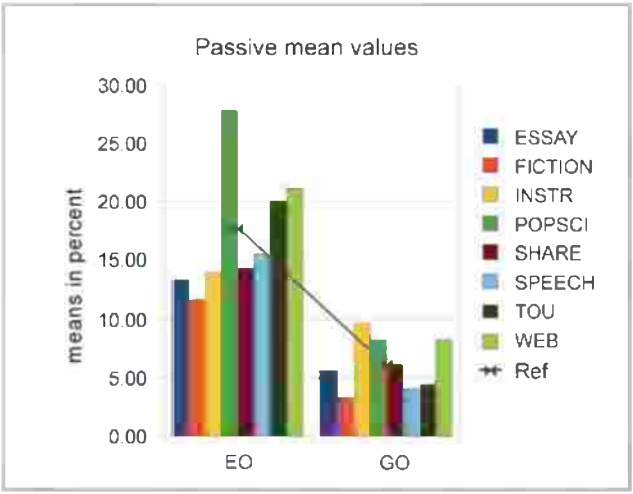


Figure 1: Mean frequencies of central passives in all registers and the reference corpora

The frequency of passive voice shows a remarkable contrastive divergence as can be gathered from the values for the reference corpora in Figure 1. The divergence can be attributed to a different range of options for expressing underspecified agency. German has a number of options in the active voice that help avoid specifying the agent of a given process such as reflexive constructions or the construction with *sich* + *lassen* + Infinitive (Götze / Hess-Lüttich 2002: 111f.). When comparing only the default option of the auxiliary in combination with the past participle – this realisation is comparable in the two languages –, the English language with its smaller range of options displays a higher number of passives, as the values for the reference corpora show. Table 1 reflects this contrastive difference (in percentage points) resulting from the differing means used in English and German in the form of persistently lower values in the German registers.

Apart from this general difference, Table 1 also reveals some register-specific features. All but two registers show a similar tendency: ESSAY, FICTION, SHARE, SPEECH, and TOU use fewer passives than the reference corpora in both languages. This is a surprising finding as all of these registers except for FICTION would be expected to express a fact-oriented view on their topic. The only register surpassing the reference values in both languages is POPSCI. However, its mean frequency is not significant in the t-test against the reference corpus in English ($t(25) = .651$, $p = .521$) nor is it in German ($t(25) = .791$,

$p = .437$). In German, the only register significantly deviating from the reference corpus is FICTION in the form of a significantly lower mean value ($t(21) = -2.545$, $p < .05$). The significance tests reported here refer to the comparison of monolingual frequencies in each register and the reference corpora only (cf. Neumann 2008).

	Passive voice	
	English	**German**
ESSAY	−11.38	−1.00
FICTION	−13.15	−3.41
INSTR	−10.80	2.95
POPSCI	3.04	1.61
SHARE	−10.41	−0.65
SPEECH	−9.27	−2.62
TOU	−4.68	−2.23
WEB	−3.68	1.61

Table 1: Relative register values for passive voice in contrastive comparison

Two registers show diverging patterns in the two languages: instruction manuals display a considerably lower relative value compared to the reference corpus in English whereas the German register contains more passives than the reference corpus. This can be explained by two diverging tendencies in this register: on the one hand, instruction manuals can be assumed to focus on direct and concrete instructions which are best expressed by concrete material processes (cf. Halliday / Matthiessen 2004) in the imperative mood (with the addressee as the assumed agent). On the other hand, passives may play a role in German instruction manuals in a more indirect way of expressing recommendations rather than direct commands. Example (1) from the CroCo Corpus is a case in point. The passive structure is combined here with the modal verb *sollten* to express the recommendation. The published translation into English is given in (1b) to illustrate the realisation by imperatives.

(1a) Falls die Fernbedienung längere Zeit nicht benutzt wird, <u>sollten</u> die Batterien <u>herausgenommen</u> und an einem kühlen, trockenem [sic] Ort <u>aufbewahrt werden</u>. (GO_INSTR_008)

(b) When the remote controller is not in use for a long time, <u>remove</u> the batteries and <u>store</u> them in a cool and dry place. (ETrans_INSTR_008)

Another explanation for this divergence is exemplified by (2). This example highlights a well-known contrastive difference between English and German where an undesirable personification of a potential non-agentive subject, in this example *Bedienungsanleitung / operating manual*, is avoided in German by a passive construction in combination with an adverbial. The English translation given in (2b) shows the inconspicuous English version, where the mapping of grammatical functions and semantic roles is more flexible (cf. Hawkins 1986). This contrastive difference should not be assumed to be heavily influenced by register constraints and should therefore be observable across all registers under investigation.

(2a) In dieser Bedienungsanleitung wird die Maximalausstattung beschrieben. (GO_INSTR_009)

(b) This operating manual describes the maximum equipment. (ETrans_INSTR_009)

A third example shows a simple active alternation in translation without additional constraints, which may be typical of the English register.

(3a) Dadurch wird sichergestellt, dass der elektrische 0-Punkt nicht unterfahren wird. (GO_INSTR_014)

(b) This ensures that the signal remains above the dead and live zero point. (ETrans_INSTR_014)

Future work will have to determine whether the contrastive divergences of the relative register values are significant. The significance testing should, however, be based on results for all functionally comparable options in both languages. Since there are open questions with respect to all findings discussed in this paper, this also applies to the following features.

4.2 Tense

Typical distributions of tense options can also give insight into which goal is pursued by a given register. In particular, past tense is often interpreted in connection with a narrative goal whereas present tense can be assumed to be the preferred tense of exposition and potentially also other goals like instruction.

A very general comparison of tense focuses on the distinction between present orientation and past orientation of the utterance. Werlich (1976: 144) argues that the axis of orientation in the continuum of time to which the sender re-

lates phenomena is either signalled by the present tense group, i.e. present tense, present perfect, future I, and future II, or by the past tense group consisting of past tense, past perfect, conditional I, and conditional II (cf. also König / Gast (2007) for a distinction between past and non-past). The query for this two-way distinction makes use of the morphological annotation where past tense is annotated, thus querying for all past tense forms as opposed to all finites. On this basis, the relative register values displayed in Table 2 can be processed in percentage points.

	Present tense	
	English	**German**
ESSAY	6.15	−0.15
FICTION	−27.13	−29.43
INSTR	11.74	10.68
POPSCI	0.71	5.56
SHARE	−3.69	0.28
SPEECH	4.82	7.23
TOU	6.43	5.64
WEB	0.83	−1.29

Table 2: Relative register values for present tense in contrastive comparison

The two languages prove to realise the register of instruction manuals quite similarly with a very strong orientation towards present tense, surpassing the respective reference values by 11.74 and 10.68 percentage points. This can be interpreted as an indication of an instructive goal. The most striking – but not surprising – finding is the under-use of present tense in FICTION as compared to the reference corpora in both languages. The register in both languages has a clear and similar preference for past tense forms in terms of relative register values. This can be interpreted as an indication of a narrative goal (cf. Neumann (2008) for a comprehensive discussion of all indicators of the different goal types).

The two registers POPSCI and SHARE exhibit a divergence. In SHARE, the relative absence of present tense in the English register as compared to the reference corpus can be explained by the report of the last business period. In the comparable German register, a cursory query of perfect tense shows a higher

frequency of this tense as compared to the perfective aspect in the English register. The more flexible use of tenses in German may result in a tendency to rely more on an orientation towards present forms as in the English register.

The higher frequency of present tense forms in German popular scientific writings as compared to the reference corpus may be more plausibly explained by slightly diverging register-specific conventions in the two languages with the German register putting almost all information in present tense.

4.3 Mood

Mood is analysed at various points in register analysis, most prominently as part of tenor of discourse. We will exemplify the subdimensions using 'social role relationship', which covers the linguistic effects of the (different) statuses of the interactants in society. If the status is not equal but hierarchical, this should have an impact on how the interactants express themselves. As shown in previous studies, the social role relationship (Steiner 2004), or power (Poynton 1985), can be detected in linguistic interaction. Subcategories such as level of authority and level of expertise have been mentioned in section 3.1. This study concentrates on the first factor of authority.

A sender with a high level of authority should be in a position to make demands at his/her discretion. This may be reflected in an above average frequency of imperatives, used to express demands for goods and services (Halliday/Matthiessen 2004). The interpretation of frequent declaratives and interrogatives is less straightforward. Declarative statements may point to either a higher level of authority on the part of the sender (she/he has the power to assert) or a lower authority (she/he is forced to provide information). The same applies to questions: the interactant posing questions may be in the higher social position allowing him/her to ask at all (this might be the case of a judge in court proceedings) or the questions posed by the interactant may reflect his/her lower social role marking deference.[3] Nevertheless, these problems do not inhibit the comparison of registers. Interrogatives can be queried combining the verbal part of speech tags with positional aspects. The frequency of declaratives as the default option is obtained by subtracting the frequencies of the other two options from the total number of finites. The relative register values are displayed in Table 3, again as percentage points.

[3] The additional difficulties connected to the analysis of addressee-related aspects in monologic written texts are discussed in Neumann (2008: 52f.).

	Declaratives		Interrogatives		Imperatives	
	English	German	English	German	English	German
ESSAY	2.83	5.42	−2.00	−0.55	−0.84	−4.86
FICTION	−2.81	-0.45	3.28	4.42	−0.46	−3.96
INSTR	−6.83	−18.57	−2.83	−2.25	9.66	20.83
POPSCI	1.04	3.86	−0.18	0.78	−0.86	−4.63
SHARE	3.52	7.24	−2.81	−2.46	−0.71	−4.76
SPEECH	2.56	4.76	−1.96	−0.78	−0.60	−3.97
TOU	−0.30	5.40	−0.63	−1.25	0.93	−4.13
WEB	−1.82	2.25	0.12	−0.28	1.70	−1.96

Table 3: Relative register values for mood options in contrastive comparison

The relative frequency of interrogatives in the eight registers follows similar patterns in the two languages. The only register showing a clear increase of interrogatives in relation to the two reference corpora is FICTION. This may be attributed to dialogues between characters in the story. Again, there is no striking difference between the two languages. The other two options, declarative and imperative, show more variation between the two languages. However, this variation is mostly similar in tendency: where the English register shows a positive deviation from the reference corpus, the German does the same and vice versa. The four registers ESSAY, POPSCI, SHARE, and SPEECH even surpass the already high frequencies of declaratives in the reference corpora in both languages. The declarative mood is clearly the option of choice in these registers.

One register shows a very marked distribution of mood options. In both languages, INSTR, the register of instruction manuals, contains a distinctly high value of imperatives at the expense of declaratives, thus containing markedly fewer declaratives than the reference corpora in both languages. This could be interpreted as a slight tendency towards a higher authority of the sender in this register. This should be viewed as some kind of technical authority rather than, say, a political one.

Summarising the findings, two registers stand out as displaying some divergences from the rest of the registers in both languages: FICTION and INSTR. This can be explained in part by the categorically different nature of literary texts and in part by similarities between ESSAY, POPSCI, SHARE, SPEECH,

TOU, and WEB, a group of registers that are similar in a tendency to provide information. A more detailed overview of differences and similarities between the registers is given in Neumann (2008).

4.4 The role of registers in contrastive studies

The findings presented in the previous sections should have shown that registers are not only relevant as a level of linguistic comparison in its own right. They also reveal patterns of variation of features – in particular in a corpus-based research design – that are not visible to a system-oriented qualitative account of language contrasts.

Findings like the ones presented here can be represented as probabilistic grammar as exemplified by Halliday (2005; in particular a reprint of Halliday / James 1993) and Nesbitt / Plum (1988). Probabilisation here refers to the distinction between choice as reflecting the qualitative (paradigmatic) options of the language system and patterns of choice reflecting the quantitative preferences in instantiation (realisation), where choice refers to possibility and patterns of choice to probability (Nesbitt / Plum 1988: 8). A first indication of how this type of representation would look with respect to the contrastive findings presented here will be shown in the following. Figure 2 displays a system network of the coarse options of the mood systems in English and German (Teich 2003), which is also the basis of the analysis presented in Section 4.3.

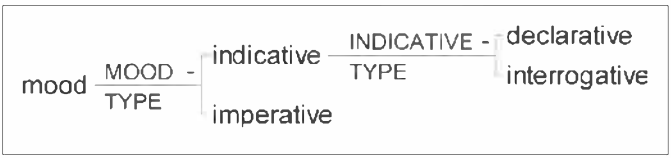


Figure 2: Coarse mood system network for English and German

We can then attach patterns of choice in the form of percentages to the options in each register. Figure 3 illustrates the resulting representations for the English reference corpus (Ref) and the registers FICTION, INSTR, and SHARE. The figure shows how differing preferences in usage can be visualised. At least one option appears blocked in each of the registers displayed: in literary texts the imperative, in instruction manuals the interrogative, and in letters to the shareholder both the imperative and the interrogative are blocked, thus leaving practically no choice in terms of mood.

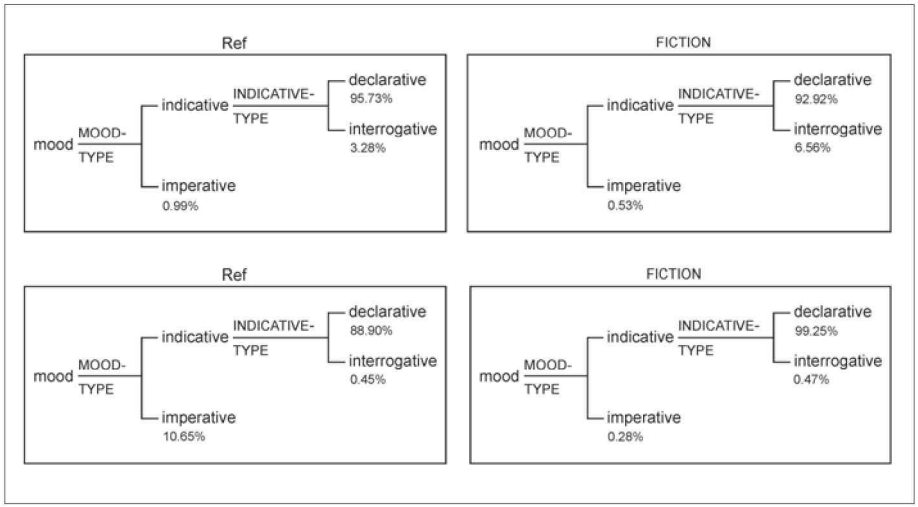


Figure 3: Mood probabilities for English subcorpora

Figure 4 gives the representation for the same subcorpora in German. Here, instruction manuals and letters to the shareholder show a similar distribution to their English registerial counterparts (see Figure 3), with INSTR displaying an even stronger reliance on the imperative mood. In FICTION, however, the imperative option cannot be interpreted as completely blocked – at least by human interpretation (significance testing would give a more reliable threshold to determine whether an option can be viewed as blocked). The register of literary texts can be assumed to vary considerably. Therefore, the findings for both English and German should be interpreted with caution: a different set of texts in the corpus might lead to quite different results.

The probabilistic representation of grammar is probably confined to the visualisation of feature distributions. Nevertheless, this in itself may be suitable for an overview of contrasts and similarities in a given language pair. It would be desirable to give a systematic account of this type for the language pair English and German as a complementary view to a qualitative system-oriented contrastive comparison as in König / Gast (2007).

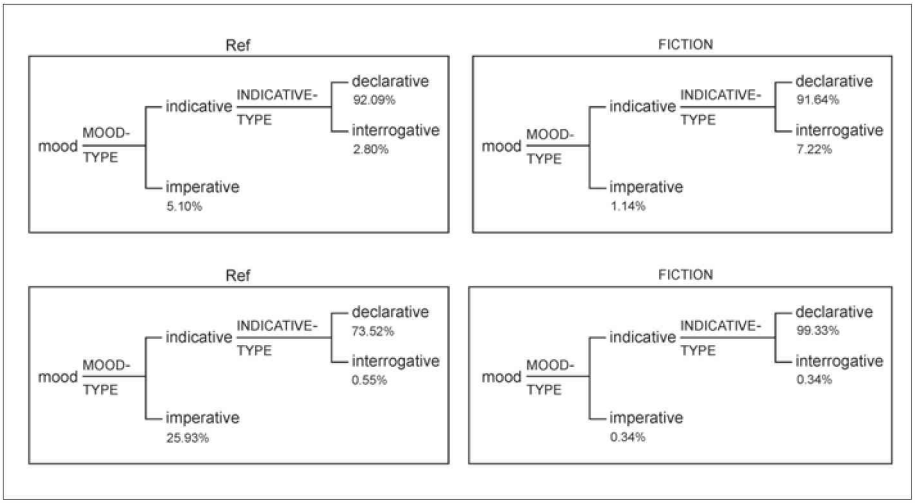


Figure 4: Mood probabilities for German subcorpora

5. Conclusion and outlook

We hope to have shown in the present paper that, in general terms, the stratification of linguistic features into registers offers a more differentiated picture of language contrasts. The quantification provides a first step towards usage-based contrastive comparisons which may ultimately take the form of a corpus-based contrastive grammar in the style of Biber et al. (1999). The method of comparing relative register values may be a way of factoring out intra-lingual systematic characteristics. This method appears promising but requires additional testing against more features and more registers. Future work also includes computing significance tests of the contrastive relative register values.

More registers as well as more texts per register are particularly needed in the reference corpora. The small size of the reference corpora limits the reliability of the current findings. Furthermore, there is room for many more fine-grained analyses building on this study. As we have shown, the analysis of voice, for instance, is still restricted to the formal correspondences and requires a more comprehensive analysis of all functional correspondences. In terms of mood, the diverging options of the imperative mood in the two languages (cf. Teich 2003) would deserve a closer investigation.

The scope of this paper only allowed a brief discussion of a probabilistic representation of grammar. A number of questions remain open in this area, for instance the question of thresholds for determining whether a certain option is blocked.

Integrating registers in the contrastive examination of languages is certainly a field that can help us understand differences but also areas of contact between languages in a more comprehensive way. Generally, there is a need in contrastive linguistics for comprehensive corpus-based descriptions. It is hoped that the present paper has provided a methodological contribution that may facilitate such descriptions.

Acknowledgements

I would like to thank the CroCo team at my former affiliation at Universität des Saarlandes, Silvia Hansen-Schirra, Erich Steiner, Oliver Čulo, Mihaela Vela, Kerstin Kunz, and Karin Maksymski, for continuous discussions of concepts and analyses and for help with all technical aspects of the study. I would also like to thank the audience at "Grammar and Corpora 2009" for their helpful comments. The research reported here was carried out as part of the DFG project no. STE 840/5-2 & HAN 5457/1-2.

References

Biber, Douglas (1988): Variation across speech and writing. Cambridge: Cambridge University Press.

Biber, Douglas (1990): Methodological issues regarding corpus-based analyses of linguistic variation. In: Literary and Linguistic Computing 5, 4: 257-269.

Biber, Douglas (1993): Representativeness in corpus design. In: Literary and Linguistic Computing 8, 4: 243-257.

Biber, Douglas (1995): Dimensions of register variation. Cambridge: Cambridge University Press.

Biber, Douglas / Johansson, Stig / Leech, Geoffrey / Conrad, Susan / Finegan, Edward (1999): Longman grammar of spoken and written English. Harlow: Longman.

Brinker, Klaus (2005): Linguistische Textanalyse. Eine Einführung in Grundbegriffe und Methoden. 6th ed. Berlin: Erich Schmidt.

Dreitzel, Hans Peter (1980): Die gesellschaftlichen Leiden und das Leiden an der Gesellschaft. Eine Pathologie des Alltagslebens. 3rd ed. Stuttgart: Enke.

Fairclough, Norman (1992): Discourse and social change. Cambridge: Polity Press.

Ghadessy, Mohsen (ed.) (1993): Register analysis. Theory and practice. London / New York: Pinter.

Götze, Lutz / Hess-Lüttich, Ernest W.B. (2002): Wahrig Grammatik der deutschen Sprache. Sprachsystem und Sprachgebrauch. 3rd ed. Gütersloh / München: Bertelsmann Lexikon.

Gregory, Michael / Carroll, Susanne (1978): Language and situation. Language varieties and their social contexts. London: Routledge & Kegan Paul.

Halliday, Michael A.K. (2005): Computational and quantitative studies. Edited by Jonathan J. Webster. (= Collected works of Michael A.K. Halliday 6). London / New York: Continuum.

Halliday, Michael A.K. / Hasan, Ruqaiya (1989): Language, context, and text: Aspects of language in a social-semiotic perspective. Oxford: Oxford University Press.

Halliday, Michael A.K. / Matthiessen, Christian M.I.M. (2004): An introduction to functional grammar. 3rd ed. London: Arnold.

Halliday, Michael A.K. / McIntosh, Angus / Strevens, Peter (1964): The linguistic sciences and language teaching. London: Longman.

Hansen-Schirra, Silvia / Neumann, Stella / Steiner, Erich (2007): Cohesive explicitness and explicitation in an English-German translation corpus. In: Languages in Contrast 7, 2: 241-265.

Hansen-Schirra, Silvia / Neumann, Stella / Steiner, Erich (forthcoming): Cross-linguistic corpora for the study of translations. Insights from the language pair English–German. Berlin: de Gruyter.

Hawkins, John A. (1986): A comparative typology of English and German. Unifying the contrasts. London: Croom Helm.

Hellinger, Marlis (1977): Kontrastive Grammatik Deutsch, Englisch. (= Anglistische Arbeitshefte 14). Tübingen: Niemeyer.

Hundt, Marianne / Sand, Andrea / Siemund, Rainer (1998): Manual of information to accompany the Freiburg-LOB Corpus of British English ('FLOB'). Freiburg. Internet: http://khnt.hit.uib.no/icame/manuals/flob/INDEX.HTM (last visited: 08 / 2010).

Johansson, Stig (1998): On the role of corpora in cross-linguistic research. In: Johansson, Stig / Oksefjell, Signe (eds.): Corpora and crosslinguistic research: Theory, method, and case studies. Amsterdam: Rodopi, 1-24.

Johansson, Stig (2007): Seeing through multilingual corpora. Amsterdam: Benjamins.

König, Ekkehard / Gast, Volker (2007): Understanding English-German contrasts. (= Grundlagen der Anglistik und Amerikanistik 29). Berlin: Erich Schmidt.

Legenhausen, Lienhard / Rohdenburg, Günter (1995): Kontrastivierung ausgewählter Strukturen im Englischen und Deutschen. In: Ahrens, Rüdiger / Bald, Wolf-Dietrich / Hüllen, Werner (eds.): Handbuch Englisch als Fremdsprache. Berlin: Erich Schmidt, 133-139.

Mair, Christian (1995): Englisch für Anglisten. Tübingen: Stauffenburg.

Matthiessen, Christian M.I.M. (1993): Register in the round: Diversity in a unified theory of register analysis. In: Ghadessy (ed.), 221-292.

Nesbitt, Christopher / Plum, Guenter (1988): Probabilities in a systemic-functional grammar: The clause complex in English. In: Fawcett, Robin P. / Young, David (eds.): New developments in systemic linguistics. Theory and application. Vol. 2. London: Pinter, 6-38.

Neumann, Stella (2003): Textsorten und Übersetzen. Eine Korpusanalyse englischer und deutscher Reiseführer. Frankfurt a.M.: Peter Lang.

Neumann, Stella (2008): Contrastive register variation. A quantitative approach to the comparison of English and German. Unpubl. habil. thesis. Saarbrücken: Universität des Saarlandes.

Neumann, Stella / Hansen-Schirra, Silvia (2005): The CroCo Project: Cross-linguistic corpora for the investigation of explicitation in translations. (= Proceedings from the Corpus Linguistics Conference Series 1). Internet: http://www.corpus.bham.ac.uk/PCLC/cl-134-pap.pdf (last visited: 08 / 2010).

Poynton, Cate (1985): Language and gender: making the difference. Victoria: Deakin University.

Steiner, Erich (2004): Translated texts: Properties, variants, evaluations. Frankfurt a.M.: Peter Lang.

Teich, Elke (2003): Cross-linguistic variation in system and text. A methodology for the investigation of translations and comparable texts. Berlin / New York: de Gruyter.

Ure, Jean (1971): Lexical density and register differentiation. In: Perren, George E. / Trim, John L.M. (eds.): Applications of linguistics: Selected papers of the Second International Congress of Applied Linguistics, Cambridge 1969. Cambridge: Cambridge University Press, 443-452.

Werlich, Egon (1976): A text grammar of English. Heidelberg: Quelle & Meyer.

THOMAS HERBST / SUSEN FAULHABER

Optionen der Valenzbeschreibung. Ein Valenzmodell für das Englische

Abstract

In diesem Aufsatz werden ein Modell zur Valenzbeschreibung für das Englische entwickelt und Grundfragen der Valenzbeschreibung diskutiert. Dabei wird von einem lexisorientierten Valenzkonzept ausgegangen, das mit formalen Satzstrukturen kombiniert wird. Daraus ergibt sich eine Unterscheidung zwischen formalen Valenzpatterns, semantisch bestimmten Partizipantenpatterns und Valenzkonstruktionen. Abschließend wird die auf dem *Valency Dictionary of English* (Herbst et al. 2004) beruhende und somit korpusbasierte *Erlangen Valency Patternbank* als Werkzeug für empirische Untersuchungen in diese Richtung vorgestellt.

1. Valenz zur Beschreibung des Englischen

Obwohl der Valenzansatz eine der gängigsten Theorien zur Beschreibung des Deutschen darstellt und auch auf die Beschreibung vieler anderer Sprachen angewandt wurde, spielt er in der englischen Linguistik noch immer eine eher untergeordnete Rolle. Zwar taucht der Terminus Valenz in letzter Zeit häufiger auch in amerikanischen und britischen Arbeiten auf: oft scheint dies aber eher einem Bestreben nach der Verwendung eines neuen Terminus als einer Rezeption der europäischen Valenztheorie geschuldet. Diese marginale Position des Valenzkonzepts in der angelsächsischen deskriptiven und theoretischen Linguistik lässt sich bis zu einem gewissen Grad wissenschaftsgeschichtlich erklären. Die dominierende Rolle des amerikanischen Strukturalismus und der generativen Transformationsgrammatik sind hierbei ebenso erwähnenswert wie eine in diesem Punkt sehr traditionell orientierte Grammatikschreibung und Lexikografie, die zum Beispiel ein mehr oder weniger weit gefasstes Phänomen *phrasal verbs* bzw. *phrasal* und *prepositional verbs* identifiziert und damit einen wichtigen Bereich aus der eigentlichen grammatischen Beschreibung als Sonderfall ausklammert. Letzteres zeigt aber auch, dass für die mangelnde Durchsetzung des Valenzkonzepts bei der Beschreibung des Englischen auch sprachimmanente Gründe verantwortlich sein mögen, etwa in dem Sinne,

dass das Modell für eine Sprache, in der Kasus eine relativ untergeordnete Rolle spielt, (zumindest auf den ersten Blick) weniger attraktiv erscheint. Von daher ist es sicherlich kein Zufall, dass viele der Arbeiten, die die Valenztheorie auf das Englische anzuwenden versuchen, im deutschsprachigen Raum entstanden sind (Emons 1974, 1978; Allerton 1982; Herbst 1983).[1] Trotz der „Kasusarmut" des heutigen Englisch stellt die Valenztheorie ein geeignetes Modell zur Beschreibung der englischen Sprache dar, und zwar

- zum einen, weil sich zeigen lässt, dass auch englische Verben, Adjektive oder Substantive über Valenzmuster verfügen, die als einzelwortspezifisch zu sehen sind und nicht hinreichend über generelle semantische Regeln oder allgemeinere Muster etwa im Sinne der Goldberg'schen (1995, 2006) *argument structure constructions* zu erklären sind (Herbst 2009, i.Dr.; Faulhaber i.Vorb.).

- zum anderen, weil das in der Standardgrammatik des Englischen, der *Comprehensive Grammar of the English Language* (Quirk / Greenbaum / Leech / Svartvik 1985) (im weiteren CGEL) von den Autoren beschrittene Verfahren, von Präpositionalverben (und nicht von Präpositionalergänzungen im Sinne der Valenztheorie) auszugehen, zwar den Vorteil hat, dass die Grammatik mit weniger Satzmustern auskommt, es gleichzeitig aber die Zahl der Lexeme der Sprache erheblich steigert, was beispielsweise nachteilige Auswirkung auf die Lemmatisierungspraxis englischer Lernerwörterbücher hat.

So würde man im Rahmen eines valenztheoretischen Ansatzes (Emons 1974, Herbst / Schüller 2008) in (1a-c) von einem Verb *invest* ausgehen;[2] CGEL hingegen identifiziert zwei Typen von *prepositional verbs* für (1b) (*type I*) und (1c) (*type II*).

(1a) *He invested money.*

(1b)$_{\text{CGEL}}$ *He invested in property.*[3]

(1c)$_{\text{CGEL}}$ *He invested his money in property.*

---

1 Eine wichtige Ausnahme stellt Somers (1987) dar.

2 Vgl. hierzu für das Deutsche beispielsweise auch Heringer (1967).

3 Die Quellen zitierter Beispielsätze werden mit Subskripten angegeben (siehe Literaturverzeichnis).

Auch wenn solche Fälle geradezu prototypische Beispiele für den von modernen Strömungen der Korpuslinguistik (Sinclair 1991, 2004) oder der Konstruktionsgrammatik betonten graduellen Übergang zwischen Lexik und Grammatik darstellen, spricht unseres Erachtens viel für die Berücksichtigung einer Perspektive in der linguistischen Beschreibung, in der in Fällen wie (1a-c) ein einziges Verb *invest* als Valenzträger angesehen wird.

Im Folgenden wollen wir einige Grundpositionen eines valenzbasierten Beschreibungsansatzes für das Englische darlegen, der im Wesentlichen die Grundlage des *Valency Dictionary of English* (VDE, Herbst et al. 2004) und der *Erlangen Valency Patternbank* darstellt. Dabei ergeben sich Unterschiede zu etablierten Modellen der Valenzbeschreibung für das Deutsche, die einerseits auf Unterschiede zwischen den beiden Sprachen, andererseits auch auf unterschiedliche theoretische Ausgangspunkte zurückzuführen sind.

2. Grundannahmen

Von seiner theoretischen Grundkonzeption her stellt der hier beschriebene Ansatz den Versuch dar, die Valenzperspektive mit Elementen der traditionellen Grammatik oder auch der Konstruktionsgrammatik zu verbinden. Dabei wird von folgenden Grundannahmen ausgegangen:

1) Valenz wird als eine Eigenschaft von Lexemen – oder genauer von *lexical units* im Sinne von Cruse (1986: 80) – gesehen.

2) Valenzträger können Verben, Substantive, Adjektive, Adverbien, Partikel (im Sinne der erweiterten Wortklasse Präposition bei Huddleston / Pullum 2002)[4] oder Mehrworteinheiten (wie etwa *at a loss*) sein.

3) Ergänzungen (im Gegensatz zu Angaben) sind Elemente, die in ihrer Form durch den Valenzträger bestimmt, in ihrer Position im Satz relativ festgelegt sind oder realisiert werden müssen, wenn der Valenzträger verwendet wird (Herbst / Schüller 2008: 22; Herbst et al.: XXIV-XXV).

4) Ein Valenzträger besitzt eine bestimmte Zahl von Valenzstellen,

   – die semantisch durch spezifische oder allgemeinere semantische Rollen (*participant roles*) charakterisiert werden,

[4] Vgl. Herbst / Schüller (2008), Pullum (2009).

- die formal durch Ergänzungen realisiert werden können,
- deren Realisierung obligatorisch, kontextuell-fakultativ oder fakultativ sein kann.

Weiterhin ist das Modell geprägt von dem Bemühen, möglichst oberflächennah vorzugehen, d.h. möglichst wenige Annahmen über möglicherweise zugrunde liegende Strukturen von Sätzen zu machen, was folgende Konsequenzen hat:

5) Kein Ergänzungstyp wird gegenüber anderen als primär angesehen.

Die Konsequenz aus dieser Maxime ist, dass unterschiedliche formale Realisierungen einer Valenzstelle als gleichberechtigte Ausdrucksformen betrachtet werden. In Beispielen wie

(2)$_{\text{NW62}}$ *They'll notice* ***the pickets****.*

(3)$_{\text{NW328}}$ *Haven't you noticed* ***that in the modern world good news comes by telephone and bad news by mail****?*

werden die fett gedruckten Elemente entsprechend als zwei verschiedene Realisierungen – Nominalphrase und *that-clause* – derselben Valenzstelle von *notice* analysiert. Dieses Verfahren ist ähnlich dem von Helbig / Schenkel (1973) eingeschlagenen, unterscheidet sich aber von dem im *Valenzwörterbuch deutscher Verben* (VALBU, Schumacher et al. 2004) bezüglich Beispielen wie (4) und (5) beschrittenen Weg.

(4)$_{\text{VALBU}}$ *Die Mehrzahl der Befragten hatte* ***einen Wahlsieg der Konservativen*** *angenommen.*

(5)$_{\text{VALBU}}$ *Realisten nehmen an,* ***dass die Grundsatzentscheidung über Atomkraft ... bis zum Sankt-Nimmerleins-Tag verschoben wird****.*

VALBU klassifiziert dabei den *dass*-Satz als „Satzförmige Ergänzung" (SE), führt ihn aber als Realisierungsform einer Akkusativergänzung (AkkE) auf, obwohl er natürlich keinerlei Kasusmarkierung trägt. Auch die Tatsache, dass es in VALBU (2004: 40) heißt, es werde angegeben, „wann neben den einfachen Ausdrucksformen eine Ergänzung als SE realisiert werden kann", deutet bereits in der Wahl der Terminologie auf eine Sonderstellung von Ka-

tegorien wie Akkusativergänzung hin. Für das Englische basiert das Valenzmodell von Emons (1974) in ähnlicher Weise auf über Kommutationsmöglichkeiten definierten Klassen.[5]

6) Es werden keine (oder möglichst wenige) Annahmen über zugrunde liegende syntaktische Strukturen oder 'fehlende' bzw. 'getilgte' Elemente gemacht.[6]

Das bedeutet zum Beispiel, dass ein Satz wie:

(6)$_{\text{VDE}}$ *We arranged to meet in two days time.*

als eine aus dem Valenzträger *arrange* und zwei Ergänzungen – [NP] und [to_INF] – bestehende Konstruktion analysiert wird und nicht – wie in der Grammatik von Quirk et al. (CGEL 1985: 16.38) – von einem *prepositional verb arrange for* ausgegangen wird, dessen Präposition bei Infinitivergänzungen getilgt wird.[7] Aus diesem Prinzip folgt ein weiteres, nämlich:

7) Der aktive Aussagesatz wird nicht als primär oder grundlegend gegenüber anderen betrachtet.

Diese Prämisse stellt wahrscheinlich den grundlegendsten Unterschied zu anderen Valenzmodellen dar. Er betrifft die Beschreibung der Ergänzungen sowie Angaben zum Grad der Optionalität. So klassifiziert etwa VALBU im Falle eines Satzes wie

(7)$_{\text{VALBU}}$ *Kinder essen gerne Pommes frites.*

---

[5] Deutlich wird die Problematik einer solchen Kennzeichnung auch im Hinblick auf die Tatsache, dass bei Satzergänzungen ein und dieselbe formale Ergänzung, beispielsweise ein *dass*-Satz, abhängig von ihrer Funktion im Satz einmal als Nominativergänzung (A) und ein anderes Mal als Akkusativergänzung (B) bezeichnet werden muss:

(A)$_{\text{VALBU}}$ [*Positiv*] *beeinflusst hat die jüngste Statistik über die Kriminalitätsrate,* ***dass 1997 weniger gestohlen wurde.***

(B)$_{\text{VALBU}}$ *Selbst der historische Wandel vom Gegeneinander zum Miteinander entschuldigt* [*nicht*], ***dass das Tauziehen nun schon sechs Jahre andauert.***

[6] Dieses Vorgehen steht in Einklang mit den Prinzipien der Linear Unit Grammar: „Earlier grammars used categories like 'ellipsis' and 'words understood' to explain this kind of phenomenon, but now that many linguists respect the actual wordings of corpora and are committed to describing the text and not some rewritten version of it, notions that there are some words missing or that the text cannot be understood as it stands are no longer tenable." (Sinclair / Mauranen 2006: 150).

[7] Vgl. CGEL (1985: 16.38): „The preposition is omitted before the infinitive clause object [...], but is present where the prepositional object is a noun phrase or, for that matter, an -ing-clause."

Kinder als obligatorische Nominativergänzung und *Pommes frites* als fakultative Akkusativergänzung.

(8)$_{\text{VALBU}}$ *Die Pommes frites waren schon nach wenigen Minuten gegessen, während die Salzkartoffeln stehen blieben.*

Dass im Passivsatz die „obligatorische" Nominativergänzung gar nicht vorkommt und die fakultative „Akkusativergänzung" als Nominativ realisiert ist, erklärt VALBU folgendermaßen: „Im Passivsatz ist die syntaktische Ausprägung der semantischen Rollen gegenüber dem entsprechenden Satz im Aktiv erheblich verändert" (2004: 55). Natürlich ist es eine Option, bei der Beschreibung der Valenz von Verben den aktiven Aussagesatz sozusagen zur Norm zu machen oder als Ausgangspunkt zu nehmen und die Valenzverhältnisse in Imperativ- oder Passivsätzen über Regeln wie „Valenzreduktion" (Helbig / Schenkel 1973, Welke 1988) zu erklären.[8] Sieht man eine Valenzbeschreibung jedoch als Aussage über das syntaktische Potenzial eines Verbs, so spricht viel dafür, die Frage, ob ein Element als fakultativ oder obligatorisch zu klassifizieren ist, unabhängig davon zu bestimmen, ob es unter den strukturellen und kontextuellen Gegebenheiten eines bestimmten Satzes oder Satztyps weggelassen werden kann oder nicht. Unter dem Gesichtspunkt der Perspektivierung erscheint es uns legitim anzunehmen, dass das Wissen darüber, welche semantischen Rollen, die ein Verb ausdrücken kann, realisiert werden müssen und welche optional sind, zum Wissen über das entsprechende Verb gehört. Aus diesem Grund definieren wir eine Valenzstelle (nicht eine Ergänzung) als obligatorisch, wenn sie realisiert werden muss, wenn das Verb verwendet wird, und als kontextuell-fakultativ, wenn sie dann nicht realisiert werden muss, wenn der durch die Rolle bezeichnete Referent im Kontext identifizierbar ist, und ansonsten als fakultativ. Dies bedingt natürlich eine Unterscheidung zwischen Notwendigkeit auf der Ebene der Valenz und einer strukturellen Notwendigkeit, in der bestimmte Elemente im Satz als obligatorisch etc. zu klassifizieren wären.[9] In dieser Sichtweise wäre *they* in:

---

8 Vgl. in diesem Zusammenhang auch den Ansatz von Allerton (1982: 45).

9 Die Frage, wie man die Optionalität von Valenzstellen oder Ergänzungen am Besten beschreibt, kann hier nicht in der gebührenden Ausführlichkeit diskutiert werden. Zu Diskussionen im Rahmen der Valenztheorie siehe z.B. Engelen (1975: 61-67), der diesbezüglich insgesamt fünf Dichotomien diskutiert („nichtweglassbar vs. weglassbar", „notwendig vs. nichtnotwendig (oder frei)", „obligatorisch vs. fakultativ", „konstitutiv vs. nichtkonstitutiv (oder frei)" und „spezifisch vs. nichtspezifisch"), oder auch Allerton (1982: 68-70), Helbig (1992: 99-107) und Herbst / Schüller (2008). Zu Fragen bezüglich der Weglassbarkeit, Fakultativität und Obligatheit von Ergänzungen siehe außerdem Ágel (2000: 247-267). Zu sehr ähnlichen Unterscheidungen vgl. auch Fillmore (2007: 144-149) oder Goldberg

(2)$_{NW62}$ ***They**'ll notice the pickets.*

aus strukturellen Gründen als obligatorisch anzusehen, weil es im vorliegenden Satz nicht weggelassen werden kann. Gleichzeitig realisiert es aber eine fakultative Valenzstelle des Verbs, die etwa im Passiv oder im Imperativ nicht realisiert werden muss.

(9)$_{BNC}$ *It was probably 2.06 a.m. when the forced window was noticed.* [HWL 1675]

(10)$_{BNC}$ *Notice that the term semantic constituent is not used to refer to a meaning only, but to a form-plus-meaning-complex.* [FAC 485]

3. Charakterisierung der Ergänzungen

Aus den oben ausgeführten Prinzipien für die Valenzbeschreibung folgt, dass die Ergänzungen von Verben in Hinblick auf folgende Eigenschaften charakterisiert werden:

1) Grundlage der Beschreibung bildet die formale Realisierung, also die Angabe des entsprechenden Phrasen- oder Nebensatztyps: [NP], [to_INF], [that_CL]. Im Rahmen des Konzepts der Partikelergänzungen, das weitgehend dem der Präpositionalergänzungen entspricht, werden Ergänzungen, die von einer traditionellen Präposition ([about_NP] oder [about_V-ing]) oder Konjunktion ([that_CL]) eingeleitet werden, subsumiert.
2) Nimmt man nicht den aktiven Aussagesatz als Basis für die Beschreibung der Ergänzungen, so bedeutet dies natürlich auch, dass für jede (betroffene) Valenzstelle anzugeben ist, welche Realisierungen sie im Aktiv- und Passivsatz haben kann. Im *Valency Dictionary of English* geschieht das dadurch, dass durch tiefgestellte Indizes die Möglichkeit des Vorkommens als Subjekt im Aktiv- bzw. Passivsatz zum Ausdruck gebracht wird, also etwa: [NP$_{act\text{-}subj}$/by_NP], [NP$_{pass\text{-}subj}$]; Ergänzungen ohne Index können nicht als Subjekt fungieren.[10]

(2006: 39f.), bei denen aber wiederum der Aktivsatz als Basis dient; dazu auch Herbst (2010). Es sei jedoch betont, dass die Tatsache, dass eine Valenzstelle als fakultativ oder kontextuell-fakultativ analysiert wird, nicht notwendigerweise bedeutet, dass diese nie realisiert werden muss. So scheint bei *notice* die Valenzstelle II nur dann kontextuell-fakultativ, wenn sie semantisch auf eine Proposition verweist – wie beispielsweise in *Outside, the night was unusually clear for Los Angeles, but Milton was in no mood to notice.* (Quelle: VDE) – nicht aber in Fällen wie in Beispielsatz (2).

[10] Vgl. Herbst/Schüller (2008: 117). Im VDE werden $_A$ und $_P$ verwendet. Für Pronomen impliziert diese Markierung auch eine Aussage über die entsprechende Kasusform. Mit diesen Indizes ist auch die Aussage über den Grad der Optionalität verbunden, wie sie für Subjekte in bestimmten Satzstrukturen gelten. Vgl. Herbst/Schüller (2008: 99).

3) Soweit formal identische Ergänzungstypen auftreten und eine übliche (thematisch unmarkierte) Reihenfolge der Ergänzungen angegeben werden kann, wird dies durch tiefgestelltes 1 bzw. 2 markiert.

Die Valenzstellen werden außer in Hinblick auf ihre formalen Realisierungsmöglichkeiten durch die aufgeführten Ergänzungen noch in Bezug auf den Grad der Optionalität und durch semantische Rollen markiert. Auf das Problem der Differenzierung semantischer Rollen soll hier nicht eingegangen werden. Für bestimmte Zwecke, etwa wenn es darum geht, die Zusammenhänge zwischen verschiedenen Valenzpatterns zu verdeutlichen, kann es sich dabei um einzelwortspezifische Partizipanten-Rollen handeln. Darüber hinaus erscheinen in manchen Fällen auch Generalisierungen im Sinne allgemeinerer Rollen wie AGENT, AFFECTED oder PREDICATIVE möglich und sinnvoll, was aber nicht bedeutet, dass behauptet werden soll, jede Valenzstelle jedes Verbs sei durch Zuweisung einer solch allgemeinen semantischen Rolle adäquat zu charakterisieren.[11]

4. Elemente der Satzanalyse

In dem von uns für das Englische entwickelten Ansatz (Herbst / Schüller 2008: 148-156) wird die Valenzbeschreibung komplementiert durch eine Reihe von Satztypen wie zum Beispiel die

- declarative-'statement'-construction;
- *wh*-interrogative-'question'-construction;
- imperative-'directive'-construction etc.

Diese Satztypen sind hinsichtlich bestimmter Eigenschaften des Subjekts und des Prädikats charakterisiert, wobei die Frage, ob das Prädikat neben dem verbalen Valenzträger Ergänzungen enthält oder nicht, keine Rolle spielt. Um

[11] Die Erfahrungen bei der Erstellung des VDE weisen eher in die gegenteilige Richtung. Dort werden Rollen nur in seltenen Fällen im Ergänzungsinventar verwendet, um formal gleiche Ergänzungen voneinander abzugrenzen. Ansonsten wird teilweise der Versuch unternommen, die semantische Funktion der Ergänzungen verbspezifisch in den semantischen Kommentaren zu den einzelnen Einträgen quasi ad hoc und nicht im Rahmen eines begrenzten Rolleninventars zu beschreiben. Zum Problem allgemeiner und spezifischer Rollen in FrameNet siehe auch Fillmore (2007: 131). Vgl. in diesem Zusammenhang insbesondere auch die Unterscheidung zwischen *participant roles* und *clausal roles* bei Herbst / Schüller (2008: 158-163) zur Erklärung des Unterschieds zwischen viel diskutierten Sätzen wie *Pat loaded the wagon with hay* und *Pat loaded the hay onto the wagon* (Beispielsätze aus Goldberg 2006: 34-37).

eine Kombination der lexikalisch bedingten Valenzperspektive mit der Analyse der strukturellen Eigenschaften der Satztypen zu erreichen, wird eine Terminologie zur Beschreibung vorgeschlagen, die beide Perspektiven kombiniert. Folgende Einheiten werden dabei angesetzt:

- **subject complement unit** (SCU): eine Einheit, die als Subjekt des Satzes fungiert und durch eine Ergänzung des regierenden Verbs realisiert wird (*you*);[12]
- **predicate complement unit** (PCU): eine Einheit, die eine Konstituente des Prädikats darstellt und durch eine Ergänzung des regierenden Verbs realisiert wird (*me* und *a reference*);
- **predicate head unit** (PHU): eine Einheit, die den Valenzträger und sog. *pre-heads* enthält (*will write*);
- **adjunct unit** (AU): Angabe (*then*);
- **linking unit** (LU): eine Einheit, die Sätze koordiniert.

(11)$_{\text{NW361}}$	*You*	*will write*	*me*	*a reference*	*then?*
	SCU	PHU	PCU1	PCU2	AU
(12)$_{\text{NW215}}$	*But*	*this year*	*the winter term*	*was*	*different*
	LU	AU	SCU	PHU	PCU

Dies ermöglicht sowohl einen hierarchischen Blickwinkel mit dem Satz als höchste Einheit und einer Unterteilung in der Form einer Konstituentenanalyse als auch eine deutliche Abgrenzung zwischen valenzgebundenen Elementen und solchen, die vom Valenzträger unabhängig sind. Überdies gestattet eine Analyse, die funktionale Konstituenten (wie *subject* und *predicate*) mit einer

[12] Der Begriff *subject complement unit* entspricht dem deutschen Begriff *Subjektergänzung* und bezeichnet damit die Ergänzung, die als das Subjekt eines Satzes fungiert. Der Begriff *subject complement **unit*** macht dabei deutlich, dass diese Kategorie nicht identisch ist mit der in der CGEL (1985: 10.8) als *subject complement* beschriebenen Kategorie. Dort wird dieser Begriff für solche Einheiten in der Satzstruktur verwendet, die sich auf das Subjekt beziehen, wie beispielsweise *different* in Satz (12). Abweichend von dieser Terminologie verwenden die meisten anderen Darstellungen hierfür nicht den Begriff *subject complement* (und analog *object complement* für Einheiten, die sich auf das Objekt eines Satzes beziehen), sondern Begriffe wie *subject attribute* (und *object attribute*) (Aarts / Aarts 1982 / 1988), *subject predicative* (und *object predicative*) (Biber et al. 1999: 126, 130) oder *predicative complement* (Huddleston / Pullum 2002: 53f.). In der hier verwendeten Terminologie entsprechen diese Einheiten PCUs mit der semantischen Rolle 'PREDICATIVE'. Vgl. hierzu auch Jespersens (1927: 355-404) Verwendung des Begriffs *predicative*.

dependenziellen Beschreibung vereint (*complement*), eine lineare Satzanalyse ohne dabei die zentrale Auffassung von Valenz als lexikalisches Phänomen aufweichen zu müssen. Der lineare Charakter der Darstellung betont überdies die wichtige Rolle von Patterns als eigene Einheiten in der Sprachbeschreibung gegenüber einer Betrachtung der Ergänzungen eines Valenzträgers in der Form eines Ergänzungsinventars.[13]

5. *valency patterns*, *participant patterns* und *valency constructions*

Unter dem Gesichtspunkt der Valenz lassen sich Patterns auf verschiedenen Ebenen feststellen:

- **Valenzpatterns** (***valency patterns***) sind Muster, die sich auf das gemeinsame Vorkommen von formalen Ergänzungen beziehen (unabhängig von der semantischen Rolle).[14]

$(13)_{NW215}$	*I*	*wouldn't call*	*myself*	*a structuralist*	
$(11)_{NW361}$	*You*	*will write*	*me*	*a reference*	*then?*
	SCU: NP	PHU: $VHC_{act:3}$	PCU1: NP	PCU2: NP	

- **Partizipantenpatterns** (***participant patterns***) sind Muster, die sich auf das gemeinsame Vorkommen von semantischen Rollen beziehen (unabhängig von der formalen Realisierung).

$(13)_{NW215}$	*I*	*wouldn't call*	*myself*	*a structuralist*
	AGENT		AFFECTED	PREDICATIVE
$(14)_{NW51}$	*Robyn*	*considered*	*herself*	*lucky to get a job for one term at one of the London colleges...*
	AGENT		AFFECTED	PREDICATIVE

- **Valenzkonstruktionen** (***valency constructions***) sind im Sinne der Konstruktionsgrammatik „form-meaning pairings", also Verbindungen aus Valenzpatterns und Partizipantenpatterns. Sie beziehen sich also auf das gemeinsame Vorkommen von formalen Ergänzungen mit denselben Rollen.

[13] Bezüglich der jeweiligen Vorzüge einer Valenzbeschreibung in der Form eines Patterninventars beziehungsweise eines Ergänzungsinventar siehe auch Herbst (2007).

[14] Siehe in diesem Zusammenhang die Unterscheidung von Satzmuster und Satzbauplan in VALBU (2004: 46f.), wobei in letzteren zwischen obligatorischen und fakultativen Ergänzungen unterschieden wird. Vgl. dazu auch Engel (1977: 180f.).

$(13)_{NW215}$	*I*	*wouldn't call*	*myself*	*a structuralist*
	NP AGENT	$VHC_{act:3}$	NP AFFECTED	NP PREDICATIVE
$(14)_{NW51}$	*Robyn*	*considered*	*herself*	*lucky to get a job for one term at one of the London colleges…*
	NP AGENT	$VHC_{act:3}$	NP AFFECTED	AdjP PREDICATIVE

Valenzkonstruktionen kommen von der theoretischen Konzeption her den *argument structure constructions* von Goldberg (1995, 2006) sehr nahe, sind aber in Bezug auf die formalen Realisierungen spezifischer. Dies ist insbesondere dann von Interesse, wenn man mögliche Parallelen zwischen den semantischen Eigenschaften von Wörtern und ihren syntaktischen Vorkommensmöglichkeiten untersucht. Obwohl Goldberg ein paar wenige Fälle von Einzelwortspezifik diskutiert, spricht sie dennoch von einer deutlichen Tendenz, dass semantisch ähnliche Verben in denselben *argument structure constructions* vorkommen.[15] Dies ist deutlich zu relativieren, wenn man sieht, dass etwa formal so unterschiedliche Ergänzungen wie die Nominalphrase in (13) und die Adjektivphrase in (14) derselben *argument structure construction* zugerechnet werden. Insofern beziehen sich die Valenzkonstruktionen in unserem Ansatz auf eine konkretere und weniger abstrakte Beschreibungsebene.

6. Valenzbeschreibungen in VDE und Patternbank

Auch wenn das eigentliche Ziel einer umfassenden Beschreibung der Valenzstrukturen des Englischen in einer Darstellung der Valenzkonstruktionen liegt, stellt die Identifikation der rein formalen Valenzpatterns einen ersten Schritt in diese Richtung dar. Mit der *Erlangen Valency Patternbank*[16] wollen wir einen Beitrag in diese Richtung leisten.

- In der augenblicklichen Fassung beruht die Patternbank auf den Daten des *Valency Dictionary of English*. Sie umfasst 511 Verben, 544 Adjektive und 274 Substantive der englischen Sprache, wobei der Anteil der dadurch abgedeckten Verbvorkommen weitaus höher ist als es zunächst den Anschein

[15] Vgl. Goldberg (2006: 58): „Semantically similar verbs show a strong tendency to appear in the same argument structure constructions." Gleichzeitig verweist Goldberg aber durchaus darauf, dass Verben hinsichtlich ihres Vorkommens in verschiedenen *argument structure constructions* bisweilen ziemlich idiosynkratisch sind: „Verbs are occasionally quite idiosyncratic in the types of argument structure patterns they appear in." (ebd.: 56).

[16] Siehe Herbst / Uhrig (2009).

hat – schätzungsweise machen die im VDE und der Patternbank erfassten Verben etwa zwei Drittel aller Verbverwendungen des British National Corpus aus.

- Die Beschreibung ist korpusbasiert – alle Beispiele des VDE basieren auf dem von John Sinclair in Birmingham entwickelten Cobuild-Korpus. Dennoch enthält das VDE eine Reihe von Informationen – etwa bei Angaben zur Passivierung oder zu Subjekten, die nicht durch Nominalphrasen realisiert werden –, die auf der Intuition der muttersprachlichen Herausgeber des VDE beruhen. Eine Erweiterung und Überprüfung der Daten anhand weiterer Korpora ist geplant.

Die Patternbank kann als erster Schritt zu einem Online-Valenzwörterbuch für das Englische gesehen werden. In der augenblicklichen Fassung enthält sie jedoch wichtige Informationstypen, die im VDE enthalten sind, nicht – nämlich Beispiele, ausführliche Bedeutungsangaben und die längeren Kommentare des VDE, in denen Angaben zur Bedeutung von Patterns bzw. zu semantischen und lexikalischen Realisierungsmöglichkeiten einzelner Ergänzungen gemacht werden, wie beispielsweise im *noteblock* des Eintrags zu *consider* (VDE: 176):

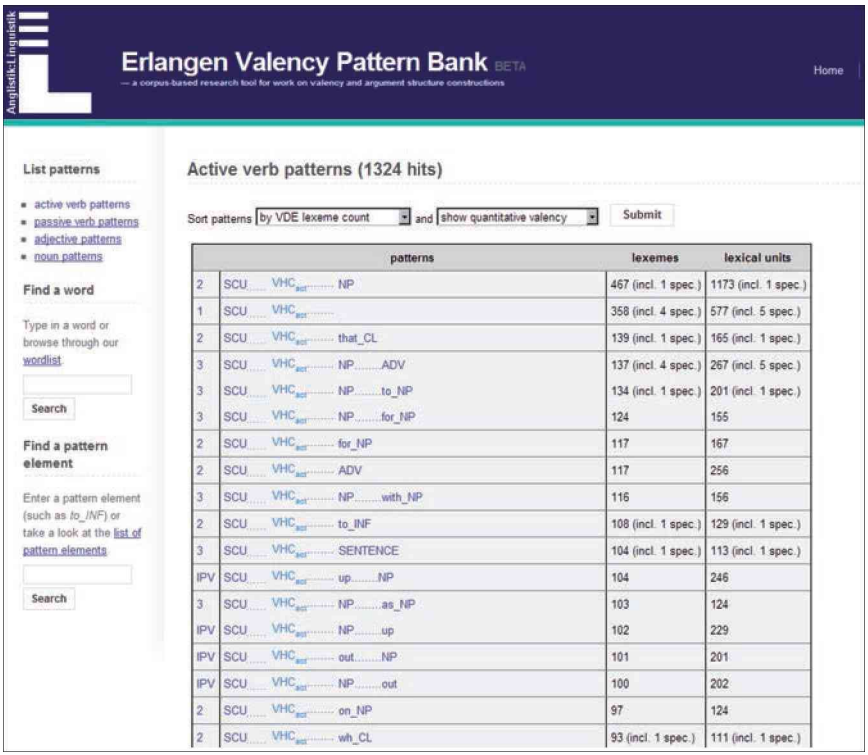
A *Consider* can mean
(i) 'think about something very carefuully'; typically used in the phrase *consider a problem*
(ii) 'think seriously about someone as a candidate or something as an option', typically used in sentences such as *She considered him for the job* or *They considered going to Amsterdam*
(iii) 'take something into account as an important factor'; used in sentences such as *You should consider your children when moving house.*
→ M Dl D2 D3 D4 D5 T4 T6
B Consider can mean 'have a certain view of someone or something'. → D3 D6 T1 T2 T3 T4 T5

Abb. 1: ***Noteblock*** **des Eintrags** ***consider*** **(VDE: 176)**

Im Wesentlichen wird durch die Patternbank eine neue Art des Zugriffs auf die VDE-Daten geschaffen. Die Abruf- und Anzeigemöglichkeiten sind vielfältig:

- **Abruf** von aktiven Verb-, passiven Verb-, Adjektiv- und Substantivpatterns;
- **Sortierung** nach Zahl der Lexeme oder Zahl der *lexical units*, bei denen das entsprechende Pattern verzeichnet ist, oder alphabetisch nach dem ersten Element des Patterns (*subject complement unit*) oder alphabetisch nach der ersten *predicate complement unit*;

- **Sortierung** nach und **Anzeige** der quantitativen Valenz;
- **Subjekte** (*subject complement units*) als SCU (für alle Subjekte, die durch NPs und möglicherweise andere Elemente realisiert werden können), [it] und [there] oder Detailansicht mit genauerer Spezifizierung;
- Anzeige der Zahl aller im VDE erfassten **Wörter** bzw. ***lexical units*** mit dem entsprechenden Pattern.



	patterns	lexemes	lexical units
2	SCU VHC$_{act}$ NP	467 (incl. 1 spec.)	1173 (incl. 1 spec.)
1	SCU VHC$_{act}$	358 (incl. 4 spec.)	577 (incl. 5 spec.)
2	SCU VHC$_{act}$ that_CL	139 (incl. 1 spec.)	165 (incl. 1 spec.)
3	SCU VHC$_{act}$ NP ADV	137 (incl. 4 spec.)	267 (incl. 5 spec.)
3	SCU VHC$_{act}$ NP to_NP	134 (incl. 1 spec.)	201 (incl. 1 spec.)
3	SCU VHC$_{act}$ NP for_NP	124	155
2	SCU VHC$_{act}$ for_NP	117	167
2	SCU VHC$_{act}$ ADV	117	256
3	SCU VHC$_{act}$ NP with_NP	116	156
2	SCU VHC$_{act}$ to_INF	108 (incl. 1 spec.)	129 (incl. 1 spec.)
3	SCU VHC$_{act}$ SENTENCE	104 (incl. 1 spec.)	113 (incl. 1 spec.)
IPV	SCU VHC$_{act}$ up NP	104	246
3	SCU VHC$_{act}$ NP as_NP	103	124
IPV	SCU VHC$_{act}$ NP up	102	229
IPV	SCU VHC$_{act}$ out NP	101	201
IPV	SCU VHC$_{act}$ NP out	100	202
2	SCU VHC$_{act}$ on_NP	97	124
2	SCU VHC$_{act}$ wh_CL	93 (incl. 1 spec.)	111 (incl. 1 spec.)

Abb. 2: Die Erlangen Valency Pattern Bank

Durch Anklicken oder spezielle Eingaben lassen sich folgende Informationen finden:

- (Bei Anklicken des Patterns): Liste aller VDE-Lemmata mit dem entsprechenden Pattern (mit Markierung von Wörtern, die etwa in entsprechenden Passivpatterns vorkommen);
- (Bei Anklicken von Wörtern in dieser Liste oder Sucheingabe des Wortes) Liste aller Patterns, mit denen das entsprechende Wort im VDE auftritt;

– (Bei Suche nach einer bestimmten Ergänzung): Anzeige aller Patterns, in der diese Ergänzung vorkommt.

Besonders erwähnenswert erscheint die Tatsache, dass Fälle, in denen besondere kontextuelle oder lexikalische Beschränkungen bestehen, eigens markiert sind. So treten zum Beispiel einige Patterns des Substantivs *loss* nur in der Verbindung *at a loss* auf:

(15)$_{\text{VDE}}$ *No wonder the mainstream parties are at a loss to know how to entice them back into the fold.*

Solche Restriktionen – oder Tatsachen wie die, dass *drop* in einem Pattern mit Adjektivergänzung nur mit dem Adjektiv *dead* vorkommt – werden in der Patternbank dadurch zum Ausdruck gebracht, dass diese Fälle als lexikalisch oder kontextuell spezifizierte Patterns unter dem Hauptpattern aufgeführt und die Lemmata in den Listen farblich hervorgehoben werden. Wir hoffen, auf diese Weise dem Übergangscharakter zwischen Grammatik und Lexis gerecht zu werden, der im Einzelfall auch zu unterschiedlichen Ansichten darüber führen kann, was als Pattern aufzufassen ist und was nicht.

Diese Art der Darstellung steht in Einklang mit dem generellen Ziel der Patternbank, eine möglichst theorieneutrale Beschreibung der Valenzverhältnisse der erfassten Wörter zu liefern. Dass dies nur bedingt möglich ist, weil zum Beispiel Entscheidungen darüber, welche Ergänzungstypen für eine Sprache anzusetzen sind, unweigerlich auch die Darstellung beeinflussen, versteht sich von selbst.[17]

Dennoch hoffen wir, mit der Patternbank ein Instrument für empirisch unterstützte korpusbasierte Forschung im Bereich von Valenzphänomenen zu bieten. Dabei ist zu berücksichtigen, dass die Daten lediglich die Auswertung des Cobuild-Korpus zum Zeitpunkt der Erstellung des VDE widerspiegeln und in mancherlei Hinsicht ergänzungsbedürftig sind, sowohl was die Anzahl der Lemmata als auch was die Vollständigkeit der erfassten Patterns angeht, was durch die dank der Zugriffsmöglichkeiten der Patternbank teils erst geschaffene Perspektive auf die VDE-Daten in einzelnen Punkten sehr deutlich wird.[18]

[17] Das betrifft zum Beispiel Entscheidungen bezüglich der Analyse einzelner Konstruktionen, etwa solcher, die als divalent oder als trivalent analysiert werden können, je nachdem, ob man einen Ergänzungstyp [NP_to_INF] zulässt oder eine Kombination aus [NP] und [to_INF] annimmt.

[18] Das gilt besonders für alle die Bereiche, die im VDE vor allem auf der Einschätzung von *native speakers* beruhen, wie zum Beispiel Passivpatterns, die Realisierungsmöglichkeiten von SCUs außer NPs oder *wh*-Realisierungen in Partikelergänzungen.

Aus diesem Grund enthält die Patternbank auch ein interaktives Internet-Portal, in dem wir um entsprechende Vorschläge bitten. Die Patternbank soll als Ausgangspunkt für Fragen zum Status von Patterns in der Sprachbeschreibung dienen. Dabei soll zum Beispiel untersucht werden, inwieweit die identifizierten Valenzpatterns mit Partizipantenpatterns korrelieren, also als Valenzkonstruktionen anzusehen sind bzw. inwieweit eine eindeutige Zuweisung der in einem Pattern auftretenden Verben zu bestimmten Partizipantenpatterns überhaupt möglich ist. Von vorrangigem Interesse ist in diesem Zusammenhang auch die hohe Zahl von nur bei einem Wort auftretenden Patterns. Soweit sich bei einer Überprüfung der Daten und Erweiterung der Datenbasis tatsächlich ein erheblicher Anteil von Patterns finden sollte, die nur bei einem einzigen oder nur bei sehr wenigen Wörtern auftreten,[19] so hätte dies erhebliche Konsequenzen für den Status von Valenzpatterns bzw. Valenzinformation im mentalen Lexikon. Die Vielfalt der in der Patternbank dokumentierten Valenzstrukturen bietet jedenfalls einen Beleg für die Notwendigkeit der Berücksichtigung einer Ebene der Valenz, die über allgemeine Strukturen wie *argument structure constructions* hinausgeht.

Das hier beschriebene Valenzmodell ermöglicht durch die separate formale, funktionale und semantische Beschreibung der einzelnen Satzelemente eine differenzierte Analyse von Sätzen. Somit werden Übergeneralisierungen – wie sie etwa durch Kategorien wie *Objekt*, die teils formal und teils semantisch definiert sind[20] – vermieden. Gleichzeitig erlaubt das Modell dadurch, dass es die Betrachtung der einzelnen Ebenen getrennt voneinander ermöglicht, tatsächliche Parallelen zwischen diesen Ebenen zu erkennen und somit einen genaueren Beschreibungsgrad in Bezug auf die Beziehungen zwischen Semantik und Form und die Postulierung möglicher Konstruktionen zu erzielen.

Literatur

Aarts, Flor / Aarts, Jan (1982 / 1988): English syntactic structures. New York / Leyden: Prentice Hall / Martinus Nijhoff.

Ágel, Vilmos (2000): Valenztheorie. Tübingen: Narr.

Allerton, David (1982): Valency and the English verb. London / New York: Academic Press.

[19] Es ist zu erwarten, dass der tatsächliche Anteil an Hapax-Patterns an der Gesamtzahl der Patterns weit geringer ist als es die Patternbank im Augenblick widerspiegelt, weil der augenblickliche Stand durch Datenlage, lexikografische Inkonsequenzen und Ziele des VDE bei der Erstellung des Wörterbuchs und die begrenzte Zahl der erfassten Wörter bedingt sein mag.

[20] Vgl. hierzu auch Aarts / Aarts (1982/1988) und Meyer (2009).

Biber, Douglas / Johansson, Stig / Leech, Geoffrey / Conrad, Susan / Finegan, Edward (1999): Longman Grammar of Spoken and Written English. Harlow: Pearson.

Cruse, D. Alan (1986): Lexical semantics. Cambridge: Cambridge University Press.

Emons, Rudolf (1974): Valenzen Englischer Prädikatsverben. Tübingen: Niemeyer.

Emons, Rudolf (1978): Valenzgrammatik für das Englische. Eine Einführung. Tübingen: Niemeyer.

Engel, Ulrich (1977): Syntax der deutschen Gegenwartssprache. (= Grundlagen der Germanistik 22). Berlin: Schmidt.

Engelen, Bernhard (1975): Untersuchungen zu Satzbauplan und Wortfeld in der geschriebenen deutschen Sprache der Gegenwart. Teilbd. 1. München: Hueber.

Faulhaber, Susen (i.Vorb.): Verb Valency Patterns. A challenge for semantics-based accounts. Berlin / New York: de Gruyter.

Fillmore, Charles J. (2007): Valency issues in FrameNet. In: Herbst / Götz-Votteler (Hg.), 129-160.

Goldberg, Adele E. (1995): Construction. A construction grammar approach to argument structure. Chicago: University of Chicago Press.

Goldberg, Adele E. (2006): Constructions at work. The nature of generalization in language. New York: Oxford University Press.

Helbig, Gerhard (1992): Probleme der Valenz- und Kasustheorie. (= Konzepte der Sprach- und Literaturwissenschaft 51). Tübingen: Niemeyer.

Helbig, Gerhard / Schenkel, Wolfgang (1973): Wörterbuch zur Valenz und Distribution deutscher Verben. 2. Aufl. Leipzig: VEB Bibliographisches Institut.

Herbst, Thomas (1983): Untersuchungen zur Valenz englischer Adjektive und ihrer Nominalisierungen. (= Tübinger Beiträge zur Linguistik 233). Tübingen: Narr.

Herbst, Thomas (2007): Valency complements or valency patterns. In: Herbst / Götz-Votteler (Hg.), 3-35.

Herbst, Thomas (2009): Valency – item-specificity and idiom principle. In: Römer / Schulze (Hg.), 49-68.

Herbst, Thomas (2010): Valency constructions and clause constructions or how, if at all, valency grammarians might sneeze the foam off the cappuccino. In: Schmid, Hans-Jörg / Handl, Susanne (Hg.): Cognitive foundations of linguistic usage patterns. Empirical studies. (= Applications of cognitive linguistics 13). Berlin / New York: de Gruyter.

Herbst, Thomas / Götz-Votteler, Katrin (Hg.) (2007): Valency. Theoretical, descriptive and cognitive issues. (= Trends in linguistics. Studies and monographs 187). Berlin / New York: de Gruyter.

Herbst, Thomas / Heath, David / Roe, Ian / Götz, Dieter (2004): A Valency Dictionary of English. Berlin / New York: de Gruyter.

Herbst, Thomas / Schüller, Susen (2008): Introduction to syntax. A valency approach. Tübingen: Narr.

Herbst, Thomas / Uhrig, Peter (2009): The Erlangen Valency Patternbank. Internet: http://www.patternbank.uni-erlangen.de (Stand: 07 / 2010).

Heringer, Hans Jürgen (1967): Wertigkeiten und nullwertige Verben im Deutschen. In: Zeitschrift für Deutsche Sprache 23: 13-34.

Huddleston, Rodney / Pullum, Geoffrey (2002): The Cambridge Grammar of the English Language. Cambridge: Cambridge University Press.

Jespersen, Otto (1927): Modern English grammar. On historical principles. Part III: Syntax. Bd. 2 Heidelberg: Carl Winter Universitätsbuchhandlung.

Meyer, Matthias L.G. (2009): Revisiting the evidence for objects in English. In: Römer / Schulze (Hg.), 211-227.

Pullum, Geoffrey (2009): Lexical categorization in English dictionaries and traditional grammars. In: Zeitschrift für Anglistik und Amerikanistik 57: 255-273.

Quirk, Randolph / Greenbaum, Sidney / Leech, Geoffrey / Svartvik, Jan (1985): A comprehensive grammar of the English language. London: Longman.

Römer, Ute / Schulze, Rainer (Hg.) (2009): Exploring the lexis-grammar interface. Amsterdam / Philadelphia: Benjamins.

Schumacher, Helmut / Kubczak, Jacqueline / Schmidt, Renate / de Ruiter, Vera (2004): VALBU. Valenzwörterbuch deutscher Verben. (= Studien zur deutschen Sprache 31). Tübingen: Narr.

Sinclair, John (1991): Corpus, concordance, collocation. Oxford: Oxford University Press.

Sinclair, John (2004): Trust the text. Language, corpus and discourse. London / New York: Routledge.

Sinclair, John / Mauranen, Anna (2006): Linear unit grammar: Integrating speech and writing. Amsterdam / Philadelphia: Benjamins.

Somers, Harold L. (1987): Valency and case in computational linguistics. Edinburgh: Edinburgh University Press.

Welke, Klaus (1988): Einführung in die Valenz- und Kasustheorie. Leipzig: VEB Bibliographisches Institut.

Quellen für Beispiele

BNC British National Corpus. Internet: http://www.natcorp.ox.ac.uk

CGEL Quirk et al. (1985)

VALBU Schumacher et al. (2004)

VDE Herbst et al. (2004)

NW Lodge, David (1989): Nice Work. Harmondsworth: Penguin (erstmals erschienen 1988).

AMIR ZELDES

On the Productivity and Variability of the Slots in German Comparative Correlative Constructions

Abstract

This paper is concerned with the syntactic productivity and semantic function of the comparative slots in the German comparative correlative construction (*je* [COMPARATIVE] ... *desto* [COMPARATIVE] and variants), i.e., how prone they are to admitting novel forms under different circumstances, and what meaning each slot is used to express in practice. Using large amounts of corpus data and quantitative productivity measures, it will be shown that comparatives in one slot behave differently from those in the other and from comparatives in general, both in terms of the lexemes they exhibit and in terms of their potential for innovation. Qualitatively, the construction is stereotypically employed to express positive or negative evaluation semantics in the *desto* clause, which depend on a spatiotemporal quantity in the *je* clause. Finally, differences are examined between cases exhibiting nominal subjects and verbal predicates in each clause and cases where these do not appear.

1. Introduction

Comparative correlative constructions (henceforth CCs) are sentences correlating two clauses with respect to comparative adjectives appearing in each clause, as in example (1):

(1) [$_{\text{CLAUSE1}}$ *Je schneller Hans rennt*], [$_{\text{CLAUSE2}}$ *desto schneller wird er müde*]
'The faster Hans runs, the faster he gets tired' (adapted from Beck 1997: 234)

(2) [$_{\text{CLAUSE1}}$ *Je früher*], [$_{\text{CLAUSE2}}$ *desto besser*]
'The sooner, the better'

Such sentences have enjoyed considerable attention, especially in the syntactic literature of recent years, yet surprisingly, little has been said about the usage of their central variable component: the comparative adjectives in each clause. This article addresses this gap for German by asking several questions about the sorts of lexemes that typically occupy each clause: how does clause1 (henceforth c1) differ from clause2 (henceforth c2) in its lexical preferences? How free are speak-

ers to innovate with the comparatives they use in c1 and c2? Are there any differences between usage in sentences like (1), with full subjects and predicates in each clause, and shorter sentences like (2) (hence short CCs), which only contain a comparative after each connector? And how do these observations fit into the syntactic and semantic analyses of CCs in literature to date?

Looking at previous work on CCs, the two most hotly debated topics so far have probably been the status of their constituent clauses as para- or hypotactic, and the question of their semantic compositionality (see McCawley 1988, Culicover/Jackendoff 1999 for the English equivalents, Beck 1997 for German, and den Dikken 2005 for a cross-linguistic account). On the one hand, the fact that many languages use symmetric forms to realize both clauses has been perceived as a syntax-semantics mismatch (see Culicover/Jackendoff 1999) since the different syntactic functions of main and subordinate clauses are expected to be reflected in the forms chosen to represent them (e.g., different connectors, as is the case in German).[1] On the other hand, it is not entirely clear how the special semantics of the correlation between the two comparatives can be derived compositionally from the two clauses, especially if both look alike, but have different semantic interpretations.[2] In (1), c1 can be interpreted as a sort of conditional to c2, i.e.: *if and when Hans runs faster, he becomes tired that much more quickly*, yet at the same time, it does not hold that *Hans runs that much faster, if and when he gets tired more quickly*.[3] For German the situation is somewhat simpler since the corresponding structure is already asymmetric on the surface, namely using the conjunction *je* for the subordi-

[1] Among the symmetric languages, the prominence of English with *the* in both clauses has played a role in leading research on CCs in this direction, e.g., in the following example from the British National Corpus (http://www.natcorp.ox.ac.uk), the structure of one clause mirrors the other:

[$_{\text{CLAUSE1}}$ *The nearer it gets*] [$_{\text{CLAUSE2}}$ *the more worried I become*] (BNC, document A4P)

The clauses may appear paratactic and symmetric on the surface, but semantics suggest the first clause is in fact subordinate. See also Abeillé/Borsley/Espinal (2006) for a discussion of symmetricity in French vs. Spanish CCs.

[2] So much so that CCs have often been used to illustrate Construction Grammar approaches as an example of a construction which requires a partially specified entry in the mental lexicon or 'construcicon', e.g., in Goldberg (2006: 5).

[3] This conditional reading has led to the occasional use of the name *comparative conditional* for these constructions. Put more formally, the relationship between the two comparatives is a unilaterally monotonous dependency, though incidentally not necessarily a proportional one. Simplifying somewhat, this corresponds to a formal structure:

$\forall x, y[g(x) > g(y) \rightarrow f(x) > f(y)]$

where g and f are the comparatives modifying their respective CC clauses (see Beck 1997: 259).

nate clause c1 and *desto* for the main clause c2, with the typical verb-second word order in the latter and verb last in the former, as is usual for main and subordinate clauses in German.[4]

Despite this, it is not usually assumed that there is any significant difference between the comparative slot in c1 and c2. The syntactic description of the phenomenon as found in Beck (1997) is of a symmetric CP dominating two externally undifferentiated CPs, each dominating a phrase DegP in their specifier (see Figure 1). Quite independently of the question regarding the analysis of the CPs, I will be concerned with the (a)symmetry of DegP, which is obligatory in all CC clauses (C' may be omitted on both sides in short CCs, or just on one side – i.e., short c1 or c2).[5] Although both DegPs are realized formally and syntactically in exactly the same way, the data will show that their usage is in fact consistently asymmetric both in the typical filling of the head Deg^0 and in the way this position is used productively with novel items, a mismatch, which to my knowledge, has received no large scale empirical study to date.

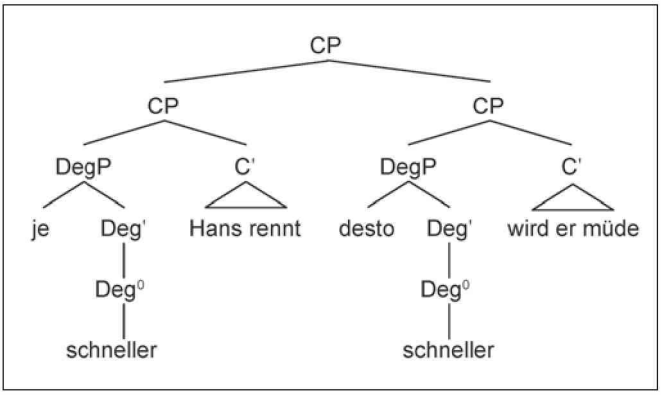


Figure 1: Syntactic analysis of German CCs, adapted from Beck (1997: 234)

4 There are of course both diachronic evidence and synchronic traces of symmetric constructions in German with both *desto … desto* and even *je … je*, but even if these are considered standard, the word order clearly distinguishes main from subordinate clause, e.g.: *Desto lauter sie sind, desto weniger werden sie selbst etwas auf die Beine stellen* 'The louder they are, the less they will get something up on its feet by themselves' (DeWaC, pos. 145401225; see Section 3.1 for information on this corpus).

5 Short CCs are sometimes considered a case of ellipsis of the subject and predicate, which has been assumed to be a copula verb (e.g., Zifonun et al. 1997: 2338). In fact the introduction of a copula would often not make a felicitous sentence, e.g., where a comparative would only fit a telic verb as in *je früher, desto besser* 'the earlier the better'. Here the sense of 'earlier' implies something happening rather than a continuous state (i.e., 'the sooner it happens' or 'the sooner you do it', but probably not 'the sooner it is'). In any case as we will see in Section 3, short CCs actually behave quite differently from other CCs in their preferences for certain lexemes and in their propensity for admitting novel lexemes.

The remainder of the article is structured as follows: Section 2 briefly presents approaches towards measuring productivity using corpus data, which will allow us to compare the usage of novel items in c1 and c2. Section 3 presents data from several large corpora on the usage of comparatives in and outside of comparative correlatives in German. Section 4 concludes by sketching the asymmetric profile of typical CC usage in German as found in the data, with some suggestions for the interpretation of the differences discussed in section 3.

2. Measuring Productivity

As the scope of this article does not permit an extensive discussion of the possible definitions for the concepts underlying the notion of linguistic productivity, this section will only attempt to give a brief overview of approaches in the empirical paradigm represented in Baayen (1993, 2001, 2009) and related work. Essentially, linguistic productivity describes the possibility of forming novel linguistic forms never before heard or produced by a speaker. Productivity is seen as a property of linguistic processes (thus it is a morphological word formation process, or a syntactic process filling an argument structure which is productive, and not affixes, verbs or other words, cf. Bauer 2001: 12-15). A prerequisite for quantitative empirical work on productivity is the view that productivity is a gradual property, and not binary or even categorical. Thus there is no dichotomy between rules of grammar which are productive and those which are not (e.g., past tense with *-ed* vs. vowel changes for weak and strong verbs in English), but rather some processes may be extremely unproductive (and indeed, novel strong past tense forms do occur; for discussion see e.g., Clahsen 1999, Ullman 1999, McClelland / Patterson 2002). A distinction between 'productive', 'unproductive' and 'semi-productive' (as made by some researchers, see Bauer 2001: 15-20) is also unhelpful in this context, both intuitively, since some processes are perceived to be more productive than others on a scale, (e.g., Dutch *ver-* vs. *-ster* in Baayen 2009: 904-907, or English deadjectival nominalization in *-ness*, *-ity* or *-cy*, cf. Plag 1999, chapter 2; see also Bauer 2001, chapters 1-2 for an in-depth discussion) and in practice, since data-based measurements lend themselves to normalized scales.

The criteria for a productive formation in most work tend to focus on novelty, regularity and transparency (see Bauer 2001: 34-58). That is to say, a process is productive if and only if it produces forms not generated or received by speakers before, which result regularly from the process and the components on

which it operates, and the resulting forms can be understood with predictable meaning in a fashion consistent with other formations from the same process. However in reality, it is impossible to directly or reliably evaluate the novelty, transparency and regularity indicative of productivity for all items associated with a process (even for one speaker it is impractical to establish whether or not she or he has produced or received a particular formation in the past, let alone for the linguistic community as a system, as Bauer (2001, 34f.) points out). Baayen (1993, 2001, 2009) therefore suggests that different aspects of productivity can be assessed, at least for a certain register, from corpus data, using the type and token frequencies of a word formation process, as well as the frequency of items appearing only once in the corpus (*hapax legomena*), which are assumed to be a superset of the neologisms therein. In particular, Baayen suggests using three different measures:

$$\text{p1: Extent of Use} = V(C,N) = \frac{Types_C}{N}$$

$$\text{p2: Hapax-conditioned Degree of Productivity} = \frac{V(1,C,N)}{V(1,N)} = \frac{Hapax_C}{Hapax_N}$$

$$\text{p3: Category-conditioned Degree of Productivity} = \frac{V(1,C,N)}{V(1,N)} = \frac{Hapax_C}{Tokens_C}$$

Although it is clearly not the case that there is a constant ratio between hapax legomena and 'true' neologisms in every corpus, these measures often seem to correspond to linguists' intuitions about productivity. p1 simply specifies how large a vocabulary the process has produced in *N* tokens of data, p2 the proportion of unique items in the corpus coming from the process and p3 the proportion of unique items within tokens belonging to the process.[6]

To illustrate the use of these measures, I use a corpus of 5 years of the German computer magazine "c't Magazin" (CT, 1998-2002, some 15 million tokens), comparing data for three adjective forming suffixes with different degrees of productivity: *-bar*, *-lich* and *-sam*, which form such adjectives as *lesbar* 'read-

[6] In Baayen's notation, $V(C,N)$ stands for the vocabulary size of a morphological category C in N tokens, or in the context discussed here, the normalized type count of the output of a process. $V(1,N)$ is the number of hapax legomena in the corpus (vocabulary types with frequency = 1, or simply $Hapax_N$), and $V(1,C,N)$ is the number of types from the relevant category C with a frequency of 1 (or $Hapax_C$). $N(C)$ is the token count for all occurrences of the category in the data.

able' (from *lesen* 'to read'), *freundlich* 'friendly' (from *Freund* 'friend') and *einsam* 'lonesome' (from *ein* 'one'). The results for these suffixes are summarized in Table 1.

	-lich	*-bar*	*-sam*
Tokens	59 472	26 865	7 691
Types	1 222	896	74
Hapax	483	354	24
p1	0.002054	0.000061	0.000005
p2	0.001356	0.000994	0.000067
p3	0.008121	0.013176	0.003120

Table 1: Productivity measures for *-lich*, *-bar* and *-sam* in the CT corpus

As the table shows, the *-sam* formation is the least productive, with few hapax legomena and the lowest score on all three measures. *-lich* exhibits a larger vocabulary than *-bar*, indicating that it has been more productive in the past, but *-bar* has a higher proportion of hapax legomena and consequently higher p2 and p3, indicating it is now easier to form novel adjectives with this suffix. This probably corresponds with most intuitive judgments (essentially the same results with a different corpus may be found in Evert / Lüdeling 2001) as *-sam* produces virtually no new forms in present-day German, and *-lich* is more restricted than *-bar*, which can form adjectives expressing potentiality from almost any transitive verb stem.

However, applying the measures to different sample sizes from each process leads to skewed results: the more words we have examined from a certain category, the more likely it becomes that the next word will not be novel (since we already 'know' more words). It is therefore necessary to compare the measures for a fixed sample size (e.g., *n* thousand samples from each process, all in the same corpus, see also Gaeta / Ricca 2006), which also allows statistical significance to be computed. The linguistic interpretation of the different measures can best be illustrated by plotting the development of vocabulary size across the corpus. This is achieved using vocabulary growth curves (VGCs, see Evert 2004), which plot the number of tokens on the x-axis and the number of types at that point on the y-axis. Thus, each newly encountered hapax legomenon raises the curve, but as more and more familiar items are encountered, it becomes gradually flat, showing that the process is approaching saturation in the data.

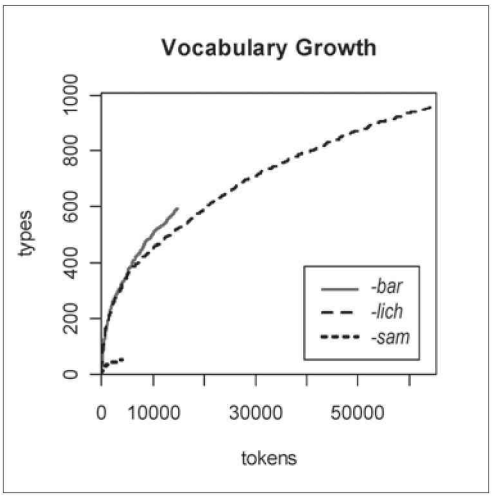


Figure 2: VGCs for *-lich*, *-bar* and *-sam* in the CT corpus

Figure 2 illustrates the higher productivity of *-bar*, which has a shorter curve (fewer types in the data, lower p1) but a steeper incline (higher proportion of hapax legomena, thus higher p3) along the curve. The *-sam* curve is very short as items with this suffix are rare, but already much flatter than the other two curves – the large majority of types from this process have already been seen, even for this small sample size. Since longer curves offer more chances for different types to occur, but with a progressively lower chance for novel hapax legomena, significance can only be evaluated based on the smallest sample size. The following section describes the corpora used in this study in more depth, followed by an analysis of differences in the productivity of the comparative formation in c1 and c2 using this methodology and a study of lexical preferences for each slot.

3. The Data for *je ... desto*

3.1 Corpora

Given the relative infrequency of CCs, a large corpus is (or several corpora are) required in order to find a sufficiently large database of clauses with *je* and *desto*. Thus, although an examination of a variety of genres would be desirable in principle, the main available sources of written language of sufficient size are newspaper language and web data. I will therefore use the controlled CT

corpus mentioned above and the largest available web corpus of German, the uncontrolled DeWaC corpus (Baroni / Kilgarriff 2006, approx. 1.7 billion tokens). In order to admit some information on usage of the construction in the spoken modality, I also use German Parliament Proceedings (GPP, from February 1996 to February 2003, totaling some 37 million tokens). As it turns out, however, the construction is rather rare in the proceedings register (about half as frequent compared to the CT corpus). For this reason, a further 27 million tokens were taken from the German version of the proceedings of the European Parliament (Europarl, Koehn 2005; partly original German sentences, partly translated from 10 other European languages), producing a dataset of about the same size for both genres. The use of translated language in this context is not optimal (though arguably expert translations forming German sentences are a valid genre in and of themselves; for a discussion see Olohan 2004, chapters 3 and 7), yet data from Europarl actually matches distributions in the proceedings of the German Parliament surprisingly closely.[7]

With tokenized and part-of-speech tagged data at hand, frequencies were extracted for all predicative / adverbial adjectives in the corpora ending in the comparative ending *-er*,[8] and the resulting 3.5 thousand lexemes were manually sorted for plausibility as a comparative adjective (filtering out both wrongly tagged non-adjectives such as *eBay-Webserver* and non-comparative, attributive cases such as *genannter* 'named'). For the remaining 2 000 or so comparative lexemes, total frequencies (potentially including wrongly tagged attributive cases) and frequencies after *je* and *desto* were extracted, as well as frequencies in the sequence *je* [COMPARATIVE], *desto* [COMPARATIVE], where the comma was optional. Using the methods introduced in Section 2, it is possible to calculate the productivity measures for the comparative formation in the available corpora, excluding those cases which follow *je* or *desto*. These results are presented in Table 2.

[7] For instance, the top 10 lexemes for comparatives after *je* match 9 out of 10 in the two corpora, with *früher* 'earlier' replacing *später* 'later' in Europarl, and 7 out of 10 after *desto*. At the same time the CT data is less conformant with both proceedings corpora in lexical choices than the latter are between themselves.

[8] The corpora were tagged using the freely available TreeTagger (Schmid 1994) and searched with the Corpus Workbench (CWB, Christ 1994) for the STTS part-of-speech tag ADJD (see Schiller et al. 1999 for the tagset).

	CT	Europarl	GPP	All Corpora
corpus tokens	14 596 537	27 317 723	36 723 139	78 637 399
corpus types	595 022	283 389	443 949	1 010 539
corpus hapax	356 075	140 730	222 221	565 020
comparative tokens	30 548	41 857	49 866	122 271
comparative types	1 149	1 160	1 383	1 969
comparative hapax	515	494	648	776
p1	0.00007872	0.00004246	0.00003766	0.00002504
p2	0.00144632	0.00351027	0.00291602	0.00137340
p3	0.01685871	0.01180209	0.01299483	0.00634656

Table 2: Corpus statistics and productivity measures for comparatives outside CCs

A direct comparison between the data for each corpus should be avoided since they are of different sizes and thus have different chances of realizing fewer or more types and hapax legomena. However, it should become clear from the vocabulary and hapax counts that CT is the richest corpus, with more types and hapax legomena than the other two corpora, despite having the smallest number of tokens. This is understandable as the magazine contains a variety of text types (reviews, editorials, readers' letters) and a high number of unique technical terms increasing both vocabulary and neologisms. For the comparative counts the situation is more moderate, but CT still has the highest type/token ratio and almost as many types as the other corpora, thus revealing again the largest variety for the smallest corpus.

3.2 Differences in Productivity for c1 and c2

Applying the measures presented in Section 2 to the slots c1 and c2 reveals differences in their productive potential to manifest new items as predicted by ratios of hapax legomena. Following Barðdal (2006, 2008), who examines the productivity of verbs governing the dative in Icelandic, and Kiss (2007), who applies Baayen's measures to the nominal slot of PPs with determinerless singular nouns in German, I will treat the filling of the comparative position in each CC clause as a productive process, with the choice of comparative paralleling the choice of a stem in a morphological process such as affixation. Figure

3a plots the vocabulary growth curve for comparatives in c1 and c2 in all corpora as compared with 3 000 randomly selected comparatives outside of CCs equally distributed between all three corpora.

It is immediately clear that non-CC comparatives (the top curve, *comp*) are more productive than CC comparatives (significant test of equal proportions at $p < .01$ for an equal sample size). Since the CC sample only covers less than 1 500 tokens, the data is extrapolated to show expected development of the curves using a finite Zipf-Mandelbrot model (FZM, see Evert 2004), which provides a good estimate of the expected divergence of the curves given more data from the same register. Results also show the c2 curve ('c2' or the extrapolation 'fzm2') to be significantly ($p < .01$) more productive than the c1 curve ('c1' or 'fzm1'). This relationship is also true for each genre separately, though splitting the corpus would result in figures rather small for a productivity study and insufficient for a significant result.

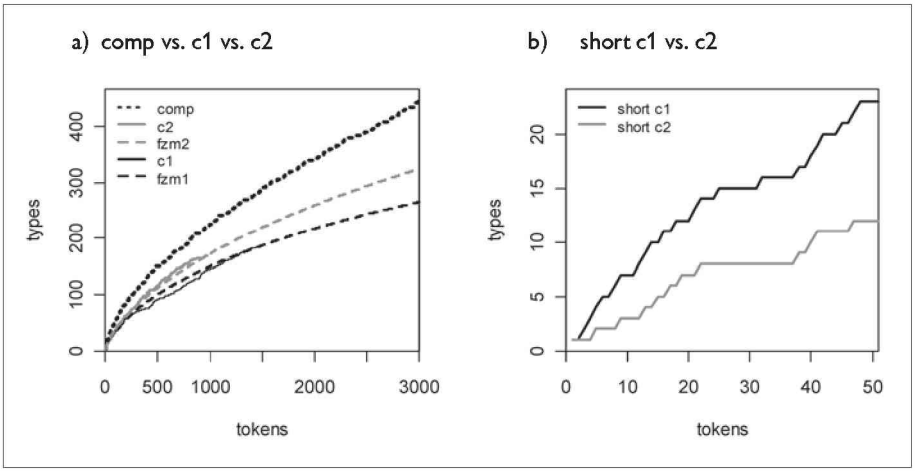


Figure 3: On the left (a), comparatives outside CCs (top curve) are more productive than c2 (middle), which is more productive than c1 (bottom). Dashed curves predict the further development of CC vocabulary based on a finite Zipf-Mandelbrot model (FZM). On the right (b), a very small sample suggests short c1 is more productive than short c2

Thus, results show that c2 is more open to lexical variation than c1. Another interesting question already raised in Section 1 is whether short CCs behave in the same way as other CCs. Surprisingly, the data exhibits a trend in the opposite direction (Figure 3b), though numbers are too small to be significant. The lexeme responsible for this situation is largely *besser* 'better', which forms

approx. 73 % of the data, or 37 matches for short c2. Since a much larger sample is needed in order to establish a meaningful trend, the experiment is repeated with DeWaC. Though uncontrolled and therefore likely more heterogeneous and possibly more productive, this dataset has the advantage of containing over 1 800 short CCs (showing the rarity of this construction: only about .0088 times per 10 000 tokens, or less than one in a million!). Results repeat the same pattern (Figure 4).

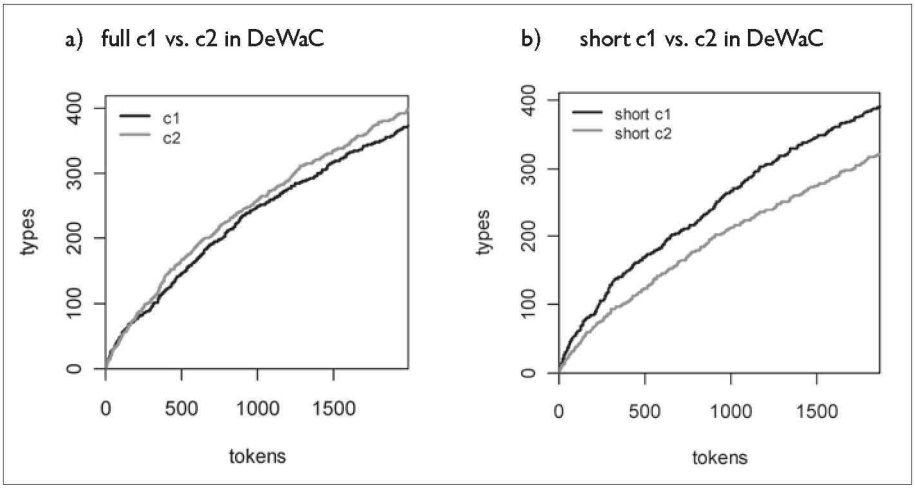


Figure 4: VGCs for same sized samples of short and long c1 and c2 in DeWaC

c2 is only a little more productive than c1 in full CCs (p3 = .028 for c1 and .031 for c2), though significantly thanks to the large sample (Figure 4a). At the same time it is considerably less productive in short CCs (see Figure 4b), in which *besser* 'better' comprises an overwhelming majority of 55% of c2 cases (1 022 matches, leading to a p3 score of .13 for c1 versus .11 for c2 and a noticeably flatter VGC). The overrepresentation of *besser* shows us that specific lexical items can be an important factor in explaining the differences in the behavior of the two slots. With these results at hand, the next section therefore turns to examine the lexemes occupying each slot in more detail.

3.3 Lexical Preferences for c1 and c2

Though it is impractical to examine each and every attested CC in the corpus, cross-sections of lexical behavior in each slot from very frequent, moderately frequent and rare items can reveal some trends. Table 3 shows the frequencies

of different comparative adjectives in each corpus and in all corpora put together in c1 and c2, as well as the total frequency for each comparative in general. Trends that are systematic and register independent should appear in all corpora, whereas mixed results might cast doubts on the presence of any observable meaningful pattern in the data.

	All Corpora			CT		GPP		Europarl	
type	**freq**	**c1**	**c2**	**c1**	**c2**	**c1**	**c2**	**c1**	**c2**
besser	18 270	70	212	19	79	32	46	19	87
später	7 983	22	5	4	1	15	2	3	2
stärker	6 844	65	56	15	26	21	13	29	17
ferner	5 975	0	0	0	0	0	0	0	0
länger	4 659	179	23	44	17	93	4	42	2
höher	3 330	179	76	109	43	55	26	27	7
größer	3 126	195	120	117	39	41	39	37	42
...									
schwieriger	1 344	7	24	4	9	1	6	2	9
kleiner	878	79	11	54	8	13	3	12	0
...									
positiver	116	0	3	0	1	0	1	0	1
dunkler	112	13	4	12	4	0	0	1	0
wahrscheinlicher	103	0	10	0	6	0	0	0	4
...									
lockerer	25	3	0	3	0	0	0	0	0
wärmer	25	1	0	0	0	0	0	1	0
mühsamer	24	0	1	0	1	0	0	0	0

Table 3: Frequencies for comparatives in and outside of CCs in each corpus (see text for translations)

An examination of the table reveals some very strong lexical preferences which are remarkably consistent across the corpora. The generally most frequent comparative, *besser* 'better', is also the most frequent lexeme in c2 in all corpora. It is not, however, the most frequent in c1 – *besser* is three times less frequent in this position in total, outranked by a large margin by otherwise considerably less frequent lexemes such as *länger* 'longer' or *höher* 'higher'.

These lexemes are in turn much more frequent in c1 than in c2, thus exhibiting the opposite asymmetry. The fourth most frequent comparative, *ferner* 'further' is not used in CCs at all, though this is unsurprising since it is almost exclusively used as a lexicalized adverb with the sense 'furthermore'.[9] Some lexemes are much more balanced, such as *stärker* 'stronger', or have a less overwhelming imbalance, such as *größer* 'bigger'. Turning to mid-frequent CC comparatives, we find consistent asymmetries yet again, where all three corpora show the same preference of some comparatives for either c1 (e.g., *kleiner* 'smaller') or c2 (e.g., *schwieriger* 'more difficult'). Finally, items that are rare or even hapax legomena in each corpus, potentially indicating less entrenched, productively formed CCs that had not been produced by speakers/writers before (see Section 2), also cluster around slots: *dunkler* 'darker', *lockerer* 'looser', and *wärmer* 'warmer' appear mostly or exclusively in c1, and *positiver* 'more positive', *wahrscheinlicher* 'more likely', and *mühsamer* 'more laborious' prefer c2.

How can these results be interpreted? A lexicalization of large lists of lexemes to prefer one position or the other seems unlikely, especially considering the evident preferences of rather infrequent items across corpora (working under the assumption that at least moderate frequency is a prerequisite for lexicalization). A more careful look at the senses of the adjectives reveals a likelier semantic explanation: c1 prefers spatiotemporal conditions, whereas c2 provides an evaluation which typically passes a subjective judgment on the favorability or likelihood associated with the increase of the condition in c1. This interpretation is evident simply by composing sentences from the most frequent c1s and c2s:

(3) *je höher, desto besser* 'the higher the better'

(4) *je länger, desto schwieriger* 'the longer the more difficult'

Such sentences also form the typical cases of short CCs (see below). Cases which appear semantically more spatiotemporal, but still appear in c2, such as *größer*, with its more balanced profile, merit a closer look. A qualitative examination of c2 sentences of this sort often reveals that such lexemes may assume a rather neutral role when used to modify a subject noun, which in turn supplies the evaluative semantics. This may appear in c2 (example 5), but also in c1 (6):

[9] A true comparative sense is possible nonetheless, e.g.: *Nichts lag aber der DDR-Diktatur* ***ferner*** *als der Frieden* 'But nothing was **further** removed from the GDR dictatorship than peace' (GPP, 6 July 2000, session 114); such cases are, however, quite rare and unattested in the CC data.

(5) *Je länger man den Rechner laufen lässt* [...] *desto größer die Gefahr, dass sich der Schaden noch vergrößert*

'The longer the computer is allowed to run [...] the greater the risk that the damage will increase even more' (CT 2000, 6: 116, segment title "Praxis: Datenrettung per Diskeditor")

(6) *Je größer der Abstand zur Vollaussteuerung* [...], *desto besser*

'The greater the distance to complete amplification [...] the better' (CT 1998, 1: 102, segment title "Prüfstand: Soundkarten")

In (5), the c1 spatiotemporal condition 'time running' is correlated with the idea of 'risk', however rather than formulating the notion adjectivally (*desto gefährlicher* 'the riskier'), *größer* 'greater' is used to modify the 'risk'. Though 'greater' basically refers to a measurable expanse (thus also spatiotemporal in an extended sense), the reading as a whole is still evaluative (risky and hence negative). Similarly in (6), *größer* does not specify the spatiotemporal semantics by itself but rather qualifies *Abstand* 'distance' (which could have also been specified by a single comparative, e.g., *weiter* 'farther'). The opposite situation, where an apparent evaluative qualifies c1, is less frequent and also turns out to blend into a larger spatiotemporal condition in most cases, as in (7):

(7) *Je besser die Komprimierung ist, um so höher fällt ohne zusätzliche Speichererweiterung die nutzbare Auflösung für größere Grafiken aus*

'The better the compression is, the higher the usable resolution turns out for bigger graphics without additional memory expansion' (CT 1999, 1: 116, segment title "Prüfstand: Laserdrucker")

Although it is unusual for *besser* to appear in c1, it qualifies a rate of compression (a sort of spatiotemporal condition), and this is in turn evaluated in terms of better print resolution. In either case, it seems that adjectives with a less specific semantic content can be used to modify CC subjects as a sort of 'light comparative', where the subject noun of the modified clause specifies the typical meaning supplied by c1 or c2. In these cases, there is therefore still a tendency for c1 to contain a spatiotemporal condition, and for c2 to contain a dependent evaluative.

Moving on to the short CCs, a more extreme set of preferences can be observed by comparison. Table 4 shows frequencies in CCs in total vs. short CCs for each lexeme.

type	freq	c1 total	c2 total	c1 short	c2 short
besser	18 270	70	212	0	37
länger	4 659	179	23	2	0
schneller	4 423	88	40	8	0
höher	3 330	179	76	0	1
größer	3 126	195	120	4	0
schlechter	1 665	10	16	0	1
früher	1 483	15	0	11	0
schlimmer	1 435	2	3	1	0
deutlicher	1 338	5	12	1	0
billiger	1 219	2	4	1	0
häufiger	1 106	11	13	0	1
kleiner	878	79	11	2	0
sicherer	524	5	7	0	1
kürzer	383	26	6	3	0
...					
ergonomischer	8	0	1	0	1
frecher	7	1	1	0	1
hilfloser	4	0	1	0	1
teurer	4	0	0	2	2
unsolider	2	1	0	1	0
reißerischer	1	1	0	1	0

Table 4: Preferences for short CCs

The data shows a stronger bias for short CCs, where the most frequent comparative *besser* is not only strongly preferred in c2, but does not occur at all in c1 (a ratio of 37 to zero, thus clauses of the type *je besser, desto* [COMPARATIVE] are entirely unattested). Conversely, the most common lexeme in c1 is *früher* 'earlier' (11 times in c1 but 0 in c2), closely followed by *schneller* 'faster' (8 and 0 respectively). Figure 5 illustrates the distribution of lexemes in each short slot.

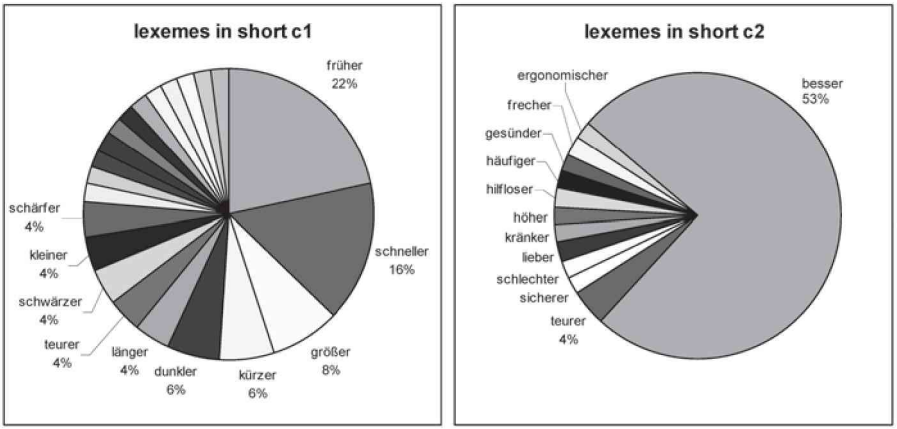


Figure 5: comparative lexemes in short c1 and short c2

The pattern is markedly different on each side. In fact, only one item appears in both slots: *teurer* 'more expensive'. While an evaluative reading of 'expensive' = 'bad' requires little explanation, an extension of its sense to a spatiotemporal reading requires an explanation. In fact, both occurrences of this comparative in c1 are subsequently evaluated in c2, namely:

(8) *je teurer, desto besser* 'the more expensive the better'

(CT 2002, 9: 6, segment title "Inhalt: 09 / 2002")

(9) *Je teurer, desto schlechter* 'The more expensive the worse'

(CT 1998, 22: 170, segment title "Software: Übersetzung")

where (9) is used sarcastically in reference to expensive machine translation software, and (8) is used negatively to say that this rule does not apply to printer inks and paper types. Although *teurer* in these latter cases is not spatiotemporal in any but the most transferred sense, the structure of relating a quantity to an evaluation still bears some resemblance to other cases.

A possible explanation for the more pronounced tendencies in short CCs is that the typical semantics of each clause can only be expressed in the comparative itself, whereas long CCs may distribute the sense of each correlate between the comparative and the subject phrase or even predicate (if both are supplied). This means that ambivalent lexemes like *größer* are read as spatiotemporal by default (notwithstanding additional meanings supplied by context and not realized overtly). At the same time the overwhelming dominance of *besser* in c2

seems to suggest the choice of a short CC is most appropriate for comparatively simple evaluations, which fits well with the fact that this slot is also the least productive. This is not to say that productivity is ruled out in c2 – only that it is less likely, much like in the case of less productive morphological affixes.

4. Summary – A Profile of German CCs in Use

The picture of CC usage in German arising from the data used in this study is of a semantically asymmetric construction, correlating a scalar, usually spatiotemporal quantity in c1 with an evaluation of the effect of a change in this quantity in c2. In c1, typical spatial examples are size and distance, such as 'bigger' or 'farther', and typical temporals are either a flexible point in time, especially using the notion of 'earlier', or durations such as 'longer' (the latter can also function spatially for distance of course). Some extended senses also found frequently are color terms (e.g., 'darker', 'brighter' or even actual colors like 'blacker', 'whiter'), where perhaps depth of color is meant, as an extension of spatial depth, and reference to prices as in 'more expensive' or 'cheaper' (though the prevalence of this category may be connected to the rather economically oriented genres examined). For this last case, a true spatiotemporal interpretation is not obvious, though it is clear why such a quantity is often correlated with an evaluation of advantageousness (this is coded in the opposite c1-c2 order in the expression *value for money*, and in the canonical CC order in the German *Preisleistungsverhältnis* 'price-benefit-ratio'). The c1 slot is less productive than c2 overall, meaning novel spatiotemporals arise somewhat less frequently.

In c2 we find both direct evaluations of quality, notably 'better' (or less frequently 'worse'), but also often evaluations of probabilities – 'more likely', 'riskier', 'more certain', and more semantically specified evaluations such as 'healthier', 'more difficult', 'more laborious'. We also find some (though fewer) spatiotemporals, notably 'greater' or 'bigger', especially when these function as a sort of semantically underspecified 'light comparative', qualifying a noun supplying the evaluative meaning. Thus, we get *desto größer die Wahrscheinlichkeit* 'the greater the probability' instead of *desto wahrscheinlicher* 'the more likely', or *desto größer die Gefahr* 'the greater the danger' instead of *desto gefährlicher* 'the more dangerous'. This slot is also significantly less productive than comparatives outside of CCs, but more productive than c1, meaning novel ways of evaluating c1 conditions are more likely than such novel spatiotemporal circumstances in full CCs.

The examination of short CCs has shown them to adhere even more closely to the lexical stereotypes, possibly since there is no more possible recourse to the semantics of other phrases (subject, predicate or other adverbials) to supply the spatiotemporal or evaluative meaning. At the same time they are also the least productive, but with the opposite internal relationship: c2 is much less productive than c1, with *besser* filling a sweeping majority of short c2s. This implies that this construction tends to be chosen precisely in cases where the message of the utterance is simply a positive judgment on some condition, leaving more variety in the expression of the condition itself; if further nuances of the evaluation are required (e.g., it is better in the sense of 'more certain', 'healthier', etc.), the short CC is apparently less preferred. Still, items other than *besser* are clearly possible, and *besser* also occurs in long CCs in c1, and then often as a 'light comparative' in much the same way as 'bigger' or 'greater'.

The theoretical status of the observations made here is not yet clear. On the one hand, it is unquestionably true that: 1. lexically, many comparatives can and do appear in both c1 and c2 which do not obey the prototypical semantics portrayed here, and 2. productively, both slots are capable of hosting novel comparatives presumably not heard before by the speaker. On the other hand, multiple, rather large datasets have shown that the properties charted here for each slot show statistically significant and consistent differences in the propensity for innovation and an unequivocal preference for certain lexemes and types of lexemes. These facts require an explanation, as they seem to suggest speakers have implicit knowledge of how to use CCs, which must be stored somehow in reference to the meaning of the construction as a whole (this brings to mind the 'constructicon' account of Construction Grammar mentioned in Section 1, as in Goldberg 1995, 2006). Accounting for such facts of usage becomes even more important if we view the emergence of grammar as a gradual codification of such 'soft constraints', which can be more or less categorical (cf. Bresnan / Dingare / Manning 2001; the soft constraints of one language, or even language stage, may be mirrored in the categorical constraints of another; see also the articles in Bybee / Hopper (eds.) 2001).

Facts of usage not touched upon here but meriting further study are the interaction between the choice of c1 and c2 (particularly likely pairs and conditional probabilities in each direction), both semantically and through preferred lexicalized orders. It is conceivable that certain frequent CCs, especially where the correlation is bilateral, form steady c1-c2 pairs in a similar way to so-called irreversible binomials (like English *black and white* but usually not

**white and black*; see Malkiel 1959, Müller 1997, Ross 1980). The difference between full CC clauses with subject and verb and those with a subject but no VP also requires a separate investigation, as well as the behavior of cases where only one clause is short. Finally, a cross-linguistic analysis to examine whether the trends in German CC data are mirrored in CCs in other languages is important for establishing whether these results reveal general semantic factors (the typical use of comparatives language-independently, or constraints imposed by world knowledge) or rather language specific preferences.[10] The study of the semantics of CCs is therefore far from complete, and offers a rich environment for comparing the usage of what seems like a single category, comparatives, but turns out to be very multifaceted depending on its embedding context.

References

Abeillé, Anne / Borsley, Robert D. / Espinal, Maria Teresa (2006): The syntax of comparative correlatives in French and Spanish. In: Müller, Stefan (ed.): Proceedings of the 13th International Conference on Head-Driven Phrase Structure Grammar. Stanford: CSLI Publications, 6-26.

Baayen, R. Harald (1993): On frequency, transparency, and productivity. In: Booij, Geert / van Marle, Jaap (eds.): Yearbook of Morphology 1992. Dordrecht: Kluwer Academic Publishers, 181-208.

Baayen, R. Harald (2001): Word frequency distributions. (= Text, Speech and Language Technologies 18). Dordrecht / Boston / London: Kluwer Academic Publishers.

Baayen, R. Harald (2009): Corpus linguistics in morphology: Morphological productivity. In: Lüdeling, Anke / Kytö, Merja (eds.): Corpus linguistics. An international handbook. Berlin: de Gruyter, 899-919.

Barðdal, Johanna (2006): Predicting the productivity of argument structure constructions. The 32nd Annual Meeting of the Berkeley Linguistics Society. Internet: http://ling.uib.no/barddal/BLS-32.barddal.pdf (last visited: 07 / 2010).

Barðdal, Jóhanna (2008): Productivity: Evidence from case and argument structure in Icelandic. Amsterdam: Benjamins.

Baroni, Marco / Kilgarriff, Adam (2006): Large linguistically-processed web corpora for multiple languages. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006). Trento, Italy. Morristown, NJ: Association for Computational Linguistics, 87-90.

[10] Preliminary results from a study of similar data in English but with different goals (Zeldes 2009) suggest that spatiotemporal and evaluative semantics are indeed typical for English CCs as well, though the particulars for some lexemes and the degrees of productivity behave in a noticeably different manner.

Bauer, Laurie (2001): Morphological productivity. (= Cambridge Studies in Linguistics 95). Cambridge: Cambridge University Press.

Beck, Sigrid (1997): On the semantics of comparative conditionals. In: Linguistics and Philosophy 20: 229-271.

Bresnan, Joan / Dingare, Shipra / Manning, Chris D. (2001): Soft constraints mirror hard constraints: Voice and person in English and Lummi. In: Proceedings of the LFG '01 Conference, Hong Kong. Internet: http://www.stanford.edu/~bresnan/lfg01.pdf (last visited: 07 / 2010).

Bybee, Joan / Hopper, Paul (eds.) (2001): Frequency and the emergence of linguistic structure. (= Typological Studies in Language 45). Amsterdam: Benjamins.

Christ, Oliver (1994): A modular and flexible architecture for an integrated corpus query system. In: Proceedings of COMPLEX'94: Third Conference on Computational Lexicography and Text Research. Budapest: Hungarian Academy of Sciences, 23-32.

Clahsen, Harald (1999): Lexical entries and rules of language: A multidisciplinary study of German inflection. In: Behavioral and Brain Sciences 22: 991-1060.

Culicover, Peter / Jackendoff, Ray (1999): The view from the periphery: The English comparative correlative. In: Linguistic Inquiry 30, 4: 543-571.

den Dikken, Marcel (2005): Comparative correlatives comparatively. In: Linguistic Inquiry 36, 4: 497-532.

Evert, Stefan (2004): A simple LNRE model for random character sequences. In: Proceedings of JADT 2004. Louvain: Presses universitaires de Louvain, 411-422.

Evert, Stefan / Lüdeling, Anke (2001): Measuring morphological productivity: Is automatic preprocessing sufficient? In: Rayson, Paul / Wilson, Andrew / McEnery, Tony / Hardie, Andrew / Khoja, Shereen (eds.): Proceedings of Corpus Linguistics 2001, UCREL Technical Paper 13 (special issue). Lancaster: Lancaster University, 167-175. Internet: http://ucrel.lancs.ac.uk/publications/cl2003/CL2001%20conference/papers/evert.pdf (last visited: 07 / 2010).

Gaeta, Livio / Ricca, Davide (2006): Productivity in Italian word formation: A variable corpus approach. In: Linguistics 44, 1: 57-89.

Goldberg, Adele (1995): Constructions: A construction grammar approach to argument structure. Chicago / London: University of Chicago Press.

Goldberg, Adele (2006): Constructions at work: The nature of generalization in language. Oxford: Oxford University Press.

Kiss, Tibor (2007): Produktivität und Idiomatizität von Präposition-Substantiv-Sequenzen. In: Zeitschrift für Sprachwissenschaft 26, 2: 317-345.

Koehn, Philipp (2005). Europarl: A parallel corpus for statistical machine translation. In: Proceedings of the Tenth Machine Translation Summit. Phuket: AAMT, 79-86.

Malkiel, Yakov (1959): Studies in irreversible binomials. In: Lingua 8: 113-160.

McCawley, James (1988): The comparative conditional in English, German and Chinese. In: Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society. Berkeley: Berkeley Linguistics Society, 176-187.

McClelland, James / Patterson, Karalyn (2002): Rules or connections in past-tense inflections: What does the evidence rule out? In: Trends in Cognitive Science 6: 465-472.

Müller, Gereon (1997): Beschränkungen für Binomialbildung im Deutschen. Ein Beitrag zur Interaktion von Phraseologie und Grammatik. In: Zeitschrift für Sprachwissenschaft 16, 1 / 2: 5-51.

Olohan, Maeve (2004): Introducing corpora in translation studies. London / New York: Routledge.

Plag, Ingo (1999): Morphological productivity: Structural constraints in English derivation. (= Topics in English Linguistics 28). Berlin: de Gruyter.

Ross, John (1980): Ikonismus in der Phraseologie. In: Zeitschrift für Semiotik 2: 39-56.

Schiller, Anne / Teufel, Simone / Stöckert, Christine / Thielen, Christine (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). Technical report, University of Stuttgart and University of Tübingen. Internet: http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-1999.pdf (last visited: 07 / 2010).

Schmid, Helmut (1994): Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the Conference on New Methods in Language Processing, Manchester, UK. Internet: http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf (last visited: 07 / 2010).

Ullman, Michael T. (1999): Acceptability ratings of regular and irregular past-tense forms: Evidence for a dual-system model of language from word frequency and phonological neighbourhood effects. In: Language and Cognitive Processes 14, 1: 47-67.

Zeldes, Amir (2009): Quantifying constructional productivity with unseen slot members. In: Proceedings of the NAACL HLT Workshop on Computational Approaches to Linguistic Creativity, June 5, Boulder, CO, 47-54. Internet: http://aclweb.org/anthology-new/W/W09/W09-2007.pdf (last visited: 07 / 2010).

Zifonun, Gisela / Hoffmann, Ludger / Strecker, Bruno et al. (1997): Grammatik der deutschen Sprache. Bd. 3. (= Schriften des Instituts für Deutsche Sprache 7.3). Berlin / New York: de Gruyter.

CYRIL BELICA / MARC KUPIETZ / ANDREAS WITT / HARALD LÜNGEN

The Morphosyntactic Annotation of DeReKo: Interpretation, Opportunities, and Pitfalls

Abstract

The paper discusses from various angles the morphosyntactic annotation of DeReKo, the Archive of General Reference Corpora of Contemporary Written German at the Institut für Deutsche Sprache (IDS), Mannheim. The paper is divided into two parts. The first part covers the practical and technical aspects of this endeavor. We present results from a recent evaluation of tools for the annotation of German text resources that have been applied to DeReKo. These tools include commercial products, especially Xerox' Finite State Tools and the Machinese products developed by the Finnish company Connexor Oy, as well as software for which academic licenses are available free of charge for academic institutions, e.g. Helmut Schmid's Tree Tagger. The second part focuses on the linguistic interpretability of the corpus annotations and more general methodological considerations concerning scientifically sound empirical linguistic research. The main challenge here is that unlike the texts themselves, the morphosyntactic annotations of DeReKo do not have the status of observed data; instead they constitute a theory and implementation-dependent interpretation. In addition, because of the enormous size of DeReKo, a systematic manual verification of the automatic annotations is not feasible. In consequence, the expected degree of inaccuracy is very high, particularly wherever linguistically challenging phenomena, such as lexical or grammatical variation, are concerned. Given these facts, a researcher using the annotations blindly will run the risk of not actually studying the language but rather the annotation tool or the theory behind it. The paper gives an overview of possible pitfalls and ways to circumvent them and discusses the opportunities offered by using annotations in corpus-based and corpus-driven grammatical research against the background of a scientifically sound methodology.

1. Introduction

This paper is inspired by a recent corpus annotation venture at the Institut für Deutsche Sprache (IDS) in Mannheim, Germany. The focus of the paper is on two separate yet related topics. On the one hand, a brief chronological overview of the linguistic annotations of DeReKo,[1] the Archive of General Refer-

[1] The name was inspired by a research project running from 1999 to 2001 in co-operation with the universities of Stuttgart and Tübingen.

ence Corpora of Contemporary Written German at the IDS, is presented, followed by a more detailed and more technical account of the planning, decision-making, deployment, and evaluation processes involved in the current annotation phase. Among other perspectives, a separate short discussion of the inter-tagger agreement between (1) Xerox' Finite State Tools, (2) the Connexor Oy's Machinese products, and (3) Helmut Schmid's Tree Tagger – as observed in the 3.75 billion sized DeReKo – is included.

On the other hand, the paper examines the potential methodological difficulties encountered when linguists include the annotated DeReKo (or any other very large annotated corpus) in their scientific research context. We conjecture that whenever linguistically challenging phenomena such as, in particular, language variation are studied, the observed annotation inaccuracy might prove to be worrisomely high, and, moreover, biased in a systematic way at that. We argue that a linguist trusting the annotations blindly would run the risk of not actually exploring the language captured in the corpus but that he or she would rather be detailing the annotation tool or the linguistic theory behind it instead. Finally, we discuss how linguists might want to circumvent such pitfall traps in order to approach very large annotated corpora in a scientifically sound way.

2. Tagging the IDS-corpora

The corpora of contemporary written German at the IDS, since 2004 called *Deutsches Referenzkorpus* – DeReKo, are one of the major resources worldwide for the study of the German language (Kupietz / Keibel 2009). The first step towards providing these corpora with access to linguistic annotation was already made in 1993, when a new version of the Corpus Search, Management and Analysis System COSMAS (IDS 1991-2009) was planned specifically in order to be capable of handling multi-layer annotations. Subsequently, the corpora were tagged several times, most notably in 1995 with the Logos Tagger and in 1999 with the Gertwol Tagger (Koskenniemi / Haapalainen 1996).

As shown in Table 1, the current annotation initiative of DeReKo also reaches back to 2002 when its start and co-ordination with the COSMAS project was incorporated into the IDS research plan 2003-2008.

2002-10	new annotation initiative incorporated in the IDS Research Plan 2003-2008
2007-01	COSMAS II confirms to support multiple stand-off annotation search by 2008
2007-10	process model for the annotation of DeReKo
2007-12	catalog of criteria for tagging tools
2008-04	start of market analysis
2008-07	25 tools → shortlist of 9
2008-08	request for evaluation versions
2008-09	start of in-depth expert study
2008-12	→ 3 tools recommended
2008-12	first annotation attempts
2009-02	filter and collation scripts developed to support XML stand-off annotations
2009-07	after 6 cpu years: 3.5 TB of annotation data produced
2009-08	DeReKo-2009-II released with full TreeTagger and Connexor annotations and partial Xerox annotations

Table 1: History of the current annotation initiative for DeReKo

2.1 Selection of tools

The deployment of the current annotation initiative was launched in late 2007, when a first coarse process model for the annotation of DeReKo with the following cornerstones was developed:

1) do not rely on judgements of a single tagger, i.e., provide multiple concurring annotations that result from different tools;
2) for every tagger include as many concurrent interpretations of each linguistic phenomenon as possible;
3) use different types of taggers to avoid systematic biases;
4) consider annotating multiple linguistic levels if appropriate tools are available;
5) invite an external expert panel to (1) carry out a market analysis in order to put up a shortlist of suitable taggers and to (2) conduct an in-depth study of the short-listed taggers in order to arrive at a final recommendation;

6) after completing the annotation phase, evaluate each annotation layer with respect to fitness for their particular intended use in linguistic research.

The next step was then to find and engage external experts in the field of computational linguistics, part-of-speech tagging, and other levels of automatic linguistic annotation to carry out an independent study, resulting in a shortlist (phase 1) and, finally, in a list of recommended tools (phase 2). Together with these experts, at first, a catalogue of linguistic, organizational, technical, and economic selection criteria was developed that can be summarized as follows:

- *linguistic:* reliability, precision, recall, disambiguation, self-assessment, tagset compatibility, types of analysis (POS, dependencies, ...), extensibility;
- *organizational:* long term perspectives, sustainability, "also applied by";
- *technical:* supported platforms, adaptability, maintainability, robustness, i / o-formats, resource requirements;
- *economic:* licensing options, license restrictions, pricing.

The subsequent market analysis was started in April 2008. Its result was a brief evaluation of about 25 tools according to the selection criteria and a shortlist of nine tools, recommended for closer investigation in the second phase of the expert study:

- GERTWOL (Lingsoft Oy)
- Machinese Tools (Connexor Oy)
- SMOR (Stuttgart University)
- TAGH (Thomas Hanneforth)
- TNT (Thorsten Brants)
- TreeTagger (Stuttgart University)
- Unsupos (Leipzig University)
- WMTrans (Canoo AG)
- Xerox FST Linguistic Suite (Xerox Company)

After an internal review of the results of the study, we decided to proceed as recommended by the expert panel. Evaluation versions of the nine tools were requested and – if granted – assessed more thoroughly in the in-depth study carried out by the experts between September and December 2008. This phase resulted in a final recommendation of three taggers:

- Machinese Tools from Connexor Oy,
- TreeTagger from Stuttgart University, and
- Xerox FST Linguistic Suite.

They are described in more detail below. Again, following the expert recommendations, we eventually decided to acquire the necessary licenses for these tools.

For our purposes, the TreeTagger was found to be the best tagger available free of charge. It employs a statistical tagging method in which transition probabilities are estimated by decision trees, hence its name (Schmid 1994). It provides disambiguated morphological and POS information in the form of STTS tags (Schiller et al. 1999). Its parameter files can be updated, i.e., the TreeTagger can be re-trained with one's own correctly tagged material. Moreover, it is continuously and actively being developed by its author Helmut Schmid at Stuttgart University.

The commercial Xerox FST Linguistic Suite from Xerox Inc., USA / France, provides very accurate tagging with rule-based POS disambiguation but no disambiguation of morphological tags. The tag set used is similar but not identical to STTS. The acquired license does not allow for a publication of tagged corpora (presumably to avoid the danger of reverse engineering).

The third recommendation consisted of the commercial products Machinese Phrase Tagger and Machinese Syntax from Connexor Oy, Helsinki. Machinese Phrase Tagger has the same functionality as Xerox FST, and, as far as we could infer from the description of these commercial products, both tools use similar techniques for this task. Machinese Syntax was the only tool tested that provides actual syntactic structures. Its analysis is based on the Functional Dependency Grammar (Tapanainen / Järvinen 1997), including, amongst other things, disambiguated POS and morphological tags. The acquired license does not allow for a publication of tagged corpora, i.e., its restrictions are comparable to those of newspaper text in DeReKo, from which only short passages may be quoted.

In the annotation cycle outlined in the following section, we only used the morphological and part-of-speech analysis components of these three tools.

2.2 The tagging process

To be able to apply the tools to the XCES-encoded DeReKo data, filters had to be developed first to mask out text that should not be annotated (e.g. metadata). In addition, for the output of the tools, postprocessors had to be developed to produce a uniform stand-off XML format. The latter was not trivial because only the Connexor tool was able to give information about character offsets of the analyzed surface forms in the original text.

Once the pre-processing and post-processing scripts worked sufficiently well on a sample of DeReKo, the process of annotating the whole DeReKo started in March 2009. As there were no versions of the tools for our default platform Solaris x86 and our provisional tests had shown that part-of-speech tagging with disambiguation and morphological tags was quite time-consuming – especially in the case of the Xerox tool –, a new (and thus untested) Linux machine with 32 cores (AMD Opteron 8356 processors at 2.3 GHz) and 256 GB RAM was borrowed from another IDS project.

The processing of DeReKo's approx. 350 XCES-files with up to 2 GB each had to be interrupted and restarted several times because new offset-linking problems with implications for the pre-processor were detected or because of hardware problems with the untested machine. In July 2009, after about 6 CPU-years (taking account of all restarts), the annotations with the TreeTagger and the Connexor tool were finished and the Xerox annotation was suspended at about 60 % to have time for a first evaluation before the DeReKo-2009-II release scheduled for August.

By then, the size of stand-off XML annotation data containing all part-of-speech and morphological analyses provided by the tools totaled 3.5 TB and, in first iteration, cornerstones 1-5 (see Section 2.1, Selection of tools) were put into action.

3. Analysis of tagging results

3.1 Methodology

In order to obtain a first impression of the reliability and usability of the tools for linguistic tasks without getting lost in 3.5 TB of annotation data, we decided to start our analysis with a superficial comparison of the outputs of the three taggers deliberately ignoring everything but POS information and everything that was not easily comparable.

3.1.1 Tag sets

To be able to perform the comparison, we first had to define a basic tag set and mappings from the original tag sets. The result was the tag set B_9 shown in Table 2 with nine part-of-speech categories. B_9 is a true subset of the Connexor base tag set, leaving out its categories for *interjection* (INTERJ) and *subordinating conjunction* (CS), which had no consistent correspondents in neither the TreeTagger nor the Xerox tag set. From the Xerox and TreeTagger tag sets the tags DATE, FM, KOKOM, KOUI, KOUS, PTKVZ, TRUNC, and XY (see Schiller et al. 1999) were not consistently mappable on common tags. According to our conservative approach, whenever a non-mappable tag was encountered in a comparison of decisions, the comparison was ignored and not counted.

Mainly in order to obtain a rough idea of how big the influence of the tag set granularity on the comparison of results was, we also defined a tentative more fine-grained base tag set B_{26} with 26 different categories derived from the STTS including some disjunctions like PRELS / PRELAT and PDS / PDAT. As the mapping was not straightforward and is in need of further inspection, results based on it will be reported in parentheses.

3.1.2 Corpus

The sample corpus we conducted the comparison on was the DeReKo-based virtual corpus (cf. Kupietz / Keibel 2009) POScomp09a with 370 million words in 1.7 million texts from mainly German newspapers from 1997 to 2009.[2]

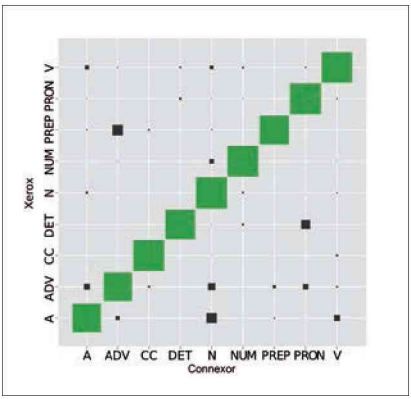
Based on the coarse base tag set B_9, we examined the average POS tag correspondence of pairs of tools. The confusion matrices in Figures 1a to d show the results. Based on the decision by the tool shown on the y-axes, each intersection point has a square with a size relative to the proportion of corresponding classifications by the tool shown on the x-axes. With no disagreement between the tools, the squares would only appear on the diagonals. Any disagreement

[2] To obtain an idea of the impact of the sample composition on tagger performance, we evaluated the significance of the factors *text source* (*publisher*), *year of publication*, *topic* (Weiss 2005), and *country of publication* on the agreement between the three tools. Tukey HSD tests on the corresponding ANOVAs showed that there were significant correlations between all factors and the degree of agreement. However, taking into account the sample size, significant correlations were not surprising and the magnitude of the influences was rather small. We concluded that our virtual corpus is suitable for the comparison and that with respect to tagger performance, DeReKo is rather homogeneous (see Giesbrecht / Evert 2009 in contrast for web corpora).

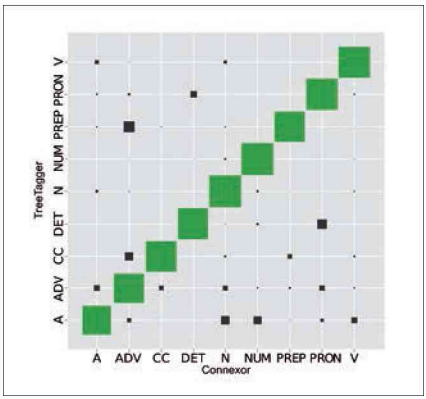
is depicted as a square outside the diagonals. For instance, as shown in Figure 1, tokens classified as adjectives by Xerox are often classified as nouns by Connexor.

B_9	Xerox	TreeTagger
A	ADJA, ADJA2, ADJA3, ADJD, ADJD2, ADJD3	ADJA, ADJD, VMPP
ADV	ADV, PTKANT, PTKCOM, PTKNEG, PTKSUP, WADV	ADV, PROAV, PTKNEG, PWAV
CC	COORD	KON
DET	ART	ART
N	NOUN	NE, NN
NUM	CARD, ORD	CARD
PREP	PREP, PTKINF	APPO, APPR, PTKZU
PRON	DEMDET, DEMINV, DEMPRO, INDDET, INDPRO, PERSPRO, POSDET, POSPRO, REFLPRO, RELPRO, REZPRO, WDET, WINV, WPRO	PDAT, PDS, PIAT, PIS, PPER, PPOSAT, PPOSS, PRELAT, PRELS, PRF, PWAT, PWS
V	VAFIN, VAINF, VINF, VMFIN, VVFIN, VVINF, VVIZU, VVPP	VAFIN, VAIMP, VAINF, VAPP, VMFIN, VMINF, VVFIN, VVIMP, VVINF, VVIZU, VVPP

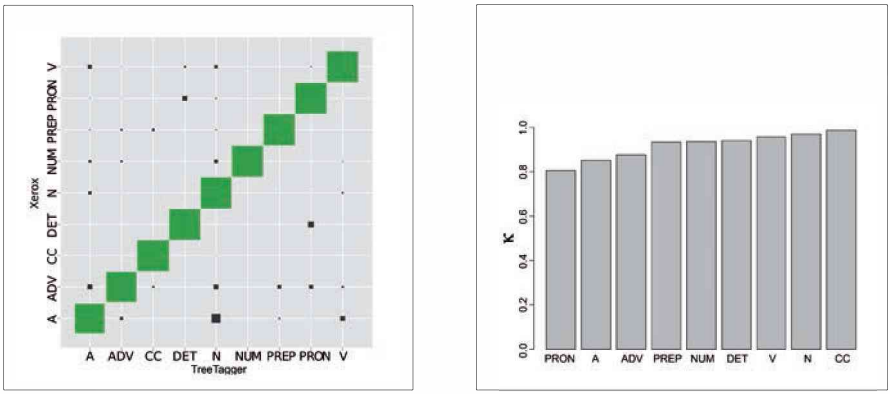
Table 2: Mapping to a coarse base tag set B_9



(a) Xerox – Connexor



(b) TreeTagger – Connexor



(c) Xerox – TreeTagger (d) by POS-tag

Figure 1: Tag correspondences and inter-tagger agreement

3.2 Assessment of reliability

Based on our coarse base tag sets, we also measured the overall agreement between the three tools with respect to single tokens. As shown in Table 3, the combination Xerox and TreeTagger had the highest percentage of agreement with 95.59 % while TreeTagger and Connexor only agreed on 93.47 % of the tokens and for the agreement of all three taggers the percentage drops to 91.57 %. To account for by-chance matches and the rather small tag set, κ coefficients were calculated additionally, shown in the last column.

ts	tagger 1	tagger 2	tagger 3	%	κ
B_9	Xerox	TreeTagger		95.59	0.947
B_{26}	Xerox	TreeTagger		(94.40)	(0.935)
B_9	Xerox	Connexor		93.86	0.926
B_9	TreeTagger	Connexor		93.47	0.921
B_9	Xerox	TreeTagger	Connexor	91.57	0.931*
STTS	TreeTagger	Gold		93-98†	

Table 3: Inter-tagger agreement for single tokens (* Fleiss' κ, † reported accuracy)

There are only a few published reliability evaluations of POS-taggers for German and generally the reported overall accuracy rates are between 93 and 98 % (cf. Giesbrecht / Evert 2009), so that taking into account our coarse tag set and a corpus of mainly 'easy' but partially very recent newspaper texts, the results are within the expected range but slightly lower then hoped for. However, the purpose of this study was not to objectively evaluate the quality of POS taggers. In that case, an inspection of the general relation between our agreement measures and a notion of *accuracy* and some deeper inspection of e.g. unknown words would have been necessary.

With respect to usability of the POS annotations in corpus-based linguistic research, the proportions of full agreement on sentences, shown in Table 4, however, look somewhat alarming. Xerox and TreeTagger only agreed on every second sentence and all three taggers agreed on less than every third sentence.

ts	tagger combination	%
B_9	Xerox, TreeTagger	50.82
B_{26}	Xerox, TreeTagger	(42.16)
B_9	Xerox, Connexor	44.74
B_9	TreeTagger, Connexor	38.92
B_9	Xerox, TreeTagger, Connxeor	31.36
STTS	TreeTagger, Gold Standard	33.67-73.85*

Table 4: Inter-tagger agreement for whole sentences with mean sentence length = 15 tokens (*estimated based on reported accuracy)

4. Pitfalls

As already indicated in the previous section, a first possible pitfall for the linguistic exploitation of automatic POS annotations can arise from differences in part-of-speech taxonomies. A perfect mapping, at least to a fine-grained common tag set, was not straightforward and in addition to consulting the documentation required the comparison of samples. The linguist is confronted with a similar situation because he / she has to find out first how the categories used by the taggers relate to categories he / she has in mind.

The other, very obvious potential pitfall is of course that the tagging tools do produce errors with regard to their intended taxonomy. Whether the observed agreement rates can be interpreted as accuracy rates or not, they are probably quite good, conservative estimates for the agreement of expected categorizations with those actually performed by the tools. That means that roughly at least every second sentence or every fourth five-word sequence will not be tagged as expected.

For linguists, however, the exact rate of errors in annotations is far less important than their possible consequences within their specific research context. This will be the topic of the following three sections.

4.1 Error types

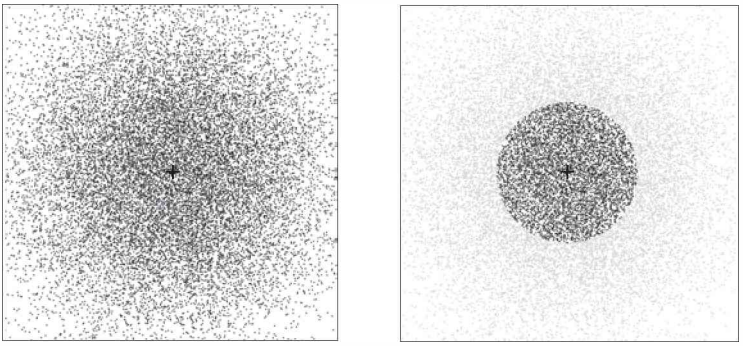
Mimicking one of the grammarian's most common lines of thought, let us assume we are looking for sentences that contain a particular sequence of parts of speech and the corresponding corpus query yielded 20 000 sentences as hits. If the accuracy of the result is roughly 75 % ($\approx 0.93^4$) and errors are distributed evenly, we are likely to have about 2 500 false hits among the 20 000: *false positives*. While this is bad news, the problem can be solved by sorting out the false hits manually. Much more problematic is that additionally we are likely to *miss* about 2 500 sentences we were actually looking for: *false negatives*. What if the unseen false negatives in fact contradict the findings based on the seen data?

As we have no access to such type II errors and consequently know nothing about them, there is a danger that without realizing we may end up not analyzing observations of language use but also the tagging tool, the theory and language model behind it, or possibly its imperfect implementation.

4.2 Error distribution

In this section, we present some considerations on the annotation error distribution from the point of view of linguistic research concerned with language variation in order to foresee and discuss possible dangers. Accordingly, our reasoning here is based on general qualitative assessments rather than hard quantitative evidence.

Let us assume we want to conduct research on language variation making use of a large collection of observations of language use events recorded in a corpus. Figure 2a gives an informal view of how the sample of language productions in our corpus is likely to be distributed around some *language core* (shown as black cross in Figures 2 to 7).

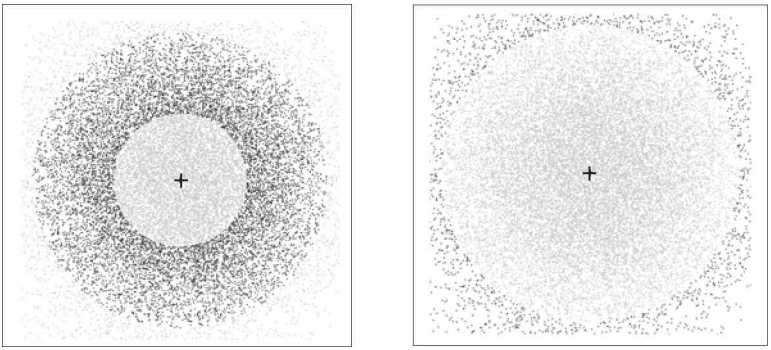


(a) Distribution of language phenomena around language core

(b) Focus on central phenomena

Figure 2: Schematic view of the distribution of language phenomena in a corpus and the focus of linguistic research

The further a point is from the centre of the plot in Figure 2a, the less familiar is the language phenomenon (LP) it stands for, e.g., it is the less frequent, the less standardized or conventionalized, the less uniformly distributed across areas, time, genres, topics, etc., the worse understood by general public, and so on. However, the central area need not represent the language core with respect to the language as a whole (which is the ambition of *representative general reference* corpora). Rather, it may refer to quite different realms of linguistic reality, it may be widely spread or tightly focused, e.g. on certain peripheral language phenomena (LPa), depending on specific sampling criteria used to build that corpus. Accordingly, we use the term *language core* to denote also the core LPa of any such – however skewed – sample of language productions throughout this paper. Since our assumed goal is to study language variation, we are likely to pay less attention to the central LPa covered by the corpus, cf. Figure 2b, and to concentrate on phenomena more distant from the corpus language core as shown in Figure 3a instead, possibly ignoring marginal phenomena highlighted in Figure 3b.

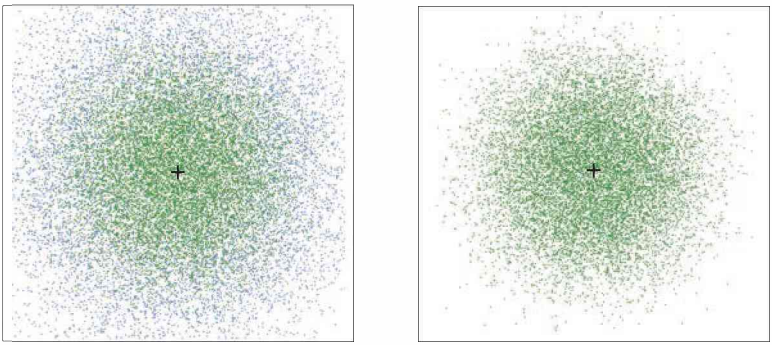


(a) Focus on language variation (b) Focus on marginal phenomena

Figure 3: Focus of linguistic research

The intuitive notion of a language phenomenon used here is a very general one. It includes any observed language event which can sensibly be captured in terms of (any combination of) lexical, syntactic, semantic, stylistic, prosodic, phonetic, pragmatic, or other linguistic characteristics. Obviously, the dimensionality of the space spanned by these characteristics is considerably high. To clarify our line of thought, we use schematic two-dimensional plots in our examples nonetheless.

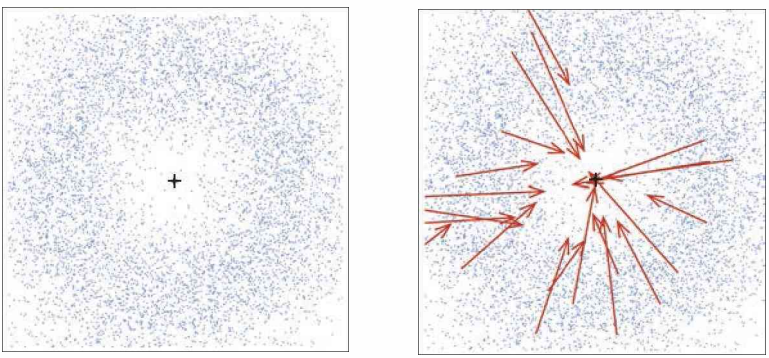
Let us also assume we have a piece of software that is capable of assigning such linguistic characteristics to observed language data, i.e., an annotation tool (tagger, parser, etc.), and we expect it to assist us in the process of exploring language variation. As is plainly evident, every annotation tool implements a notion of language core of its own, and, to keep our argumentation as simple as possible, we assume this language core is identical with that of our corpus. However, no annotation tool is perfect. While some LPa documented in our corpus are likely to lie within the linguistic scope of the annotation tool, others are not. It seems reasonable to expect that with increasing distance from the language core, the probability of a LP being within the scope of the annotation tool decreases. Figure 4a shows an assumed typical distribution of LPa within and outside the linguistic scope of an annotation tool.



(a) All phenomena (b) Within-the-scope phenomena

Figure 4: Distribution of language phenomena and the linguistic scope of the annotation tool (green: LP lies within the scope of the annotation tool; blue: LP lies outside the scope of the annotation tool)

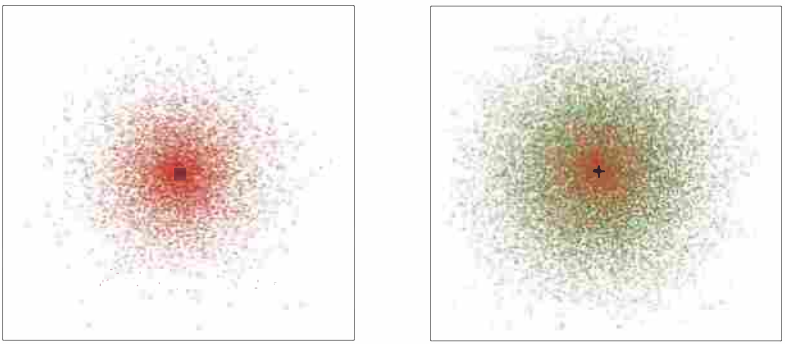
Let us now try to assess how accurate the annotation of these two groups of LPa is likely to be. Because the green dot phenomena (cf. Figure 4b) lie within the scope of the annotation tool, it is reasonable to assume that they are more likely to be annotated correctly than the blue dot phenomena shown in Figure 5a



(a) 'True' distribution (b) Drift towards language core

Figure 5: Language phenomena that lie outside the linguistic scope of the annotation tool and their distribution around language core

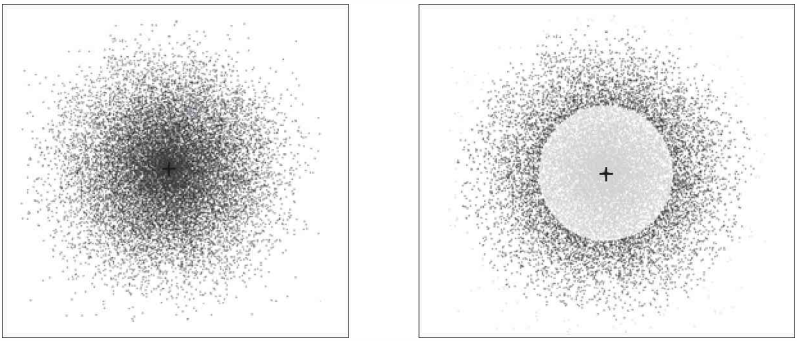
which lie outside the linguistic scope of the annotation tool. However, if the outside-the-scope LPa are more prone to annotation errors, the following question arises: can a sensible judgement be made about whether or not there is a systematic bias in the annotations actually attached to the outside-the-scope LPa by the tool as compared with their 'true' (or 'correct') linguistic characteristics? We suggest that random noise *plus a systematic drift* towards language core is introduced by the annotation tool into the data for the blue dot LPa, as indicated by red arrows in Figure 5b, provided the tool tends to assign some 'best guess' attribute to an LPa it is unaware of, which is what most annotation tools do.



(a) Annotations of outside-the-scope LPa (b) All annotations

Figure 6: Distribution of annotations around language core (green: the annotated LP lie within the linguistic scope of the annotation tool; red: the annotated LP lie outside the scope of the annotation tool)

Thus, the 'true' distribution of the LPa outside the scope of the annotation tool as shown in Figure 5a is mapped by the annotation tool to a biased distribution of – partially wrong – corpus annotation tags plotted as red dots in Figure 6a. In Figure 6b, this annotation tag distribution is plotted together with the distribution of the – predominantly correctly assigned – tags of the LPa shown previously in Figure 4b. Consequently, the resulting overall distribution (Figure 7b)



(a) Distribution of annotations (b) Focus on language variation

Figure 7: Distribution of annotations around language core and the focus of linguistic research concerned with language variation

of annotation tags in the space spanned by the 'true' linguistic characteristics is also biased towards language core with respect to the overall distribution of LPa in the corpus (Figure 2a). Especially the language variation phenomena we are actually interested in (cf. Figure 3a) tend to be misleadingly annotated as being more consistent with the linguistic scope of the annotation tool than they in fact are. It is due to this systematic annotation drift that our corpus data – if inspected indirectly, i.e., through the annotation layer – appear partially drained of linguistic variation (see Figure 7b).

4.3 Ways around

Unlike many technological (e.g. NLP or IR) projects, pure linguistic research is expected to be meticulously concerned with the theoretical status of its data and particularly of any annotation data included in language corpora. In linguistics, it is in general of crucial importance to construe annotation data as mere assessments (or 'opinions') of (human or automatic) annotators rather than as straightforward observations of language use. Consequently, these 'opinions' might sometimes turn out to be incomprehensible because of unclear or unfamiliar terminology (e.g., part-of-speech taxonomy), sometimes objectively wrong, and sometimes just grossly incompatible with one's own judgement.

A first remedy is to consider using more than a single source of such 'opinions', particularly when rare or non-standard linguistic phenomena are the object of research. However, this might still prove to be insufficient to avoid type II er-

rors. For instance, it is not unlikely that all tools are biased in the same or similar way, as sketched in the previous section. Thus, to increase the recall even more, it might be advisable to take into account not only those interpretations that were regarded as most plausible by the annotation tools but rather to consider all interpretations that were regarded as possible. While such an approach is supported by the current DeReKo annotation initiative, it is obvious that in extreme cases this might either still not be enough or that the resulting number of concurrent interpretations might render the use of the annotation data impractical.

There are no universal solutions, neither to the problem of uncertainty associated with relevant but unseen false negatives nor to the precision vs. recall trade-off. We suggest that a possible general strategy for using annotated corpora in linguistic research in a scientifically sound way is to adhere to the following 'safety' guidelines:

1) start with a general corpus query that aims at maximum recall;
2) apply a filter to remove the largest group of *false positives*;
3) cross-check manually (based on a random sample, if appropriate) if your filter has any undesired side-effects with respect to both *false negatives* and *false positives*;
4) adapt the filter and incorporate it into your corpus query;
5) repeat until *false positives* can be handled (e.g. filtered out) manually;
6) include the final query and a detailed error discussion in your publication.

Two of the authors have applied these guidelines in the course of two months' worth of research on German subject infinitival clauses with and without *zu*. A concise result of their endeavor has been published in Kubczak / Konopka (2008: 258f.).

5. Conclusions and Outlook

It is an inevitable fact that automatic linguistic annotations on a scale between *observation* and *interpretation* clearly have the status of interpretations if they depend on a norm, such as a POS taxonomy, which cannot be derived from the data. And in addition, annotations are typically erroneous with respect to such norms. Nevertheless, the use of, e.g., POS annotations can undoubtedly

make corpus query tasks easier. However, to achieve scientifically sound results based on automatic annotations, the adoption of a careful, potentially time-consuming strategy is indispensable. In general, such a strategy will be needed to avoid uncontrollable errors, as in search tasks most notably type II errors (false negatives). Such a strategy will typically involve an initial maximization of recall and many, eventually manual, iterations of filtering out false positives. In addition, to take account of error-prone interpretational status, the adoption of good practices from other sciences, as for example a strict experimental design and the discussion of possible sources of errors is particularly important.

While the necessity for such a careful modus operandi will never disappear, there is also ample room for improving the current annotation of DeReKo and its usability in linguistic research. Because the mission of the IDS is to conduct basic and applied linguistic research rather than to pursue language technology, we will not try to improve tagging tools ourselves, but, for instance, we consider providing a *maximum recall* and a *maximum precision* annotation layer to simplify the search and filter tasks respectively. Other techniques to enhance the benefits of annotations generated by third-party NLP tools in our research context might be to include auxiliary morphological analyzers to perform regular lexicon updates, or to apply the *ensemble of classifier* approach, where applicable.

References

Giesbrecht, Eugenie / Evert, Stefan (2009): Part-of-speech tagging – a solved task? An evaluation of POS taggers for the Web as corpus. In: Iñaki, Alegria / Leturia, Igor / Sharoff, Serge (eds.): Proceedings of the 5th web as corpus workshop (wac5). San Sebastian, Spain. Internet: http://purl.org/stefan.evert/PUB/GiesbrechtEvert2009_Tagging.pdf (last visited: 07 / 2010).

IDS = Institut für Deutsche Sprache (ed.) (1991-2009): COSMAS I / II Corpus Search, Management and Analysis System. Internet: http://www.ids-mannheim.de/cosmas2/ (last visited: 07 / 2010).

Koskenniemi, Kimmo / Haapalainen, Mariikka (1996): GERTWOL – Lingsoft Oy. In: Hausser, Roland (ed.): Linguistische Verifikation. Dokumentation zur Ersten Morpholympics 1994. (= Sprache und Information 34). Tübingen: Niemeyer, 121-140.

Kubczak, Jacqueline / Konopka, Marek (2008): Grammatical variation in near-standard German: A corpus-based project at the Institute for the German Language (IDS) in Mannheim. In: Štícha, František / Fried, Mirjam (eds.): Selected contributions from the conference Grammar and Corpora, Sept. 25-27, 2007, Liblice, Czech Republic. Prague: Academia, 251-260.

Kupietz, Marc / Keibel, Holger (2009): The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In: Minegishi, Makoto / Kawaguchi, Yuji (eds.): Working papers in corpus-based linguistics and language education No. 3. Tokyo: Tokyo University of Foreign Studies (TUFS).

Schiller, Anne / Teufel, Simone / Stöckert, Christine / Thielen, Christine (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical Report. Universität Stuttgart: Institut für maschinelle Sprachverarbeitung.

Schmid, Helmut (1994): Probabilistic part-of-speech tagging using decision trees. In: International conference on new methods in language processing. Manchester, UK, 44-49.

Tapanainen, Pasi / Järvinen, Timo (1997): A non-projective dependency parser. In: Proceedings of the 5th conference on applied natural language processing. Washington DC. Morristown, NJ: Association for Computational Linguistics, 64-71.

Weiss, Christian (2005): Die thematische Erschließung von Sprachkorpora. Technical Report. Mannheim: Institut für Deutsche Sprache. Internet: http://www.ids-mannheim.de/kl/projekte/methoden/te/te.pdf (last visited: 07 / 2010).

IV. Einblicke in die aktuelle Forschung / Insights into current studies

FRANTIŠEK ŠTÍCHA

Der kommunikative und der systembezogene Status grammatischer Phänomene mit niedriger Häufigkeit

Abstract

Nach einer kurzen Übersicht über die möglichen Gründe für die (sehr) niedrige Frequenz, die gewisse sprachliche Phänomene in großen elektronischen Korpora aufweisen, werden zwei solcher grammatischer Erscheinungen im Deutschen und Tschechischen en detail diskutiert: (a) der perfektive Aspekt von tschechischen Adjektiven, die vom Partizip Präsens abgeleitet sind (***probudící*** *se dítě*) und die (mögliche) perfektive Lesart der analogen Wörter und Phrasen im Deutschen (*ein* ***aufwachendes*** *Kind*); (b) die analytische Form des Komparativs bei Adjektiven, Typ *mehr schön*: Im Deutschen wird diese Form gelegentlich benutzt, aber nur bei einer eng begrenzten Gruppe von Adjektiven, nämlich solchen mit einem verbalen Stamm (oder Wurzel) und dementsprechender Semantik (z.B. ***mehr abhängig***). Im Tschechischen ist demgegenüber diese Form des Komparativs deutlich gebräuchlicher (z.B. ***více barevný*** – 'mehr farbig'). Allerdings gehört sie auch in der tschechischen Grammatik eher zur Peripherie.

1. Allgemeines

Erst seit relativ wenigen Jahren ermöglichen große elektronische Korpora der Linguistik, systematisch nach Formen und Strukturen der Sprache zu suchen oder aber zufällig jene zu entdecken, die weder häufig, noch in jedem Satz oder auf jeder Seite eines Textes, und auch nicht in jedem Buch oder bei jedem Autor etc. vorkommen.

Formen oder Strukturen, die auch in einem großen Korpus mit mehreren Milliarden Wortformen sporadisch bis extrem selten bzw. nur vereinzelt vorkommen, können einen sehr unterschiedlichen sprachlichen und / oder kommunikativen Status haben. Es kann sich handeln um:

- Fehler (Tippfehler, Leichtsinnsfehler, Fehler wegen ungenügender „Ausdruckspflege" oder Fehler aufgrund mangelnder Sprachbeherrschung); Fehler können allerdings nicht immer von absichtlichen Abweichungen unterschieden werden;

- exhibitionistische Absicht, die vor keiner Durchbrechung einer festen Sprachnorm zurückschreckt, nur dem Humor dient, und unter Umständen (d.h. nicht in jedem Text) negativ beurteilt wird;
- stilistische Exklusivität in der Kunstprosa oder Poesie;
- einen systembezogenen Ausdruck eines sporadischen bis extrem seltenen Kommunikationsbedarfs in einem Standardtext;
- individuelle Zufallswahl einer zwar akzeptablen, potenziell existierenden und der Standardkommunikation dienenden, jedoch (vielleicht) nicht empfehlenswerten Form.

Da die ersten drei Typen der sprachlichen Seltenheit einerseits klar erkennbar, andererseits extrem facettenreich und veränderlich erscheinen, werden sie hier außer Acht gelassen. Im Folgenden sollen also nur die letzten zwei Kategorien behandelt werden.

2. Zwei Beispiele grammatischer Phänomene mit niedriger Häufigkeit und ihr sprachlich-kommunikativer Status

An zwei Beispielen unterschiedlicher grammatischer Phänomene sollen nun die eher nicht plausiblen Kategorien der Seltenheit eines sprachlichen Ausdrucks dargestellt werden.

Allgemein betrachtet handelt es sich dabei um die Kategorien Aspekt sowie Komparation bei Adjektiven und Adverbien. Bei der Kategorie Aspekt werde ich mich konkret mit dem semantischen Unterschied zwischen perfektiv und imperfektiv bei attributiven Präsenspartizipien im Deutschen und Tschechischen befassen und bei der Komparation mit dem Status der analytischen Form in beiden Sprachen.

Es handelt sich also erstens um die Aspektbedeutung von Strukturen wie:

*ein **aufwachendes** Kind*

und zweitens um potenzielle Strukturen wie

***mehr farbiges** Bild* (für 'farbigeres Bild')

die im Tschechischen häufiger real vorkommen, im Deutschen dagegen nur bei wenigen Adjektiven zu finden sind.

2.1 Beispiel 1: Der perfektive Aspekt des attributiven Partizip I im Deutschen und seine Entsprechung im Tschechischen

Zuerst soll dargestellt werden, wie 'perfektiver Aspekt' hier verstanden wird. Da dabei theoretische Ausführungen (vgl. z.B. Quintin 1994, Katny (Hg.) 2000, Povejšil 1987, Riecke 2000, Uhrová / Uher 1977, 1984) nur wenig helfen können, bediene ich mich – um Missverständnissen vorzubeugen – einer Gegenüberstellung des Deutschen und des Tschechischen. Im Tschechischen sind der perfektive und der imperfektive Aspekt in den meisten Fällen leicht durch unterschiedliche morphologische Verbformen zu unterscheiden. Wir wollen von Satz (1) ausgehen:

(1) Das Kind wacht auf.

Bei diesem vom Kontext isolierten Satz wird die Bedeutung des Prädikats in der Regel der Bedeutung des **Imperfektivs** im Tschechischen entsprechen:

Dítě	se	**probouzí** (imperfektiv).
Kind	refl. Morphem	wacht auf (imperfektiv).

Die Imperfektivität des Vorgangs des Aufwachens kann umschrieben werden mit 'befindet sich im Prozess des Aufwachens'. Sobald das Verb jedoch in einen größeren Kontext eingebettet ist, wie es in (2) der Fall ist, wird das Prädikat in der Regel **perfektiv** verstanden:

(2)	Eine Frau	**wacht auf** und weiß nicht, wo sie ist.	
	Žena se	**probudí** (perfektiv)	a neví, kde je.
	Frau refl. Morphem	wacht auf (perfektiv)	und weiß nicht, wo sie ist.

Man könnte zwar im Tschechischen auch in diesem Kontext die imperfektive Verbform (***probouzí*** statt ***probudí***) benutzen und im Deutschen die imperfektive Bedeutung wahrnehmen, sprachlich würde die Zeit des Aufwachens dann allerdings zugleich als die des Nichtwissens dargestellt werden. Dies widerspricht jedoch der üblichen Wahrnehmung ähnlicher Sachverhalte: zuerst wacht jemand auf, und erst nachdem man aufgewacht ist, nimmt man etwas wahr (oder ruft die Polizei etc.).

Aspektuell wird natürlich auch **das attributive Partizip** gebraucht. Im Folgenden wird das attributiv gebrauchte Partizip I (kurz: **AP I**) behandelt. Über das AP I wird (in den Grammatiken des Deutschen und auch in der sonsti-

gen einschlägigen Literatur) behauptet, es sei immer imperfektiv. Diese Annahme entbehrt jedoch einer Logik und widerspricht auch der Erfahrung. Was die sprachliche Logik betrifft, ist dies ziemlich leicht einzusehen, wenn wir das AP I immer und nur als Ergebnis einer kommunikativ-funktionalen Nominalisierung auffassen. Zuerst wird dies in einer einfachen Übersicht dargestellt:

Aspekt	Verbale Darstellung	Attributive Darstellung durch Nominalisierung
Imperfektiv bei Gegenwartsbezug	*das / ein* ***Kind****, das eben* ***aufwacht***	*das / ein eben* ***aufwachende****(s)* ***Kind***
Perfektiv bei Zukunftsbezug	*das / ein* ***Kind****, das bald* ***aufwacht***	*das / ein bald* ***aufwachende(s)*** ***Kind***

Tab. 1: Aspekt und attributives Partizip I

Es gibt keinen Grund anzunehmen, dass die Nominalisierung nur beim imperfektiven Aspekt möglich ist. Wenn der imperfektive Gebrauch des AP I stark überwiegt und der perfektive Gebrauch bei manchen Partizipien praktisch nicht zu finden ist, liegt das nur am selten vorliegenden kommunikativen Bedarf nach dieser Art der Nominalisierung. Die folgenden Belege zeigen klar die perfektive Bedeutung der Prädikate *aufwacht* und *erwacht*.

(3) Auch wenn der Trainer des Spitzenreiters, Volker Finke, die Mannheimer als „schlafende Riesen" bezeichnet. „Ich hoffe, dass Volker Recht hat, und der Riese **bald aufwacht** [...]". (Mannheimer Morgen, 4.10.2002)

(4) Um 1920 sucht der kranke Komponist Gustav von Aschenbach (Dirk Bogarde) Erholung in Venedig. Künstlerisch wie privat fühlt er sich ausgelaugt. Doch **bald erwacht** wieder die Unruhe in ihm [...]. (St. Galler Tagblatt, 17.11.2001)

Unter den vielen Belegen des entsprechenden AP I *aufwachend-*, *erwachend-* können wir jedoch kaum eine klare Lesart für die perfektive Bedeutung finden: *eine aufwachende Frau* oder *eine erwachende Natur* werden in der Regel eine Frau oder eine Natur sein, die im Prozess des *Auf-* oder *Erwachens* begriffen sind.

Warum ist der perfektive Gebrauch des AP I (ausführlicher dazu siehe Štícha 2009) so selten? Eine Erklärung wäre, dass die Perfektivität gegenüber der Imperfektivität mehr prozessuelle Dynamik bedeutet, da ein Vorgang hierbei be-

grifflich als rasche Veränderung vom Start zum Ziel aufgefasst wird. Diese begriffliche Dynamik kongruiert nicht mit dem begrifflichen Bild einer statischen attributiven Nominalisierung. Außerdem ist der perfektive Gebrauch eines AP I nur dann funktional, wenn es sich um einen expliziten oder impliziten Zukunftsbezug handelt. Dies sind Bedingungen, denen der Sprachbenutzer aber nur selten begegnet. Hinzu kommt noch, dass die Seltenheit einer sprachlichen Form oder Struktur schon selbst ein Hindernis zur Benutzung derselben darstellt. Trotzdem finden sich perfektiv gebrauchte AP I etwa bei Heinrich Böll oder Uwe Tellkamp (siehe die Belege weiter unten).

Um einen Zukunftsbezug kann es sich z.B. im folgenden Satz (5) handeln:

(5) Das bald **aufwachende** Kind wird bestimmt Hunger haben.

Die Subjektphrase

*das bald **aufwachende** Kind*

im Satz (5) kann als systembezogene und funktionale Entsprechung zu folgender Phrase mit Relativsatz aufgefasst werden:

*das Kind, das bald **aufwacht***

Im Folgenden sollen einige Belege der (höchstwahrscheinlich) perfektiven Bedeutung präsentiert werden, die dem Mannheimer Korpus „W-öffentlich" (alle öffentlichen Korpora des Archivs W), dem Prager Parallelkorpus „Intercorp" und dem Roman „Der Turm" von Uwe Tellkamp entstammen:

(6) Bei Abweichungen davon greift das ESP sofort ein. Dem Fahrer wird dies durch eine **aufleuchtende** Kontrolleuchte signalisiert [...]. (Frankfurter Rundschau, 1.2.1997)

(7) Im Wasser **ertrinkende** Insekten versorgen die Ananas mit Eiweiß. (Mannheimer Morgen, 21.8.1999)

(8) Durch ein offenes Fenster stieg in der Nacht zum Donnerstag Günther G. (37) in eine Wohnung in der Innsbrucker Sterzingerstraße ein. Die **aufwachende** Besitzerin verständigte die Polizei. (Neue Kronen-Zeitung, 12.8.1994, S. 12)

(9) [...] goß ihm schließlich eine unversehens aus einer bisher unsichtbaren Tür **auftauchende** Sekretärin Tee ein. (Böll, Gruppenbild mit Dame, Prager Parallel-Korpus „Intercorp")

In den Sätzen (6) und (7) entspricht das perfektive AP I dem perfektiven Prädikat in (6a) und (7a):

(6a) Wenn die Kontrolleuchte **aufleuchtet** (perfektiv), wird dies dem Fahrer signalisiert.

(7a) Wenn die Insekten im Wasser **ertrinken** (perfektiv), versorgen sie die Ananas mit Eiweiß.

Problematisch ist allerdings Satz (8). Die imperfektive Lesart des AP *aufwachende* verträgt sich kaum mit dem Sinn dieses Satzes: Man kann kaum jemanden verständigen, wenn man sich erst im Prozess des Aufwachens befindet. Der zu erwartenden Vorzeitigkeit bei einem Vergangenheitsbezug widerspricht wiederum die partizipiale Präsensform.

In Satz (9) von Böll könnte jedoch auch die subjektive Zeitperspektive eine Rolle spielen: Der Autor will vielleicht durch das AP I die 'auftauchende Sekretärin' vor den Augen des Lesers erscheinen lassen und sie nicht durch das AP II (*aufgetauchte*) in die Vergangenheit verdrängen.

Die folgenden Belege sind dem Roman *Der Turm* von Uwe Tellkamp entnommen:[1]

(10) … jäh **aufblitzende** Fahrradspeichen … (433)

(11) … die **aufblitzende** Wut im Gesicht des Blauuniformierten, die **hochschnellende** Hand … (806)

(12) … die **aufglimmende** Zigarette konnte eine Täuschung sein … (570)

(13) … wie eine giftige, auf Hitzetiegeln rasch **aufsiedende** Flüssigkeit … (426)

Im Tschechischen ist der perfektive Gebrauch des entsprechenden attributiven Partizips noch viel seltener als im Deutschen. Die relevanten Verhältnisse kann man in Tabelle 2 leicht überblicken:

[1] Suhrkamp Verlag, Frankfurt am Main 2008. In einer Rezension wird auch der Gebrauch der attributiven Partizipien erwähnt: „Das Überbordende der obsessiv verwendeten Adjektivpartizipien macht einen fertig." (Julia Encke, Uwe Tellkamp: Der Turm. Das geheime Land. FAZ.NET, http://www.faz.net/s/Rub1DA1FB848C1E44858CB87A0FE6AD1B68/Tpl~Ecommon~SThemenseite.html (Stand: 5/2010)).

	Deutsch		Tschechisch	
	Verbale Darstellung	Gebrauch	Verbale Darstellung	Gebrauch
Imperfektiv	*das Kind, das eben* ***aufwacht***	üblich	*dítě, které se právě* ***probouzí***	üblich
Perfektiv	*das Kind, das bald* ***aufwacht***	üblich	*dítě, které se brzy* ***probudí***	üblich
	Attributive Darstellung		Attributive Darstellung	
Imperfektiv	*das eben* ***aufwachende*** *Kind*	üblich	*právě se* ***probouzející*** *dítě*	üblich
Perfektiv	*das bald* ***aufwachende*** *Kind*	selten	*brzy se* ***probudící*** *dítě*	extrem selten

Tab. 2: Häufigkeit der Aspektrealisierungen im Deutschen und Tschechischen

Bisher ist nicht bekannt, warum die perfektive, vom Präsens-Stamm abgeleitete attributive Partizipialform im Tschechischen seit mindestens hundert Jahren völlig außerhalb des Sprachgebrauchs steht und im gegenwärtigen Tschechischen fast normwidrig ist (Štícha 2008). Trotzdem kann man im Internet auch in ernstzunehmenden Texten verschiedener Stilebenen vereinzelt Belege finden, in denen diese Form verwendet wird. Dies verdeutlicht die funktionale Kraft des Nominalisierungsbedarfs (ebd.), der gelegentlich über den Zwang des gängigen Sprachgebrauchs Oberhand gewinnt.

2.2 Beispiel 2: Der analytische Komparativ

Der analytische Komparativ ist aus dem Englischen gut bekannt, wie z.B. in

a ***more coloured*** *picture*

Im Französischen ist dies sogar die einzige Möglichkeit zur Bildung des Komparativs (mit Ausnahme von drei Adjektiven mit synthetischer Form):

une image ***plus colorée***

Im Tschechischen steht diese Möglichkeit dagegen eher am Rande des Sprachsystems. Bei manchen Adjektiven ist der analytische Komparativ durchaus akzeptabel und keine Seltenheit, wenn auch viel weniger üblich als der synthetische Komparativ. So ist z.B. am tschechischen

více	***barevný***	*obraz*
mehr	farbiges	Bild

nichts auszusetzen. Es gibt jedoch im Tschechischen Nationalkorpus nur vier Belege dieses analytischen Komparativs gegenüber etwa 300 Belegen für die synthetische Form *barevnější* (farbiger).

Bei einem semantisch-morphologischen Typ der Adjektive, die den deutschen partizipialen Adjektiven auf *-end* entsprechen, gibt es im Tschechischen allerdings nur die Möglichkeit des analytischen Komparativs, z.B. bei den tschechischen Entsprechungen von *aufregender*, *überraschender*, *beunruhigender*.

Ob diese Art der Komparation auch im Deutschen gebraucht wird, wird in den deutschen Grammatiken allerdings nicht erwähnt. Nur Helbig / Buscha (2005) schreiben, dass in beschränktem Umfang auch die analytische Komparationsform mit *mehr* möglich ist, besonders bei partizipialen und längeren zusammengesetzten Adjektiven. Eine Suche im Korpus zeigt allerdings ziemlich bald, dass dies nicht stimmt.

Die analytische Komparativform '*mehr* + Adjektiv' (z.B. *mehr farbig*), die die synthetische Form ersetzen können müsste, ist im Korpus bei zahlreichen, wahrscheinlich den meisten Adjektiven überhaupt nicht zu finden. Nur bei einigen wenigen Adjektiven gibt es einen oder zwei Belege gegenüber hunderten oder tausenden Belegen mit synthetischem Komparativ.

Unstrittig ist, dass der analytische Komparativ im Deutschen keine andere kommunikative Funktion hat als der synthetische Komparativ – von speziellen Strukturen wie *mehr laut als schön klingen* abgesehen. Folglich sind die äußerst seltenen Belege des analytischen Komparativs nicht als systembezogene Strukturen zu betrachten, sondern man kann sie im Deutschen zur individuellen Zufallserscheinung der Parole erklären – vielleicht mit Ausnahme einiger Adjektive, die eine abstrakte Relationseigenschaft benennen, z.B. *abhängig von*. Solche Adjektive sollten in jeder großen Grammatik des Deutschen aufgelistet werden. Okkasionell wird auch eine hybride Struktur mit dem synthetischen Komparativ gebraucht, der mit dem Adverb *mehr* verbunden wird.

Einige Belege aus den IDS-Korpora:

(14) Fragen, die vom einleitenden „Wann ist es Ihnen das erste Mal zu Bewußtsein gekommen, daß Sie kein normales Kind waren?" bis zum abschließenden „Mir kommt vor, daß Sie jetzt **mehr intensiv, mehr konzentriert** arbeiten wie je." (Die Presse, 18.11.1992)

(15) Doch die Zeiten änderten sich. „Der Ökologie wurde ein immer höherer Stellenwert innerhalb des Bad Homburger Baurechts eingeräumt", erzählt Berg. Das zwang die Golfer dazu, sich bei jedem Bauteil **mehr intensiver** mit diesem Thema und den dafür Zuständigen auseinanderzusetzen, um die Forderungen des Gesetzbuches zu erfüllen. (Frankfurter Rundschau, 22.8.1998)

(16) Ludwig Feuerbach versuchte die Stellung des Menschen in seiner Tätigkeit **mehr konkreter** zu fassen. (http://de.wikipedia.org: Wikipedia, 2005)

Die Verhältnisse in beiden Sprachen sind in den Tabellen 3 und 4 angedeutet:

Synthetischer Komparativ	Belege	Analytischer Komparativ	Belege
schöner *als*	2875	***mehr schön*** *als*	0
intensiver *als*	1050	***mehr intensiv*** *als*	0
aufregender *als*	241	***mehr aufregend*** *als*	0
farbiger *als*	36	***mehr farbig*** *als*	0
konzentrierter *arbeiten*	59	***mehr konzentriert*** *arbeiten*	1
bekannter *als*	660	***mehr bekannt*** *als*	10
abhängiger *von*	86	***mehr abhängig*** *von*	36

Tab. 3: Synthetischer und analytischer Komparativ: Deutsch

Synthetischer Komparativ	Belege	Analytischer Komparativ	Belege
krásnější (***schöner***)	1667	*více krásný*	1
intenzivnější (***intensiver***)	898	*víc(e) intenzivní*	3
koncentrovanější (***konzentrierter***)	144	*víc(e) koncentrovaný/-án*	17
známější (***bekannter***)	2068	*víc(e) známý*	122
závislejší (***abhängiger***)	133	*víc(e) závislý*	192

Tab. 4: Synthetischer und analytischer Komparativ: Tschechisch

3. Fazit

Ich habe in meinem Beitrag gezeigt, dass es in der Sprache selten bzw. äußerst selten vorkommende Strukturen gibt, die trotzdem nicht offensichtlich gegen die standardsprachliche Norm verstoßen. Diese Strukturen können in ihrem sprachlichen Status jedoch unterschiedlich sein:

- Die perfektive Bedeutung (Lesart, Sinn) des attributiven AP I kann als systembezogen und kommunikativ leistungsfähig bewertet werden. Man kann sie als verborgene sprachliche Kraft ansehen, die in seltenen kommunikativen Augenblicken nützlich werden kann.
- Der analytische Komparativ als Ersatz der synthetischen Form erscheint im Deutschen dagegen als Resultat individueller Zufallswahl, die zwar nicht als etwas Falsches abzulehnen, dennoch – der Normwidrigkeit wegen – keineswegs empfehlenswert ist. Es ist jedoch damit zu rechnen, dass eine individuelle Zufallswahl der Gegenwart die Norm der Zukunft werden kann.

Literatur

Helbig, Gerhard / Buscha, Joachim (2005): Deutsche Grammatik. Berlin / München: Langenscheidt.

Katny, Andrzej (Hg.) (2000): Aspektualität in germanischen und slawischen Sprachen. Poznan: Wydawnictwo naukowe UAM.

Povejšil, Jaromir (1987): Zur Opposition perfektiv – imperfektiv. In: Explizite Beschreibung der Sprache und automatische Textbearbeitung. Bd. XIV. Probleme und Perspektiven der Satz- und Textforschung. Prag: Fakultät für Mathematik und Physik, Karls-Universität, 27-30.

Riecke, Jörg (2000): Über die Darstellung der Aktionsarten in den Grammatiken des Deutschen. In: Brünner Beiträge zur Germanistik und Nordistik R 5: 19-36.

Quintin, Hervé (1994): Zur morphosyntaktischen und semantischen Einordnung von deutschen Partizipien und Partizipialsätzen. In: Bresson, Daniel / Dalmas, Martine (Hg.) (1994): Partizip und Partizipialgruppen im Deutschen (= Eurogermanistik 5). Tübingen: Narr, 91-107.

Štícha, František (2008): Uzuálnost, funkčnost a systémovost jako kritéria gramatičnosti: K jednomu typu morfologické derivace (*udělajíc – udělajíci*). In: Slovo a slovesnost, 69: 176-191.

Štícha, František (2009): Das attributive Partizip I und der Aspekt. In: Acta facultatis philosophicae universitatis ostraviensis, Studia Germanistika 4: 81-94.

Uhrová, Eva / Uher, František (1977): Zur Interpretation der Aktionsarten im Deutschen und Tschechischen. In: Brünner Beiträge zur Germanistik und Nordistik I: 45-73.

Uhrová, Eva / Uher, František (1984): Deutsche und tschechische resultative Verbalpräfixe in der Theorie der Aspektualität. In: Brünner Beiträge zur Germanistik und Nordistik IV: 57-75.

SAID SAHEL

Monoflexion als Erklärung für Variation in der Nominalphrasenflexion des Deutschen

Abstract

In diesem Beitrag werden drei Variationsphänomene aus dem Bereich der Nominalphrasenflexion des Deutschen anhand aktueller Zeitungsbelege untersucht. Es wird argumentiert, dass sich die im untersuchten Korpus beobachtete Variation auf die Tendenz zur Monoflexion zurückführen lässt.

1. Einleitung

Neben dem Prinzip der Kongruenz unterliegt die strukturelle Organisation von Nominalphrasen (NPs) im Neuhochdeutschen der Tendenz zur Monoflexion. In den Referenzgrammatiken sowie in den Lehrbüchern für Deutsch als Fremdsprache wird die Tendenz zur Monoflexion an der 'Kooperation' zwischen Determinierer und attributivem Adjektiv bezüglich der morphologischen Markierung der grammatischen Kategorien der NP festgemacht.

Die „monoflexivische Kooperation" (Weinrich 2003: 487) zwischen Determinierer und Adjektiv wird durch die Formflexibilität des Adjektivs ermöglicht. Denn das Adjektiv weist zwei Flexionsmuster auf: ein so genanntes starkes Flexionsmuster mit stärker differenzierten Formen und ein sogenanntes schwaches Flexionsmuster, dessen Formen weniger differenziert sind. Die Wahl eines dieser beiden Flexionsmuster hängt von der Existenz bzw. der morphologischen Beschaffenheit des Determinierers ab. Demnach gilt der Grundsatz, dass entweder der Determinierer oder das Adjektiv das starke Flexiv trägt. Die Grundregel für die monoflexivische Kooperation zwischen diesen zwei Komponenten der NP wird in der Duden-Grammatik seit der vorletzten Auflage (Duden 2005) auf die folgende einfache Formel gebracht:

> Wenn dem Adjektiv ein Artikelwort mit Flexionsendung vorangeht, wird das Adjektiv schwach flektiert, sonst stark. (Duden 2009: 363)

Tatsächlich ist es aber so, dass die monoflexivische Kooperation nicht nur auf den Determinierer und das Adjektiv beschränkt ist, sondern auch andere

Komponenten der NP erfasst, wie es an der Flexion des Adjektivs in NPs ohne vorangestellten Determinierer deutlich wird (*im Sommer* ***vergangenen Jahres***). Entgegen der obigen Regel trägt das Adjektiv nicht das starke Flexiv *-es*, obwohl ihm kein Artikel vorangeht. Daran zeigt sich, dass auch zwischen dem Adjektiv und dem Substantiv eine monoflexivische Kooperation besteht. Das Substantiv enthält bereits das starke Flexiv *-es* und übernimmt somit die Markierung der grammatischen Kategorien der NP; bei der Flexion des Adjektivs wird im Sinne des Prinzips der Monoflexion auf das schwache Flexionsmuster zurückgegriffen.

Vor dem Hintergrund des sprachhistorischen Faktums, dass sich die Variante *-en* als Adjektivflexiv in Konstruktionen wie *im Sommer vergangenen Jahres* erst im 18. Jahrhundert endgültig gegenüber der Variante *-es* durchgesetzt hat, und unter der Annahme, dass diese Entwicklung auf die Tendenz zur Monoflexion zurückzuführen ist (vgl. Admoni 1985: 1541), stellt sich die berechtigte Frage, ob es seitdem weitere Entwicklungen im Bereich der Nominalphrasenflexion gegeben hat und gibt, die auf das Weiterwirken dieses Prinzips hindeuten.

Tatsächlich lässt sich im gegenwärtigen Standarddeutschen eine Variation in der Nominalphrasenflexion ausmachen, deren Ursache u.a. in der Tendenz zur Monoflexion zu suchen ist. Im Folgenden werden Korpusdaten zu drei ausgewählten Variationsphänomenen vorgestellt, an denen gezeigt werden soll, dass die bereits in frühen Stufen des Hochdeutschen einsetzende Entwicklung, was die strukturelle Gestaltung der Nominalphrase nach dem monoflexivischen Prinzip angeht, noch nicht abgeschlossen ist. Im ersten Fall geht es um die Variation der Flexion des zweiten Adjektivs bei doppelter Adjektivattribuierung, wenn die NP bei fehlendem Determinierer im Dat Sg Mask / Neutr steht (Beispiel 1):

(1a) Sie ist aus weichem rot***em*** Leder (Berliner Zeitung, 19.03.2005)

(b) Ihnen fehlt es an erfahrenem politisch***en*** Personal (Berliner Zeitung, 23.02.2004)

Im zweiten Fall betrifft die Variation die Flexion des Determinierers in der Kollokation *dieses Jahres* bzw. *diesen Jahres* (Beispiel 2):

(2a) Anfang *dies**es** Jahres* soll der Senat der neuen Bauordnung zustimmen (Berliner Zeitung, 04.01.2005)

(b) Noch im Laufe *dies**en** Jahres* sollen es 7 000 sein (Berliner Zeitung, 22.02.2005)

Der dritte Variationsfall manifestiert sich als Kasusalternation zwischen Genitiv und Dativ in NPs im Plural, wenn die Rektion durch die Präposition *trotz* determiniert ist (Beispiel 3):

(3a) Wir haben trotz ***der*** *Steuerausfälle* unseren Kernhaushalt [...] hingekriegt (Berliner Zeitung, 1.8.2003)

(b) Gebreselassie war trotz Problem***en*** an der Achillessehne angetreten (Berliner Zeitung, 21.8.2004)

Auch wenn die Variation in jedem der drei Fälle je eine andere Komponente der Nominalphrase betrifft: das Adjektiv (Beispiel 1), den Determinierer (Beispiel 2) bzw. das Substantiv (Beispiel 3), sodass die drei Phänomene auf den ersten Blick unterschiedlich gelagert zu sein scheinen, lässt sie sich auf einen gemeinsamen Nenner bringen: die Tendenz zur Monoflexion.

2. Die Korpusstudie

2.1 Das Korpus

Das Ausmaß der Variation und die Häufigkeitsverteilung der Varianten im gegenwärtigen Standarddeutschen werden anhand von Belegen aus dem Berliner-Zeitung-Korpus, das Teil einer Sammlung mehrerer elektronischer Korpora der Berlin-Brandenburgischen Akademie der Wissenschaft ist (www.dwds.de), untersucht. Das Korpus enthält alle Ausgaben der Berliner Zeitung, die zwischen Januar 1994 und Juni 2005 erschienen sind, und zählt ca. 252 Millionen laufende Wortformen. Das Korpus ist online öffentlich zugänglich, lemmatisiert, mit Wortartinformationen versehen und mit einer Suchmaschine abfragbar.

2.2 Ergebnisse

Variation am Adjektiv: Doppelte Adjektivattribuierung Dat Sg Mask / Neutr

Die Ausgaben der Berliner Zeitung aus den Jahren 2004 und 2005 (die zwei jüngsten Jahrgänge im Korpus) wurden nach Nominalphrasen im Dat Sg Mask / Neutr durchsucht, in denen dem Kopfnomen zwei (nicht durch ein Komma getrennte) attributive Adjektive ohne vorangestellten Determinierer vorangehen (siehe Beispiel 1). Es wurden insgesamt 205 Belege extrahiert und in die Auswertung einbezogen. Die Auswertung ergab die folgende Verteilung der starken und schwachen Flexion des zweiten Adjektivs:

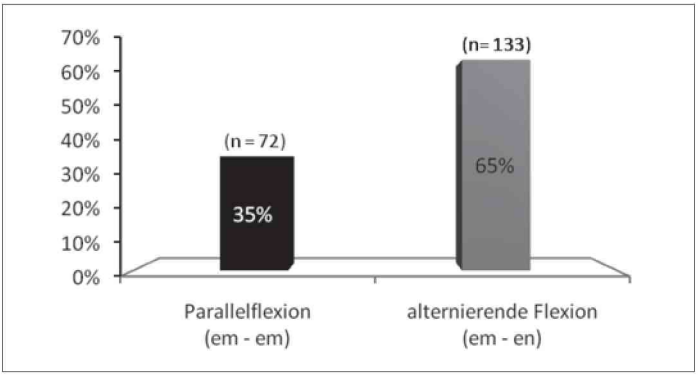


Abb. 1: Verteilung der Parallel- und alternierenden Flexion bei doppelter Adjektivattribuierung

In zwei Dritteln der Belege (65 %) trägt das zweite Adjektiv das schwache Flexiv *-en*, während in lediglich einem Drittel die starke Flexion, also das Flexiv *-em*, vorliegt. Die Verteilung zugunsten der alternierenden Flexion überrascht insofern, als die Parallelflexion im Neuhochdeutschen allgemein als die standardsprachliche Regel gilt. Dass im Dat Sg Mask/Neutr das zweite Adjektiv auch schwach flektiert wird, wird zwar in den meisten Referenzgrammatiken erwähnt; die meisten Angaben der Grammatiken lassen sich aber, was das Vorkommen und die Verteilung der zwei Varianten angeht, nicht immer mit dem realen Sprachgebrauch in Einklang bringen. Jung (1990: 296) hält die alternierende Flexion bei mehrfacher Adjektivattribuierung für ein Randphänomen, dem man „manchmal" in der Literatursprache begegnet. Nach Helbig/Buscha (2001: 276f.) ist die alternierende Flexion nur dann möglich, wenn es sich bei dem ersten Adjektiv um „ein unbestimmtes Zahladjektiv oder ein ähnlich klassifizierendes Adjektiv [handelt]".[1] In der jüngsten Auflage der Duden-Grammatik steht zwar, dass „in einer Abfolge mehrerer starker attributiver Adjektive vom zweiten Adjektiv an die Endung *-en* stehen [kann]", im anschließenden Satz wird jedoch angemerkt, dass „Parallelflexion (also überall die Endung *-em*) [...] in der Standardsprache vorgezogen [wird]" (Duden 2009: 959). Die Verteilung der zwei Varianten in den untersuchten Zeitungsbelegen legen aber nahe, dass die alternierende, nicht die Parallelflexion in der Standardsprache präferiert wird. Zumindest gilt diese Präferenz für die Zeitungssprache. Die beobachtete Verteilung kann nicht als zufällig oder als eine Besonderheit des untersuchten Korpus angesehen werden. Eine ähnliche Verteilung zugunsten der alternierenden Flexion findet sich auch in der überregio-

[1] In keinem der hier untersuchten 205 Belege handelt es sich bei dem ersten Adjektiv um ein unbestimmtes Zahladjektiv.

nalen Zeitung *Die Zeit.* In 54 Belegen aus den Jahren 2004 und 2005 werden die attributiven Adjektive in 61 % (n = 37) alternierend und in nur 39 % (n = 17) parallel flektiert. Wenn man die Verteilung der zwei Varianten in der Berliner Zeitung und in Der Zeit als repräsentativ für die Zeitungssprache erachtet und sie mit den Ergebnissen der Untersuchung von Moulin-Fankhänel (2000) vergleicht, stellt man einen Anstieg der alternierenden Flexion fest. Die Autorin untersuchte Zeitungsbelege, über deren Quelle sie zwar keine genauen Angaben macht, die aber „größtenteils aus den achtziger und neunziger Jahren [stammen]" (ebd.: 84). In ihrer Untersuchung überwog noch die Parallelflexion mit einer Quote von 56 %.

Der deutlich höhere Anteil von Fällen mit alternierender Flexion kann hauptsächlich auf die Tendenz zur Monoflexion zurückgeführt werden. Durch die alternierende Flexion wird eine mehrfache Markierung der grammatischen Kategorien der NP vermieden, indem analog zu NPs mit vorangestelltem Determinierer nur das erste Element der NP mit einem starken Flexiv versehen wird. Die schwache Flexion des zweiten Adjektivs kann demnach als Reanalyse des ersten Adjektivs als Determinierer bzw. klammereröffnendes Element erklärt werden.

Variation am Determinierer: *dieses Jahres* vs. *diesen Jahres*

Die Ausgaben der Berliner Zeitung aus dem Jahr 2005, dem jüngsten Jahrgang im Korpus, wurden nach den Kollokationen *dieses Jahres* bzw. *diesen Jahres* durchsucht. Es wurden insgesamt 621 Belege extrahiert und in die Auswertung einbezogen. Abbildung 2 zeigt die Verteilung der Varianten im untersuchten Korpus:

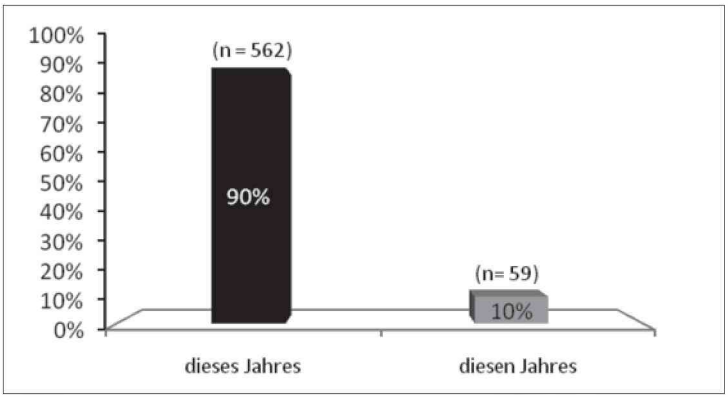


Abb. 2: Verteilung der Varianten *dieses Jahres* und *diesen Jahres*

Die Verteilung der starken und schwachen Flexion des Demonstrativpronomens *dies-* in der festen Wendung *dieses Jahres* bzw. *diesen Jahres* fällt mit 90 % zu 10 % deutlich zugunsten der starken Flexion aus. Unabhängig von der Verteilung der Varianten ist es interessant zu beobachten, dass Pronomina wie *dies-*, die ursprünglich nur ein Flexionsmuster kannten (nämlich das pronominale), anfangen, wie Adjektive zu flektieren, also sowohl stark als auch schwach. Eisenberg (1989: 199) spricht in diesem Zusammenhang davon, dass „Pronomina sich hier den Adjektiven an[gleichen], den Gen Sg auf **en** bilden [Hervorh. i. Orig.]" und dass diese Angleichung „wohl auf dem generell engen Zusammenhang zwischen pronominaler und adjektivischer Flexion [beruht]". Trägt das Demonstrativpronomen *dies-* das schwache Flexiv *-en*, verhält es sich also syntaktisch wie ein attributives Adjektiv und nimmt in der NP die Position des Adjektivs und nicht die des Determinierers ein. Tatsächlich ist es aber so, dass die schwache Flexion des Demonstrativpronomens *dies-* (noch) auf Kollokationen beschränkt ist, „in denen es eine tendenziell eher attributive als demonstrative Funktion aufweist" (Stenschke 2007: 80).

Die variable Flexion von *dies-* lässt eine ähnliche Entwicklung erkennen, die das Adjektiv schon im 18. Jahrhundert abgeschlossen hatte, als sich im Gen Sg Mask / Neutr die Variante *-en* (*im Sommer vergangen**en** Jahres*) nach einer langen Variationsphase endgültig gegenüber der Variante *-es* (*im Sommer vergangen**es** Jahres*) durchgesetzt hat. Vor diesem Hintergrund ist zu vermuten, dass die Entwicklung bei *dies-* noch weiter fortschreiten wird. Diese Annahme wird durch eine eigene Untersuchung im Zeit-Korpus, einem Teilkorpus der Berlin-Brandenburgischen Akademie der Wissenschaft (www.dwds.de), gestützt. In diesem Korpus zeigt sich von 1946 bis 2009 eine kontinuierliche Zunahme der Variante *diesen Jahres*: Während in den Ausgaben von 1946-1960 diese Variante kein einziges Mal vorkommt, steigt ihr Anteil in den Jahrgängen 1990-2000 und 2000-2009 auf 2,1 % bzw. 5,6 %. Stützen lässt sich die Annahme einer fortschreitenden Entwicklung beim Pronomen *dies-* durch die viel weiter fortgeschrittene Entwicklung beim Pronomen *jed-*. Die Variante *jeden* lässt sich „im Dudenkorpus fast halb so häufig" (Duden 2009: 969) belegen wie die Variante *jedes* (*die Pflicht jed**en** / jed**es** Schülers*) und überwiegt sogar in festen Wendungen wie *Menschen jed**en** / jed**es** Alters*.

Damals beim Adjektiv wie heute beim Pronomen lässt sich die Variation auf den Einfluss des Substantivs, das bereits das für den Gen Sg Mask / Neutr eindeutige Flexiv *-(e)s* trägt und somit die schwache Flexion beim Adjektiv bzw. Pronomen bewirkt, zurückführen.

Variation am Substantiv: *trotz* + Gen vs. *trotz* + Dat

Die Ausgaben der Berliner Zeitung aus dem Jahr 2003 wurden nach Belegen durchsucht, in denen die Präposition *trotz* eine NP im Plural regiert, wobei die regierte NP entweder aus einem Determinierer und dem Kopfnomen (trotz_Det_N) oder einzig aus dem Kopfnomen (trotz_N) besteht. Für jede Konstellation wurden die ersten 50 Treffer gewählt. Alle in den 50 Treffern der zweiten Konstellation enthaltenen Substantive weisen im Dativ Plural eine Form auf, die von den anderen Kasusformen des Plurals morphologisch unterscheidbar ist, sodass bei der Auswertung immer eindeutig war, welche Kasusrektion vorlag. Die Belege wurden danach ausgewertet, welchen Kasus *trotz* in der jeweiligen strukturellen Konstellation regiert. Abbildung 3 zeigt die Verteilung der regierten Kasus im untersuchten Korpus:

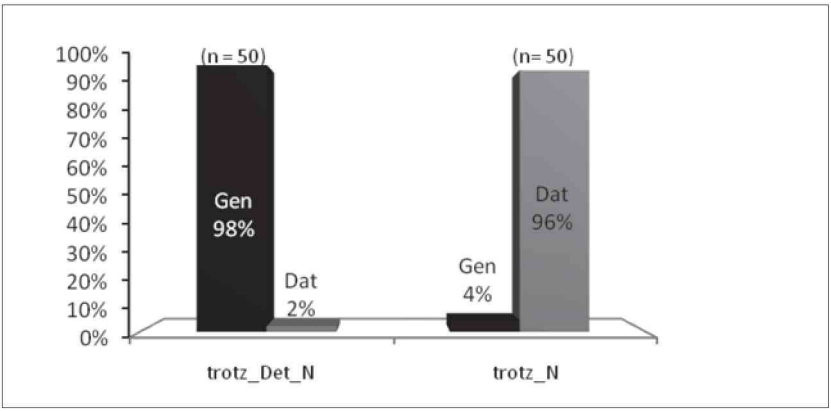


Abb. 3: Verteilung von Genitiv und Dativ nach *trotz* in NPs mit Determinierer (trotz_Det_N) bzw. ohne Determinierer (trotz_N)

Es liegt ein spiegelbildliches Verhältnis der Verteilung von Genitiv und Dativ in den beiden untersuchten Konstellationen vor. Genitivrektion überwiegt zu einem Anteil von 98 % in NPs mit vorangestelltem Determinierer (Beispiel 3a), während in NPs mit dem Kopfnomen als einziger Komponente zu 96 % Dativrektion vorliegt. Die Befunde bestätigen die allgemein bekannte Kasusvariation nach *trotz*. Sie legen aber auch nahe, dass diese Kasusvariation nicht willkürlich ist, sondern in diesem Fall von der Struktur der regierten NP gesteuert wird. Dieser Befund ist keine Einzelbeobachtung. Di Meola (2000: 214f.), der das Rektionsverhalten mehrerer Präpositionen mit Kasusalternation untersuchte, fand einen ähnlichen Zusammenhang zwischen der Verteilung der beiden Kasus und der Struktur der NP.

Diese nahezu komplementäre Verteilung von Genitiv und Dativ kommt dadurch zustande, dass jeweils der Kasus gewählt wird, bei dem im Sinne der Tendenz zur Monoflexion eine mehrfache Markierung in der NP vermieden wird. In NPs mit vorangestelltem Determinierer wird der Genitiv gewählt, da der Genitiv im Plural nur einmal markiert wird, nämlich am Determinierer. Der Dativ Plural wird hingegen am Deteminierer und am Substantiv markiert. Die Wahl des Dativs würde hier zur Polyflexion führen und wird daher vermieden. In Ein-Wort-NPs führt die Wahl des Dativs nicht zu mehrfacher Markierung der NP und bringt noch den Vorteil mit sich, dass der Kasus morphologisch sichtbar wird.

3. Diskussion

Die Datenanalyse und die Interpretation der Befunde haben gezeigt, dass Variationsphänomene, die auf den ersten Blick unterschiedlich gelagert zu sein scheinen, auf einen gemeinsamen Nenner gebracht werden können.[2] Die Variation in den untersuchten Fällen wurde auf die Tendenz zur Monoflexion zurückgeführt, die seit dem Althochdeutschen eine wesentliche Rolle bei der strukturellen Organisation von Nominalphrasen mit kongruierenden Komponenten spielt. Die synchron auftretende Variation in diesem Bereich kann als Fortführen dieser noch nicht abgeschlossenen Entwicklung und somit als 'Sprachwandel' angesehen werden.

Diese Auffassung erscheint umso plausibler vor dem Hintergrund der Annahme, dass Sprachveränderungsprozesse nicht zufällig, sondern gerichtet sind (vgl. Mattheier 1984: 769). Die Richtung der Veränderung kann wiederum im Sinne des natürlichen grammatischen Wandels von Wurzel (1994) als Abbau bzw. Reduktion von Markiertheit verstanden werden: Markierte Konstruktionen werden durch weniger markierte ersetzt. Die im Bereich der Nominalphrasenflexion zu beobachtende Entwicklung kann in diesem Zusammenhang als Abbau von Polyflexion zugunsten von Monoflexion erachtet werden. Die Markiertheit polyflexivischer gegenüber monoflexivischen Konstruktionen ergibt sich wiederum aus dem Prinzip der „Systemangemessenheit" von Wurzel (1994). Es gibt diachron wie synchron genügend Anhaltspunkte dafür,

[2] In den Referenzgrammatiken wurden bisher die hier untersuchten Variationsphänomene individuell erklärt und auf unterschiedliche Ursachen zurückgeführt. Erst in der jüngsten Auflage der Duden-Grammatik wird ein expliziter Zusammenhang zwischen diesen (und weiteren) Variationsphänomenen und der Tendenz zur Monoflexion hergestellt (Duden 2009: 947).

monoflexivische Konstruktionen als die systemangemesseneren anzusehen. Zum einen wird seit dem Althochdeutschen Polyflexion in Nominalphrasen kontinuierlich zugunsten von Monoflexion abgebaut, zum anderen deuten die hier untersuchten Fälle von Variation in der Nominalphrasenflexion im gegenwärtigen Sprachgebrauch darauf hin, dass die Tendenz zur Monoflexion fortgesetzt wird.

Dadurch, dass die in der Nominalphasenflexion beobachtete Variation auf die Tendenz zur Monoflexion zurückgeführt wird, wird keineswegs ausgeschlossen, dass sie auch durch weitere systemimmanente Faktoren gesteuert bzw. begünstigt wird, die sich von Fall zu Fall unterscheiden können.

Literatur

Admoni, Wladimir (1985): Syntax des Neuhochdeutschen seit dem 17. Jahrhundert. In: Besch, Werner / Reichmann, Oskar / Sonderegger, Stefan (Hg.) (1985): Sprachgeschichte: Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung. Bd. 2. Berlin u.a.: de Gruyter, 1538-1555.

Di Meola, Claudio (2000): Die Grammatikalisierung deutscher Präpositionen. (= Studien zur deutschen Grammatik 62). Tübingen: Stauffenburg.

Duden (2005): Duden – Die Grammatik. 7., völlig neu erarb. u. erw. Auflage. Mannheim u.a.: Dudenverlag.

Duden (2009): Duden – Die Grammatik. 8., überarb. Auflage. Mannheim u.a.: Dudenverlag.

Eisenberg, Peter (1989): Grundriß der Deutschen Grammatik. 2., überarb. u. erw. Auflage. Stuttgart: Metzler.

Helbig, Gerhard / Buscha, Joachim (2001): Deutsche Grammatik: ein Handbuch für den Ausländerunterricht. Neubearb. Aufl. Berlin u.a.: Langenscheidt.

Jung, Walter (1990): Grammatik der deutschen Sprache. 10. Aufl. Mannheim u.a.: Bibliographisches Institut.

Mattheier, Klaus J. (1984): Sprachwandel und Sprachvariation. In: Besch, Werner / Reichmann, Oskar / Sonderegger, Stefan (Hg.) (1984): Sprachgeschichte: Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung. Bd. 1. Berlin u.a.: de Gruyter, 768-779.

Moulin-Fankhänel, Claudine (2000): Varianz innerhalb der Nominalgruppenflexion. Ausnahmen zur sogenannten Parallelflexion der Adjektive im Neuhochdeutschen. In: Germanistische Mitteilungen 52: 73-97.

Stenschke, Oliver (2007): 'Ende diesen Jahres': Die Flexionsvarianten von Demonstrativpronomina als ein Beispiel für Degrammatisierung. In: Deutsche Sprache 35, 1: 63-85.

Weinrich, Harald (2003): Textgrammatik der deutschen Sprache, 2. revidierte Aufl. Hildesheim u.a.: Olms.

Wurzel, Wolfgang Ullrich (1994): Grammatisch initiierter Wandel. Bochum: Brockmeyer.

Korpus

Das Digitale Wörterbuch der deutschen Sprache des 20. Jahrhunderts (DWDS) (www.dwds.de).

GEORG ALBERT

Innovative Sprachverwendungen: Verbreitung und Kontext

Abstract

In der Sprachwissenschaft werden Neuerungen im Zusammenhang mit allgemeinen Sprachwandeltheorien diskutiert. Manche Abweichungen von sprachlichen Normen werden als interessante Innovationen wahrgenommen und in der Folge womöglich als neue Ausdrucksmöglichkeiten etabliert. Andere Abweichungen dagegen werden von vorneherein als Fehler oder Lapsus eingestuft, wieder andere Sprachverwendungen funktionieren nur in bestimmten Kontexten und haben eine kurze Lebensdauer. Es stellt sich also die Frage, wie und warum sich bestimmte Abweichungen als Innovationen in der Sprache durchsetzen.

Das hier vorgestellte Dissertationsprojekt mit dem Arbeitstitel „Bedingungen und Kontexte innovativer Sprachverwendungen" soll Trends und Entwicklungen unter Berücksichtigung komplexer Zusammenhänge von Textproduzent, Kommunikationssituation, Lebensstil, Textsorte und Medium beschreiben.

Methodisch ist ein hypotheseninspiriertes, korpusgestütztes Verfahren geplant. Konkrete Phänomene sollen auf ihre Verteilung über verschiedene Textsorten und Domänen hin untersucht werden. Das Ziel ist dabei, systematische Zusammenhänge zwischen innovativen Sprachverwendungen und den Texten, in denen sie auftreten, erkennen zu lassen.

1. Theoretischer Hintergrund

Innovation ist kein innerhalb der Sprachwissenschaft etablierter und definierter Terminus, auch wenn er gelegentlich – meist im Zusammenhang mit lexikalischen Entwicklungen – in sprachwissenschaftlichen Arbeiten erscheint. *Innovation* soll deshalb auch aus wirtschafts- und sozialwissenschaftlicher Perspektive betrachtet und für die Sprachwissenschaft erst operationalisiert werden. Das bedeutet, dass *Innovation* von *Neuerung*, *Abweichung* und *Fehler* abzugrenzen ist (vgl. Cherubim 1980).

Innovativität ist eine Qualität, die einem Phänomen von außen zugeschrieben wird, und geht über bloße Neuheit hinaus. Innovationen werden als neu und dabei zugleich auch als relevant wahrgenommen. Auch Fehler können etwas

Neues sein, werden im Gegensatz zu Innovationen aber negativ sanktioniert. Aus Sicht des oder der Produzenten kann eine Innovation daher als eine motivierte Abweichung verstanden werden.

Innovationen lassen sich in Hinblick auf die Motive ihrer Produzenten und auf ihre Effekte betrachten. Damit hängen Prozesse der Diffusion von Innovationen zusammen. Auf Sprache bezogen meint das die Ausbreitung einer innovativen Form oder eines innovativen Musters innerhalb verschiedener Texte, die sozial, areal, stilistisch etc. beschränkt sein kann, im Extremfall aber bis hin zur Standardisierung der Form bzw. des Musters reicht. Mit einer solchen Diffusion können Änderungen sozial-stilistischer Charakteristika einer Form und / oder semantische Wandelprozesse einhergehen, die dann jeweils zu beschreiben wären.

Innovationen sind mit Kosten und Risiken verbunden, in der Wirtschaft sind dies Entwicklungs- und Produktionskosten sowie das Risiko eines Misserfolgs am Markt. Allerdings ist die Qualität „Innovation" im Erfolgsfall an einen Gewinn gebunden (vgl. Schulz et al. 2000: 56), der Kosten und Risiko rechtfertigt. Sprachliche Innovationen erfordern kommunikative Kompetenz, Kreativität, sprachliches Reflexionsvermögen und einen gewissen kognitiven Aufwand, sie bergen das Risiko des kommunikativen und / oder sozialen Scheiterns. Ihr Nutzen ergibt sich innerhalb spezifischer Kontexte in Form eines stilistischen Effekts und der Reaktion anderer Akteure, wenn diese eine Abweichung oder Neuerung als Innovation positiv sanktionieren.

2. Hypothesen und Fragestellungen

Aus dem Gesagten ergeben sich folgende Fragestellungen im Hinblick auf innovative Verwendungen von Sprache: Welche stilistischen Markierungen gehen mit innovativen Sprachverwendungen einher? Wird eine Form aufgegriffen und breitet sie sich aus, oder kann sie sich auf dem sprachlichen Markt nicht durchsetzen, bleibt sie eine vorübergehende Mode? Welche formalen und semantischen Entwicklungen begleiten die Ausbreitung? Lassen sich Faktoren identifizieren, die eine Verbreitung innovativer Sprachverwendungen begünstigen? Dabei ist zu beachten, in welchen sozial-stilistischen Kontexten und in welchen Textsorten die Innovationen auftauchen. Ressourcen für innovative Sprache können beispielsweise Dialekte oder Szenesprachen sein, wobei deren jeweiliges Prestige die Diffusion unterschiedlich beeinflusst. Grundsätzlich soll unterstellt und dementsprechend auch am Material belegt werden,

dass mit neuen Formen durchaus noch mehr als nur „der Bedarf an Neubenennungen in einer Kommunikationsgemeinschaft befriedigt [wird]" (Herberg / Kinne / Steffens (Hg.) 2004: XI). Es wird die Hypothese aufgestellt, dass sich aus den expressiven sowie sozial-distinktiven Bedürfnissen von Sprecherinnen und Sprechern ein ständiger Innovationsbedarf ergibt, so wie dies auch im Rahmen von Grammatikalisierungstheorien als Faktor im Sprachwandel beschrieben wird (vgl. Girnth 2000: 59f.).

3. Das Korpus

Den bisherigen Untersuchungen liegt ein Korpus von Chatprotokollen des Anbieters *SpinChat* aus den Jahren 2002, 2004, 2007 und 2009 zugrunde. Diese Protokolle enthalten jeweils die Chattexte, die innerhalb einer Woche während je zwei Stunden am Abend (19:00 Uhr bis 21:00 Uhr) sowie während eines ganzen Tages (00:00 Uhr bis 24:00 Uhr) entstanden sind. Dabei handelt es sich um Texte aus verschiedenen Chatrooms bzw. so genannten Channels, die teilweise regional differenzierende Namen tragen. Konkret wurden die Texte aus den Chatrooms „Bayern", „Berlin-Brandenburg", „Einfach nur Chatten" / „30-40-50-Chat-Set",[1] „Hessennetz" und „Pälzer unner sich" untersucht. Insgesamt kommen so knapp 300 000 Zeilen gechatteten Textes zusammen, wobei die durch das Chatprogramm automatisch generierten Zeilen mit den Mitteilungen *entered* bzw. *left*[2] nicht mitgezählt sind.

4. Verwendungen von *können* in verschiedenen syntaktischen Umgebungen

Ein Beispiel für Sprachgebrauch, der sich an den dynamischen Grenzen zwischen Standard und Nonstandard bewegt, sind verschiedene Verwendungsweisen von *können*. *Können* tritt als Vollverb, häufiger jedoch als modales Auxiliar auf. Für den Gebrauch als modales Auxiliar gilt im Standard in intrasubjektiver Verwendung das Muster *können* + Infinitiv als üblich (vgl. Zifonun et al. 1997: 1255), allerdings kann der Infinitiv bei „Ausdrücken, die eine Bewegungsrichtung oder eine intellektuelle Fähigkeit eindeutig bezeichnen" (Weinrich 2003: 299), fehlen (siehe (1) und (2), Beispiele nach Weinrich).

---

[1] Hier wurde das Material aus „30-40-50-Chat-Set" um Material aus dem vergleichbaren Chatroom „Einfach nur Chatten" ergänzt, weil von ersterem kein Material von 2002 vorliegt.

[2] Diese erscheinen, wenn ein Nutzer den Chatroom „betritt" bzw. „verlässt".

(1) Ich habe meine Schlüssel verloren, jetzt kann ich nicht nach Hause.

(2) Ich werde mich um diese Stelle bewerben, ich kann ja recht gut Französisch.

Im Chat-Korpus wird *können* ohne Infinitiv in ähnlicher Verwendung mit NPs verschiedenster Art (vgl. (3) und (4)) kombiniert, semantische Restriktionen, wie die von Weinrich beschriebenen, haben in diesem standardfernen Raum offensichtlich wenig Gültigkeit. Darüber hinaus finden sich dann sogar einige Belege mit der Kombination von *können* mit einer adverbialen Adverb- oder Adjektivphrase (vgl. (5) bis (7)). Solche Verwendungsweisen zeigen die Kreativität und das innovative Potential der Domäne Chat, wo die Akteure ihre eigenen Konventionen und Sanktionen entwickeln.

(3) himbeerbrombeer: ~Shania~: kannste mal ebenklar text
(30-40-50-Chat-Set, 11.01.2007)

(4) nelly.van.nilla: ich kann nicht kochen, pallmallblue ;) [...] nicht um diese zeit *erklär [...] da kann ich nur canapés..*g
pallmallblue: das hatten wir heute schon nelly.van.nilla im mom kann ich nur noch tiefkühlkost [...]
(30-40-50-Chat-Set, 22.01.2009)

(5) zwiebelfisch: soll ich die vorspeise machen the_untouchable *grins
the_untouchable: eher den nachtisch bitte zwiebelfisch
zwiebelfisch: öööööh the_untouchable ich kann süß nicht so gut *schäm
(30-40-50-Chat-Set, 11.01.2007)

(6) MaMo: hihi bigwife also chatteufel mag es nicht schnell*gggg
bigwife: hey MaMo ..... kannst du mal für chatteufel langsam und so ????
(30-40-50-Chat-Set, 11.01.2007)

(7) ass: aber trotzdem kann man 4, wenn man vier anschluesse hat oda?
(Berlin-Brandenburg, 11.01.2007)

Die Verteilung der verschiedenen Verwendungsweisen von *können* im Gesamtkorpus wurde berechnet und soll noch mit Korpora verglichen werden, die andere Arten von Texten enthalten. Tabelle 1 zeigt die Verteilung im Chat-Korpus:

Verwendung	mit Infinitiv	Anapher/ Anadeixis	mit NP	mit Richtungsadverbiale	mit sonstiger Adverbiale	*etw. abkönnen*
Anteil	85,81%	7,43%	5,41%	0,72%	0,37%	0,26%

Tab. 1: Verwendungsweisen von *können* im Chat-Korpus

5. *namd, nabend, nabends*: Varianten einer Begrüßungsformel

Das folgende Beispiel zeigt die Variation einer Begrüßungsformel und die zunehmende Häufigkeit einer bestimmten innovativen Variante in den Chattexten. Die univerbierte Form *nabend* ist in allen Chatrooms belegt und steigt in „30-40-50-Chat-Set" zwischen 2004 und 2009 stark an. Auch die noch stärker reduzierte Form *namd* kommt in allen Räumen außer „Pälzer unner sich" vor, ist allerdings in den „30-40-50-Chat-Set"-Texten von 2009 besonders häufig. Interessant ist die Form *nabends*, deren Entwicklung in Abbildung 1 dargestellt ist.

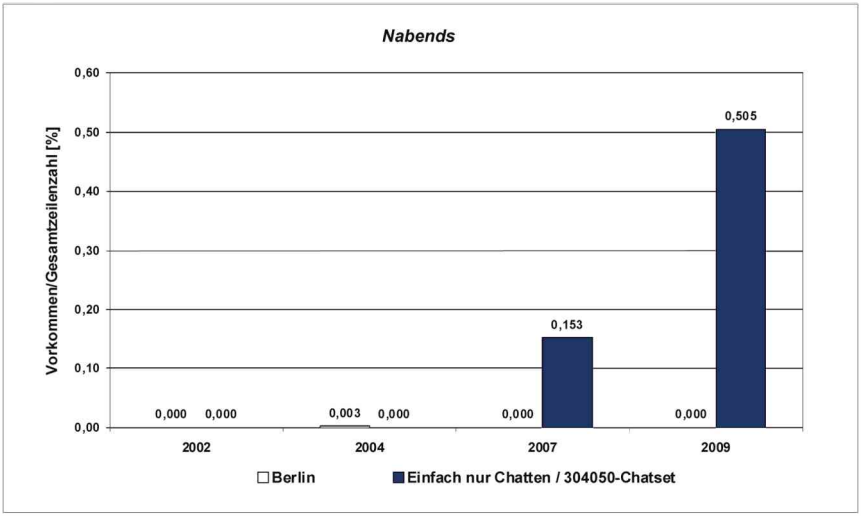


Abb. 1: Entwicklung der Bergrüßungsformel *nabends*

Auch *nabends* ist ein Fall von Univerbierung mit *n-* als proklitischem Element. Die besondere Innovativität dieser Formenbildung liegt in dem Element *-s*, das hier nicht als Morphem zur Bildung eines Adverbs interpretiert werden kann, da *nabends* ausschließlich – und nicht im idiosynkratischen Gebrauch nur einer Chatterin oder eines Chatters – als Begrüßungsformel belegt ist. Die zunehmende Vorkommenshäufigkeit lässt sich als Aufbau einer chatinternen Konvention deuten, mithin als eine Mode 'in statu nascendi'. Stichprobenartige Suchabfragen mit *Google* haben gezeigt, dass diese Form auch in anderen Internetforen oder Chats vorkommt, allerdings steht eine systematische Vergleichsuntersuchung noch aus.

6. Aufbau und Auflösung chatinterner Konventionen: Wurzelverben

Die folgenden Beispiele zeigen den Abbau bzw. Aufbau chatinterner Konventionen unmittelbar am Vorkommen einzelner Verben in unflektierter Form. Die in der Literatur als 'Inflektive' oder 'Wurzelverben' bezeichneten Formen dienen dem Aufbau eines virtuellen oder fiktiven Szenarios, sie bezeichnen Handlungen, die nur virtuell ausgeführt und im Rahmen einer Chatkommunikation imaginiert werden. Sie konkurrieren mit den so genannten Emoticons, ikonischen Zeichen, die aus Interpunktionszeichen (Doppelpunkt, Klammer etc.) zusammengesetzt oder als grafische Animationen in den Text eingefügt werden. Sowohl die Emoticons als auch die Wurzelverben gelten als charakteristisch für Chatkommunikation. Syntagmen mit Wurzelverben wurden im Chat über einen längeren Zeitraum hinweg zunehmend konventionalisiert verwendet und konnten der Schreiberin oder dem Schreiber als Ausweis der eigenen Kompetenz im Umgang mit Chattexten dienen (vgl. Henn-Memmesheimer 2004). In manchen Fällen kann durch den Vergleich über die vier zur Verfügung stehenden Jahrgänge eine zunehmende Anpassung an Standardformen beobachtet werden, d.h. die Verbformen mit Flexionsendung nehmen zu. Für alle untersuchten Verben gilt, dass ihr Gebrauch über die Jahre alles andere als stabil ist. Die Beispiele zeigen, dass die explizite Nennung von Handlungen zeitlich variiert. Exemplarisch sei die außergewöhnliche Entwicklung von *wink* (Abb. 2) dargestellt.[3] Im Fall von *wink* ist – anders als bei den meisten vergleichbaren Verben – die Form mit einer 3. Pers. Sg.-Markierung sehr viel häufiger als die Wurzelform. Dies erklärt sich aus der stark kon-

[3] Untersucht wurden außerdem *kicher, knuddel, knutsch, lach, nick, schmunzel* und *zwinker*.

ventionalisierten Abschiedsformel *winkt zum Abschied*, die unmittelbar vor Verlassen des Chatrooms geschrieben wird. Zwischen 2007 und 2009 sinkt der Gebrauch allerdings rapide ab.

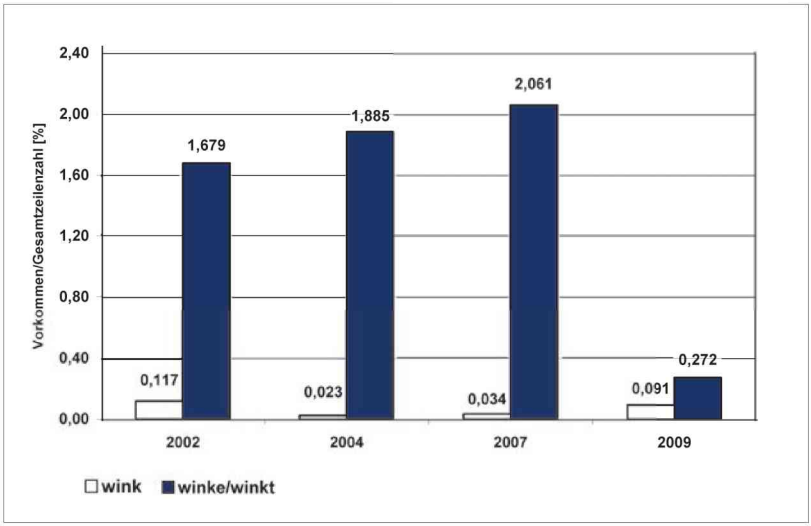


Abb. 2: Entwicklung von *wink-*

Zu interpretieren wäre dies als Abbau (oder in anderen Fällen als Aufbau) von Konventionen, von Moden. Wir haben es offenbar mit kulturellen Zeichen zu tun, die wie andere, nicht-sprachliche Zeichen Moden unterworfen sind. Sie sind als Elemente offener Codes zu interpretieren, die sich im Gebrauch verändern.

7. Ausblick

Geplant ist die Untersuchung weiterer Verben wie *brauchen*, *machen*, *geben* oder *tun*. Eventuelle innovative Verwendungen jenseits des in Standardgrammatiken beschriebenen Gebrauchs sollen vor dem Hintergrund ihrer Entstehungskontexte und ihrer Ausbreitung beschrieben werden. Theoretisch und systembezogen soll der Frage nachgegangen werden, ob solche Innovationen als Ausgangspunkt von Grammatikalisierungen bzw. ihre Diffusion als Grammatikalisierungspfade beschrieben werden können.

Ein weiteres Beispiel für Innovationen im Nonstandard-Raum ist der Genussynkretismus beim Indefinitartikel *nen*. Androutsopoulos (1998: 275f.) beschreibt den Kasuszusammenfall bei *nen*, im Chatkorpus finden sich Belege wie z.B. (8):

(8) zimtsternchen: aber, du bist nen girl (Berlin-Brandenburg, 01.04. 2002)

Schließlich sollen Vergleichsanalysen mit anderen Korpora durchgeführt werden, auf deren Basis eine qualitative Analyse sprachlicher Moden und ihrer Kontexte erfolgen kann.

Literatur

Androutsopoulos, Jannis K. (1998): Deutsche Jugendsprache. Untersuchungen zu ihren Strukturen und Funktionen. Frankfurt a.M.: Lang.

Cherubim, Dieter (1980): Abweichung und Sprachwandel. In: Cherubim, Dieter (Hg.): Fehlerlinguistik. Beiträge zum Problem der sprachlichen Abweichung. Tübingen: Niemeyer, 124-152.

Girnth, Heiko (2000): Untersuchungen zur Theorie der Grammatikalisierung am Beispiel des Westmitteldeutschen. Tübingen: Niemeyer.

Henn-Memmesheimer, Beate (2004): Syntaktische Minimalformen: Grammatikalisierungen in einer medialen Nische. In: Patocka, Franz / Wiesinger, Peter (Hg.): Morphologie und Syntax deutscher Dialekte und Historische Dialektologie des Deutschen. Beiträge zum 1. Kongress der Internationalen Gesellschaft für Dialektologie des Deutschen, Marburg / Lahn, 5.-8. März 2003. Wien: Edition Präsens, 84-118.

Herberg, Dieter / Kinne, Michael / Steffens, Doris (Hg.) (2004): Neuer Wortschatz. Neologismen der 90er Jahre im Deutschen. Berlin / New York: de Gruyter.

Schulz, Klaus-Peter et al. (2000): Wie lernt man Innovationen zu managen? Die Realisierung eines Lehrkonzepts für Innovationsmanagement. In: io management 11: 56-65.

Weinrich, Harald (2003): Textgrammatik der deutschen Sprache. 2. Aufl. Hildesheim u.a.: Olms.

Zifonun, Gisela / Hoffmann, Ludger / Strecker, Bruno et al. (1997): Grammatik der deutschen Sprache. Bd. 2. (Schriften des Instituts für deutsche Sprache 7.2). Berlin / New York: de Gruyter.

EVA BREINDL / MAIK WALTER

Kausalverknüpfungen im Deutschen

Eine korpusbasierte Studie zum Zusammenspiel von Konnektorbedeutung, Kontexteigenschaften und Diskursrelationen

Abstract

Der Beitrag dokumentiert die Ergebnisse einer korpusbasierten Studie zum Gebrauch von 16 Kausalmarkern des Deutschen (*also*, *aufgrund*, *da*, *daher*, *darum*, *denn*, *deshalb*, *deswegen*, *folglich*, *nämlich*, *schließlich*, *sodass*, *wegen*, *weil*, *weshalb*, *weswegen*), zeigt Korrelationen mit Kontextmerkmalen und interpretiert die Ergebnisse auf der Folie von Diskurstheorien.

1. Gegenstand und Zielsetzung

Das Deutsche verfügt wie die meisten Sprachen über eine ganze Reihe sprachlicher Mittel, um einen Kausalzusammenhang zwischen zwei Sachverhalten auszudrücken. Neben Ausdrücken des Nennwortschatzes wie *Ursache*, *Folge*, *bedingen*, *auslösen*, *mit sich bringen* usw. spielen hier Ausdrücke des grammatischen Wortschatzes, nämlich Konnektoren, die wichtigste Rolle. Deren semantische Gemeinsamkeit lässt sich als spezifische, eben „kausale" Zuweisung der semantischen Rollen ANTEZEDENS und KONSEQUENS zu den syntaktisch definierten Relata des Konnektors, nämlich internes und externes Konnekt, modellieren (in Anlehnung an das Konzept der Abbildung von Partizipantenrollen wie THEMA, AGENS, PATIENS etc. an die Komplemente des Verbs im einfachen Satz). Es ergeben sich dann zwei semantische Subklassen kausaler Konnektoren: „ANTEZEDENS-Marker" wie *weil*, *denn*, *nämlich*, *wegen*, bei denen der Konnektor sein internes Konnekt als ANTEZEDENS markiert – sie werden oft als „kausal im engeren Sinn" bezeichnet – und „KONSEQUENS-Marker" wie *sodass*, *weshalb*, *darum* und *deshalb*, bei denen der Konnektor sein internes Konnekt als KONSEQUENS-Argument markiert, traditionell konsekutiv genannt.

Jenseits dieses die semantische Klasse konstituierenden Musters der Rollenzuweisung unterscheiden sich Kausalkonnektoren jedoch beträchtlich – und nicht nur in ihren syntaktischen Eigenschaften. Auch bei Anpassung der syntaktischen Struktur ihrer Konnekte lassen sich Kausalkonnektoren nicht beliebig wechselseitig ersetzen, wie die folgenden Beispiele zeigen.

(1a) **Da** die Schneeschmelze einen überdurchschnittlichen Wasserabfluss zur Folge haben wird, könnten in den kommenden Monaten intensive Regenfälle weitere Hochwasser auslösen. (St. Galler Tagblatt, 19.5.1999)

(1b) In den kommenden Monaten könnten Regenfälle weitere Hochwasser auslösen, **denn schließlich** wird die Schneeschmelze einen überdurchschnittlichen Wasserabfluss zur Folge haben.

(2a) Anwendungsbereiche haben im modernen Verständnis von Naturwissenschaften eine sehr hohe Bedeutung. **Denn schließlich** soll das erworbene Wissen in Lebenssituationen außerhalb der Schule genutzt werden können. (http://bundesumweltministerium.de/files/pdfs/allgemein/application/pdf/wasser)

(2b) #Das erworbene Wissen soll in Lebenssituationen außerhalb der Schule genutzt werden können, **sodass** Anwendungsbereiche im modernen Verständnis von Naturwissenschaften eine sehr hohe Bedeutung haben.

In (1b) scheint gegenüber (1a) die Darstellung des Kausalzusammenhangs stärker von einer objektiven, nachrichtentauglichen Berichterstattung in Richtung einer aus der subjektiven Sprecherperspektive getragenen Begründung verschoben, die in einer berichtenden Textsorte fehl am Platz erschiene. (2b) wirkt gegenüber der Begründungsrelation in (2a) geradezu semantisch abweichend.

2. Theoretischer Hintergrund

Auf die Frage, was hinter solchen Gebrauchsunterschieden steht, – am prominentesten ist diese bei der Kausaltrias *da*, *denn*, *weil* – haben Grammatiken bis heute keine befriedigende und die linguistische Spezialliteratur keine einheitliche Antwort gegeben; Allgemeinwörterbücher kapitulieren hier ohnehin meist. Die erfolgversprechendsten Beiträge haben in den letzten beiden Jahrzehnten hierzu drei Forschungsstränge geleistet:

1) Das *Handbuch der deutschen Konnektoren* (Pasch et al. 2003) hat mit seinen detaillierten Beschreibungen der Syntax der Konnektoren die Basis für semantische Differenzierungen gelegt, die in zahlreichen Publikationen im

Umfeld des Handbuchs vertieft wurden.[1] Eine entscheidende Rolle spielt dabei die Anbindung informationsstruktureller Unterschiede in den Verknüpfungen an die syntaktische Subklassenzugehörigkeit des Konnektors und die strukturellen Eigenschaften der Konstruktion.

2) Die Theorie, dass Kausalkonnektoren auf unterschiedlichen „Ebenen" bzw. in unterschiedlichen „Domänen" verknüpfen können und dass es diesbezüglich lexikalische Restriktionen gibt. Als prominenteste Vertreterin ist hier Sweetser (1990) zu nennen. Sie unterscheidet eine Ebene der realweltlichen Zusammenhänge (Sachverhaltsebene oder propositionale Ebene), eine epistemische Ebene der Sprecherannahmen und -einstellungen und eine illokutive Ebene. Sweetsers Theorie wurde auch im *Handbuch der deutschen Konnektoren* aufgegriffen und hat sich z.B. in der Beschreibung von *denn* als einem „nicht-propositionalen Konnektor" niedergeschlagen. Sweetser selbst reklamiert für ihre kognitiv begründete Unterscheidung eine sehr allgemeine Erklärungskraft weit jenseits von kausalen Verknüpfungen; in der Literatur wird der Ansatz aber meist mit sogenannten reduktiven Schlüssen wie Beispiel (3) illustriert, das zwei Lesarten hat: eine auf der Sachverhaltsebene, bei der Ilses Abwesenheit die Ursache für Erwins Depressionen ist, und eine epistemische, bei der Ilses Abwesenheit für den Sprecher ein Indiz für seine Schlussfolgerung auf die Annahme ist, dass Erwin Depressionen hat. Mit *denn* ist diese epistemische Lesart immer möglich, mit *weil* wohl nur in einer prosodisch desintegrierten Form mit zwei Intonationsphrasen.

   (3a) Erwin hat wieder Depressionen, weil seine Freundin Ilse nie zuhause ist.

   (3b) Erwin hat wieder Depressionen, denn seine Freundin Ilse ist nie zuhause.

3) Im Rahmen von Theorien der Diskursrelationen wie der RST (Rhetorical Structure Theory, vgl. Mann/Thompson 1988) können Unterschiede zwischen Konnektoren als unterschiedliche Diskursrelationen erfasst werden. Sweetsers drei Ebenen finden sich bei Knott/Sanders (1998) und in weiteren Arbeiten im Umfeld eines Projekts an der Universität Utrecht in dichotomischer Form als Differenzierung „semantischer" von „pragmatischen"

[1] Eine Übersicht über die Publikationen findet sich auf der Projekthomepage: www.ids-mannheim.de/gra/konnektoren/#pubs. Der zweite Handbuchband wird die Ergebnisse zur Semantik der Konnektoren in systematischer Form darstellen (Breindl/Volodina/Waßner i. Vorb.).

Relationen. Im Bereich der kausalen Relationen kommt als Differenzparameter das Merkmal „Volitionalität" hinzu, das grosso modo als Unterschied zwischen Ereignisursachen vs. Handlungsbegründungen beschrieben werden kann.

Der vorliegende Beitrag geht auf eine korpusbasierte Studie am IDS Mannheim zurück (Breindl / Walter 2009),[2] die sich als Versuch einer Validierung zentraler Aussagen der genannten Theorien und ihrer Reichweite verstanden wissen will.

Als Ausgangspunkt für die Studie dienten vier Hypothesen, die wir aus den oben genannten Forschungssträngen abgeleitet haben:

1) Grad der syntaktischen Integration der Konstruktion: Die Wahrscheinlichkeit einer pragmatischen (epistemischen oder illokutiven) Lesart steigt mit dem Grad der Desintegration der verknüpften Sätze.

2) Grad der Satzförmigkeit der Argumente: Je weniger satzförmig die Konnekte einer Kausalverknüpfung kodiert sind, desto geringer ist ihr Spielraum, die Sprechereinstellung ausdrücken zu können und desto unwahrscheinlich ist eine pragmatische Relation.

3) Grad der Subjektivität: Je „subjektiver" eine Kausalverknüpfung formuliert wird, desto wahrscheinlicher ist eine pragmatische, die Sprechereinstellung involvierende Lesart.

4) Thematische Rolle des Subjekts: Agentische Subjekte im KONSEQUENS-Konnekt können ein Hinweis auf das Vorliegen der Relation VOLITIONAL CAUSE sein und machen eine Relation NONVOLITIONAL CAUSE unwahrscheinlich, EXPERIENCER-Subjekte im KONSEQUENS machen – als potenzielle Indikatoren von Sprecherinvolvierung – eine pragmatische Lesart wahrscheinlicher.

Zur Validierung dieser Hypothesen am Korpus wurden definiert:

a) ein Set von vier Diskursrelationen, das sich an Knott / Sanders (1998) anlehnt und über Merkmale definiert ist: Die beiden semantischen Relatio-

[2] Die Studie ist Ergebnis des von der DFG geförderten Gemeinschaftsprojekts „Kausalitätsmarker als Kohärenzmittel und ihre Formalisierung für die automatische Textanalyse" (BR 3463 / 1-1 und STE 733 / 7-1; 1.2) unter Leitung von Eva Breindl (Mannheim) und Manfred Stede (Potsdam). Gegenstand des Projekts war die korpusbasierte Beschreibung kausaler Verknüpfungen als einer Schnittstelle zwischen Grammatik und Diskurs und die Formalisierung der Ergebnisse in Hinblick auf automatische Analyseverfahren.

nen VOLITIONAL CAUSE und NONVOLITIONAL CAUSE, die pragmatische Relation PRAGMATIC CLAIM, die in Ermangelung eines einheitlichen Objektivierungsverfahrens in PRAGMATIC CLAIM I (Begründung von Annahmen) und PRAGMATIC CLAIM II (Begründung von Sprechakten) differenziert wurde, und die bezüglich semantisch vs. pragmatisch unspezifische Relation PURPOSE;

b) ein Set von insgesamt 13 formalen und 5 funktionalen Merkmalen, denen wir potenziell diagnostische Kraft unterstellten, und zwar in der folgenden Zuordnung:[3]

- Grundkorrelation (1): als diagnostisch unterstellten wir jeweils eine bestimmte Ausprägung der Merkmale INTEGRATION, POSITION, KORRELATKONSTRUKTION, MEHRFACHES VORKOMMEN VON MARKERN, LINEARISIERUNG, VERBSTELLUNG, SATZMODUS;
- Grundkorrelation (2): als diagnostisch unterstellten wir jeweils eine bestimmte Ausprägung der Merkmale SATZFORM, UMFANG des internen Konnekts von Präpositionen;
- Grundkorrelation (3): als diagnostisch für eine pragmatische Lesart unterstellten wir die positive Belegung der Merkmale NONFAKT IM KONSEQUENS, KONJUNKTIV, FREMDPERSPEKTIVE, 1. PERSON, WERTENDER AUSDRUCK IM KONSEQUENS, SPRECHEREINSTELLUNG;
- Grundkorrelation (4): als Hinweis auf das Vorliegen der Diskursrelation VOLITIONAL CAUSE testeten wir AGENS IM KONSEQUENS; als Hinweis für das Vorliegen der Diskursrelation PRAGMATIC CLAIM testeten wir EXPERIENCER IM KONSEQUENS.

3. Die Durchführung der Korpusstudie

Das Deutsche Referenzkorpus DeReKo wurde für die Studie herangezogen.[4] Es enthält 7 022 872 Texte mit 1 828 805 828 Wörtern in 2 086 Dokumenten und besteht aus Zeitungen, Sach- und Fachtexten sowie belletristischer Literatur aus Deutschland, Österreich und der Schweiz von den Jahren 1772 bis 2008. Einen Schwerpunkt bilden Zeitungstexte aus den letzten beiden Dekaden.

[3] Die Diskursrelationen werden ausführlich erläutert in Breindl / Walter (2009: 89-97), die formalen und funktionalen Merkmale ebd. (52-88).

[4] Das Korpus ist online verfügbar: www.ids-mannheim.de/kl/projekte/korpora/.

Für 16 Kausalmarker wurde mit Hilfe der Funktion 'Zufallsauswahl' des Korpusrecherche- und -analysesystems COSMAS II eine Stichprobe mit je 200 Belegen im Kontext von je zwei Vorgänger- und Nachfolgesätzen zusammengestellt. Nicht kausale und nicht analysierbare Fälle wurden im Anschluss ausgefiltert. 75 % der Belege (N = 2410) konnten ausgewertet werden. Dazu bestimmten wir zunächst das interne und das externe Konnekt (bzw. ANTEZEDENS und KONSEQUENS) und annotierten jeden Beleg nach den bereits angeführten formalen und funktionalen Merkmalen. Daraufhin bestimmten wir mit Hilfe von Paraphrasen holistisch die Diskursrelationen und ermittelten abschließend auf der Basis der oben dargestellten Hypothesen signifikante Korrelationen zwischen den Kausalmarkern, den Diskursrelationen und ausgewählten Merkmalen (vgl. Breindl / Walter 2009: 43-51).

4. Ergebnisse

In Bezug auf die **Korrelation zwischen Konnektoren und Diskursrelationen** ließen sich zwar Präferenzen, aber in keinem Fall eine 1:1-Zuordnung ableiten; Anders als in der Literatur häufig behauptet, ist also kein Kausalkonnektor lexikalisch auf eine bestimmte Relation oder Ebene festgelegt (siehe Abb. 1).

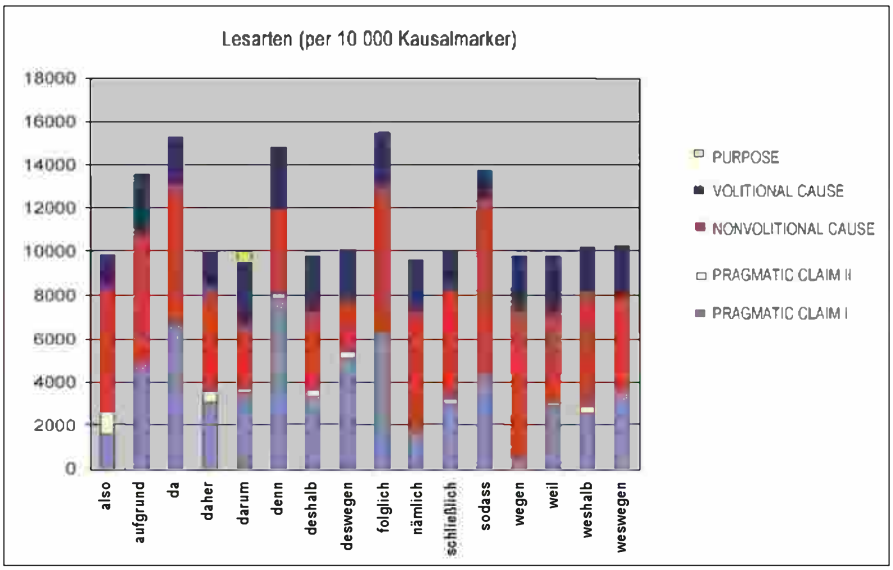


Abb. 1: Korrelation zwischen Konnektoren und Diskursrelationen[5]

[5] Die hohe Zahl von Lesarten ergibt sich durch Berücksichtigung von Fällen, die als ambig gewertet wurden.

Zur Ermittlung der Präferenzen wurde in Breindl / Walter (2009) ein dreistufiges Filterverfahren vorgeschlagen (siehe Tab. 1). Im ersten Schritt wurden für jede Relation die drei häufigsten Konnektoren bestimmt (TOP3-Analyse). Danach ließen sich Konnektoren mit einer Präferenz für den semantischen Relationstyp (*darum*, *aufgrund*, *sodass*, *wegen*) von solchen mit Präferenz für den pragmatischen Relationstyp (*da*, *also*, *daher*) unterscheiden. Die drei Konnektoren *denn*, *folglich*, *deshalb* waren zwar häufig, traten aber in beiden Relationstypen auf, präferierten also keinen der beiden Typen. Eindeutige Vertreter wurden fett markiert und in Tabelle 1 in einer eigenen Zeile zusammengefasst. Im zweiten Schritt wurde jeder Konnektor danach ausgewertet, ob er häufiger in der Diskursrelation PRAGMATIC CLAIM I oder aber NONVOLITIONAL CAUSE erscheint (diese beiden Relationen stellen den Löwenanteil der Belege). Im dritten Schritt wurden die Ergebnisse der beiden Analysen kombiniert. Die Tabelle fasst die Ergebnisse zusammen: Nur *da* präferiert die pragmatische Relation, nämlich PRAGMATIC CLAIM I. Die semantische Verknüpfung hingegen präferieren die Kausalmarker *sodass*, *wegen* und *aufgrund*, wobei lediglich *sodass* und *wegen* eine bestimmte Relation, nämlich NONVOLITIONAL CAUSE, präferieren (vgl. Stede / Walter i.Dr.).

		<PRAGMATIC>	**<SEMANTIC>**
1	TOP-3-Konnektoren pro Relation	<PRAGMATIC CLAIM I> *denn*, ***da***, *folglich*	<VOLITIONAL CAUSE> ***darum***, *denn*, ***aufgrund***
		<PRAGMATIC CLAIM II> ***also***, ***daher***, *deshalb*	<NONVOLITIONAL CAUSE> ***sodass***, ***wegen***, *folglich*
			<PURPOSE> ***darum***, *deshalb*, ***wegen***
	Konnektoren mit Präferenz für einen Relationstyp	*da*, *also*, *daher*	*darum*, *aufgrund*, *sodass*, *wegen*
2	Präferenz bezogen auf Diskursrelationen	<PRAGMATIC CLAIM I> vor <NONVOLITIONAL CAUSE> *denn*, *da*, *deswegen*, *darum*	<NONVOLITIONAL CAUSE> vor <PRAGMATIC CLAIM I> *deshalb*, *weil*, *weswegen*, *daher*, *schließlich*, *weshalb*, *nämlich**, *also*, *aufgrund*, *folglich*, *wegen**, *sodass*
3	Konnektoren mit eindeutiger Präferenz	***da***	***sodass***, ***wegen***, (***aufgrund***)

Tab. 1: Das dreistufige Filterverfahren bei der Ermittlung der Relationsspezifik von Konnektoren

Als noch weniger gebrauchsdifferenzierend erwies sich das Merkmal 'Volitionalität'. Anders als etwa im Niederländischen (vgl. Stukker 2005) erwies sich im Deutschen mit Ausnahme des präferent VOLITIONAL CAUSE-kodierenden *sodass* keiner der Marker in dieser Hinsicht spezifisch. Es zeigte sich hier aber eine signifikante **Korrelation zwischen funktionalen Kontextmerkmalen und Relationen**: VOLITIONAL-CAUSE-Verknüpfungen haben fast durchweg ein AGENS-Subjekt im KONSEQUENS-Argument. Signifikante Zusammenhänge zeigten sich auch zwischen bewertenden Aussagen im ANTEZEDENS und pragmatischen Relationen sowie zwischen nonfaktischen KONSEQUENS-Konnekten und einer Klassifikation als PRAGMATIC CLAIM (siehe Abb. 2).

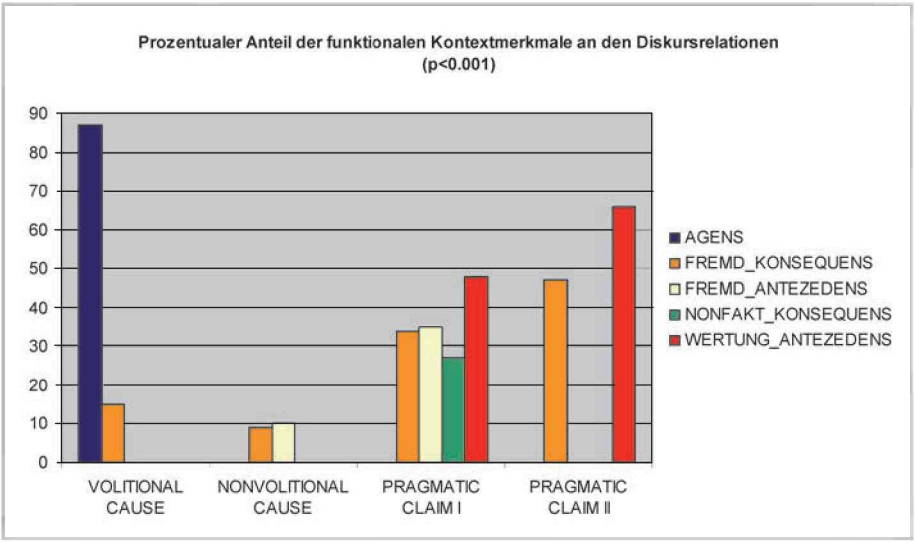


Abb. 2: **Hochsignifikante Korrelationen zwischen einigen funktionalen Merkmalen** (AGENS, FREMDPERSPEKTIVE IM KONSEQUENS, FREMDPERSPEKTIVE IM ANTEZEDENS, NONFAKTISCHES KONSEQUENS, WERTENDE ELEMENTE IM ANTEZEDENS) **und Diskursrelationen**

Bei den hochsignifikanten Korrelationen zwischen formalen Merkmalen und Diskursrelationen fällt auf, dass – mit Ausnahme der als Indikator für semantische Relationen zu wertenden Korrelatkonstruktion – eher pragmatische Relationen eine besondere formale Ausprägung erhalten. Signifikante Korrelationen gibt es zum Vorliegen von Sprecherrolle oder Ausdrücken der Sprechereinstellung im KONSEQUENS oder zu Desintegrationskonstruktionen (siehe Abb. 3).

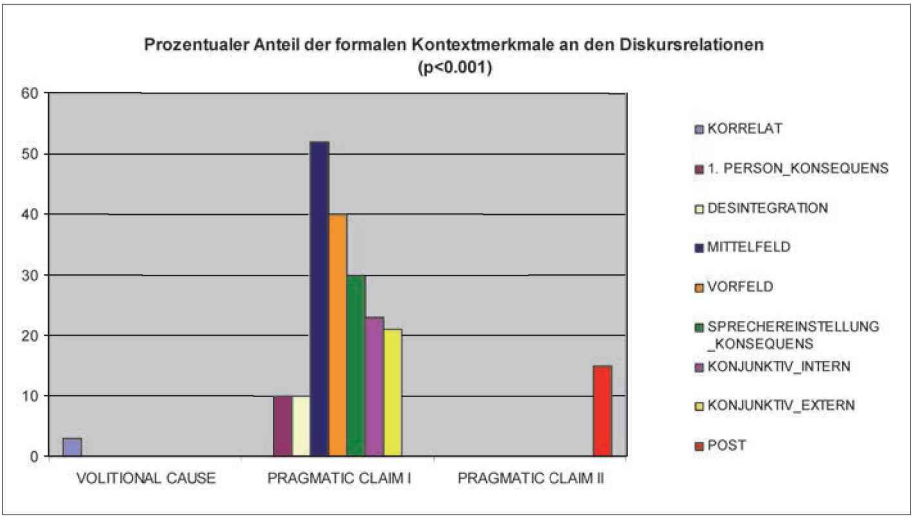


Abb. 3: Hochsignifikante Korrelationen zwischen formalen Merkmalen (KORRELAT, AUSDRUCK DER 1. PERSON IM KONSEQUENS, DESINTEGRATION, AUFTRETEN DES KONNEKTORS IM MITTELFELD ODER VORFELD, AUSDRUCK DER SPRECHEREINSTELLUNG IM KONSEQUENS, KONJUNKTIV IM INTERNEN BZW. EXTERNEN KONNEKT, POSTPONIERTES INTERNES KONNEKT) **und Diskursrelationen**

Literatur

Breindl, Eva / Walter, Maik (2009): Der Ausdruck von Kausalität im Deutschen. Eine korpusbasierte Studie zum Zusammenspiel von Konnektoren, Kontextmerkmalen und Diskursrelationen. (= amades 38). Mannheim: Institut für Deutsche Sprache.

Breindl, Eva / Volodina, Anna / Waßner, Ulrich H. (i.Vorb.): Handbuch der deutschen Konnektoren. Teil 2: Semantik. Berlin / New York: de Gruyter.

Knott, Alistair / Sanders, Ted (1998): The classification of coherence relations and their linguistic markers: An exploration of two languages. In: Journal of Pragmatics 30: 135-175.

Mann, William / Thompson, Sandra (1988): Rhetorical structure theory: Towards a functional theory of text organisation. In: Text 8: 243-281.

Pasch, Renate / Brauße, Ursula / Breindl, Eva / Waßner, Ulrich H. (2003): Handbuch der deutschen Konnektoren. Linguistische Grundlagen der Beschreibung und syntaktische Merkmale der deutschen Satzverknüpfer (Konjunktionen, Satzadverbien und Partikeln). (= Schriften des Instituts für Deutsche Sprache 9). Berlin / New York: de Gruyter.

Stede, Manfred / Walter, Maik (i.Dr.): Zur Rolle der Verknüpfungsebenen am Beispiel der Kausalkonnektoren. In: Breindl, Eva / Ferraresi, Gisella / Volodina, Anna (Hg.): Satzverknüpfungen: Zur Interaktion von Form, Bedeutung und Diskursfunktion. (= Linguistische Arbeiten 534). Berlin / New York: de Gruyter

Stukker, Ninke (2005): Causality marking across levels of language structure: A cognitive semantic analysis of causal verbs and causal connectives in Dutch. Utrecht: LOT.

Sweetser, Eve E. (1990): From etymology to pragmatics. Metaphorical and cultural aspects of semantic structure. Cambridge u.a.: Cambridge University Press.

MANFRED STEDE / UWE KÜSSNER

Kausale Konnektoren in der Automatischen Textanalyse[1]

Abstract

Wir schlagen vor, als Teilproblem des „rhetorischen Parsing" – d.h. der Zuordnung einer Baumstruktur aus Diskursrelationen zu einem Text – die Aufgabe der „lokalen Kohärenzanalyse" systematisch zu untersuchen: die Identifikation eines Konnektors im Text, den durch ihn verbundenen Textsegmenten und die Zuweisung einer einzelnen Diskursrelation. Wir beschreiben zunächst eine als Grundlage durchgeführte Korpusstudie, dann den Aufbau unseres Konnektorenlexikons und schließlich unser Verfahren der automatischen lokalen Kohärenzanalyse.

1. „Rhetorisches Parsing"

Eine auch in der Computerlinguistik seit nunmehr 20 Jahren recht populäre Theorie zur Beschreibung der Struktur von Texten ist die Rhetorical Structure Theory (RST; Mann / Thompson 1988). Sie geht davon aus, dass ein Text sich in elementare Diskurseinheiten zerlegen lässt, die über sog. Diskursrelationen zueinander in Beziehung gesetzt werden: Das Inventar dieser (bei Mann / Thompson etwa 20) Relationen umfasst verschiedene kausale, temporale und adversative Relationen, aber auch 'inhaltsärmere' (semantisch weniger belegte) Relationen wie Elaboration oder Background. Die RST postuliert, dass durch eine solche Relation einerseits benachbarte minimale Diskurseinheiten, dann aber rekursiv auch größere Abschnitte verbunden werden können (siehe dazu auch Stede 2007, Kap. 7). Dabei dürfen keine Segmente unverbunden bleiben und keine „kreuzenden Kanten" entstehen, wodurch insgesamt als Deskription eines kohärenten Textes eine Baumstruktur entstehen soll.

Erste Ansätze zur Automatisierung der RST-Analyse wurden etwa von Corston-Oliver (1998) und Marcu (2000) vorgelegt. Das Thema ist aber bis heute populär, eine aktuelle Arbeit zum „Rhetorischen Parsing" durch maschinelles

[1] Die hier vorgestellte Arbeit basiert auf Ergebnissen des DFG-Projekts „Kausalitätsmarker als Kohärenzmittel und ihre Formalisierung für die Automatische Textanalyse" (2006-2008), einem Gemeinschaftsprojekt zwischen dem IDS Mannheim (Dr. Eva Breindl) und der Universität Potsdam (Prof. Dr. Manfred Stede).

Lernen aus entsprechend annotierten Korpora wurde von duVerle / Prendinger (2009) präsentiert. Grundsätzlich gilt für diese Arbeiten, dass ein inhaltliches 'Verständnis' des Textes in keiner Weise angestrebt wird, sondern der Versuch unternommen wird, durch explizite Signale an der Textoberfläche den RST-Baum zumindest teilweise zu bestimmen. Die wichtigsten Signale sind naturgemäß Konnektoren; es gibt aber auch Versuche, beispielsweise eine adversative Relation anhand der Inhaltswörter in den Segmenten zu 'erraten', wiederum durch maschinelles Lernen aus Korpora (Soricut / Marcu 2003).

Wir halten die Bearbeitung der Aufgabe des „rhetorischen Parsing" aus zwei Gründen gegenwärtig für nur bedingt aussichtsreich. Zum Einen ist die Aufgabe des Erkennens einer kompletten Textstruktur allein auf der Grundlage von Oberflächensignalen außerordentlich schwierig und sollte u.E. durch geeignete Zerlegung in kleinere Teilaufgaben bearbeitet werden. Zum anderen halten wir eine RST-artige Baumstruktur auch aus textlinguistisch-deskriptiver Sicht nicht für eine optimale Beschreibung; Gründe für diese Zweifel sind in Stede (2008) dokumentiert. Wir schlagen daher vor, zunächst eine kleinere Aufgabe in das Zentrum der Aufmerksamkeit zu rücken: die gründliche Behandlung der einzelnen Verbindungen von Diskurseinheiten auf der Basis von Konnektoren.

2. Lokale Kohärenzanalyse

Wesentlich bescheidener als das rhetorische Parsing konzentriert sich die lokale Kohärenzanalyse auf die einzelnen 'Verbindungsstellen' zwischen Diskurssegmenten. Zudem berücksichtigen wir nur solche Verbindungen, die durch einen Konnektor explizit markiert sind. Die Aufgabe besteht dann aus drei Teilschritten:

1) Konnektoren identifizieren
2) Segmente identifizieren und den Konnektoren zuordnen
3) (ggf.) Lesart des Konnektors bestimmen

Wir verdeutlichen die Aufgabe anhand des in Abbildung 1 gezeigten Beispieltexts. Kursiv sind diejenigen Wörter gesetzt, die potenziell Konnektoren sein können – aber nicht müssen. Dipper / Stede (2006) haben gezeigt, dass von den von ihnen untersuchten 135 häufigen Konnektoren des Deutschen etwa ein Drittel auch eine Nicht-Konnektor-Lesart haben (vor allem als Abtönungsparti-

kel oder als lokatives / temporales Adverb). Im Beispieltext sind *denn* und *doch* zwei Wörter, die eine automatische Analyse zunächst als potenziellen Konnektor identifizieren würde, dann aber wieder ausscheiden müsste: *denn* ist hier Fragepartikel, *doch* eher Abtönungspartikel als kontrastiver Konnektor.

> Nein, ich kannte sie nicht, der gemeinsame Geburtsort ist purer Zufall. Ich wurde jetzt *auch* von Grünen-Politikern gefragt: Wie sieht das *denn* aus, *wenn* zwei Migrantinnen aus der Türkei jeweils in die andere Partei wechseln, das macht euch *doch* total unglaubwürdig. *Aber* es geht hier nicht um persönliche Rivalitäten zwischen zwei Türkinnen, das ist *schließlich* keine Provinzposse. Das muss allen mal klar werden.
>
> (Bilkay Öney, *Tagesspiegel*-Interview 13.5.09. Hervorhebung durch die Verf.)

Abb. 1: Beispieltext für die lokale Kohärenzanalyse

Der zweite Schritt besteht in der Identifikation der von den Konnektoren jeweils verbundenen Segmente und der Zuordnung der jeweiligen Rolle (deren Bezeichnung von der Diskursrelation abhängt). Für das *wenn* im Beispieltext müsste als Antezedens *zwei Migrantinnen aus der Türkei jeweils in die andere Partei wechseln* und als Konsequens *Wie sieht das denn aus* markiert werden. Diese Aufgabe ist vor allem für adverbiale Konnektoren häufig schwierig (nicht nur für Computer, sondern auch für Menschen).

Im dritten Schritt schließlich muss für jeden identifizierten Konnektor seine semantisch / pragmatische Lesart bestimmt werden. Die ist für eindeutige Konnektoren wie *obwohl* trivial, für einen unterbestimmten Konnektor wie *und* natürlich schwierig. Einige Konnektoren sind ambig zwischen gut voneinander unterscheidbaren Lesarten – wie in unserem Beispiel *schließlich*: es kann temporal oder argumentativ gelesen werden, im Beispieltext ist die letztere Lesart richtig.

3. Korpus-Annotation: Kausale Konnektoren und Segmente

Nachdem wir die Aufgabe der lokalen Kohärenzanalyse zunächst allgemein beschrieben haben, konzentrieren wir uns im Folgenden auf die Gruppe der kausalen Konnektoren, die in unserem Projekt im Vordergrund des Interesses stand. Um eine Datengrundlage zu erhalten, auf deren Basis die Implementierung der automatischen Kohärenzanalyse (siehe Kap. 5) vorgenommen wer-

den kann, haben wir in einem Korpus von Texten sämtliche Kausalkonnektoren und die verbundenen Segmente manuell annotiert. Die Daten wurden einer Produktbesprechungs-Website (doyoo.de) entnommen und enthalten insbesondere Erfahrungsberichte und Bewertungen von Hotels. Dieses Genre versprach eine relativ hohe Frequenz von Kausalkonnektoren, sowohl in eher semantischen (Reisende beschreiben, was sie aus welchen Gründen getan haben) als auch in eher pragmatischen (Reisende argumentieren für ihre Meinungen / Empfehlungen) Lesarten. Annotiert wurden insgesamt 200 Texte, indem zunächst automatisch alle potenziellen Kausalkonnektoren vormarkiert wurden; dies geschah aufgrund einer von unseren Projektpartnern im IDS Mannheim erstellten Liste von etwa 70 Konnektoren. Es wurden dann manuell diejenigen Wörter wieder ausgesondert, die nicht als Konnektor verwendet wurden (s.o.). Die im Korpus verwendeten, mithin annotierten, kausalen Konnektoren sind: *also*, *aufgrund*, *da*, *dadurch*, *daher*, *damit*, *darum*, *dementsprechend*, *demnach*, *denn*, *deshalb*, *deswegen*, *doch*, *durch*, *halber*, *nämlich*, *schließlich*, *so*, *somit*, *so dass*, *so* X *dass*, *um*, *wegen*, *weil*, *weshalb*.

Anschließend erfolgte die Annotation der Segmente mit dem Software-Werkzeug *MMAX2* (mmax2.sourceforge.net). Des weiteren wurden anhand ausführlicher Richtlinien auch der illokutionäre Status der Segmente sowie etwaige Fokuspartikeln annotiert; dies spielt für den hier beschriebenen Teil des Vorhabens allerdings keine Rolle. Insgesamt wurden ca. 1 100 Konnektoren mit 2 300 Segmenten annotiert; das zugrunde liegende Schema ist in Peldszus et al. (2008) genauer dokumentiert. Zur Illustration hier ein Beispiel, welches auch zeigt, dass mitunter mehrere Gründe zu einer Folge in Beziehung zu setzen sind (auch der umgekehrte Fall kann auftreten):

[Wir buchten All Inklusive]$_{\text{Folge}}$ **da** [wir uns keinen Kopf ums Geld machen wollten]$_{\text{Grund1}}$ und **da** [es nicht viel mehr kostete]$_{\text{Grund2}}$.

Abb. 2: Beispiel zur Analyse des Hotelbewertung-Korpus (Peldszus et al. 2008)

4. Das Konnektorenlexikon

Die Grundidee des Ansatzes ist, als zentrale Wissensquelle ein Lexikon der kausalen Konnektoren zu verwenden, in dem in deklarativer Form alle Informationen vorgehalten werden, die für die automatische Analyse zu einem bestimmten Konnektor relevant sind. Dies geht auf ältere Arbeiten am „Diskursmarker-Lexikon (DiMLex)“ zurück (Stede / Umbach 1998), das nun speziell für die Kausalkonnektoren erheblich ausgebaut wurde.

Zur syntaktischen Information legen wir als theorieneutrale Repräsentation Merkmale über das Auftreten in topologischen Feldern gemäß Pasch et al. (2003) ab; diese Repräsentation kann dann – manuell oder automatisch – in Suchmuster im Format eines bestimmten syntaktischen Parsers übersetzt werden. Des weiteren enthält ein Eintrag für ein ambiges Wort gewichtete Desambiguierungsregeln in Form regulärer Ausdrücke über Lexemen und Wortarten im Kontext, im Anschluss an die Arbeiten von Dipper / Stede (2006). Die Gewichtungen wurden zum Teil anhand der o.g. Korpusstudie vorgenommen, zum Teil auch anhand der empirischen Studie von Breindl / Walter (2009). Sie enthält auch einige Beobachtungen zur Unterscheidung zwischen semantischen und pragmatischen Lesarten bestimmter Konnektoren, die wir wiederum in den Desambiguierungsregeln für Diskursrelationen verwenden. Diese Regeln verwenden syntaktische Merkmale der an der Konnexion beteiligten Sätze. Schließlich enthalten die Lexikoneinträge noch verschiedene semantische und pragmatische Merkmale (Kookkurrenz mit Fokuspartikeln, Stilebene etc.), die für die hier beschriebene Arbeit aber keine Rolle spielen.

5. Implementierung der lokalen Kohärenzanalyse

Um eine automatische lokale Kohärenzanalyse zu betreiben, greifen wir auf zwei Vorverarbeitungsschritte zurück: Die Zuweisung von Wortarten (part-of-speech tags) zu den Wörtern sowie die Ergebnisse eines Dependenzparsings für die einzelnen Sätze, wofür wir auf den *Connexor* Parser (www.connexor.com) zurückgreifen. Die Verknüpfung dieser Module ist innerhalb unserer Dokumentverarbeitungs-Werkbank *MOTS* relativ komfortabel möglich. Das Modul für die lokale Kohärenzanalyse liest zunächst das Konnektorenlexikon ein und bearbeitet dann den Text, indem zunächst (anhand der Lexikoneinträge) alle potenziellen Konnektoren markiert werden. Wenn nötig, werden dann die Desambiguierungsregeln geprüft, um etwaige Nicht-Konnektor-Lesarten zu erkennen. Die Bestimmung der verbundenen Segmente erfolgt anhand der syntaktischen Analyse des Dependenzparsers; sie ist für Präpositionen und Konjunktionen relativ zuverlässig möglich, während für Adverbiale nur 'geraten' werden kann: Die beiden unmittelbar nebeneinander stehenden Sätze (von denen im zweiten der adverbiale Konnektor steht) werden als die Segmente behandelt, doch ob auch weitere Nachbarsätze noch hinzuzufügen wären, lässt sich automatisch nicht feststellen.

Die Implementierung ist in einer ersten Version fertiggestellt und dient derzeit zum Testen und Weiterentwickeln der Desambiguierungsregeln; in naher Zukunft soll dann eine Evaluierung des Moduls vorgenommen werden, und anschließend ist eine Ausweitung auch auf andere (nicht-kausale) Arten von Konnektoren geplant.

Literaturverzeichnis

Breindl, Eva / Walter, Maik (2009): Der Ausdruck von Kausalität im Deutschen. Eine korpusbasierte Studie zum Zusammenspiel von Konnektoren, Kontextmerkmalen und Diskursrelationen. (= amades 38). Mannheim: Institut für Deutsche Sprache.

Corston-Oliver, Simon (1998): Computing representations of the structure of written discourse. Ph.D. diss. Univ. of California, Santa Barbara.

Dipper, Stefanie / Stede, Manfred (2006): Disambiguating potential connectives. In: Online-Proceedings der Konferenz zur Verarbeitung natürlicher Sprache KONVENS-06, Konstanz. Internet: http://ling.uni-konstanz.de/pages/conferences/konvens06/konvens_files/konvens06-proc.pdf (Stand: 11 / 2010).

duVerle, David A. / Prendinger, Helmut (2009): A novel discourse parser based on support vector machine classification. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Singapur, September 2009. Singapur: Association for Computational Linguistics, 665-673. Internet: http://www.aclweb.org/anthology/P/P09/P09-1075 bzw. http://nlp.csie.ncnu.edu.tw/~shin/acl-ijcnlp2009/proceedings/CDROM/ACLIJCNLP/pdf/ACLIJCNLP075.pdf (Stand: 11 / 2010).

Mann, William C. / Thompson, Sandra A. (1988): Rhetorical structure theory: Toward a functional theory of text organization. In: Text 8, 3: 243-281.

Marcu, Daniel (2000): The theory and practice of discourse parsing and summarization. Cambridge, MA: MIT Press.

Pasch, Renate / Brauße, Ursula / Breindl, Eva / Waßner, Ulrich Hermann (2003): Handbuch der deutschen Konnektoren. Linguistische Grundlagen der Beschreibung und syntaktische Merkmale der deutschen Satzverknüpfer (Konjunktionen, Satzadverbien und Partikeln). (= Schriften des Instituts für Deutsche Sprache 9). Berlin / New York: de Gruyter.

Peldszus, Andreas / Herzog, André / Hofmann, Florian / Stede, Manfred (2008): Zur Annotation von kausalen Verknüpfungen in Texten. In: Storrer, Angelika / Geyken, Alexander / Siebert, Alexander / Würzner, Kay-Michael (Hg.): Proceedings der Konferenz zur Verarbeitung natürlicher Sprache KONVENS-08, Ergänzungsband. Berlin, 71-83. Internet: www.ling.uni-potsdam.de/~stede/Forsch/Kausal-IDS/KK/ko08.pdf (Stand: 11 / 2010).

Soricut, Radu / Marcu, Daniel (2003): Sentence level discourse parsing using syntactic and lexical information. In: Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT / NAACL-2003). Edmonton, Canada: Association for Computational Lingusitics 149-156. Internet: http://acl.ldc.upenn.edu/N/N03/N03-1030.pdf (Stand: 11 / 2010).

Stede, Manfred (2007): Korpusgestützte Textanalyse. Tübingen: Narr.

Stede, Manfred (2008): Disambiguating rhetorical structure. In: Journal of Research in Language and Computation 6, 3: 311-332

Stede, Manfred / Umbach, Carla (1998): DiMLex: A lexicon of discourse markers for text generation and understanding. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. Bd. 2. Montréal: Association for Computational Linguistics, 1238-1242. Internet: www.aclweb.org/anthology/P/P98/P98-2202.bib (Stand: 11 / 2010).

JULIA RICHLING

Diachrone Analyse eines Newsgroup / Webforum-Korpus

Abstract

Diskussionsforen weisen eine breite Varianz hinsichtlich der Konzeption ihrer Beiträge auf. Diese können lange Texte oder auch nur eine knappe Bestätigung enthalten und decken damit das gesamte Spektrum zwischen ausformulierten Fachtexten und einzelnen Wortfetzen ab.

Die vorliegende Arbeit untersucht entsprechende Korpora anhand charakteristischer Merkmale und identifiziert Zeit, technische Basis und Diskurssituierung als relevante Einflussfaktoren.

1. Einleitung

Die Analyse internetbasierter Kommunikation (IBK) hat im Rahmen der Untersuchung der Entwicklung sprachlicher Gewohnheiten innerhalb der letzten zehn Jahre an deutlicher Beliebtheit gewonnen. Im Gegensatz zu den in vielen Korpora vertretenen Printmedien werden diese medial schriftlichen Erzeugnisse von der Abwesenheit äußerer Korrekturinstanzen bestimmt. Diese sprachliche Freiheit begünstigt eine große Varianz und Aufgeschlossenheit gegenüber sprachlichen Mitteln, die vom Standardsprachgebrauch abweichen. Im Vergleich zur mündlichen Sprache können die Daten mitunter durch eine noch größere Natürlichkeit bestechen, werden sie doch in einer nicht-experimentellen Umgebung erhoben.

Im Kontrast zu der Euphorie, die man als Sprachwissenschaftler angesichts der sich damit eröffnenden Möglichkeiten empfindet und die selbst vor einer schnellen Recherche mit Hilfe von *Google* zur Überprüfung der Akzeptanz interessanter Konstruktionen nicht Halt macht, steht die Heterogenität der Texte sowie der Umstand, dass die Forschung internetbasierter Kommunikation in ihren Anfängen steckt. Dies wird nicht zuletzt daran deutlich, dass es gerade einmal eine Handvoll frei verfügbarer deutscher IBK-Korpora gibt (Beißwenger / Storrer 2008). Nach Wissen der Autorin ist keines von ihnen zuverlässig wortartenannotiert.

Die vorliegende Arbeit widmet sich der Aufgabe der Beschreibung von IBK mit Hilfe des Modells der konzeptionellen Mündlichkeit / Schriftlichkeit nach Koch / Österreicher (1985). Dieses Modell stellt ein etabliertes Instrument dar, sich der Frage, wie die IBK im Gefüge schriftlicher und mündlicher Kommunikation zu verorten sei, zu nähern. Es trennt die phonische / grafische Realisierung – als mediale Mündlichkeit / Schriftlichkeit bezeichnet – von der Konzeption der sprachlichen Beiträge. Die Ausprägungen dieser Konzeption sind auf einer Skala angesiedelt, die durch die beiden Pole Mündlichkeit und Schriftlichkeit beschränkt werden.

Die Positionierung sprachlicher Beiträge auf dieser Mündlichkeits-Schriftlichkeits-Skala wird nach Burger (2005: 143) von folgenden Variablen bestimmt:

- „formal / informell" (betrifft Kontext und Stil)
- „schriftsprachlich / umgangssprachlich" (betrifft Stil, berührt sich mit arealen Aspekten)
- „spontan / vorbereitet" (betrifft Sprachproduktion)

Der Untersuchung der Netzsprache geht die Feststellung voraus, dass es die Netzsprache nicht gibt, d.h. nicht in der Form eines homogenen Gefüges, über das verallgemeinernde Aussagen getroffen werden können. Dies macht es erforderlich, weitere Kategorien zu finden. So erweitert Dürscheid (2003) das Schema nach Koch / Österreicher um den Aspekt der Synchronität.

Während synchrone und quasi-synchrone Kommunikationsumgebungen durch ihren technischen Funktionsumfang die Gesprächsartigkeit der sprachlichen Äußerungen fördern und asynchrone Kommunikationsplattformen wie Blogs in einer textaffinen Tradition stehen, liegt das Augenmerk dieser Arbeit auf Internetforen, die eine Art Hybrid darstellen.

2. Untersuchungsgegenstand

Bei Internetforen handelt es sich auf technischer Seite um eine asynchrone Kommunikationsform, allerdings deutet bereits die alternative Bezeichnung 'Diskussionsforen' darauf hin, dass ihre Funktion unter anderem im gesprächsartigen Austausch von Informationen und Meinungen verstanden wird. Auf technischer Seite fördert die Darstellungsweise (z.B. in Form einer expliziten Darstellung der baumartigen Struktur aus Beiträgen und deren Antworten) häufig diese Funktion.

Während in einem Chat nur selten bildschirmfüllende Texte gefunden werden können und in einem Blog ein Beitrag, der ausschließlich die Worte *Vielen Dank, jetzt klappt es wirklich!* enthält, sehr ungewöhnlich wäre, sind in Foren beide Extremfälle nicht nur möglich, sondern auch üblich.

Diese große Spanne an möglichen Beitragsgestaltungen erhärtet die Frage, an welchem der beiden Pole im Mündlichkeits-Schriflichkeits-Kontinuum sich die Beiträge im Durchschnitt orientieren und durch welche Faktoren die Konzeption der Beiträge beeinflusst werden können.

2.1 Thesen

Drei potenzielle Faktoren sollen im Folgenden näher betrachtet werden:

Zeit: Wie bereits in der Einführung erwähnt, gilt die IBK als für Sprachwandel besonders empfänglich. Vom Standard abweichende sprachliche Formen werden von Communities und Individuen verwendet, um sich von anderen abzugrenzen, Trends werden gebildet und verworfen und neu entstehende Kommunikationsplattformen fördern das Aufkommen darauf angepasster Konventionen, die wiederum nicht selten in anderen Umgebungen übernommen werden. Zusätzlich dazu hat sich die Nutzerstruktur während der letzten Jahre (siehe ARD-ZDF Onlinestudie – http://www.ard-zdf-onlinestudie.de/) deutlich verändert.

> These 1: Im mikrodiachronen Verlauf werden die Beiträge zunehmend mündlicher und IBK-typischer formuliert werden.

Kommunikationsumgebung: Bei asynchronen Diskussionsforen handelt es sich nicht um eine einheitliche Kommunikationsumgebung, sondern vielmehr um eine Vielzahl davon, die alle einem einheitlichen Zweck dienen. Beispiele für diese Kommunikationsumgebungen sind Newsgroups und Webforen. Newsgroups existierten bereits vor dem WWW und waren über entsprechende News-Clients zugänglich. Newsgroups stehen somit für ein Kommunikationsinstrument der 'alten Schule'. Viele Nutzer aus den Prä-WWW-Zeiten sind ihm treu geblieben, wogegen es den neueren Nutzergenerationen weitgehend unbekannt ist. Diese bevorzugen Webforen, die leichter zugänglich und häufig in komplexeren Kommunikationsumgebungen eingebettet sind. Einige Webforen verfügen über Funktionen, die eine tendenziell quasi-synchrone Verwendung erlauben.

These 2: Im Vergleich zwischen Webforen und Newsgroups weisen Webforen die höhere konzeptionelle Mündlichkeit auf.

Diskursstruktur. In Diskussionsforen werden die verschiedenen Beiträge in Gesprächsfäden (Threads) untergliedert. Ein Nutzer, der eine Frage oder ein Anliegen hat, kann einen solchen Gesprächsfaden eröffnen. Holmer (2008) nennt diese Beiträge „Seeds" oder „Isolated", wenn auf diese Beiträge keine Antworten erfolgen. Da mit solchen einleitenden Beiträgen häufig eine Bitte oder ein zu erläuterndes Anliegen verbunden ist, ist zu vermuten, dass diese Beiträge besonders sorgfältig geschrieben werden.

These 3: Thread-initiale Forenbeiträge sind schriftlicher konzipiert als ihre Antworten.

2.2 Korpora

Um sicher zu gehen, dass es neben den gewählten Variablen keine weiteren Störfaktoren gibt, wurden die beiden Vergleichsgruppen so gewählt, dass sie bezüglich der Thematik und der Nutzerstruktur eine möglichst große Übereinstimmung aufweisen. Bei der Newsgroup gab es die zusätzlichen Kriterien, dass sie eine möglichst lange Laufzeit mit einer relevanten monatlichen Beitragszahl haben sollte. Bei dem Webforum wurde darauf geachtet, dass es einen möglichst kleinen Funktionsumfang besitzt, es z.B. nicht möglich ist, bereits verfasste Einträge zu editieren.

Es wurden folgende Korpora untersucht:

- Newsgroup de.rec.motorrad,[1] 24.11.1991 - 30.4.2007, 55 620 Beiträge, 4 392 059 Tokens
- Parsimony-Webforum E30,[2] 27.6.2000 - 30.4.2007, 1 087 419 Beiträge, 27 624 296 Tokens
- E30-Forum (initiale Beiträge), 27.6.2000 - 30.4.2007, 204 189 Beiträge, 8 439 760 Tokens

3. Analyse

Um die Konzeption der Beiträge auszuwerten, wurden in Analogie zu den Variablen nach Burger folgende Merkmalsgruppen ausgewählt:[3]

[1] http://groups.google.com/group/de.rec.motorrad.

[2] Der Forenanbieter *Parsimony.net* hat 2008 seinen Betrieb eingestellt.

[3] Für die Betrachtung weiterer Merkmale sei auf Richling (2008) verwiesen.

– **Elemente, die Formalität signalisieren:** In vielen Aspekten ist die Formalität der IBK auf einem sehr niedrigen Niveau angesiedelt. Die Ausnahme bilden Begegnungen, bei denen eine anfängliche Anonymität nicht gegeben ist. Beispiele dafür lassen sich u.a. in den Experten-Chats des Dortmunder Chatkorpus finden. Im Allgemeinen herrscht der Eindruck vor, dass die Umgangsformen unabhängig von der Kommunikationsform von der virtuellen Gemeinschaft bestimmt werden. So ist es in vielen journalistischen Foren (z.B. das Forum von Spiegel Online: http://forum.spiegel.de/) üblich, sich zu siezen, wobei eine gleichzeitige Abwesenheit von Begrüßungen und Verabschiedungen auffallend ist. In Freizeitforen hat sich dagegen häufig das Du durchgesetzt. Im Gegenzug legen die Mitglieder mancher dieser Foren (wie z.B. unter http://www.chefkoch.de/forum/) sehr viel Wert auf Begrüßungs- und Verabschiedungsformeln. Diese sind in der Regel relativ informell (*hi*, *tachchen*, *servus*, *tschüss*, *CU*), stellen jedoch im Vergleich zur Abwesenheit eines solchen Grußes die formellere Alternative dar.

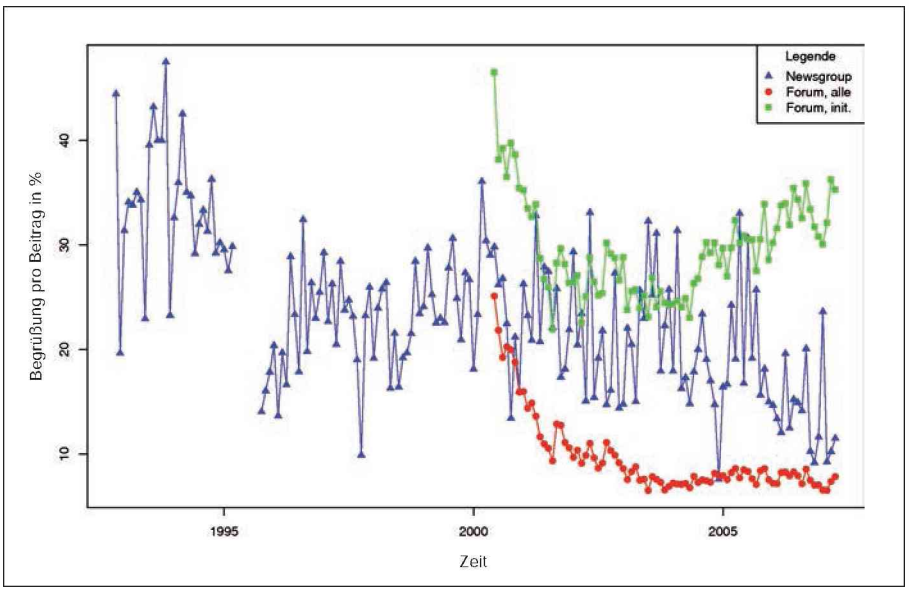


Abb. 1: Begrüßung pro Beitrag in %

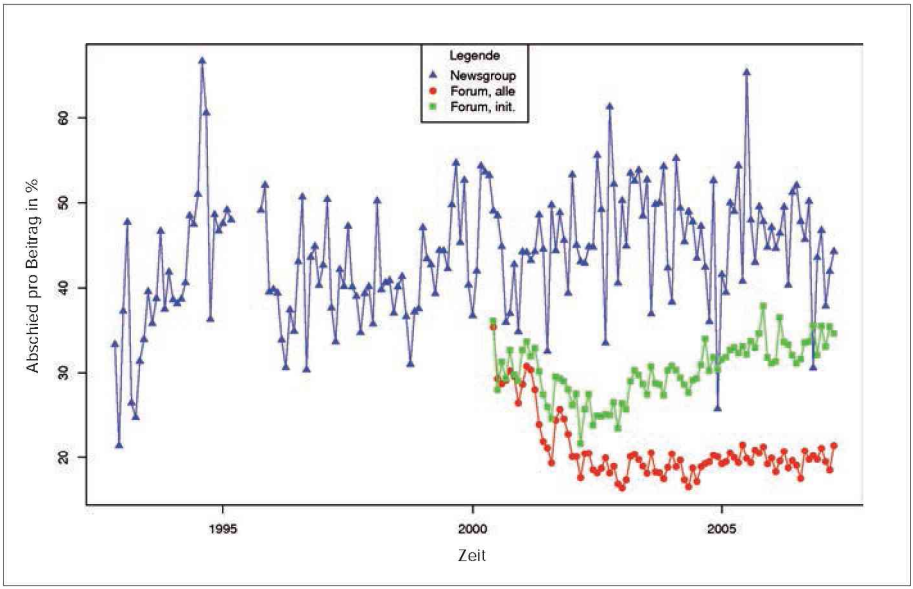


Abb. 2: Verabschiedung pro Beitrag in %

Die Ergebnisse der drei Vergleichsgruppen (Abb. 1und 2) zeigen ein sehr differenziertes Bild. Während in den frühen Neunzigern in der Motorrad-Newsgroup Monate auftraten, in denen über 40 % der Beiträge Begrüßungen enthielten, nahm die Häufigkeit dieser Formeln im diachronen Verlauf deutlich ab. Interessant ist, dass die Verabschiedungen dieses Verhalten nicht zeigen, wogegen sich die Verlaufskurven der Forenbeiträge und initialen Forenbeiträge hinsichtlich Begrüßungen und Verabschiedungen ähneln. Zusammenfassend lässt sich sagen, dass sichtbare, jedoch uneinheitliche Veränderungen im mikrodiachronen Verlauf auszumachen sind, die Newsgroup in diesen beiden Elementen formaler geprägt ist und die initialen Beiträge deutlich häufiger Begrüßungen und Verabschiedungen enthalten als ihre Antworten.

– **Elemente, die auf graphematischer Ebene phonische, umgangssprachliche Sprechweise nachbilden:** Diese Elemente sind keinesfalls IBK-spezifisch, sondern wurden unter anderen in Briefen des 19. Jahrhunderts und Werbeslogans („Schreibste mir – schreibste ihr – schreibste auf M.-K.-Papier.“) verwendet. Auch hier sind deutliche Unterschiede zwischen einzelnen Gemeinschaften festzustellen. Während diese Elemente in journalistischen Foren häufig in Zitaten oder auf andere Weise markierten Kontexten zu finden sind, sind sie in vielen Freizeitforen etabliert und werden biswei-

len scheinbar unbewusst eingesetzt: „Wenn ich doch vorm Rechner sitze, hab ich doch auch die Zeit, die Wörter eben auszuschreiben, oder?“ (Zitat aus einer Antwort bei www.gutefrage.net).

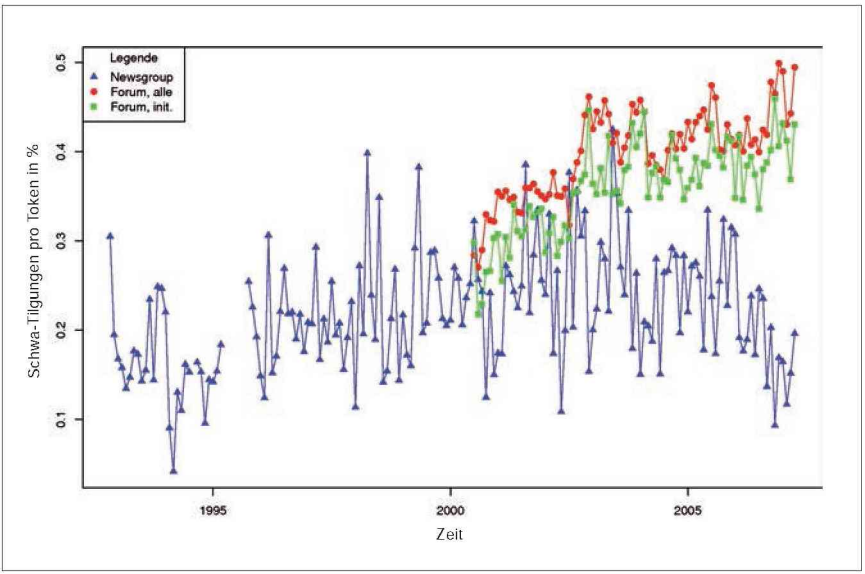


Abb. 3: Schwa-Tilgung pro Token in %

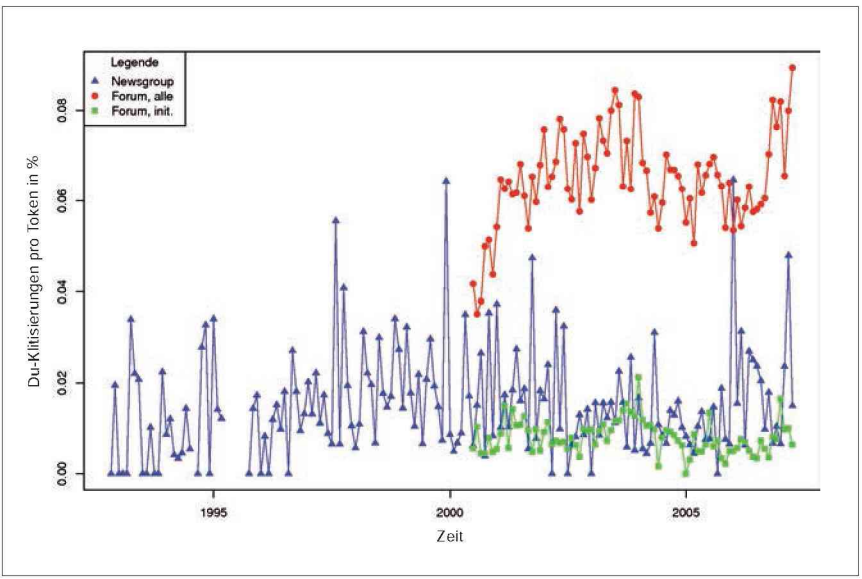


Abb. 4: *Du*-Enklise pro Token in %

Die hier untersuchten Elemente – Schwa-Tilgung im Auslaut (*ich hab*, *ich sag*) und Enklise des Personalpronomens der 2. Person Sg. (*haste, sagste*) zeigen auf, dass dieser Unterschied auch zwischen der Newsgroup und dem Webforum besteht. Im mikrodiachronen Verlauf (Abb. 3 und 4) kann innerhalb der zweiten Gruppe ein Anstieg der Enklisen festgestellt werden. Der Häufigkeitsverlauf der *Du*-Enklisen zeigt hingegen auf, dass dieses Ergebnis mit Vorsicht zu betrachten ist, solange es nicht mit der Frequenz zugrundeliegender Strukturen – hier die Anrede in der zweiten Person singular und die entsprechenden Verben – in Beziehung gesetzt wird. Es ist davon auszugehen, dass diese Werte mit den Vorkommen des entsprechenden Personalpronomens korrelieren. Dennoch kann festgehalten werden, dass das Phänomen der simulierten Mündlichkeit in dem untersuchten Webforum hinsichtlich der untersuchten Elemente stärker auftritt als in der Newsgroup.

– **Elemente, die Spontanität oder lange Vorbereitung anzeigen:** Da die sogenannte Spontansprache auf schriftlicher Ebene ein Produkt bewusster Überlegung sein kann, wurde hier vornehmlich die Beitragslänge betrachtet.

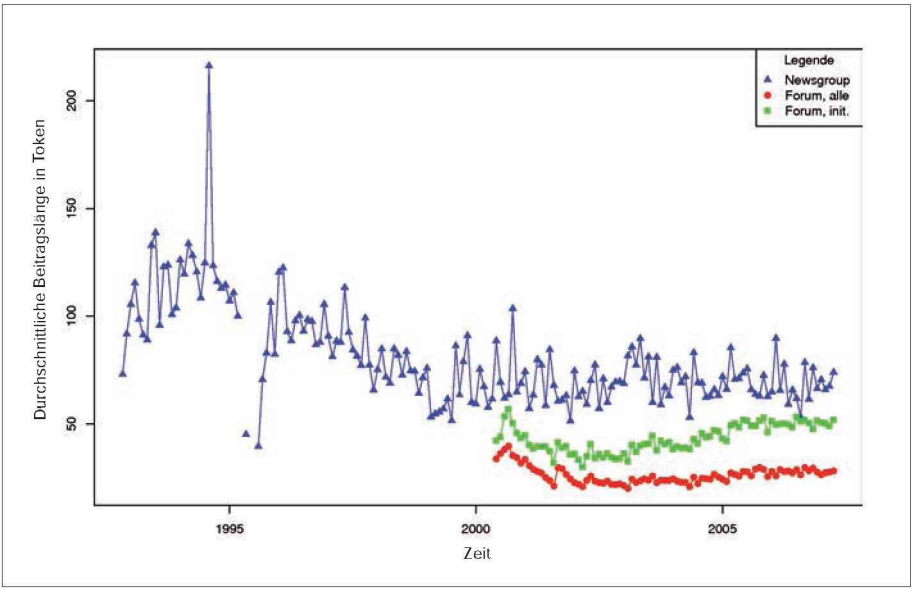


Abb. 5: Durchschnittliche Länge

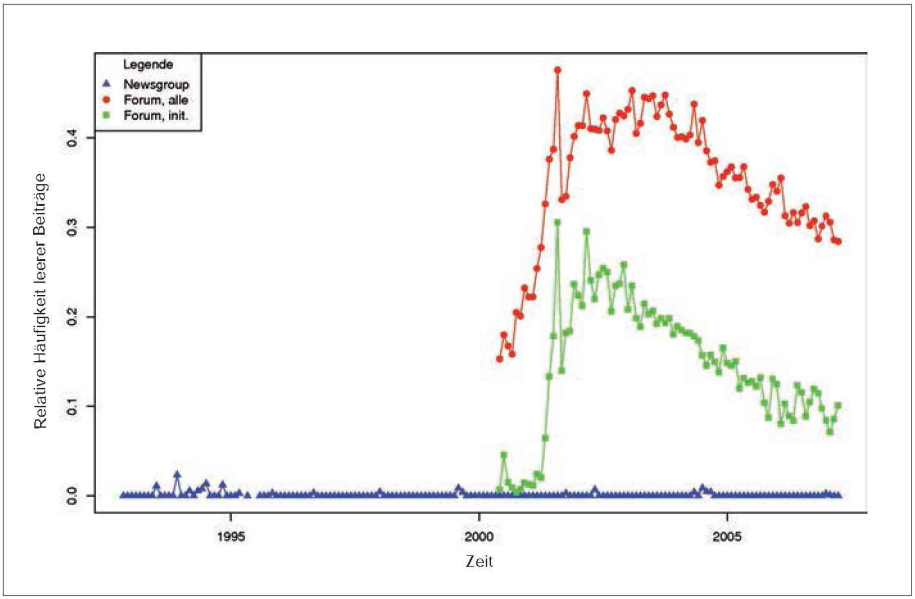


Abb. 6: Leere Beiträge in %

Hier ist bei der Newsgroup eine anfängliche Zunahme und eine anschließende Abnahme zu beobachten (Abb. 5). Ab 2000 bleibt die durchschnittliche Länge bei etwa 75 bis 80 Tokens konstant. Während die durchschnittliche gesprächsfadeneröffnende Nachricht in dem Webforum etwa 40 Tokens lang ist, erreicht die Gesamtheit der Webforenbeiträge gerade einmal eine durchschnittliche Länge von 25 Tokens. Zusätzlich wurde untersucht, wie hoch der Prozentsatz der Beiträge war, deren Inhalt vollkommen leer ist, bzw. ausschließlich durch die Betreffzeile vermittelt wird (Abb. 6). Während dieses Vorgehen bei Newsgroups aus Darstellungsgründen abgelehnt wird, beträgt der Anteil dieser inhaltsleeren Beiträge im Webforum mitunter über 40 %. Die Frequenz ist innerhalb der einleitenden Beiträge niedriger, erscheint jedoch in Hinsicht auf ihre Funktion und den damit verbundenen Erwartungswert erstaunlich hoch.

4. Zusammenfassung und Ausblick

Es konnte in Bezug auf die zweite These gezeigt werden, dass die Newsgroup – hinsichtlich der ausgewählten Merkmale – gegenüber dem Webforum eine ausgeprägtere konzeptionelle Schriftlichkeit aufweist. Ebenfalls konnte die dritte These bestätigt werden, dass die initialen Beiträge im Vergleich zu ihren

Antworten schriftlicher konzipiert sind. Hinsichtlich der ersten These kann festgestellt werden, dass sich die relative Häufigkeit der Merkmale im mikrodiachronen Verlauf verändert – mitunter sogar in einem hohen Grad. Jedoch ist dabei keine eindeutige Tendenz hinsichtlich einer Vermündlichung zu erkennen. Diese ließe sich höchstens durch eine künstliche Beschränkung des Zeitrahmens konstruieren.

Auf Basis dieser Ergebnisse ist es wünschenswert, einen genaueren Blick auf die grammatikalischen Strukturen dieser Beiträge werfen zu können. Ein detailliertes Vorgehen erfordert jedoch weitere Entscheidungen hinsichtlich der Annotation und Auswertung. So sieht das Stuttgart-Tübingen-Tagset (STTS), das im Bereich der POS-Annotation als de-facto Standard gilt, die Annotation klitischer Formen oder IBK-typischer Elemente wie Emoticons nicht vor. Daher ist es vorgesehen, entsprechende Annotationen zu entwickeln und mit Hilfe adäquat aufbereiteter IBK-Korpora das Zusammenspiel individueller Freiheit, gemeinschaftsgebundener Regeln und den daraus resultierenden Entwicklungen genauer zu untersuchen.

Literatur

Beißwenger, Michael / Storrer, Angelika (2008): Corpora of computer-mediated communication. In: Lüdeling, Anke / Kytö, Merja (Hg.): Corpus linguistics. An international handbook. Bd. 1. (= Handbücher zur Sprache und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science 29.1). Berlin / New York: de Gruyter, 292-308.

Burger, Harald (2005): Mediensprache. Eine Einführung in Sprache und Kommunikationsformen der Massenmedien. Berlin: de Gruyter.

Dürscheid, Christa (2003): Medienkommunikation im Kontinuum von Mündlichkeit und Schriftlichkeit. Theoretische und empirische Probleme. In: Zeitschrift für Angewandte Linguistik 38: 37-56.

Holmer, Torsten (2008). Discourse structure analysis of chat communication. Language@ Internet 5. Internet: www.languageatinternet.de/articles/2008/1633 (Stand: 11 / 2010).

Koch, Peter / Österreicher, Wulf (1985): Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgebrauch. In: Romanistisches Jahrbuch 36: 15-43.

Richling, Julia (2008): Die Sprache in Foren und Newsgroups. Saarbrücken: Verlag Dr. Müller.

TOMAS BY

The Prolog version of the Tiger Dependency Bank

Abstract

Like the PARC 700 dependency bank, the Tiger dependency bank has lemmatised tokens and no representation of the word order, which means that multiple occurrences of the same word type in one sentence are ambiguous. In the Tiger dependency bank, around 9% of the word tokens are ambiguous in this way, compared to 15% in PARC 700. The following describes a Prolog version of the Tiger dependency bank that represents the word order explicitly, uses the surface forms of the words rather than base forms, makes a clear distinction between words and empty nodes, and stores the token attributes in one place rather than spreading them out across the file. Together with the data in Prolog format, graphical representations of the dependency trees are provided in PDF format.

1. Introduction

The Tiger dependency bank (Forst et al. 2004) is a conversion of 1865 sentences from the Tiger Treebank (Brants et al. 2002) to a dependency format based on the one used in the PARC 700 dependency bank (By 2007: 261f.). This format is intended for parser evaluation, by converting the parser output to a set of dependency relations between strings representing the base forms of some of the words in the sentence and comparing these to the supposedly correct ones in the dependency bank. As pointed out in By (ibid.: 263), this fails to take into account that a sentence may contain more than one instance of the same word type, which may lead to comparisons against the wrong dependency relations. In the PARC 700 dependency bank around 15% of the tokens are ambiguous in this way (ibid.: 275-277), and in the Tiger dependency bank it is about 9% (By 2009: 120-122). Other problems with the PARC 700 format, inherited by the Tiger dependency bank and not shared by the Prolog format described here, include a spurious distinction between full tokens and attribute tokens, the absence of a distinction between words and empty nodes in the dependency tree, and an unnecessarily distributed representation of the token attributes. The rest of the paper contains four sections. One on how the tokenisation in the Tiger dependency bank differs from the Tiger treebank,

one that explains the problems with the PARC 700/Tiger dependency bank data format, and one on the conversion into Prolog data format. A summary of findings then concludes the paper.

2. Tokenisation

Probably the most obvious application for the Tiger dependency bank is for the evaluation of a parser designed for the Tiger treebank. In this situation there is an issue of differences in tokenisation. In contrast to the PARC 700 dependency bank (By 2007: 278), the tokenisations in the Tiger case are quite similar, as shown in Table 1. The main difference are the 'multi-word' tokens, which seem to be mainly compound words in the treebank that have been split up in the dependency bank, and hyphenated words that are individually tokenised in the treebank, but are one token in the dependency bank. The 'no correspondence' tokens are mostly punctuation. The difference between the 'empty string' tokens in Table 1 and 'empty nodes' (of which there are 1 161 in the Tiger dependency bank) is that the former have a position in the word order since they are parts of the representation of fused words (cf. By 2009: 121, 123).

	Treebank	**Dependency bank**	**(both)**
Identical			26 923
Same, except for capitalisation			1 220
Morphology			9
Extra punctuation in the Treebank token			45
Space character in DB but not in TB			9
Tokens that correspond directly	28 206	28 206	28 206
Multi-word tokens	2 554	262	
Components of multi-word tokens	587	5 379	
Tokens with no correspondence	4 969	5	
Empty strings (fused words)	–	563	
Total number of tokens	36 316	34 415	

Table 1: Comparison of the tokenisations of the Tiger corpora

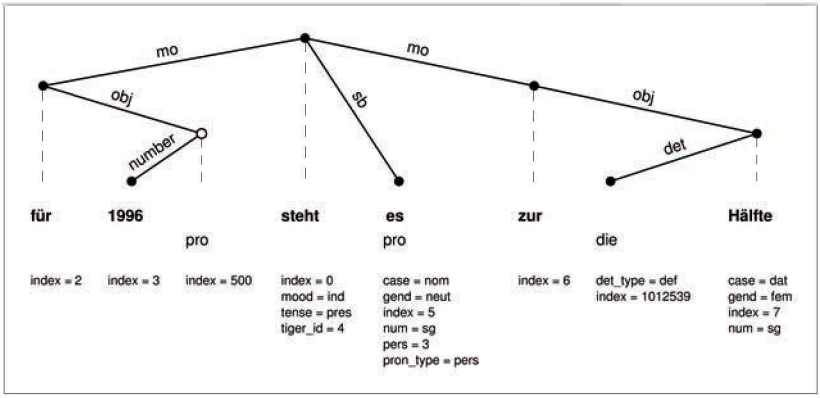


Figure 1: Tiger sentence 9 438

3. Problems with the PARC 700 / Tiger dependency format

In the PARC 700 dependency format, which is also used in the Tiger dependency bank, multiple occurrences of the same word in a sentence cannot be distinguished, since there is no explicit representation of the word-order (By 2007: 265f., By 2009: 120f.). This affects around 9 % of the word tokens in the Tiger dependency bank and about 15 % in the PARC 700 dependency bank. When the data is used for parser evaluation, it would seem likely for this ambiguity to cause inaccuracy in the measurements, since wrong identification of the words can mean comparison against the wrong dependency links.

Both of the dependency banks make use of symbolic tokens (for example, *es* in sentence 9 438 / Figure 1, which is encoded as *pro* in the data) and 'empty nodes' that do not correspond to words in the sentence (such as the first *pro* token, and the *die* token, in Figure 1). In the file format there is no indication as to whether a token corresponds to a word or not, so this is a further source of ambiguity. The string *pro* occurs as a word once in PARC 700 (By 2007: 278) and three times in the Tiger dependency bank (By 2009: 120). This problem is particularly acute in the case of ellipsed words in conjunctions in the Tiger dependency bank. Figure 2 shows an example. In sentence 9 572 there is only one occurrence of the word *schulden*, but in the tree there are two tokens with this name. There is no indication in the data that only one of them corresponds to a word in the sentence and that the other one is an empty node. The decision to show them in Figure 2 as first the real word and then the empty node, was arbitrary. Besides the obvious problem of having two items in the 'gold standard'

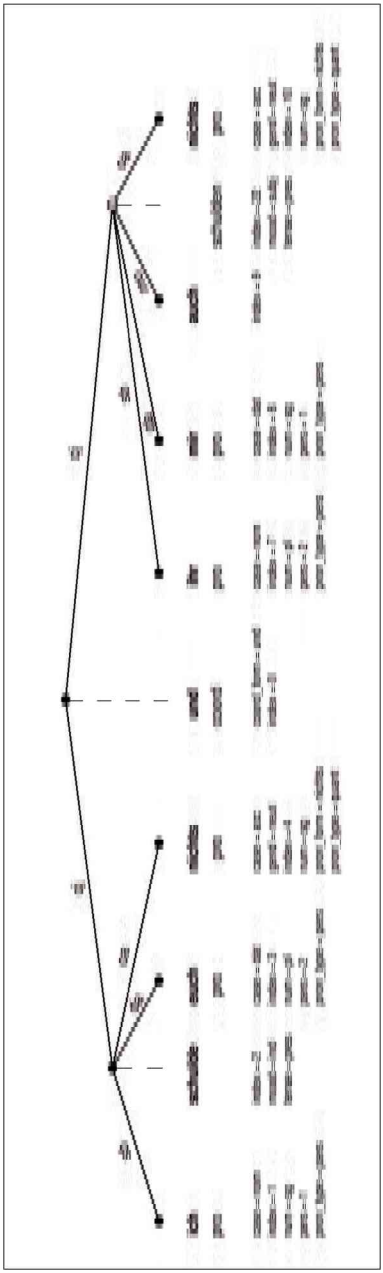


Figure 2: Tiger sentence 9572

that look like two occurrences of the same word while there is only one occurrence in the sentence, it seems unlikely that any parser evaluation using this 'dependency-based' method will be reliable for sentences with ellipsed verbs. Unless the parser (or a post-parsing data conversion) creates empty nodes in the right places in the tree, those links will not match, and even if the trees are isomorphic, false positives are still possible because of the ambiguity.

Another type of construction that is problematic in the Tiger dependency bank is the representation of 'fused words', such as the *zur* in Figure 1. It is a contraction of *zu* and *der*, which are represented separately in the Tiger data. The problem, of course, is that there is only one word in the sentence and that the two nodes are not directly related in the tree. The somewhat similar contractions of negations in English (e.g. *wasn't*) are tokenised as two separate units, both in the Penn Treebank and in PARC 700, and they are directly connected in the tree.

The Prolog format also stores all the attributes together with the token that they belong to. In the PARC 700/Tiger dependency bank format they are spread out individually in the file, and only connected to the token via the token id number. This is merely a technicality, as there is no great difficulty involved in searching through the file and collecting all the attributes of a particular token. On the other hand, with the Prolog representation there is no need to do that as all the attributes are immediately available together with the token string.

In the PARC 700, but only to a limited extent in the Tiger dependency bank, there is a spurious distinction between 'full' tokens and words that are encoded as attributes of other tokens. There is no example of this phenomenon in Figure 1 or 2, but the infinitive markers are treated this way in the Tiger dependency bank (By 2009: 124). It seems that this is a legacy of the particular parser technology used to produce the data from which the PARC 700 dependency bank was created (By 2007: 263).

4. Prolog conversion

The conversion of the Tiger dependency bank from PARC 700 data format to Prolog data format is done in two steps (By 2009: 124, 126f.). First, the original sentence from the 'sentenceform' field is tokenised and matched against the dependency bank tokens. This is in effect a search procedure that terminates when a complete tokenisation of the sentence string is found that contains all the tokens that occur in the dependency bank representation of the sentence. The second step of the conversion is a scoring procedure that computes a cost for each ambiguous token, based on the distance of the dependency links. In the (common) situation where there are multiple dependency links to and from the ambiguous tokens, the average length of the links per word is used instead (and some weights are added). Some practical experimentation was involved in the design of this scoring function. While the total number of ambiguous tokens is lower in the Tiger dependency bank than in PARC 700, the frequency of highly ambiguous sentences is higher. This seems to be mainly because of the high number of articles in German. In the Tiger dependency bank many of these have the exact same representation (either *die* or *pro*). During the conversion of PARC 700 (By 2007: 269) only three sentences were too difficult or time-consuming for the automatic procedure, and had to be hard-coded. For the Tiger dependency bank this number was about fifty (out of three times as many sentences). In about thirty cases where sentences typically contained between five and ten articles each, the software failed to produce a result in a reasonable time so the disambiguation was hard-coded. In another hundred or so, the resulting disambiguation was found to be wrong, so these were also corrected by hand (By 2009: 126f.).

The Prolog data format has a number of advantages. The tokens have exactly the same sequence of alphabetical characters as in the original text (punctuation is included only when it is part of the dependency structure in the original Tiger dependency bank files). This means that linking the output of any parser to the

tokens in the dependency bank is a (relatively) simple matter of looping from left to right and comparing the alphabetical characters. The word order is encoded explicitly in the Prolog data, eliminating the 9% token ambiguity in the Tiger dependency bank. There is a clear distinction between words and empty nodes, and the attributes are stored in one place together with the token and not spread out individually in the file. Additionally, data in Prolog format can be loaded into a Prolog interpreter database with a single command and then explored very easily by querying the database, or printed in other formats using short bits of code. This combination of human-friendly syntax for data and an interpreter with a database is unique to Prolog (and some closely related languages).

The converted and slightly corrected Tiger dependency bank, in Prolog format, can be downloaded from the following web page:

http://www.basun.net/homepages/tomas/papers/tiger/ (last visited: 10/2010)

It should be noted that no linguistic transformations or modifications occurred in the data during the conversion, apart from corrections of a small number of errors (By 2009: 125).

5. Summary

In the PARC 700 data format that the Tiger dependency bank uses, the combination of lemmatised tokens and the lack of explicit encoding of the word-order means that sentences with more than one occurrence of token types with the same base form are ambiguous. The number of tokens involved in this type of ambiguity is about 9% in the Tiger dependency bank.

The Prolog version of the Tiger dependency bank avoids these problems by representing the word order explicitly and using the surface forms of the words rather than the base forms. It also makes a clear distinction between word tokens and empty nodes, and stores the token attributes together with the token instead of spreading them out individually in the file, as in the PARC 700 format.

References

Brants, Sabine / Dipper, Stefanie / Hansen, Silvia / Lezius, Wolfgang / Smith, George (2002): The TIGER treebank. In: Hinrichs, Erhard / Simov, Kiril (eds.): Proceedings of the first workshop on treebanks and linguistic theories. Sozopol, 24-41.

By, Tomas (2007): Some notes on the PARC 700 dependency bank. In: Natural Language Engineering 13, 3: 261-282.

By, Tomas (2009): The TiGer Dependency Bank in Prolog format. In: Kłopotek, Mieczysław A. / Przepiórkowski, Adam / Wierzchoń, Sławomir T. / Trojanowski, Krzysztof (eds.): Recent advances in intelligent information systems. Warsaw: EXIT, 119-129.

Forst, Martin / Bertomeu, Núria / Crysmann, Berthold / Fouvry, Frederik / Hansen-Schirra, Silvia / Kordioni, Silvia (2004): Towards a dependency-based gold standard for German parsers – The TiGer Dependency Bank. In: Hansen-Schirra, Silia / Oepen, Stephan / Uszkoreit, Hans (eds.): Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora. Geneva: COLING, 31-38.

SILKE SCHEIBLE / RICHARD JASON WHITT / MARTIN DURRELL / PAUL BENNETT

Investigating diachronic grammatical variation in Early Modern German. Evidence from the *GerManC* corpus

Abstract

The aim of the *GerManC* project is to compile a corpus of German texts for the early modern period (1650-1800). It is intended to be representative and to this end it will consist of 2 000 word extracts from nine text-types, with an equal number of each from five major regions and from three sub-periods of fifty years (1650-1700, 1700-1750, 1750-1800), giving about a million words. In addition, analytical tools are being developed concurrently to tag and lemmatise the corpus with a view to annotating it more fully by the end of the project.

One objective of the corpus is to document the process of standardisation in German, and the texts which have been digitised already provide interesting data in this respect. In particular, they demonstrate clearly the development of grammatical variation and in many cases its gradual elimination during this period, for example

1) Norms changing in line with prestige models:

   In non-literary texts, the ending *-e* is initially the most frequent in the weak adjective in the nominative and accusative plural, e.g., *die starke Könige*. However, by the third sub-period the ending *-en* has become dominant, e.g., *die starken Könige*, and differences in usage between text-types indicate that this reflects the more general adoption of a prestige variant from literary texts.

2) Elimination of variation:

   In the early part of this period there is still variation in the relative position of the auxiliary verbs and main verb infinitives in subordination, but by the end of the period this variation is almost unknown.

3) Reduction in regional variants:

   In the first sub-period regional – especially southern – variants are widely retained, for example gender forms of the numeral *zwei* or inflections in the singular of weak feminine nouns (e.g., *auf der Gassen*). By the third sub-period these are no longer encountered.

1. The *GerManC* corpus project: background and aims

The *GerManC* project is a corpus project based at the University of Manchester with the aim of compiling a representative corpus of German for the period 1650-1800. It is intended to fill a gap in the provision of historical corpora of German and to parallel historical corpora of English for this period, notably the *ARCHER* corpus and the Helsinki corpus (cf. Meyer 2002: 20-22). Like these, it will be as representative as possible of language usage in this period, including a wide range of registers with each register being represented by a sample of equal size. To achieve this in an optimal fashion, the corpus will not consist of complete texts, which could mean that one text type (e.g., long novels) would be overrepresented, but of relatively short extracts of 2 000 words, following the model of the *ARCHER* and other English corpora.

Given differences in the availability of suitable texts, the specific registers selected are not identical to those of the *ARCHER* corpus. Those which are well attested enough to offer a satisfactorily representative range of material for inclusion consist of four which are more 'oral' in character, in the sense of being closer to the spoken language (dramas, newspapers, personal letters, and sermons), and five which share the typical characteristics of 'written' language (legal documents, scientific writing, narrative or biographical writing, and works of scholarship in the humanities). Chronological representativeness is also desirable in a diachronic corpus, and this will be achieved by following the model of the Bonn Corpus of Early New High German (1350-1650) – the period immediately preceding that of *GerManC* (cf. Hoffmann / Wetter 1987) – by taking equal numbers of texts for each register from a succession of fifty year sub-periods, i.e., 1650-1700, 1701-1750, 1751-1800.

Unlike English, there is still considerable regional variation in German during this period, although it diminishes markedly over the period in question as the more northerly standard originating in the Central German area was gradually adopted in the South. In order to enable this development to be traced systematically in the *GerManC* corpus, it must include a representative selection of texts from the major regions of the German speech area, i.e., North German, West Central German, East Central German, West Upper German (the South-West, including Switzerland) and East Upper German (the South-East, including Austria). The *GerManC* corpus foresees in principle the selection of three

2 000-word text samples for each subdivision in terms of register, sub-period, and region. Given nine registers, three chronological sub-periods, and five regions, the complete final corpus will consist of nearly a million words.

A pilot project of twelve months was supported by a grant from the Economic and Social Research Council (ESRC). This involved the compilation of a partial corpus of a single genre (newspapers) and the evaluation of a number of tools and was completed in spring 2007. This newspaper corpus has been deposited at the *Oxford Text Archive* and is freely available for academic study. Further details about this corpus, and examples of the kind of data it provides, are given in Durrell / Ensslin / Bennet (2007, 2008a, 2008b).

Funding for the completion of the project over three years was granted in 2008 jointly from the ESRC and the Arts and Humanities Research Council (AHRC), and work was commenced in September 2008. Most of the texts involved are printed in *Fraktur* (black letter font), and the experience of the pilot project demonstrated that scanning such texts with OCR was impractical and prone to error. All texts are thus being keyed in twice and the results compared electronically to eliminate mistakes ('double-keying'). Texts will then receive structural mark-up according to TEI5 Lite guidelines.[1] In order to facilitate a thorough linguistic investigation of the data, the corpus will be annotated in terms of tokens, sentence boundaries, parts of speech, and morpho-syntactic categories, and the corpus will be comprehensively lemmatised.

Because of the degree of variability in lexis, morphology, syntax, and orthography characteristic of this period in the development of German, and the additional variation introduced by the three variables of genre, region, and time, automatic annotation of the texts poses a major challenge. In the course of the project it is therefore intended to carry out a systematic evaluation of current corpus annotation tools and assess their robustness across the various sub-corpora, identify procedures for building and improving annotation tools for historical texts, and incorporate them in a historical text processing pipeline. Fuller information about the corpus, with details on progress and previous publications, is available from the *GerManC* website: http://www.tinyurl.com/germanc.

[1] For details, see http://www.tei-c.org/Guidelines/Customization/Lite/ (last visited: 10 / 2009).

2. Data on diachronic grammatical variation in Early Modern German

By early September 2009 first inputting was nearly complete for four further genres (drama, narrative prose, humanities texts, and scientific texts), and for about 50 % of legal texts and sermons. This material, added to the completed pilot corpus of newspapers, already provides interesting data illustrating the process of standardisation in German, which is one of the main objectives of the project. With the proviso that the present incomplete state of the corpus does not guarantee full representativeness, a selection of data is presented below in order to illustrate the kind of information about the development of German during this period which the corpus will offer.

2.1 Shifts in norms in line with prestige models – the 'weak' adjective

At the beginning of the Early Modern German period the endings *-e* and *-en* are in competition in the nominative and accusative singular of the weak adjective, cf. Solms / Wegera (1991: 175-184) and Ebert et al. (1993: 189-199), e.g.:

(a) die gut***e*** Kinder (b) die gut***en*** Kinder

The variant *-e* is frequent in all regions **except** East Central Germany in the first period 1650-1700, and it is particularly common in less prestigious genres (newspapers, scientific writing).[2] However, *-en* is dominant in all genres in East Central Germany, whose written language (*Meißnisch*) has high prestige, **and** in the prestigious literary genres. In the course of the 18th century the variant *-en* becomes dominant in all genres and all regions, and by 1800 it has become the definitive codified norm. The following tables illustrate this for the genres where relatively full data are already available:

[2] For a detailed account of this variable in the pilot newspaper *GerManC* corpus, see Durrell / Ensslin / Bennet (2008a).

	1650-1700		1701-1750		1751-1800	
	-e	*-en*	*-e*	*-en*	*-e*	*-en*
Newspapers	135 (74%)	48 (26%)	71 (48%)	78 (52%)	25 (15%)	141 (85%)
Scientific texts	69 (60%)	46 (40%)	56 (49%)	58 (51%)	5 (4%)	122 (96%)
Humanities texts	34 (39%)	54 (61%)	21 (18%)	99 (82%)	12 (11%)	91 (89%)
Narrative prose	21 (31%)	46 (69%)	13 (22%)	46 (78%)	7 (11%)	55 (89%)
Drama	9 (16%)	48 (84%)	4 (10%)	35 (90%)	4 (9%)	40 (91%)
Total	**268 (53%)**	**242 (47%)**	**165 (34%)**	**316 (66%)**	**53 (11%)**	**449 (89%)**

Table 1: Distribution of the variant endings by genre

	1650-1700		1701-1750		1751-1800	
	-e	*-en*	*-e*	*-en*	*-e*	*-en*
North German	54 (49%)	56 (51%)	20 (18%)	92 (82%)	5 (4%)	111 (96%)
West Central	53 (54%)	46 (46%)	41 (41%)	58 (59%)	9 (12%)	66 (88%)
East Central	19 (20%)	76 (80%)	16 (15%)	94 (85%)	2 (2%)	110 (98%)
West Upper	76 (78%)	21 (22%)	46 (48%)	49 (52%)	30 (27%)	82 (73%)
East Upper	66 (61%)	43 (39%)	42 (65%)	23 (35%)	7 (8%)	80 (92%)
Total	**268 (53%)**	**242 (47%)**	**165 (34%)**	**316 (66%)**	**53 (11%)**	**449 (89%)**

Table 2: Distribution of the variant endings by region

2.2 Elimination of variation: clause-final verb clusters

In Middle High German and Early New High German the relative position of the elements in the clause-final verb complex was subject to considerable variation, cf. Ebert et al. (1993: 438-440). With modal verbs we find, for example, the following orders with some frequency:

(a) [...] FINITE + NON-FINITE: [...], *dass du es heute* [...] ***sollst machen***

(b) [...] FINITE [...] + NON-FINITE: [...], *dass du es* [...] ***sollst*** *heute* [...] ***machen***

(c) [...] NON-FINITE + FINITE: [...], *dass du es heute* [...] ***machen sollst***

By 1650, though, most variation had been eliminated in these simple clusters with two verbs, and order (c), which is the norm in modern standard German, had already been established as the dominant variant. Table 3 contains

data for clusters with the modal *sollen* in the *GerManC* corpus to date and shows that the other variants are sparsely attested, and hardly at all after 1750:

	1650-1700	1701-1750	1751-1800
(a)	12 (4%)	13 (6%)	
(b)	13 (5%)	6 (3%)	3 (2%)
(c)	256 (91%)	197 (91%)	166 (98%)

Table 3: Clause final clusters with *sollen* + infinitive

However, variation prevails for longer in more complex verb clusters. Thus, in cases with a finite modal verb and a passive infinitive we find that variant (x), with the finite verb first, is relatively frequent until 1750, and only after this has variant (y), which represents the modern standard, become clearly dominant, as shown on Table 4.

(x) FINITE+PAST PART.+INFINITIVE: [...], *dass es* [...] ***soll gemacht werden***

(y) PAST PART.+INFINITIVE+FINITE: [...], *dass es* [...] ***gemacht werden soll***

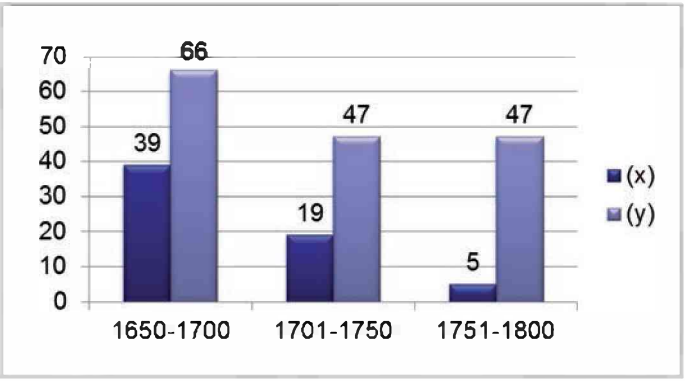


Table 4: Clause final clusters with *sollen* + passive infinitive

A further type, (z), with PAST PART+FINITE+INFINITIVE is attested three times in the earliest period 1650-1700, e.g.:

(z) [...], *dass es* [...] ***gemacht soll werden***

2.3 Elimination of regional variants: the declension of 'weak' feminine nouns

Middle High German has a class of 'weak' feminine nouns, which like the modern German 'weak' masculine nouns have the suffix -(*e*)*n* in the oblique cases of the singular, e.g.:

Nom.: *diu erde* Acc.: *die erden* Gen.: *der erden* Dat.: *der erden*

This suffix begins to be eliminated from the 16th century, cf. Ebert et al. (1993: 176-178) and Wegera (1987: 110-148), so that feminine nouns now have no case endings in the singular, except in fossilised phrases like *auf Erden*. However, this process is protracted and *-en* is still frequent in oblique cases of the singular in the 17th century, and well into the 18th century in some regions, especially East Upper German, so that both the following paradigms occur in our data:

(a) Nom. *die Erde* Acc. *die Erde* Gen. *der Erde* Dat. *der Erde*

(b) Nom. *die Erde* Acc. *die Erden* Gen. *der Erden* Dat. *der Erden*

In our material 62 feminine nouns occur with the suffix -(*e*)*n* in oblique cases of the singular, though many are attested with it only once. The most frequent of these is *Erde*, with the distribution of the variant endings in the relevant cases given on table 5:[3]

	1650-1700		**1701-1750**		**1751-1800**	
	-e	*-en*	*-e*	*-en*	*-e*	*-en*
North German	7	2	9	4	28	
West Central	2	1	1	3	7	
East Central	5	12	3	5	6	
West Upper		7	29	8	9	
East Upper	1	13	8	12	14	
Total	**15 (30%)**	**35 (70%)**	**50 (61%)**	**32 (39%)**	**64 (100%)**	

Table 5: Declension of *Erde* in the singular

[3] Note that Table 5 excludes occurrences of the phrase *auf Erden*, which is always found in this form.

Table 6 gives all occurrences of -(*e*)*n* in oblique singular cases of feminine nouns in our material. It is still widespread in the 17th century, but later it is most common in Upper German, and becomes scarce after 1750.

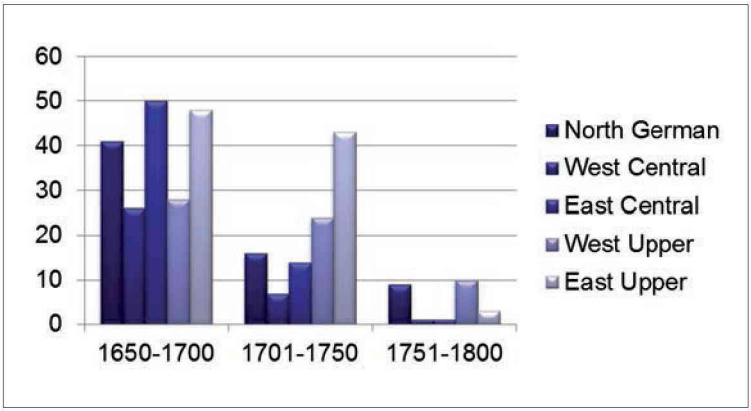


Table 6: Occurrences of -(e)*n* in oblique singular cases of feminine nouns

References

Durrell, Martin / Ensslin, Astrid / Bennet, Paul (2007): GerManC. A historical corpus of German 1650-1800. In: Sprache und Datenverarbeitung 31: 71-80.

Durrell, Martin / Ensslin, Astrid / Bennet, Paul (2008a): Zur Standardisierung der Adjektivflexion im Deutschen im 18. Jahrhundert. In: Czachur, Waldemar / Czyżewska, Marta (eds.): Vom Wort zum Text. Studien zur deutschen Sprache und Kultur. Festschrift für Professor Józef Wiktorowicz zum 65. Geburtstag. Warszawa: Instytut Germanistyki Uniwersytetu Warszawskiego, 259-267.

Durrell, Martin / Ensslin, Astrid / Bennet, Paul (2008b): Zeitungen und Sprachausgleich im 17. und 18. Jahrhundert. In: Zeitschrift für deutsche Philologie 127 (special issue), 263-279.

Ebert, Robert Peter / Reichmann, Oskar / Solms, Hans-Joachim / Wegera, Klaus-Peter (1993): Frühneuhochdeutsche Grammatik. Tübingen: Niemeyer.

Hoffmann, Walter / Wetter, Friedrich (eds.) (1987): Bibliographie frühneuhochdeutscher Quellen. Ein kommentiertes Verzeichnis von Texten des 14.-17. Jahrhunderts (Bonner Korpus). 2nd ed. Frankfurt a.M.: Lang.

Meyer, Charles F. (2002): English Corpus Linguistics. An Introduction. Cambridge: Cambridge University Press.

Solms, Hans-Joachim / Wegera, Klaus-Peter (1991): Grammatik des Frühneuhochdeutschen. Vol. 6: Flexion der Adjektive. Heidelberg: Winter.

Wegera, Klaus-Peter (1987): Grammatik des Frühneuhochdeutschen. Vol. 3: Flexion der Substantive. Heidelberg: Winter.

CHRISTOPHER COX

Quantitative perspectives on syntactic variation: Investigating verbal complementation in a corpus of Mennonite Plautdietsch[1]

Abstract

This study presents a quantitative analysis of syntactic variation in the ordering of verbal constituents in a series of bipartite verbal complementation constructions in Canadian Mennonite Plautdietsch. The alternation in verb-final contexts between v_1-v_2 and v_2-v_1 complement orders attested in corpus data is rendered statistically through the application of generalized linear mixed effects modelling. Such quantitative, corpus-based methods, this study concludes, might serve not only in evaluating the simultaneous effects of multiple hypothesized predictors of variation, as is undertaken with the Plautdietsch corpus data, but also in permitting a greater number of linguistic data and potentially competing predictors to bear upon the task of analysis than would otherwise be possible.

1. Introduction

A common problem in linguistic description lies in analyzing grammatical variation – in seeking to derive perspicuous and yet empirically adequate accounts of variable phenomena in language. While the threads of variation might be perceived to run throughout the fabric of grammar, the theoretical and practical issues such variability poses are felt with perhaps particular acuteness in contemporary syntactic research, where analyses commonly adopt introspection and categorical judgements as their primary bases of argumentation. As Bresnan et al. (2007) contend, however, introspective methods applied without systematic consideration of empirically-attested variation in both attestation and judgement may risk unintentionally underestimating the range and complexity of licit variation, and thus ultimately provide only partial

[1] Thanks are due to R. Harald Baayen, Grzegorz Kondrak, John Newman, Sally Rice, and members of the Department of Linguistics at the University of Alberta for their helpful comments on this research. The support of the Trudeau Foundation, the Killam Trusts, and the Social Sciences and Humanities Research Council of Canada (SSHRC) is gratefully acknowledged here.

accounts of the phenomena of interest. On this view, an active, analytical engagement with empirically-attested variation is essential to empirically-adequate grammatical description.

Corpora may play a natural role in this analytical process, inasmuch as they provide contextualized examples of grammatical variation which corpus-based techniques might then assist in understanding. The present paper pursues one such quantitative, corpus-based account of grammatical variation, concentrating upon verb placement alternations in West Germanic verbal complementation constructions (VCC). The remainder of this paper is organized as follows: after introductory consideration of the typological status of verbal complementation in the Continental West Germanic languages, specific attention is given to variation in VCC attested in Mennonite Plautdietsch (ISO 639-3: pdt). A quantitative analysis is then conducted upon instances of naturally-occurring variation in such constructions drawn from a corpus of Canadian Mennonite Plautdietsch. Both the results and the methodology used to obtain them are briefly discussed in the conclusion, with special reference to the larger problems which such variable constructions pose for syntactic analysis generally.

2. Verbal complementation

Verbal complementation (VC) refers here to the introduction of verbal elements v_{i+1} as a syntactic argument of another verb v_i. From a typological perspective, VC represents a common strategy for the analytical construction of complex predicates, both globally (cf. Dixon 2006) and in the Continental West Germanic languages specifically (cf. Wurmbrand 2004, 2005; Zwart 2005). In the latter group of languages, verbal complementation commonly finds use in the expression of aspectual, epistemic, and deontic information relating to the predicate under construction, as is the case with lexical verbs introduced as the complements of auxiliary (in perfective constructions, expressing aspectual distinctions) and modal (in epistemic and deontic constructions) verbs.

Despite cognate patterns of morphological marking in and the considerable frequency of use of verbal complementation constructions across the Continental West Germanic languages, acceptable complement orders differ notably between individual constructions and speech communities. As examples (1a-c) demonstrate, even within relatively simple, bipartite modal-infinitive constructions in verb-final contexts, striking variation exists in the reported grammaticality of different verbal complement orders:

Modal-infinitive construction, verb-final context

(1a) *...weil er das Buch* ***muss**$_1$ **kaufen**$_2$ / **kaufen**$_2$ **muss**$_1$* [*1-2 / ✓2-1]
'because he had$_1$ to buy$_2$ the book.' (Standard German: Wurmbrand 2004: 74)

(b) *...da Valère dienen boek* ***wilt**$_1$ **kuopen**$_2$ / ***kuopen**$_2$ **wilt**$_1$* [✓1-2 / *2-1]
'that Valère wants$_1$ to buy$_2$ this book.' (West Flemish: Wurmbrand 2005: 331)

(c) *...dat Jan het boek* ***kan**$_1$ **lezen**$_2$ / **lezen**$_2$ **kan**$_1$* [✓1-2 / ✓2-1]
'that Jan can$_1$ read$_2$ the book.' (Standard Dutch: Wurmbrand 2005: 324)

Such variation in the relative acceptability of even these common constructions is relatively unexpected, given the close genetic and geographic relationship between most of these languages, and raises questions as to the ultimate motivations of this variability, which neither consistently identifies the same logically-possible linear orders as being grammatical or ungrammatical cross-linguistically (1a-b), nor presents a categorical delimitation between such alternants in all cases (1c).

3. Analyzing verbal complementation in Mennonite Plautdietsch

Such non-categorical variation in verbal complement orders is also attested in verb-final contexts in Mennonite Plautdietsch (MP), one such Continental West Germanic language (cf. Epp 1993, Kaufmann 2005). The corpus sentences presented here as (2a) and (2b) mirror the variability noted in (1c), with both v_1-v_2 and v_2-v_1 orders being attested:

Modal-infinitive construction, verb-final context

(2a) *Mi heat sikj daut soo aus wann doa waut opp´em Hoff* ***mucht**$_1$ **senne**$_2$*.
'It sounds to me as if there might$_1$ be$_2$ something on the yard.' (RE1972.S0000881)

(b) *Aus mien Hollända daut een Stoot toojehorcht haud, fruag he mi meteenst waut fer,ne Sproak soont* ***senne**$_2$ **mucht**$_1$*.
'Once my Dutchman had listened to that for a while, he suddenly asked me what (kind of a) language that might$_1$ be$_2$.' (RE1972.S0000929)

Formal statistical modelling of this ordering alternation is adopted in this study to assess the relevance of a number of factors proposed in the literature on VCC to bear upon the acceptability of particular alternants in Continental West Germanic languages. A sample of bipartite verbal complementation constructions occurring in verb-final contexts, such as (2), was drawn from a small, orthographically normalized and POS-tagged corpus of Mennonite Plautdietsch (approx. 124 000 tokens), representing the recent works of two prominent Canadian Mennonite authors, Reuben Epp (RE) and Jacob M. Fehr (JMF).

Queries of this corpus via POS tags were able to retrieve 4 711 unique instances of VC. Subsequent inspection and manual annotation of each such construction revealed 1 382 instances of VC involving a bipartite complementation structure in a verb-final context. These data were then coded for four distinct classes of predictors:

1) *Textual and authorial predictors:* author identity (i.e., RE or JMF), text identity, text genre (i.e., verse or prose), text translation status (i.e., composed in Plautdietsch or translated into Plautdietsch from another language)

2) *Morphosyntactic predictors:* morphological regularity of participant verbs, presence or absence of separable verbal prefixes, tense and subject agreement features of the finite verb, clause passivity, clause type (e.g., causal, temporal, etc.), complementizer identity, constructional context (e.g., modal-infinitive, auxiliary-participle), presence or absence of coordination involving the complement verb

3) *Lexical and semantic predictors:* verb lemma, verb class (i.e., lexical, modal, auxiliary), verb-pair mutual information scores

4) *Processing-related predictors:* verb frequency (of tokens and lemmata), Kullback-Leibler divergence from the ordering patterns of other verbs appearing in the same construction (serving as a measure of the degree to which a given order is unexpected in a particular constructional context), syntactic weight (via clause length in orthographic words), structural parallelism (via sentence distance from the immediately preceding instance of VC in the same text)

The alternation between v_1-v_2 and v_2-v_1 order was rendered statistically through generalized linear mixed-effects modelling (cf. Baayen 2008), treating the observed verb order as a binary dependent variable and all of the above factors as potential predictors. As well, adjustments for idiosyncratic differences be-

tween individual verbs (e.g., in their frequency and distribution across constructional contexts) were made automatically within the model by including both v_1 and v_2 as random effects.

4. Results and discussion

Summarized results of this statistical modelling are given in Table 1, presenting those fixed effects from the list given above which emerged as statistically significant predictors of verbal order.

Fixed Effect	Estimate	Std. Error	z-Value	p	
(Intercept)	5.12994	0.83442	6.148	7.85e-10	***
Author=RE	0.70619	1.58228	0.446	0.655372	
Syntagm=Modal-Inf	-0.10564	0.41579	-0.254	0.799446	
TextGenre=Verse	-1.31984	0.39968	-3.302	0.000959	***
V1ProportionInV1V2Order	-8.71853	1.05551	-8.260	< 2e-16	***
poly(ClauseLength, 2)1	-16.63070	3.60994	-4.607	4.09e-06	***
poly(ClauseLength, 2)2	10.83093	2.89583	3.740	0.000184	***
Coordination=CoordinatedV2	-0.98559	0.45916	-2.146	0.031835	*
V1ProportionalUseByRE	21.99279	6.56676	3.349	0.000811	***
V2Prefix=Separable	-1.58729	0.35311	-4.495	6.95e-06	***
V1Tense=Present	0.72030	0.26833	2.679	0.007376	**
StructuralParallel=NoPrecedent	-0.58674	0.26330	-2.228	0.025855	*
Author=RE : Syntagm=Modal-Inf	2.63046	0.66376	3.963	7.40e-05	***
Author=RE : Clause=Temporal	4.04043	1.71708	2.353	0.018619	*
Author=RE : TextGenre=Verse	-3.41530	0.72884	-4.686	2.79e-06	***

Table 1: Summarized fixed effects structure of the generalized linear mixed-effects model of bipartite VCC in verb-final contexts (n = 1382). Negative estimates favour v1-v2 order

Several observations might be made on the basis of this model. Significant differences appear to exist between the two authors represented in the corpus sample as regards their ordering preferences in particular syntagms (v_2-v_1 order is favoured by RE in modal-infinitive constructions), clause types (RE prefers v_2-v_1 order in temporal clauses), and text genres (verse commonly features v_1-v_2 order with both authors, although to a greater extent with RE).

These strong author-specific effects notwithstanding, it is still possible to observe more general effects of coordination (sentences with coordinated v_2 favour v_1-v_2 order), syntactic weight (longer clauses overwhelmingly favour v_1-v_2 order), complement verb prefixation (v_1-v_2 order is more likely to be observed if v_2 features a separable prefix), and, perhaps most unexpectedly, finite verb tense (present tense favours v_1-v_2 order, albeit weakly). Taken together, these results suggest that the represented speakers of MP are sensitive to specific constructional, morphological, lexical, and processing factors in their use of this alternation – a hypothesis which might be tested by extending the present model to further speakers, speech communities, and proposed predictors.

Indeed, while some caution must necessarily be exercised in interpreting results based upon such a limited linguistic sample, the quantitative, corpus-based approach taken in the analysis of these constructions might itself be advanced as one viable means of assessing the empirical adequacy of hypothesized motivations for VC ordering variation, insofar as these proposed predictors are capable of being consistently formalized and applied to the coding of corpus data. In the case of the present MP corpus data, such methods were effective not only in discerning even comparatively subtle effects (such as that of finite verb tense) upon verbal constituent ordering, suggesting directions for further research, but were able to do so given a relatively limited amount of data and a comparatively wide range of possible predictors. The extension of quantitative, corpus-based analysis to further speakers and speech communities might be hoped to render possible the simultaneous consideration of more extensive collections of linguistic data and more detailed sets of predictors than could receive comparable treatment by hands and eyes alone – and thus ultimately contribute to the perspicuity and empirical adequacy of the resulting analyses.

References

Baayen, R. Harald (2008): Analyzing linguistic data. A practical introduction to statistics using R. Cambridge: Cambridge University Press.

Bresnan, Joan / Cueni, Anna / Nikitina, Tatiana / Baayen, R. Harald (2007): Predicting the dative alternation. In: Boume, Gerlof / Krämer, Irene / Zwarts, Joost (eds.): Cognitive foundations of interpretation. Amsterdam: Royal Netherlands Academy of Science, 69-94.

Dixon, Robert M.W. (2006): Complement clause types and complementation strategies in typological perspective. In: Dixon, Robert M.W. / Aikhenvald, Aleksandra Y. (eds.): Complementation. Oxford: Oxford University Press, 1-48.

Epp, Reuben (1993): The story of Low German & Plautdietsch. Tracing a language across the globe. Hillsboro, Kansas: Reader's Press.

Kaufmann, Göz (2005): Der eigensinnige Informant. Ärgernis bei der Datenerhebung oder Chance zum analytischen Mehrwert. In: Lenz, Friedrich / Schierholz, Stefan J. (eds.): Corpuslinguistik in Lexik und Grammatik. (= Stauffenburg Linguistik 37). Tübingen: Stauffenberg, 61-95.

Wurmbrand, Susi (2004): West Germanic verb clusters: The empirical domain. In: É. Kiss, Katalin / van Riemsdijk, Henk (eds.): Verb clusters. A Study of Hungarian, German, and Dutch. Amsterdam: Benjamins, 43-85.

Wurmbrand, Susi (2005): Verb clusters, verb raising, and restructuring. In: Everaert, Martin / van Riemsdijk, Henk (eds.): The Blackwell Companion to Syntax. Malden, MA: Blackwell, 227-341.

Zwart, Jan-Wouter (2005): Continental West Germanic languages. In: Cinque, Guglielmo / Kayne, Richard S. (eds.): The Oxford handbook of comparative syntax. New York / Oxford: Oxford University Press, 903-946.

SILVIA HANSEN-SCHIRRA

Empirical profiling of LSP grammar

Abstract

This paper presents the analysis of LSP grammars in a three-dimensional way: first of all, the horizontal layer of LSP texts is reflected using sub-corpora from various disciplines. Secondly, the vertical layer is investigated by comparing scientific with popular-scientific discourse. Finally, the last dimension is the interlingual comparison between English and German – including English-German translations.

The multidimensional corpus is annotated for parts-of-speech, phrases structure and terminology and typical LSP register features are extracted. The findings are interpreted against the backgrounds of language typology, degree of expertise and interdisciplinary variation. Additionally, translations from the language pair English-German are evaluated against the originally produced language in terms of specific properties of translated text. It is tested whether translations are more explicit and easier to understand than originals in the source and target language.

1. Motivation

Languages for special purposes (LSP) are structured in a multi-dimensional way: first of all, the horizontal layer of LSP reflects the language usage of different disciplines, mainly focusing on the lexicon (e.g. Mayer / Reisen (eds.) 1999, Arntz / Mayer / Reisen (eds.) 2003 or Schmitt 2003). Secondly, the vertical layer defines the degree of expertise from a lexico-grammatical perspective (e.g. Gläser 1990 for English, Göpferich 1995 for natural sciences and Niederhauser 1997 for (popular-)scientific writing). Finally, the last dimension is the interlingual comparison between LSP texts (e.g. Trumpp 1998 or Biber 1995).

This paper presents a corpus-based investigation of the three LSP dimensions. Using multi-layer annotation, the grammar of LSP texts can be examined in addition to the lexicon. Based on empirical evidence, the LSP features of the sub-corpora are contrasted against the backgrounds of language typology, degree of expertise and diverging disciplines. Additionally, translations from the language pair English-German are evaluated against the originally produced language in terms of specific properties of translated text. The hypothesis here is that the translations are more explicit or simplified in terms of

LSP features compared to originals in the source language as well as in the target language (cf. Baker 1996). Moreover, it is tested whether the typical register features of the source language 'shine through' in the translations.

2. Multidimensional corpus design

The corpus under investigation has to fulfill many requirements. The first aim of the corpus analysis is to describe the differences between popular-scientific and scientific texts (focusing on the vertical layer of LSP). Secondly, different disciplines are compared to each other in terms of linguistic complexity and density (focusing on the horizontal layer of LSP). In a third step, the specificity of translated LSP text is compared to English and German originals – do the translations reflect the typical patterns of the source language (showing interference) or of the target language (showing standardization)?

Thus, the corpus contains English and German popular-scientific and scientific texts for the following disciplines: archaeology, astronomy, biology, chemistry, maths, medicine, physics, and technology. Additionally, English-German and German-English popular-scientific translations are included in the corpus. We collected 15 000 words for each discipline in each language and degree of expertise – altogether, the corpus comprises 540 000 words. This multidimensional corpus is processed in the following way: part-of-speech tagging, phrase chunking and terminology extraction were carried out for all sub-corpora. On this basis, the texts were investigated for the following typical and untypical LSP register features (e.g. Fluck 1997, Halliday / Martin 1993, Beneš 1981, Sager / Dungworth / McDonald 1980): term density; verbal vs. nominal style; frequency, length, and complexity of syntactic phrases; frequency, length, and complexity of sentences; nominalisations; compounding; pre- and post-modifications; frequency of pronouns; lexical density, type token ration, etc. Relevant results are discussed in the following chapters.

3. The role of the lexicon

The area of terminology has been described exhaustively with regard to the lexicon of LSP (e.g. Arntz / Picht 1991 or Mayer / Schmitz / Zeumer (eds.) 2002 for terminology management; Schmitt 2003 for lexicography and Reinke 1999 for the practical use of terminology within translation memories). Within this context, it is a commonly accepted technique to extract terms on the basis of LSP corpora (e.g. Heid et al. 1996 for monolingual lexicography; Bernhard

2006 for multilingual work). Even the extraction of bilingual terminology from aligned corpora has proven to be successful (e.g. Vintar 1999, Carl / Rascu / Haller 2004a). The output can then be used as lexical input for machine translation systems or as term banks for human translators (as a preprocessing step for terminology management). Moreover, parallel term extraction serves as a basis for the development of cross-linguistic ontologies (cf. Volk / Vintar / Buitelaar 2003) or for text mining in LSP (cf. Vintar et al. 2003).

The following examples describe how the corpus introduced above is used for terminology extraction and how this terminology in turn construes the grammar of LSP. We used the IAI terminology tool (cf. Haller 2006, Carl et al. 2004b) for extracting German and English terms from the popular-scientific translation corpus. The following table shows an excerpt from the resulting English and German term lists focusing on the term *hydrogen* (in German: *Wasserstoff*).

German term list	**English term list**
Wassermolekül	*hydrogen*
Wasserstoff	*hydrogen-bonding*
Wasserstoffatom	
Wasserstoffbindung	
Wasserstoffbrückenbildung	
Wasserstoffbrückenbindung	
Wasserstoffhalogenid	
Wasserstoffverbindung	

Table 1: Term extraction for English and German

The lists show that more terms are found for German than for English; this is, however, caused through typological differences between the languages involved: for German, all compounds including a terminological component are listed whereas the English multi-word units (see also Dias et al. 2000) are split up and only the terminological component can be found in the list (the other components are either not classified as terms or occur in other positions on the alphabetical list). These multi-word units can, however, be found by using n-grams, i.e. statistically calculating co-occurring terms (see Brants 2000 for a description of the methodology). On the basis of statistical processing, the following English multi-word terms including *hydrogen* are listed:

bi-grams	tri-grams
hydrogen atom	*normal hydrogen bond*
hydrogen bond	*strong hydrogen bond*
hydrogen bonding	*weak hydrogen bond*
hydrogen-bonding framework	*normal hydrogen bonding*
hydrogen chloride	*strong hydrogen bonding*
hydrogen fluoride	*weak hydrogen bonding*

Table 2: Statistically calculated multi-word terms for English

The n-gram lists show that terminological collocations are used in English where we can find compounds in German. The combination of term extraction and word alignment (which is included in the translation corpus described above) shows that the translation for *Wasserstoffatom* is, for instance, *hydrogen atom* and the translation for *Wasserstoffbrückenbindung* is *hydrogen bond*.

The examples discussed in this chapter show that terminological multi-word units, which can automatically be extracted using corpus linguistics and statistical techniques, are part of the nominal style which is preferably used in LSP (cf. Fluck 1997, Biber 1995, Sager / Dungworth / McDonald 1980, Beneš 1981, etc.). It is important to take complex terminological expressions into account when analyzing LSP grammar since they are the basis for construing nominal style and grammatical metaphor in expert communication (see Halliday / Martin 1993).

4. LSP grammar

As explained above, the lexicon plays an important role for coining the typical nominal style of LSP texts. For this reason, lexical density (LD), type token ratio (TTR), term density and term TTR are investigated in addition to the following lexico-grammatical, syntactic, and cohesive features: nominal vs. verbal style,[1] sentence length and cohesive reference. The hypothesis is that the findings for the following features are higher in scientific texts than in popular-scientific texts: LD, TTR, term density, term TTR, sentence length and nominal style. In contrast, verbal style and cohesive reference should be more frequently used in popularizations.

[1] For this distinction, parts-of-speech which are typically used in nominal phrases (nouns, adjectives, prepositions) are contrasted with verbal ones (verbs, adverbs, conjunctions).

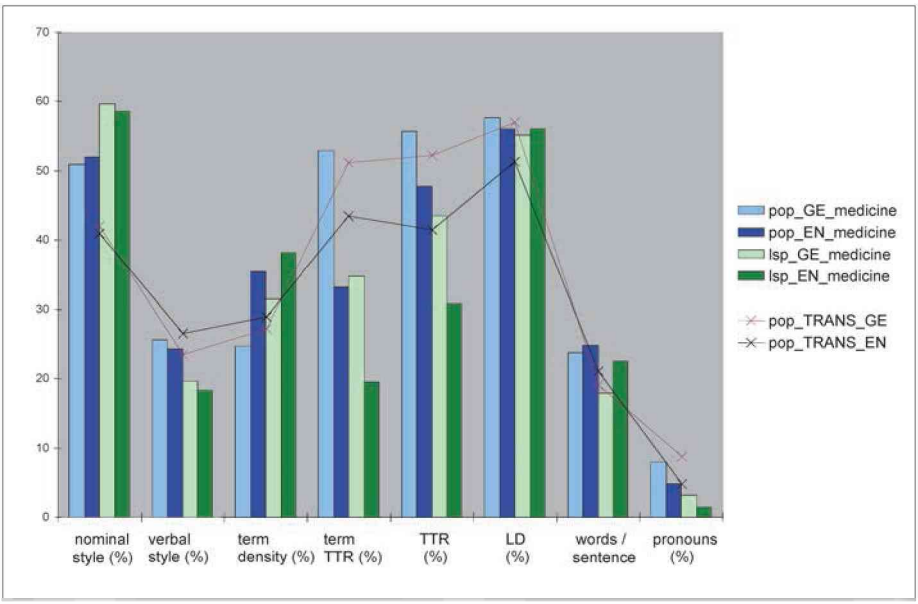


Figure 1: Differences in the vertical layer

Figure 1 shows the differences between medical scientific (abbreviated with "lsp") and popular-scientific texts (abbreviated with "pop"). The hypothesis formulated above concerning the vertical layer of LSP can be confirmed as follows: for both languages, scientific texts are characterized by higher values for nominal style and term density as well as lower values for verbal style and pronouns. The figures for sentence length, TTR, and term TTR contradict the hypothesis in such a way that they are higher in popular-scientific texts than in scientific texts. The reason for the high TTRs lies in the frequent use of paraphrases or metaphors in popular-scientific texts which explain the relevant terms by using other or less specialized terms. The following example from the English technology corpus of popular-scientific texts shows the increase in terms by using an explanatory metaphor:

> The atom is like a solar system, with electrons whirling around the nucleus like planets orbiting a star. (pop_EN_technology)

Moreover, contrastive differences can also be found: term density is higher in English popular-scientific and scientific texts than in German popular-scientific and scientific texts respectively. This may be explained by the problem displayed in Tables 1 and 2: during the automatic term extraction, English multi-word terms are separated into their single constituents. This distortion

can be clarified by taking the statistical n-gram analysis into consideration (see Chapter 3). Furthermore, TTR and term TTR are higher in German popular-scientific texts and in German scientific texts than in English popular-scientific and scientific texts respectively. This difference may also be explained by the above-mentioned analysis problem of German compounds and English multi-word units because each German compound is counted as an individual type whereas English nouns may be repeated in several multi-word units. Again, this distortion can be dealt with by including n-grams. In order to detect 'real' differences in language typology, a register-neutral reference corpus has to be used as tertium comparationis. LSP features can then be differentiated from colloquial language use on this basis.

Simplification can be found for the following features when comparing popular-scientific translations with comparable medical (popular-)scientific texts in the target language (see Figure 1): nominal style, TTR, and sentence length in English and German translated text (with lower values in the translations) and verbal style, term density, and LD in English translations only (with higher values for verbal style and lower values for term density and LD). The translations are less complex as far as these LSP features are concerned. The simplification strategy is more common in German-English translations than in English-German ones. However, whether these simplified translations are also easier to understand and how they are received in the target language has to be examined through psycholinguistic experiments.

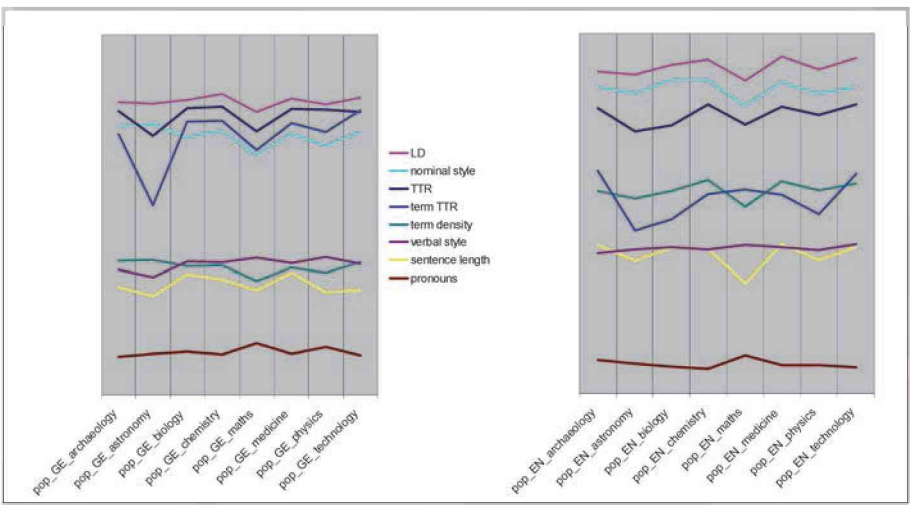


Figure 2: Differences in the horizontal layer

Figure 2 demonstrates that some disciplines use LSP features less frequently than others. This is shown by comparing the horizontal layer of LSP in English and German. In both languages, the popular-scientific sub-corpora of astronomy, maths and physics display the lowest values for the typical LSP features (LD, nominal style, (term) TTR, term density, sentence length) and in most cases higher values for the untypical LSP features (verbal style and pronouns). The following two examples from English popular-scientific writing show the differences in information-packing when physical texts are compared to medical ones (the former being less densely packed than the latter):

> New Theory Explains the Physics of Foam
>
> Light and fluffy foam is actually serious business. To ensure that canned draughts have predictable, lasting heads, for example, Guinness developed the Widget, a plastic ring that releases nitrogen into the beer as it is opened; it won the Queens Award for Technological Advancement in 1991. (pop_EN_physics)
>
> Men and women may metabolize fructose differently
>
> Short-term high fructose intake among young men resulted in increased blood triglycerides (fats) and increased insulin resistance, factors associated with an elevated risk for cardiovascular disease and type 2 diabetes, report Dr. Luc Tappy and colleagues. (pop_EN_medicine)

On the basis of these results it would be interesting to test the semantic affinity of different disciplines based upon their lexico-grammatical similarity. The results discussed here suggest that astronomy, maths and physics are characterized by a certain level of similarity – irrespective of the language. However, this assumption has to be corroborated by further investigations taking lexical, grammatical and semantic features into account.

5. Conclusions and outlook

This paper presented a corpus-based investigation of the horizontal, vertical and interlingual LSP dimensions. Typical and untypical LSP features could be detected on the basis of multi-layer annotation of the multi-dimensional corpus. The results were interpreted against the backgrounds of language typology (including translation specificity), degree of expertise and diverging disciplines.

As mentioned in Chapter 4, the product-based perspective introduced through the corpus study should be complemented by psycholinguistic experiments to test the comprehensibility of popular-scientific texts on the one hand and the

perception of translations on the other. Research questions can arise within this context. Are popular-scientific texts easier to understand than scientific ones? Do they meet the needs of a lay public? How are translations perceived in the target language? Do they fulfill the standards of the target language or do interference effects disturb the perception? Finally, the development of metrics for the classification of translated vs. original and expert vs. lay communication as well as the differentiation of languages and disciplines can be seen as the / a long-term goal of this study.

References

Arntz, Reiner / Picht, Heribert (1991): Einführung in die Terminologiearbeit. Hildesheim: Olms.

Arntz, Reiner / Mayer, Felix / Reisen, Ursula (eds.) (2003): Terminologie in Gegenwart und Zukunft. Ausgewählte Beiträge der DTT-Symposien 1989-2000. Köln: DTT.

Baker, Mona (1996): Corpus-based translation studies. The challenges that lie ahead. In: Somers, Harold (ed.): Terminology, LSP and translation. Amsterdam: Benjamins, 175-186.

Beneš, Eduard (1981): Die formale Struktur der wissenschaftlichen Fachsprachen in syntaktischer Hinsicht. In: Bungarten, Theo (ed.): Wissenschaftssprache. München: Fink, 185-212.

Bernhard, Delphine (2006): Multilingual term extraction from domain-specific corpora using morphological structure. In: Proceedings of the EACL. (=11th Conference of the European Chapter of the Association for Computational Linguistics). Trento, 171-174. Internet: http://acl.ldc.upenn.edu/eacl2006/companion/pd/22_bernhard_11.pdf (last visited: 01/2011).

Biber, Douglas (1995): Dimensions of register variation: A cross-linguistic comparison. New York: Cambridge University Press.

Brants, Thorsten (2000): TnT – a statistical part-of-speech tagger. In: Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000. (April 29-May 4). San Francisco, CA: Morgan Kaufmann Publishing Inc. Internet: http://dx.doi.org/10.3115/974147.974178 (last visited: 01/2011).

Carl, Michael / Rascu, Ecaterina / Haller, Johann (2004a): Using weighted abduction to align term variant translations in bilingual texts. In: Proceedings of LREC 2004. Lisbon, 1973-1976.

Carl, Michael / Hernandez, Maryline / Preuss, Susanne / Enguehard, Chantal (2004b): English Terminology in CLAT. In: Proceedings of the Workshop on Computational & Computer-assisted Terminology, Lisbon. Internet: http://www.iai.uni-sb.de/carl/papers/wsh.pdf (last visited: 01/2011).

Dias, Gaël / Guilloré, Sylvie / Bassano, Jean-Claude / Pereira Lopes, José Gabriel (2000): Combining linguistics with statistics for multiword term extraction: A fruitful association? In: Proceedings of Recherche d'Informations Assistee par Ordinateur 2000. Paris, France (12.-14. April), 1-20. Internet: http://www.di.ubi.pt/~ddg/publications/riao2000.pdf (last visited: 01/2011).

Fluck Hans-Rüdiger (1997): Fachdeutsch in Naturwissenschaft und Technik. Einführung in die Problematik der Fachsprachen und die Didaktik / Methodik eines fachsprachlichen Deutschunterrichts. Heidelberg: Groos.

Gläser, Rosemarie (1990): Fachtextsorten im Englischen. (= Forum für Fachsprachen-Forschung 13). Tübingen: Narr.

Göpferich, Susanne (1995): Textsorten in Naturwissenschaft und Technik: Pragmatische Typologie – Kontrastierung – Translation. (= Forum für Fachsprachen-Forschung 27). Tübingen: Narr.

Haller, Johann (2006): AUTOTERM – automatische Terminologieextraktion Spanisch-Deutsch. In: Gil, Alberto / Wienen, Ursula (eds.): Multiperspektivische Fragestellungen der Translation in der Romania. (= Sabest 14). Frankfurt a.M.: Peter Lang, 229-242.

Halliday, Michael A.K. / Martin, James R. (1993): Writing science: Literacy and discursive power. London / Washington, D.C.: The Falmer Press.

Heid, Ulrich / Jauß, Susanne / Krüger, Katja / Hohmann, Andrea (1996): Term extraction with standard tools for corpus exploration – Experience from German. In: Proceedings of the TKE '96 International Conference on terminology and Knowledge Engineering. Frankfurt a.M.: Indeks. Internet: http://zaunkoenig.ims.uni-stuttgart.de:65042/cocoon/imsbib/query.html?headline=Publications%20of%20the%20CorpLex%20group&language=en&format=html&keyword=CorpLex (last visited: 01/2011).

Mayer, Felix / Reisen, Ursula (eds.) (1999): Deutsche Terminologie im internationalen Wettbewerb. Akten des Symposiums des Deutschen Terminologie-Tags e.V., 24.-25. April 1998. Köln: DTT.

Mayer, Felix / Schmitz, Klaus-Dirk / Zeumer, Jutta (eds.) (2002): eTerminology. Professionelle Terminologiearbeit im Zeitalter des Internet. Akten des Symposiums des Deutschen Terminologie-Tags e.V., 12.-13. April 2002. Köln: DTT.

Niederhauser, Jürg (1997): Das Schreiben populärwissenschaftlicher Texte als Transfer wissenschaftlicher Texte. In: Jakobs, Eva-Maria / Knorr, Dagmar (eds.): Schreiben in den Wissenschaften. (= Textproduktion und Medium 1). Frankfurt a.M.: Peter Lang, 107-122.

Reinke, Uwe (1999): Überlegungen zu einer engeren Verzahnung von Terminologiedatenbanken, Translation Memories und Textkorpora. In: LDV-Forum 16, 1/2: 64-80.

Sager, Juan C. / Dungworth, David / McDonald, Peter F. (1980): English special languages. Principles and practice in science and technology. Wiesbaden: Brandstetter.

Schmitt, Peter A. (2003): Fachlexikographie in der Internet-Ära: Vom PC zum polytechnischen Großwörterbuch. In: Lebende Sprachen 3: 97-113.

Trumpp, Eva Cassandra (1998): Fachtextsorten kontrastiv. Englisch – deutsch – französisch. (= Forum für Fachsprachen-Forschung 51). Tübingen: Narr.

Vintar, Špela (1999): A parallel corpus as a translation aid: Exploring EU terminology in the ELAN Slovene-English parallel corpus. Proceedings of the 34th Colloquium of Linguistics. Frankfurt a.M.: Peter Lang, 839-848.

Vintar, Špela / Todorovski, Ljupčo / Sonntag, Daniel / Buitelaar, Paul (2003): Evaluating context features for medical relation mining. In: Proceedings of the 14th ECML / 7th PKDD. Workshop on data mining and text mining for bioinformatics. 22-26. September 2003. Cavtat-Dubrovnik. Internet: http://muchmore.dfki.de/pubs/ecml_final_new.pdf (last visited: 01/2011).

Volk, Martin / Vintar, Špela / Buitelaar, Paul (2003): Ontologies in cross-language information retrieval. In: Proceedings of the Workshop Ontologie-basiertes Wissensmanagement (WOW2003). Luzern. Internet: http://muchmore.dfki.de/pubs/w-onto-w.pdf (last visited: 01/2011).

OLGA O. BORISKINA

Noun Cryptotype Analysis as an Approach to Corpus-driven Modelling of N+V Collocations

Abstract

The paper presents a possible corpus-based approach to the study of N+V collocations traditionally referred to as metaphorical. The approach is both the continuation of Cognitive Metaphor Theory with reference to folk categories and the elaboration of B.L. Whorf's criteria for classifying lexicon in *cryptotype* vs. *phenotype* grammatical categories. The approach employs corpora data analysis and The Method of Nomina Abstracta Cryptotype Distribution (MoNACD). The method is to help explore the cognitive organization of the least unitary category of the lexicon (cryptotype), conceptual abstractions and their metaphorical representations.

1. Introduction

The corpora-related and lexical grammar-oriented project, currently in progress at the Department of Theoretical and Applied Linguistics (Voronezh State University, Russia), suggests an approach to advancing the problem of capturing *V+N combination regularities.* The combination regularities depend, to a certain extent, on the classification background of nouns. Grammatical research in pre-Corpus Linguistics was concerned chiefly with the study of grammatical categories which bear *morphological marking*, i.e., *outward representation in morphemes*, called *phenotypes* in the terminology of Whorf (1956: 68-72). The field studies of covert linguistic classes (*cryptotypes*) in world languages did not have a remarkable impact on Linguistic Theory. Corpus Linguistics has broadened the horizons of grammatical research to make it open to exploring, describing and comparing *cryptotypes* of world languages.

The main objectives of the project are to recognize the *cryptotype* noun classes in Contemporary English and to model the *distribution* of abstract nouns in such cryptotypes. Basically, MoNACD is applicable to the research of collocational propensity of nouns by drawing inferences from the use of nouns in language corpora which help to propose hypotheses about cryptotype intentions and collocational arbitrariness of abstract nouns. Since MoNACD is corpus-informed, the properties of nouns (Cryptotype activeness, collocation se-

lectivity and Radius Index) are grounded on the quantitative analysis of corpora. The complexity of qualitative characteristics and maxims of noun cryptotype distribution modelling allows the retracing and accounting for the stabilization and variation of cryptotype-driven behaviour of abstract nouns. The description of such distribution can throw light on the construction of our knowledge in mind and its metaphoric mapping in a language. So, on the one hand, the method is targeted at the better understanding of metaphors in modern languages; on the other, we believe that the description of cryptotype modelling of abstract nouns can contribute to the computational modelling of their combinability.

2. Defining a cryptotype

> A covert linguistic class may not deal with any grand dichotomy of objects, it may have a very subtle meaning, and it may have no overt mark other than certain distinctive "reactances" with certain overtly marked forms. It is a submerged, subtle, and elusive meaning, corresponding to no actual word, yet shown by linguistic analysis to be functionally important in the grammar. (Whorf 1956: 72).

The definition of a cryptotype provided by B.L.Whorf has been specified in research practice to meet the objectives of the project. A noun *cryptotype* is defined as a covert word class which lacks overt morphological marking. How is it recognized?

The first objective of the project is to recognize noun cryptotypes in Contemporary English. We attempted to attain the objective via verb capacity to classify (sort out) nouns as well as noun capacity to select verbs to co-occur with. Theoretically, a verb can project syntactical positions for nouns, and thus, the syntactical valency of a verb can be regarded as the classification principal of nouns (Apresjan 1967, 2008; Melčuk 1988). Accordingly, verb syntactical valency and semantics can be the key to noun cryptotype recognition (Kretov 1987). On the whole, a noun cryptotype is marked by *lexical selection*. The verb stems are regarded in the project as *classifiers* of nouns which can relate nouns to cryptotypes whereas nouns are considered to be apt to *select* the verbs to go with in accordance with their *cryptotype intention* to occur in a certain syntactical position the verb projects. It is plausible, therefore, to classify nouns on the grounds of their *realized valency* or *realized cryptotype intention.*

Consider the following examples:

> When everything is going fine for us, and God is blessing us, when our **life is flowing** smoothly, it's easy to have faith, it's easy then to trust – He wants **to pour His life** into our life so that our life can grow strong. – Now I'm going **to stream** Uncle Manfred's **life** for you, but before I do, here are some questions. As **life flooded back** into their daughter, Linda and Junki went limp with relief. – there are many other ways in which this technology could be used to **sprinkle life** into Chile's arid zones. – 'From his canvasses, **life spills out**'. – Antonio's family had been mortified by the way his **love life was splattered** across the papers. – Oakley's pale as a maiden, **the life's leaking from** him. – global superstar Justin Timberlake, has led the low-key actress's **private life** so that it **could be splashed** onto the tabloids.

The noun *life* is classified by the subject or object valency of the verb stems *to flow*, *to pour*, *to ooze*, *to stream*, *to leak*, *to splash*, *to splatter*, *to sprinkle*, *to flood* and thus related to the EL cryptotype LIQUIDUS. The conceptual abstraction which is represented by the noun *life* appears to be categorized in the English-speaking culture as a liquid, with features typical of the bodies of water. The corpus data is regarded as discourse evidence of the noun's relation to the cryptotype.

3. How is a noun cryptotype formulated?

According to Potebnja (1976) "the meaning of a word is subject to change, while its *inner form* remains". Because the *inner form* conserved in a word generated the word combinatory potential, it tends to influence the word combinability in modern languages. The strategy we implement is clustering verb classifiers with respect to their *inner form* retraced in the Oxford English Dictionary, Version 3.1. Clustering verbs of similar semantics appears to be a challenging task, which is feasible owing to the English verb analysis in Sinclair et al. (eds.) (1996). Additionally, the research in linguistic classification of the basic elements (fire, water, air, and earth) conducted at Voronezh State University (Boriskina / Kretov 2003) contributed to verb clustering.

Lexico-semantic verb clusters are formed on the basis of the cognitive and communicative relevance of a semantic feature (attribute) the verbs represent.

Below are six verb clusters with cryptotypes to which they attribute nouns:

1) The verbs with the inner form 'be capable of moving' would ideally be clustered in '*the verbs of motion*' with 52 representative lexemes (*to go*, *walk*,

come, *travel*, *follow*, *approach*, etc.). Thus, nouns which occur as their subjects are attributed to the cryptotype MOVENS, e.g., the noun *crisis* belongs to the cryptotype. Cf., *Then came the oil crisis.*

2) The verb stems representing acts of possession are clustered due to the semantic attribute 'able to own' (e.g., *to take*, *grab*, *hold*, *give*). These verbal classifiers relate nouns that occur as their subjects to the cryptotype HOMO TENENS. Cf., *This recession has taken a fragile sector and has made it even more fragile.*

3) The verbs which represent speech acts (e.g., *to say*, *answer*, *suggest*) relate nouns that occur as their subjects to the cryptotype HOMO LOQUENS. Such conceptual abstraction as law is categorized in English as being 'able to talk'. Cf., *there is an unwritten law which says sports cars for the "enthusiast" market have rear-wheel drive.*

4) The nouns that occur as objects of verbs representing acts of possession (e.g., *to take*, *grab*, *hold*, *give*) are attributed to the cryptotype RES PARVA. Such conceptual abstractions as effort, experience and growth are categorized in English as objects 'small enough to fit the hand, to be transportable with a hand'. Cf., *I give the same effort to every class, but the results don't always match; her support enabled many talented men to obtain important field experience. You have taken your personal growth and packaged it in a way that benefits thousands of other people.*

5) The verbs cluster representing the acts of penetrating as a sharp-pointed object does (*to prick*, *stab*, *stick*, *pierce*, *thrust*, *spear*, *pin*, *puncture)* relate nouns to the cryptotype RES LONGA. Such conceptual abstractions as pain and light appear to be categorized in English as sharp-pointed. Cf., *Pain stabbed at her chest, twisting like a barbed snake; pricking pain; pain punched him in the gut; he felt the pain spearing through him again; a sudden spasm of pain pierced her heart, subduing her anger; it wasn't long before the pain of loss began to penetrate the anaesthetic of crowds; a white light stabbed my eyes; bright neon club light thrusting into the darkened back room; a sudden light pierced the darkness; the light from the kitchen window punched out into the early-morning darkness; sunlight penetrates down into clear ocean at most about 200 meters.*

6) The verbs which represent the way liquid exists (*to flow*, *flood*, *pour*, *leak*, etc.) relate nouns which occur as their subjects and objects to the cryptotype LIQUIDUS. Cf., *images and memories flooded his mind in a brilliant whirl.*

4. Corpus analysis of verb-noun collocations

Within the current project the possible noun-verb co-occurrences are examined in the *Corpus of Contemporary American English* and the *British National Corpus*. Extraction of collocations from corpora requires each noun to be tested on its occurrence in the subject and object syntactical positions of each of the classifiers (lemmas) of six above-mentioned noun cryptotypes. The results of the corpus query have been stored in MS Word format for the further generation and maintenance of the example subcorpus (Bank of *cryptotype-bound* V-N collocations).

5. The Method of Nomina Abstracta Cryptotype Distribution

The other objective of the project is to study the cryptotype distribution of nouns within each cryptotype and across them. Cryptotype distribution of nouns is the result of linguistic categorization in terms of metaphoric mapping of our knowledge. Presumably, categorization is not only the act of attributing an element to a set, but primarily the naïve and at the same time sophisticated process of formation of the sets (categories) on the basis of similarity and analogy among the essential attributes of entities. The cryptotypes under consideration within the project framework are defined by *global analogies*: we have taken the challenge of ignoring the 'surface semantic' differences of words to find similar *global internal semantic attributes* (analogies), which reflect the cognitive and communicative background of mythological thinking the nouns share. Thus, the cryptotype incorporates nouns of diverse semantics and themes which bear combinatory resemblance. Members of a cryptotype differ in the degree to which they are typical of a category in as much as they differ in the degree to which all members must share the class attribute. The graded membership of nouns in a cryptotype can be a useful form of vagueness as it helps to better capture the idea of typicality. It explains why a noun can be attributed to more than one cryptotype.

6. Why study cryptotypes of English nouns?

First of all, the cryptotype noun classes in English and other languages of deficient morphology can appear to be overt (morphologically marked) in other languages. Secondly, the noun cryptotypes of the English language are culturally constructed. So, by studying the cryptotype distribution of abstract nouns,

we learn about the cultural implications of a diverse language community such as the English-speaking world. Finally, the study of *cryptotype* can bring recurrent *word combination regularities* to light. Native English speakers' awareness of these regularities is "sensed rather than comprehended", which makes them non-applicable for the non-native speakers' discourse production or human-computer interaction. Thus, a collocation matrix has to be introduced, which could be used for foreign language learning purposes and, ideally, would be systematically detected by computer.

From a computational perspective MoNACD allows:

1) the formalizing of the metaphorical potential of abstract nouns;
2) the study of the cryptotype-driven syntactical behaviour of nouns by comparing their intra-cryptotype and cross-cryptotype profiles;
3) the modelling of the noun combinatory dynamics in terms of the word tendency to alter its collocational scope;
4) the implementation of a computational model of *cryptotype-bound* V+N metaphorical collocations of the English language.

7. Conclusion

On the whole, the proposed approach to the study of V+N collocations allows one to view the process of categorization of conceptual abstractions through the lens of cognitive computer science. Apparently, MoNACD is one of the viable directions of corpus-based methodologies for linguistic description because the study of word-classes marked by *lexical selection* appears credible in corpus-based and corpus-informed research. It is expected to lead to interesting insights into the cultural matrix of knowledge representation. The real mission of the Noun Cryptotype Analysis is the computer implementation of the cryptotype distribution of nouns, which should be relevant to the demands of text processing, so that the computer could conform to the co-occurrence regularities of recurrent words recognized by the English-speaking community, and follow them in speech synthesis. The cryptotype distribution of nouns determines, to a large extent, the combinatory potential of nouns, and can, therefore, be applied for Natural Language Processing as well as Natural Language Generation purposes.

References

Apresjan, Jurij D. (1967): The experimental study of the meaning of the Russian verb. Moscow: Nauka.

Apresjan, Jurij D. (2008): Systematic lexicography. Oxford: Oxford University Press.

Boriskina, Olga O. / Kretov, Alexei A. (2003): Linguistic categorization: Mind through the prism of cryptotype. Voronezh: Voronezhskij universitet.

British National Corpus. Internet: http://corpus.byu.edu/bnc/ (last visited: 01 / 2010).

Corpus of Contemporary American English. Internet: http://www.americancorpus.org/ (last visited: 01 / 2010).

Kretov, Alexei A. (1987): The role of metaphor in lexico-semantic organization. In: Aspect 2: 61–70.

Melčuk, Igor (1988): Semantic description of lexical units in an explanatory combinatorial dictionary: Basic principles and heuristic criteria. In: International Journal of Lexicography V.1, 3: 165-188.

Potebnja, Alexandr A. (1976): Thought and language. Moscow: Nauka.

Sinclair, John et al. (eds.) (1996): Collins COBUILD Grammar Patterns 1: Verbs. London: HaperCollins.

Whorf, Benjamin Lee (1956): Language, thought and reality: Selected writings of Benjamin Lee Whorf. Cambridge, MA: MIT Press.

SIAW-FONG CHUNG / YU-WEN TSENG

Learning Prepositions: A Corpus-based Study in Taiwan EFL Contexts[1]

Abstract

This preliminary corpus-based study aims to explore how the English preposition *to* is used by Taiwanese undergraduate learners of English. Through analyzing the collocations of *to* on the immediate left (L1) and right (R1) in a learner corpus built from essays collected from undergraduate (sophomore) learners of English, we counted their total occurrences in the corpus and the frequency at which a particular word appears on the L1 or R1 of *to*. We also observed that the collocates most frequently combined by the learners are used correctly, as learners' errors only occur with words appearing in lower frequency (e.g. **become to, *enter to*). In addition, five hundred instances were randomly selected to further scrutinize the distributions of sense and errors committed, from which the influence of language transfer and some patterns of how preposition *to* is used are examined.

1. Introduction

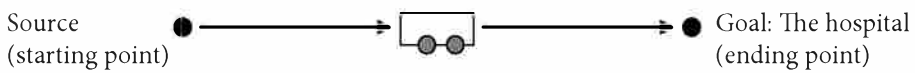
Language transfer (cf. Gass / Selinker (eds.)1983, Kellerman / Perdue 1992) is an issue widely discussed in second language acquisition, and preposition is one of the areas in which cases of language transfer can be found. Language transfer, in a basic sense, refers to the phenomena of transferring the structural patterns of native language to that of the target language. Selinker (1983) discussed three possible kinds of language transfer. Positive transfer is a predominant pattern of one of two alternative linguistic items in the native language which is a non-error compared with the target language; negative transfer is identified as a process in which the predominant pattern of one of two alternative linguistic items in the native language is an error in target language. The third type, neutral language transfer, differs from the previous two in that there is no predominance of either of two alternative linguistic items in the native language. On the

[1] We would like to thank the NCCU Foreign Language Learner Corpus Project and the National Science Council project of Siaw-Fong Chung (Project Number: NSC 97-2410-H-004-001) for supporting this work. We would also like to thank the members from Corpus Research Group, the anonymous reviewers and the participants from the Grammar and Corpora 2009 conference for their comments on this work.

other hand, Odlin (1989) further elaborated on the classifications of learners' language performance on negative transfer. It could be underproduction, rendering avoidance strategy due to unfamiliarity with certain structures, overproduction of particular language forms to avoid using unfamiliar structures, substitution of the target language forms with a native language one, calque, which refers to a word in a language as a translation of a word in another language, and alternations of structures. For language avoidance, Schachter (1974) proposed that in the analysis of learner language, in addition to focusing on the language produced by learners, those linguistic structures or expressions that are consistently avoided in usage are worth tending to as well.

Another issue which is usually related to preposition is the image schema theory (Johnson 1987, Lakoff 1987). It is claimed that a large number of metaphorical expressions are instantiated from some basic image-schemas, as the spatial constructs provided by Johnson (1987), including CONTAINER, PATH, POINT, SURFACE, etc., from which more abstract concepts are mapped and metaphors projected. There is a schematic pattern in which an internal structure is present. For example, the PATH schema involves a starting point, an end-point, and a sequence of locations that link the starting point to the end-point. When we begin at the starting point and arrive at the end-point, we would pass through these locations that connect these points. Expressions with the preposition *to* correspond to the PATH schema, as illustrated in example (1):

(1) Your brother <u>drives</u> his car **to** the hospital fast. (S008_010.txt)



In example (1), *your brother* does the motion of driving from a starting point to reach the goal of this path, or the end-point, *the hospital.* Throughout this path to the hospital, *the car* passes through many different locations connected by this motion.

From these studies one can see that prepositions have been examined in different facets, but not many studies have examined prepositions used by learners. This study attempts to fill this gap by examining the preposition *to* found in EFL students' writing. More specifically, this paper intends to answer the following questions: How do the senses of *to* pattern in a learner corpus? What is

the frequency of the collocates of *to* and what does the frequency imply? Furthermore, we also attempt to answer what might cause the occurrences of errors with regard to the use of *to*, if any.

2. Methodology

In order to observe how Taiwanese EFL learners use the preposition *to* in English, essays were collected from undergraduate (sophomore) learners of English at National Chengchi University, Taiwan. The purpose is to observe the collocational patterns of *to* and the possible errors which occur due to several causes. A corpus was built based on this purpose. Both Wordsmith (Scott 1999) and AntConc (Anthony 2005) tools were used for the analyses discussed here. One hundred and fourteen essays were collected and they have amounted to 69 407 tokens with 6 295 word types. Its type-token ratio, an index used to indicate the variedness of vocabulary in texts, is about 9.07%.[2] On average, each essay constitutes about 609 words. For the preposition *to*, 2 289 instances were found and this shows a coverage of 3.30% from all the words in the whole corpus.

3. Results

The results will be discussed in terms of the sense distributions of *to*, the analysis of the collocates of *to*, and an investigation into learners' errors.

3.1 Sense Distributions

In order to scrutinize the data further, five hundred instances of *to* were randomly selected from the total 2 289 for further analyses. As a precaution not to select similar data from the same text or from the same learners, these 500 random examples were selected based on the re-sorting of a list of random integers generated by Excel in which every concordance line was assigned a random integer. They were first grouped according to the senses categorized in Merriam-Webster's online dictionary (2009), as in Table 1. If an error is found such as (2), the error will be singled out as a separate category.

(2) *The idea that one **to** have talks with his inner self is valued... (S009_004.txt)

[2] See also http://www.speech-therapy-information-and-resources.com/lexical-density.html (last visited: 10/2010 for lexical density). We are grateful to Professor Hengsyung Jeng from National Taiwan University for pointing out this reference to the first author.

	SENSE	COUNTS	%
1.	Movement, direction, or contact	66	13.2
2.	Purpose, tendency, or result	104	20.8
3.	Indication of position or relation in time	0	0.0
4.	Addition, connection, belonging, or response	29	5.8
5.	The extent or degree	22	4.4
6.	Comparison, correspondence, or proportion	23	4.6
7.	Application of an adjective or noun, receiver, the relation of a verb to its complement	100	20.0
8.	Infinitive	138	27.6
9.	Errors	18	3.6
	TOTAL	500	100

Table 1: Sense Distributions of the Preposition *to* (According to Major Categories in Merriam-Webster)

The preposition *to* has the highest frequency used as an infinitive (27.6%), indicating that the to-infinitive combinations are frequently used by the learners, as in (3).

(3) ... the purpose that people <u>want</u> **to** improve their appearances is unchangeable. (S010_004.txt)

The preposition *to* denoting intention usually involves the volition of the subject and what the subject intends to do. In this example, *improve their appearances* acts like the goal that lays in the future point of the path. The second highest of all is the sense representing purpose, tendency, or result (20.8%), exemplified below. In this example, *a huge sum* acts as a goal on this path that *the tiny extra tax* could possibly attain.

(4) ... the tiny extra tax could finally <u>accumulate</u> **to** a huge sum. (S009_003.txt)

From the randomly selected five hundred instances, eighteen errors were found in the current analysis, constituting 3.6% of errors from all the instances analyzed. Among these, eight instances (44.5%) seem to express the infinitive sense, one (5.5%) denotes a movement in time, four (22.2%) show the purpose or result, and the remaining five (27.8%) indicate the application of a noun or

an adjective. The instances for errors appearing in the last three senses (movement, purpose and application respectively) are shown in (5). (The infinitive sense has been listed in (2).)

(5) (a) ?Then, we finally understood what Nash equilibrium is **to** the end. (S007_008.txt)

(b) *...once they don't follow the trend, they are **to** be lagged behind. (S009_004.txt)

(c) *...we will begin to have excessive hair loss, and our abilities **to** memory will decrease. (S005_004.txt)

Example (5a) shows the movement in time, which will be discussed in the error analyses section. In (5b) the learner uses *to* to express the result of change, while in (5c) the meaning of the noun before *to* is related to the following one.

3.2 Collocates Analysis

For all these instances of *to*, we analyzed their most frequently appearing collocates on the immediate left (L1) and right (R1), as given in Table 2. For each collocate, its total occurrence in the corpus is given (e.g., 472 for *have* on the L1 of *to*). From these total instances, the frequency at which this word appears in the L1 or R1 position is also given (e.g., 91 out of 472 instances of *have* (thus, 19.27%) appear after *to*).

WORD	TOTAL	L1	%	WORD	TOTAL	R1	%
have	472	91	19.27	*smoke*	201	54	26.86
want	104	72	69.23	*do*	210	52	24.76
need	75	40	53.33	*tell*	112	41	36.60
go	75	33	44.00	*make*	127	40	31.49
used	40	21	52.50	*get*	121	38	31.40
according	**38**	**38**	**100.00**	*have*	472	34	7.20
tend	**20**	**20**	**100.00**	*go*	75	30	40.00
going	**20**	**20**	**100.00**	*learn*	76	28	36.84
choose	42	16	38.09	*live*	83	27	32.53
start	25	16	64.00	*grow*	61	26	42.62

Table 2: Top Ten Collocates on the First Left (L1) and First Right (R1) of the Keyword to

Some collocates in Table 2 have 100% usage with *to*: *according*, *tend*, *going* (in bold), as shown in example (6).

(6) According **to** US Food and Drug Administration, "healthy" products are certified with some criteria for limiting amount of fat or cholesterol... (S003_004.txt)

Learners knowingly use these collocates before *to* and they usually learn them as chunks in class (cf. Lewis 2002). Comparing all the collocates on the left and right of *to* in Table 2, it can be seen that most of the collocates in the L1 position have generally higher percentages than those in the R1 position. For *have*, which appears on both the L1 and R1 of *to*, the percentage is still higher in the L1 position (19.27%) than in the R1 position (7.20%). Thus, *have to* is higher in percentage than *to have*.

In addition, only a small number of words appearing in the L1 position (shaded in Table 2) has percentages lower than 50, while all the words appearing in the R1 position have percentages lower than 50. This result shows that the types of collocates appearing before *to* (L1) are less varied than those after *to* (R1). For example, one can say *have to*, *want to*, and *need to* with a combination of collocates after *to* (*have / want / need to smoke*, *do*, *tell*, and *make*). Therefore, our results in Table 2 not only reflect this linguistic phenomenon, but also provide a methodology to examine the varied forms appearing before and after a target word.

We also found that most of the combinations in Table 2 are correctly used by our learners, and learners' errors only occur with words appearing in lower frequency (not in Table 2). Two examples are *become to* (as in **become **to** leave home*) and *enter to* (as in **to enter **to** college*) in the corpus. From these examples, it is possible that learners sometimes try to avoid using structures they are unsure about and this is reflected in the examination of specific linguistic forms in corpora. Kleinmann (1977) suggested that linguistic structures avoided by learners are those patterns predicted to be more difficult in contrastive analysis. Moreover, Kleinmann also agreed that this avoidance was connected to the affective state of the learner. For example, it is related to one's confidence to produce the pattern. In the section that follows we look further into the errors made by learners.

3.3 Error Analyses

As discussed, only eighteen instances of erroneous combination with *to* were found, further proving that they tend to occur in lower frequency. Among these cases, a large number of them seem to be the evidence of negative language transfer from the learners' native language, Mandarin, in which the equivalents of the preposition *to* are varied. It can be translated as *dao*4 'to reach a place; up to (a figure or time)', *dui*4 'for or as regards', and *rang*4 'to cause or to make someone do something'. Therefore, learners might use *to* to express similar meaning, as shown in (7).

(7)(a) ?Then, we finally understood what Nash equilibrium is **to** the end. (S007_008.txt)

(b) *Another reason of their individual performing styles is their unique interpretation **to** the roles they plays. (S007_002.txt)

In example (7a), *the end* functions as an indication of time instead of the goal of this process, so it should be *"in" the end*. The expression with *dao*4 in Mandarin *dao*4 *zui*4 *hou*4 (*to the end) might be the cause of confusion between the preposition *to* and *in*. In (7b), *the roles they plays* is the main issue to be interpreted (in a unique way), so learners included the preposition *to* to indicate this relation, which is usually expressed with *dui4* in Mandarin *dui*4 *jiao*3 *se*4 *de quan*2 *shi*4 (**interpretation to the roles*). It could thus be a possible reason of where this error comes from.

In addition to cross-linguistic influence, there are some cases which might be the results of an overgeneralization of the syntactic structure *VERB to VERB*, as in example (8).

(8)(a) *When we grow older and older, we <u>become</u> **to** <u>leave</u> home and make our own social life... (S005_002.txt)

(b) ?I <u>regretted</u> **to** <u>tell</u> her the truth because it has made her preoccupied for a long time, and blocked me from life routine. (S010_010.txt)

Learners applied the *VERB to VERB* construction to whatever verbs they use regardless of the different meanings and syntactic structure implied in a specific verb. For the verb *become*, it is usually followed by a description of what the subject will be like instead of an action that he or she is going to do. For the verb *regret*, there is a distinction between the forms of its complement that tend to be confusing for learners.

Two instances of misusing the infinitive are found, as in example (9), which could be explained by employing the image-schema theory.

(9)(a) *... you have better **to** choose online shopping. (S005_005.txt)

(b) *Besides, I also have lots of fun **to** read the articles that the BBS users have written everyday. (S004_004.txt)

In (9a), the learner used *to* as a connection that brings the end-point *choose online shopping* of this motion; however, the collocation *have better* does not entail a path for this movement. The same goes for *to* in (9b) as the motion *read the articles* is what the agent has been doing already, rather than the goal the agent intends to carry out. These analyses reveal that learners mistakenly conceived these events or processes on a path as something or someone that might change from one point to another or appear at the future end. Although misuses occurred in lower frequency, they provide evidence that learners were unsure whether it is correct to involve this path when using *to* with certain collocation.

4. Conclusion

This study, though carried out as a pilot study, will be expanded when larger amounts of corpora data are collected. However, based on the analyses above, some problems remain unsolved which will be our suggestions for future studies. For the analysis of sense enumeration for *to*, we found that it is sometimes impossible to distinguish the semantic meaning from the syntactic meaning. For example, one possible suggestion for this is to apply image schema theory, through which we might have different explanations of the preposition *to*. In addition to the senses, we also observed that some uses of *to* are affected by the meanings of the words around it; thus, the concept of coercion (Pustejovsky 1995) may become useful in explaining this phenomenon. Moreover, though this study is advantaged in its use of corpora data for error analyses, only a few errors were found in the two-word analysis. If we further examine three- or four-word combinations of *to*, more errors may be found, which will be worth exploring in the future.

References

Anthony, Lawrence (2005): AntConc: Design and development of a freeware corpus analysis toolkit for the technical writing classroom. In: Proceedings of the International Professional Communication Conference (IPCC), Limerick, Ireland. Limerick: International Professional Communication Conference 2005, 729-737.

Gass, Susan / Selinker, Larry *(eds.)* (1983): Language transfer in language learning. Rowley, MA: Newbury House.

Johnson, Mark (1987): The body in the mind. Chicago / London: The University of Chicago Press.

Kellerman, Eric / Perdue, Clive (1992): Special issue in cross-linguistic influence in second language acquisition. New York: Pergamon.

Kleinmann, Howard H. (1977): Avoidance behavior in adult second language acquisition. In: Language Learning 27: 93-107.

Lakoff, George (1987): Women, fire and dangerous things: What categories reveal about the mind. Chicago: University of Chicago Press.

Lewis, Michael (2002): The lexical approach. Boston: Thomson Heinle.

Merriam-Webster Online (2009): Internet: http://www.merriam-webster.com/dictionary/to (last visited: 15.06.2009).

Odlin, Terence (1989): Language transfer. Cambridge: Cambridge University Press.

Pustejovsky, James (1995): The generative lexicon. Cambridge, MA: MIT Press.

Schachter, Jacqueline (1974): An error in error analysis. In: Language Learning 24: 205-214.

Scott, Michael (1999): Wordsmith tools Version 3. Oxford: Oxford University Press.

Selinker, Larry (1983): Language transfer. In: Gass / Selinker (eds.), 33-53.

SVETLANA SAVCHUK

The *Russian National Corpus* as a Tool for Research on Grammatical Variability*

Abstract

The paper presents the *Russian National Corpus* (RNC) as a tool for research on grammatical variability. The RNC has all the necessary quantitative and qualitative characteristics to provide an adequate set of examples for various types of linguistic research. Being a representative collection of texts, the RNC reflects Russian language usage in two dimensions: 'horizontally' (in functional varieties) and 'vertically' (from a historical perspective).

Based on the corpus data, the problem of grammatical variation can be divided into three aspects:

1) the setting of the correlation of variants in contemporary language usage,
2) the study of the development that occurred during a certain period,
3) the comparison of these findings with the recommendations found in dictionaries and grammar manuals in order to evaluate the adequacy of these recommendations in real usage.

The capabilities of the RNC for linguistic investigation are demonstrated by means of corpus-based analyses of variants of genitive plural forms of masculine nouns (the names of those belonging to some branches of the armed forces have been chosen as an example).

1. Introduction

The corpus-based approach appears extremely useful and fruitful when researching grammatical variability and correlation between normative recommendations and real language usage. The following three aspects should be discussed in more detail:

1) If the focus is on the synchronic aspect, the co-existing variants should be studied concerning their distribution among different spheres of functioning, as well as social, professional variants of the language, etc.

* This work was supported by the Fundamental Research Program in the Department of History and Philology at the Russian Academy of Sciences (RAS) "Text in sociocultural environment: levels of historical, literary and linguistic interpretation" and by the Fundamental Research Program of the Presidium of the RAS (project "Russian language of the 18th century: corpus-based study of lexical and morphological variability").

2) When dealing with the problem in its diachronic aspect, some changes in usage, appearance and disappearance of the variants should be taken into consideration, as well as the changing of correlation between several variants, increasing and declining trends, and so on.

3) This aspect concerns the evaluation of the different variants with regard to a standard, or so-called 'codified norm' fixed in dictionaries and grammar manuals. A variant can be codified in the literary language or remain uncodified. There is a natural discrepancy: some variants can fluctuate over a long period of time while normative estimations of these variants change continually. Consequently, the evaluation of normative recommendations should also be regarded from synchronic and diachronic points of view, and in relation to real usage.

As far as statistical values are widely accepted as objective indicators of the current distribution of language phenomena, modern large text corpora seem to be a reliable tool for the research of linguistic norms and variation. Naturally, the reliability of this research depends on the parameters of the corpus: its capacity, coverage, and the linguistic information represented in it. The Russian National Corpus has all necessary quantitative and qualitative characteristics to provide an adequate set of examples for various types of linguistic research.

2. The Russian National Corpus: main parameters

A large group of specialists from Moscow, St. Petersburg, Voronezh and other Russian university centers have been creating the Russian National Corpus (RNC) within the program of the Russian Academy of Sciences since 2003. Although the project is still in progress, the corpus is already being used for research and educational purposes.

The RNC meets all the requirements for large contemporary text corpora, such as:

1) Large size
2) Representativeness
3) Linguistic annotation
4) Query tools

1) As far as the size of the corpus is concerned, the RNC contains approx. 170 million tokens at present (as of 2009).

2) The RNC is a representative corpus. It only includes complete texts of different language forms (spoken, written, and electronic) and of different functional spheres: fiction, journalism, memoirs, academic writing, administrative documents, religious texts, poetry, everyday dialogues, TV-programmes, broadcasts, etc.

3) The RNC is an annotated corpus: all texts are supplied with different types of linguistic annotation.

   Metatextual annotation refers to the text as a whole and includes information regarding the author's name, sex, age or date of birth, text characteristics (date of origin, functional sphere, text type, genre, domain), etc.

   Morphological annotation is performed automatically by a parser developed for modern Russian texts and based on Zaliznyak's *Grammatical dictionary of Russian* (1977 / 2003). The morphological information consists of four groups of tags:

   a) Lexeme (the lemma and the part of speech to which it belongs);
   b) Grammatical features of the lexeme (e.g., gender for nouns and transitivity for verbs);
   c) Grammatical features of the word-form (e.g., case for nouns and number for verbs);
   d) Information concerning non-standard forms of the lemma, orthographic variations, etc.

   Semantic annotation is performed automatically by 'Semmarkup', a software program by A. E. Poliakov which uses the semantic dictionary of the corpus. There are three groups of tags assigned to words:

   a) Class ('proper name', 'reflexive pronoun', etc.);
   b) Lexical and semantic features (thematic class of the lexeme, indications of causality or assessment, etc.);
   a) Derivational features ('diminutive', 'adjectival adverb', etc.).

   As a result, most words in a text are tagged with a number of semantic and derivational parameters such as 'person', 'substance', 'space', 'diminutive', 'verbal noun', etc.

   Sociological annotation is only specific to corpora of spoken language. It is assigned to different speakers' utterances and characterizes a word usage with regard to the sex and age of a speaker (if this information is available).

Sociological annotation allows a user to create his/her own sub-corpora by various parameters or their combinations: by a speaker's sex, age, or year of birth (this option is only available for movie transcripts), etc.

Accentological annotation is used in the Accentological Corpus. According to this annotation, each word is supplied with stress marks making it possible to carry out different kinds of search requests and retrieve data concerning stressed or unstressed word-forms in combination with grammatical and semantic features.

4) The corpus is available for all users at the following site: http://ruscorpora.ru. The search system is provided by the 'Yandex' server. Users can create their own subcorpora based on particular metatextual parameters and then run queries for words, grammemes and semantic features in various combinations, receiving contexts as query results.

3. The Russian National Corpus: composition and capabilities in the research of variation

The Russian National Corpus consists of the following subcorpora (as of 2009):

- Corpus of modern written texts (1950-2008): 97.4 million words
- Corpus of spoken language (1930-2008): 8.5 million words
- Corpus of written texts (18th century to the first half of the 20th century): 68 million words of which:
  - 26 million words are texts from the 19th century
  - 40 million words are texts from the first half of the 20th century
  - 2.6 million words are texts from the 18th century
- Poetry corpus: 3.2 million words
- Accentological corpus: 5.3 million words
- Dialect corpus: c. 200 000 tokens
- Parallel aligned corpus: c. 5.3 million words

Being a representative collection of texts, the RNC reflects Russian language usage in two dimensions: 'horizontally' (in functional varieties) and 'vertically' (from a historic perspective). Based on the corpus data, the problem of grammatical variation can be divided into three aspects:

1) the distribution of the correlation of variants in contemporary language usage;
2) the development of variants within a certain period;
3) the comparison of these findings with the recommendations of dictionaries and grammar manuals in order to evaluate the adequacy of these recommendations in real usage.

The capabilities of the RNC in linguistic investigations are demonstrated by two examples of corpus-based variant analyses of one of the 'weak points' of grammatical norm.

4. A corpus-based study of variants of genitive plural forms

4.1 Variants of genitive plural forms of masculine nouns

In modern literary Russian there are three variants of genitive plural masculine endings: *-ov*, *-ej*, zero (-Ø). The principle of selection, which was discovered by Jakobson (1956/1984: 135-140), has been adopted by grammarians and is used in grammatical descriptions: if there is a -Ø ending in the nominative singular, there is a non-zero ending in the genitive plural, and vice versa, a non-zero ending in the nominative singular involves a -Ø ending in the genitive plural.

According to this, generic forms with the ending *-ov* are standard for most masculine nouns with stems ending in a hard consonant or [j], and forms with the ending *-ej* are standard for masculine nouns with stem-final soft consonant or *ж*, *ш* (Shvedova (ed.) 1980: 498, Andrews 2001: 34). According to Zaliznyak (1967: 219), 97.3 % of masculine nouns have standard genitive plural forms with non-zero endings (Zaliznyak 1967: 219).

Genitive plural forms of masculine nouns with the -Ø ending are the exceptions to this rule because they have the same ending as nominative singular forms. According to Graudina (1976/2000), there are about 200 nouns with the -Ø ending in contemporary written and spoken language, which belong to several semantic groups:

1) Names of people as members of different associations – ethnic, military, political: gen. pl. *грузин* 'Georgians', *бурят* 'Buryats', *румын* 'Romanians'; *гусар* 'hussars', *драгун* 'dragoons', *кадет* 'Cadets'.

2) Names of some paired items: gen. pl. *чулок* 'stockings', *ботинок* 'boots', *брюк* 'trousers'.

3) Names of some measurement units in combination with numerals: gen. pl. *300 грамм* '300 grammes', *40 мегабайт* '40 megabytes', *20 рентген* '20 roentgens'.

4) Names of some fruit and vegetables: gen. pl. *баклажан* 'aubergines', *помидор* 'tomatoes', *гранат* 'pomegranates', *апельсин* 'oranges' (these forms are allowed as variants in colloquial speech).

The genitive plural of some nouns allows either the *-ov* or -Ø ending: *грамм-ов* and *грамм-Ø*, *помидор-ов* and *помидор-Ø*, *кадет-ов* and *кадет-Ø*, *гардемарин-ов* and *гардемарин-Ø*. This group is especially interesting for the study of variants because it includes words for which the process of variant competition is still in progress.

Contrary opinions exist on the correlation of variants in this 'weak point' of language norm. According to Markov (1992), genitive *-ov* forms have been gradually displacing -Ø forms since the 12th century and the process still is going on. According to another point of view, -Ø forms have become more active since the end of the 20th century and for this reason are considered to be the dominant variants in the observed group of nouns (Glovinskaya 2008).

The corpus-based study of the correlation of variants within the mentioned subgroups and hereafter within the whole group of nouns is likely to shed new light on the matter.

4.2 A corpus-based study of *-ov* and -Ø variants of genitive plural forms

As an example, the names of ranks belonging to some branches of the armed forces have been chosen because they form a finite list including 14 nouns. The variants of genitive plural forms of all these nouns were examined in written texts dating from four periods: the 18th century, the 19th century, and the first and the second half of the 20th century. Table 1 shows the total number of the relevant word-form and Table 2 demonstrates its frequency (items per million tokens).

Gen.pl. variant	18th cent.	19th cent.	20th cent. (1st half)	20th cent. (2nd half)
солдат 'soldier'	75	>1000	>2500	>4000
солдатов	1	5	11	6
партизан 'partisan'	0	7	>300	>350
партизанов	0	23	5	4
рекрут 'recruit'	23	96	5	1
рекрутов	1	26	20	26
*кадет*1 'cadet' (military)	1	40	54	10
*кадетов*1	8	9	13	25
*кадет*2 'Cadet' (party)	0	0	18	3
*кадетов*2	0	0	193	48
гренадер 'grenadier'	10	37	34	9
гренадеров	1	22	33	13
гардемарин 'midshipman'	0	1	13	2
гардемаринов	0	15	10	12
гусар 'hussar'	11	130	50	23
гусаров	1	38	15	10
карабинер 'carabineer'	5	4	0	0
карабинеров	2	6	7	11
драгун 'dragoon'	6	79	38	9
драгунов	0	18	3	1
кирасир 'cuirassier'	0	20	30	9
кирасиров	0	14	3	7
улан 'uhlan'	0	43	26	7
уланов	0	27	5	9
янычар 'janissary'	9	23	7	14
янычаров	1	2	5	0
рейтар 'rider'	0	4	0	7
рейтаров	0	1	2	8

Table 1: Total number of variants of gen.pl. word-forms in different subcorpora

Gen.pl. variant	18th cent.	19th cent.	20th cent. (1st half)	20th cent. (2nd half)
солдат 'soldier'	28.8	>38.5	>62.5	>41.1
солдатов	0.38	0.19	0.28	0.06
партизан 'partisan'	0	0.27	>7.5	>3.6
партизанов	0	0.88	0.13	0.04
рекрут 'recruit'	8.9	3.7	0.13	0.01
рекрутов	0.39	1	0.5	0.27
*кадет*1 'cadet' (military)	0.39	1.5	1.35	0.1
*кадетов*1	3.1	0.35	0.33	0.26
*кадет*2 'Cadet' (party)	0	0	0.45	0.03
*кадетов*2	0	0	4.8	0.49
гренадер 'grenadier'	3.8	1.42	0.85	0.09
гренадеров	0.39	0.85	0.83	0.13
гардемарин 'midshipman'	0	0.39	0.33	0.02
гардемаринов	0	0.58	0.25	0.12
гусар 'hussar'	4.2	5	1.25	0.23
гусаров	0.39	1.46	0.38	0.1
карабинер 'carabineer'	1.9	0.15	0	0
карабинеров	0.77	0.23	0.18	0.1
драгун 'dragoon'	2.3	3.03	0.95	0.09
драгунов	0	0.69	0.08	0.01
кирасир 'cuirassier'	0	0.77	0.75	0.09
кирасиров	0	0.54	0.08	0.07
улан 'uhlan'	0	1.65	0.65	0.07
уланов	0	1.03	0.13	0.09
янычар 'janissary'	4.2	0.88	0.18	0.14
янычаров	0.39	0.08	0.13	0
рейтар 'rider'	0	0.15	0	0.07
рейтаров	0	0.04	0.05	0.08

Table 2: Frequency (items per million tokens) of variants of gen.pl. word-forms in different subcorpora

The words on the above list (except *солдат* and *партизан*) are not frequent in modern texts and mainly belong to the passive vocabulary. The whole group can be subdivided into 3 subgroups according to the relation between *-ov* and -Ø variants of the genitive plural.

Subgroup 1 only includes two frequently used words: *солдат* 'soldier' and *партизан* 'partisan'. Each of them has only one codified variant with the -Ø ending. The variants with *-ov* (*солдатов*, *партизанов*) are sub-standard and mainly used in fiction for stylistic purposes.

Subgroup 2 includes the words *рекрут* 'recruit', *кадет*1 '(army) cadet', *кадет*2 'Constitutional Democrat, Cadet', *гренадер* 'grenadier', *гардемарин* 'midshipman', *гусар* 'hussar', *карабинер* 'carabineer'. The competition of variants has been continuing during the past three centuries, the process being especially active in the 20th century. The rates for each variant (defined as the ratio of the variants to the total number of genitive plural forms of each noun over different periods of time) are displayed in Figures 1 and 2.

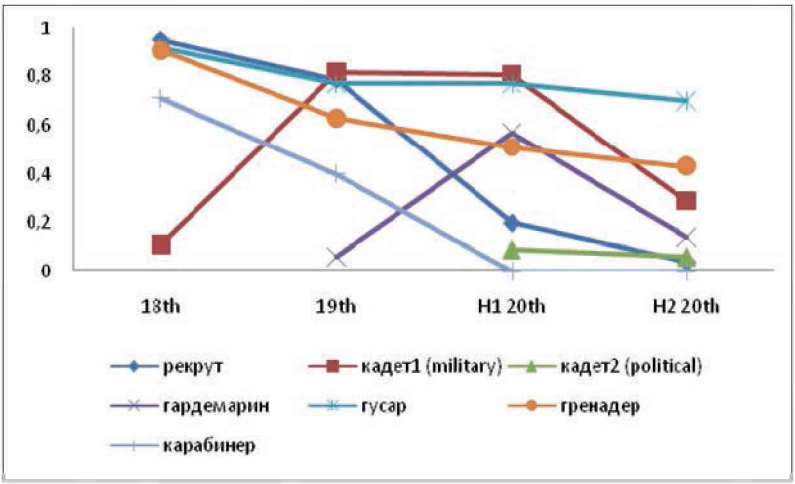


Figure 1: Variant ratio of genitive plural forms with the -Ø ending of masculine nouns in Subgroup 2

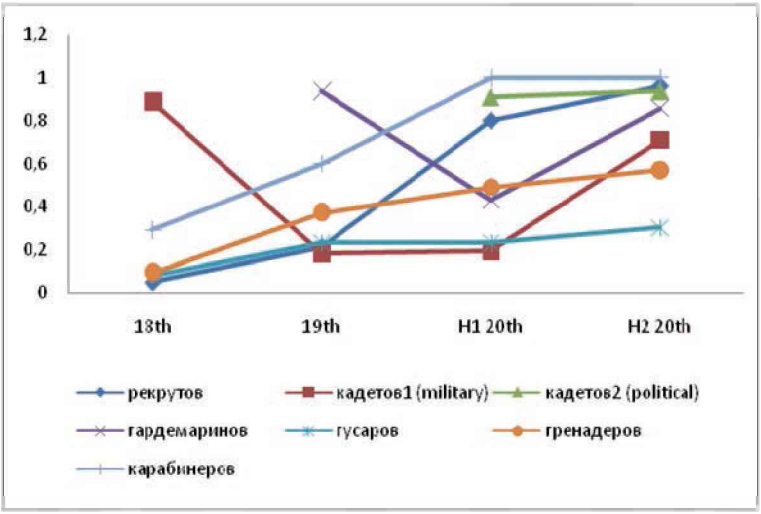


Figure 2: Variant ratio of genitive plural forms with the *-ov* ending of masculine nouns in Subgroup 2

As the diagrams show, the nouns in the second group have one point in common: the proportion of their variants with the -Ø ending dropped towards the end of the 20th century whilst the proportion of *-ov* variants increased.

Subgroup 3 includes the nouns *драгун* 'dragoon', *кирасир* 'cuirassier', *улан* 'uhlan', *рейтар* 'rider', *янычар* 'janissary'.

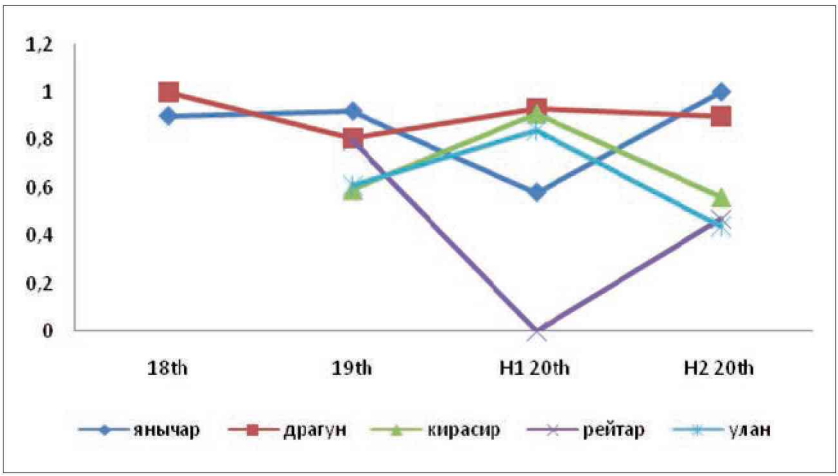


Figure 3: Variant ratio of genitive plural forms with the -Ø ending of masculine nouns in Subgroup 3

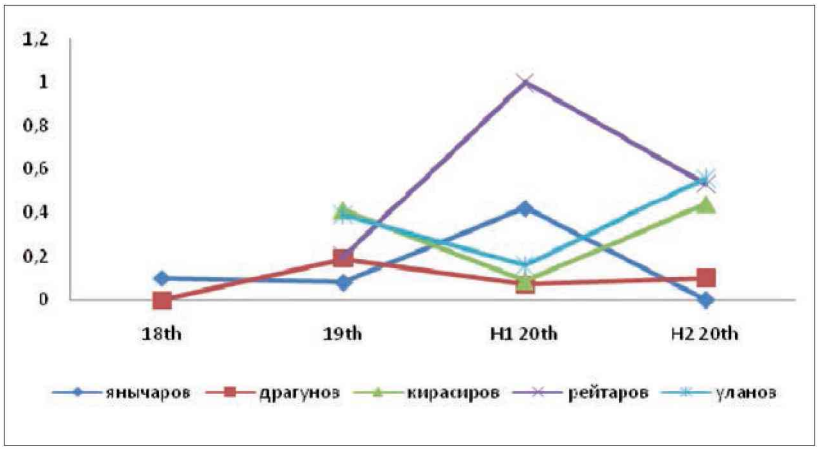


Figure 4: Variant ratio of genitive plural forms with the *-ov* ending of masculine nouns in Subgroup 3

On the whole, the nouns of the third subgroup are characterized by a lower rate of forms with the *-ov* ending and a higher rate of those with the -Ø ending. The quantitative relation between the two forms may be contrasting (as *янычар – янычаров*, *драгун – драгунов*) or may almost be equal (*кирасир – кирасиров*, *улан – уланов*). The second common peculiarity of nouns from Subgroup 3 (except for *рейтар/ рейтаров*) is that the proportion of their variants with -Ø and *-ov* endings remains nearly static for the period under study.

The grammatical classification of the studied nouns based on the gen.pl. variants ratio and its dynamics correlates with their lexico-semantic classification. Subgroup 1 includes two nouns from the active vocabulary. The gen.pl. variant competition of the word *солдат* was resolved by the 18th century; of the word *партизан*, by the beginning of the 20th century. The word *партизан* was naturalized in Russian in the 18th century with the meaning 'a strong supporter of a party, cause, or person'; the genitive plural is formed with the *-ov* ending. In contemporary Russian this meaning is considered outdated. The second meaning ('a member of an armed group formed to fight secretly against an occupying force') arose during the Patriotic War against Napoleon in 1812 and became commonly used; its genitive plural word-form has the -Ø ending.

Subgroup 2 includes words that are relatively infrequent in comparison with previous periods, but which are not out of use in contemporary language. Some of them, for one reason or another, even became more common at the end of the 20th century. This applies, for instance, to the following words:

a) *кадет*1 (cadet) 'a pupil of a military middle school' (due to the reestablishment of cadet corps in Russia);
b) *кадет*2 (Cadet) 'Constitutional Democrat' (from its abbreviated name for members of the Constitutional-Democratic Party founded in 1905)
c) *гардемарин* 'midshipman' and *гусар* 'hussar' which have become heroes of literature and cinema ("Midshipmen, forward" by Svetlana Druzhinina, "Hussar Ballad" by Eldar Ryazanov, or "Squadron of Flying Hussars" by Stanislav Rostotsky are very popular films), and so on.

The predominant form of the genitive plural in this group is the *-ov* form although in the 18th century, forms with the -Ø ending were prevalent (except for the noun *кадет*). In current usage, we may observe some regularity: if the word becomes more common, or a new sense develops, or a very old sense is revived in new contexts, the variant with the *-ov* ending is used. An exception to this rule is the noun *гусар* 'hussar': the -Ø form is still prevailing nowadays, although the gap between the frequency ratios of the two variants has become smaller.

Subgroup 3 includes words referring to the passive vocabulary. They fell out of use in the first half of the 20th century because the corresponding military branches were reorganized. Now these words are only used in historical contexts. A dictionary should be consulted when using these obsolete words that is why the preference of -Ø variants within this group remains rather constant during the whole period.

4.3 Corpus data vs. normative recommendations

Normative recommendations differ in different publications. The most authoritative sources are as follows:

Rozental' (1952 / 1977): *Practical stylistics of Russian* – -Ø variant for all nouns is recommended.

Graudina et al. (1976 / 2004): *Stylistic Dictionary of Variants* – -Ø variant for *гардемарины*, *гренадеры*, *рейтары*, *солдаты*, *уланы* and semantic rules of variant choice for the other words. The form *кадет* must be used for 'pupils of a military middle school', but *кадетов* when referring to 'members of the Constitutional Democratic Party'; and the forms *драгун*, *кирасир*, *янычар* with collective nouns – 'detachment', 'brigade', 'squadron', etc., but *драгунов*, *кирасиров*, *янычаров* when referring to individuals.

Zaliznyak (1977/2003): *Grammatical Dictionary of Russian*; Yes'kova (1994): *Short Dictionary of grammar difficulties* – for subgroup 1, only the variant with the -Ø ending is recommended. For subgroups 2 and 3, both variants are acceptable.

As can be seen from the corpus data, a slight expansion of the inflection *-ov* can be observed within the investigated group of nouns since the 18th century. This inflection is perceived as dominant for genitive plural forms of masculine nouns with stems ending in a hard consonant, which in turn leads to the unification of the plural case paradigm, in which masculine forms with the *-ov* ending are distinguished from all other forms with the -Ø ending (feminine, neutral, pluralia tantum).

Against the background of these data, the recommendations for the -Ø variant by Rozental' (1952/1977) and Graudina et al. (1976/2004) seem to be out of date: they are not supported by current usage and the corpus data. The semantic differentiation between 'collective' and 'individual' meaning seems to be too narrow, thus the underlying rules relying on them were violated as early as the 18th and 19th centuries, e.g., "[...] сам же взял с собою [...] суздальских шестьдесят *гренадеров*, сто мушкетеров, [...] и тридцать шесть воронежских *драгун*" (Suvorov 1786), "Миних послал вперед к Яссам Кантемира с трехтысячным отрядом волохов, *драгунов* и *гусар*, а сам следовал за ним"(Kostomarov 1862-1875).

As can be seen from the above, the permissive remark in Zaliznyak's *Grammatical Dictionary* (1977/2003) is most acceptable for contemporary normative manuals and grammar books. According to Zaliznyak, the variant with the *-ov* ending is generally recommended whilst the variant ending with -Ø is regarded as an option and appropriate, for instance, in archaized speech.

For more examples for using the Russian National Corpus as a tool for research on grammatical variability see Savchuk (2007), Savchuk/Grishina (2008), Kiseleva et al. (eds.) (2009).

5. Conclusions

The corpus approach for the study of variants in synchronic and diachronic aspects enables us to carry out a qualitative and quantitative analysis of units and constructions; to reveal trends in the relation between competing variants; to trace the development of new phenomena, and to amend lexicological descriptions and normative recommendations.

References

Andrews, Edna (2001): The Russian Reference Grammar. Internet: www.seelrc.org: 8080/grammar/mainframe.jsp?nLanguageID=6 (last visited: 11 / 2010).

Glovinskaya, Marina J. (2008): Aktivnyje protsessy v grammatike. In: Krysin, Leonid P. (ed.): Sovremennyj russkij jazyk. Aktivnyje protsessy na rubezhe XX i XXI vekov. Moskva: Jazyki Slavjanskich Kul'tur JSK.

Graudina, Ljudmila K./Ickovič, Viktor A./Katlinskaja, Lija P. (1976 / 2004): Grammaticheakaya pravilnost' russkoj rechi: Stilisticheskij slovar' variantov. Moskva: AST; Astrel'.

Jakobson, Roman (1956 / 1984): The relationship between genitive and plural in the declension of Russian Nouns. In: Waugh, Linda R. / Halle, Morris (eds.): Russian and Slavic Grammar: Studies 1931-1981. Berlin: Mouton, 135-140.

Es'kova, Natalja A. (1994): Kratkij slovar' trudnostej: Grammaticheskije formy. Udarenije. Moskva: Russkij jazyk.

Kiseleva, Ksenija et al. (eds.) (2009): Korpusnyje issledovanija po russkoj grammatike. Moskva: Probel-2000.

Markov, Vitalij M. (1992): Istoricheskaja grammatica russkogo jazyka: Imennoje sklonenije. Izhevsk: Izd-vo Udmurtskogo universiteta.

Rozental', Ditmar E. (1952/1977): Prakticheskaya stilistika russkogo jazyka. Moskva: Vysshaya shkola.

Savchuk, Svetlana (2007): Corpus-based investigation of language change: the case of RNC. In: Davies, Mark / Rayson, Paul / Hunston, Susan / Danielsson, Pernilla (eds.): Proceedings of the Corpus Linguistics 2007 July 27-30, University of Birmingham, UK. Internet: http://www.corpus.bham.ac.uk/corplingproceedings07/paper/181_Paper.pdf (last visited: 11/2010).

Savchuk, Svetlana / Grishina, Elena (2008): Variation in Russian. Dictionary project. In: Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" Issue 7, 14. Moskva: RGGU [= Russian State University for the Humanities], 466-474.

Shvedova, Natalija Yu. (ed.) (1980): Russkaja grammatika. Moskva: Nauka.

Zaliznyak, Andrej A. (1977 / 2003): Grammaticheskij slovar' russkogo jazyka. Moskva: Russkije slovari.

Zaliznyak, Andrej A. (1967): Russkoje imennoje slovoizmenenije. Moskva: Nauka.

RUSKA IVANOVSKA-NASKOVA

Italian-Macedonian parallel corpus

Abstract

The aim of this study is to present an ongoing project at the Department of Romance Languages at the Ss. Cyril and Methodius University in Skopje, Macedonia, namely, the compilation of an Italian-Macedonian parallel corpus. The specific aim of this paper is to describe the methodology and the selection of texts adopted in the creation of the corpus and to discuss its application in the teaching of Italian as a second language.

1. Introduction

The availability of adequate bi- and multilingual language resources such as dictionaries, databases and corpora is vital for the work of linguists, translators and teachers of foreign languages. While there is a relatively vast number of resources that involve at least one language spoken worldwide, such as English or Spanish, there is a limited number of resources that deal with rare language pairs and, in general, with less frequently used languages, such as Macedonian. This paper presents in brief a small project whose aim is to contribute to the creation of such language tools for less frequently used languages. It describes the initial phase of the compilation of the Italian-Macedonian parallel corpus, possible ways of applying it, and future work.

2. General considerations and methodology

The idea for the Italian-Macedonian parallel corpus derives from a small community of linguists and teachers of Italian at the Faculty of Philology in Skopje, Macedonia, who often experience difficulties in their work due to the lack of appropriate tools combining the two languages. There are only two general dictionaries (Kitanovski 1996, 2003) and no phraseological or specialized dictionaries. The aim of this project is to provide a parallel corpus which can be used as a complementary tool in both applied and theoretical linguistic research.

In the preparatory phase of the compilation several complex models of bilingual parallel corpora were considered, such as the English-Norwegian and the English-Portuguese parallel corpus which include translated texts of various genres in both directions and allow several types of comparison (Johansson 2007, Frankenberg-Garcia 2006, Borin 2002). However, these complex models are unachievable for rare language combinations, such as Italian and Macedonian, due to the lack of translated texts. Therefore, a relatively simple, monodirectional model for this parallel corpus was adopted (Laviosa 2002: 37).

3. Selection of the texts and tools

One general problem regarding the acquisition of any parallel corpus is the availability of translated texts which are not always easy to come by. Many difficulties were encountered in finding translated texts from Italian into Macedonian. The accessible translations mainly encompassed one genre: fiction prose. For these reasons, only texts belonging to this group were included in the initial phase of the compilation. In particular, three contemporary novels were considered: *Il deserto dei tartari* (Dino Buzzati), *Palomar* (Italo Calvino) and *Uno, nessuno e centomila* (Luigi Pirandello).

The corpus was compiled with *ParaConc* (Barlow 2003, version 1.0 269). This Windows-based programme was mainly chosen because it is flexible and easy to use. One of the main features of this programme is that it can perform an automatic alignment based on the Gale-Church algorithm (Gale / Church 1993). The output can then be manually checked and corrected within the same programme, which simplifies the alignment procedure. Furthermore, *ParaConc* allows different types of investigations: basic types of search for words, regular expressions and more complex investigations which involve tags, etc. The concordance window allows the filtering of the concordances and parallel observation. The programme also identifies the collocations of the words and gives word lists of the texts together with the word frequencies.

4. Compilation of the corpus

The compilation of the corpus consisted of several steps: the collection of the raw texts, the automatic alignment, and a manual validation of the aligned units.

Since most of the texts were not available in digital format, they had to be digitalized. During this phase of the acquisition of the corpus, the texts were scanned and the output was transferred into an OCR programme. The errors were corrected manually and the texts were stored as plain text files with Unicode (UTF 8) encoding. The text pairs were aligned automatically at sentence level within the programme. Although the automatic alignment showed a high level of accuracy as far as the mapping was concerned, there was still the need for a manual check and correction of the output. The following figure shows a sample of aligned text in the corpus:

Figure 1: Aligned text sample

At the moment the corpus is comprised of 283 158 words (129 561 words of Italian texts and 153 597 words of translated texts in Macedonian).

5. Examples of investigations and possibilities for use

The corpus can be searched for words and regular expressions. For example, a simple search of the Italian text connector *dunque* in the corpus gives the following output:

Figure 2: Search results for *dungue*

The upper section of the window presents the extracted units of the Italian corpus followed by the matching sentences in Macedonian. The translated sentences can also be centred if a possible translation of the particular word is given. The programme enables one to reach the broader context of the sentence, which is sometimes crucial for some types of research. In this case, several possible translations of the connector *dunque* are easily identified: *значи*, *според тоа*, *оттаму*, *тогаш*, two of which (*оттаму*, *тогаш*) are not present in the standard Italian-Macedonian dictionary. Furthermore, the corpus can be also searched for regular expressions such as **ndo*, which identifies all occurrences of the gerund in the Italian texts. The examples show various types of use of the gerund: instances of "perifrasi progressiva" ("Non sta *contemplando*, perché per la contemplazione ci vuole ..."), examples of various types of dependent clauses with a gerund ("*volendo* evitare le sensazioni vaghe, egli si ...", "Però, – pensa *andando* avanti e ..."). The translated sentences in Macedonian show a whole range of corresponding structures: the use of the present tense for the "perifrasi progressiva" ("*Не набљудува* зашто за набљудување е потребен ..."), the verbal adverb that ends in *-јќи* for many dependent clauses with a gerund ("*сакајќи* да избегне неодредени впечатоци, тој ...", "Меѓутоа,

– си мисли тој *продолжувајќи* да оди ...”), the verbal noun that ends in *-ење* preceded by various prepositions (“*Ritornando* dalla sua passeggiata, Palomar ripassa ...” – “*На враќање* од прошетката, Паломар повторно минува ...”). There are also examples of sentences where the gerund used in dependent clauses is substituted using finite dependent clauses in Macedonian (“la gobba dell'onda *venendo* avanti s'alza in un punto ...” – “Врвот на бранот *кој пристигнува* се извишува во една точка ...”).

These kinds of queries of the parallel corpus can be used in the beginning by teachers of Italian who can obtain a large number of examples from authentic texts together with the translated sentences (Botley et al. 2000). The data-driven approach favours the use of various corpora for the design of teaching materials and underlines the importance of the inductive approach in language acquisition (Gavioli 2005: 27-29). The parallel corpus can be a valuable tool, especially for university students who have acquired some metalinguistic competence and can take advantage of this corpus-based approach. Various types of language exercises can be performed with the parallel corpus: comparison of grammatical patterns, idiomatic expressions, etc.

Furthermore, the corpus may be used by linguists for extracting examples for various types of contrastive research. Lexicographers can also exploit the corpus for making bilingual dictionaries and other types of lexical databases. Translators could also take advantage of this tool for researching possible solutions for translation problems (Erjavec 2003, Fachinetti 2007).

6. Future work

One of the main disadvantages of the corpus is its limited size. We are also aware that, unfortunately, the question of representativeness cannot be addressed at this stage of the compilation because of the lack of translated texts from various domains. There could be few or no examples for certain types of language phenomena.

Therefore, the first future task concerns the enlargement of the corpus database with different texts: non-fiction prose, legal, scientific or technical texts, etc. The diversity of texts will raise the representativeness and the quality of the corpus. Also, translations of Macedonian texts into Italian will be included in order to achieve a bidirectional model of parallel corpora. In a later phase the

corpus will be annotated at POS level and lemmatized, thus allowing a broader spectrum of investigations. Nevertheless, we believe that even at this stage the Italian-Macedonian parallel corpus can be a valuable tool for some types of empirically-grounded research that involve these two languages.

References

Barlow, Michael (2003): Paraconc: A concordancer for parallel texts. Internet: www.athel.com/paraconc.pdf (last visited: 01 / 2010).

Botley, Simon McEnery / Tony Wilson, Andrew (2000): Multilingual corpora in teaching and research. Amsterdam / Atalanta: Rodopi.

Borin, Lars (2002): Parallel corpora, parallel worlds: Selected Papers from a Symposium on Parallel and Comparable Corpora at Uppsala University, Sweden, 22-23 April, 1999. Amsterdam / Atalanta: Rodopi.

Erjavec, Tomaž (2003): Compilation and exploitation of parallel corpora. In: Journal of Computing and Information Technology 11, 2: 93-102.

Fachinetti, Roberta (2007): Corpus Linguistics: 25 years on. Amsterdam / Atalanta: Rodopi.

Frankenberg-Garcia, Ana (2006): Using a parallel corpus in translation practice and research. Actas da Contrapor 2006, 1a Conferência de Tradução Portuguesa, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 11-12 September 2006: 142-148.

Gale, William A. / Church, Kenneth W. (1993): A program for aligning sentences in bilingual corpora. In: Computational Linguistics 19 N.1: 75-102.

Gavioli, Laura (2005): Exploring corpora for ESP learning. (= Studies in Corpus Linguistics 21). Amsterdam / Philadelphia: Benjamins.

Johansson, Stigg (2007): Seeing through multilingual corpora: on the use of corpora in contrastive studies. (= Studies in Corpus Linguistics 26). Amsterdam / Philadelphia: Benjamins.

Kitanovski, Naum (1996): Dizionario Italiano-Macedone. Skopje: Euroklient.

Kitanovski, Naum (2003): Dizionario Macedone-Italiano. Skopje: Euroklient.

Laviosa, Sara (2002): Corpus-based translation studies: Theory, findings, applications. Amsterdam / New York: Rodopi.