



**Korpuslinguistik und interdisziplinäre
Perspektiven auf Sprache**

**Corpus Linguistics and
Interdisciplinary Perspectives on Language**

Bd./Vol. 5

Herausgeber/Editorial Board:

Marc Kupietz, Harald Lungen, Christian Mair

Gutachter/Advisory Board:

Heike Behrens, Mark Davies, Martin Hilpert,
Reinhard Köhler, Ramesh Krishnamurthy, Ralph Ludwig,
Michaela Mahlberg, Tony McEnery, Anton Näf,
Michael Stubbs, Elke Teich, Heike Zinsmeister

Jost Gippert / Ralf Gehrke (eds.)

Historical Corpora

Challenges and Perspectives

narr |
VERLAG

Bibliografische Information der Deutschen Nationalbibliothek

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

© 2015 · Narr Francke Attempto Verlag GmbH + Co. KG
Dischingerweg 5 · D-72070 Tübingen

Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt.
Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne
Zustimmung des Verlages unzulässig und strafbar. Das gilt insbesondere für
Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und
Verarbeitung in elektronischen Systemen.
Gedruckt auf chlorfrei gebleichtem und säurefreiem Werkdruckpapier.

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

Redaktion: Melanie Steinle, Mannheim
Layout: Andy Scholz, Essen (www.andyscholz.com)
Printed in Germany

ISSN 2191-9577
ISBN 978-3-8233-6922-6

Contents

Preface	9
Martin Durrell: ‘Representativeness’, ‘Bad Data’, and legitimate expectations. What can an electronic historical corpus tell us that we didn’t actually know already (and how)?.....	13
Karin Donhauser: Das Referenzkorpus Altdeutsch. Das Konzept, die Realisierung und die neuen Möglichkeiten	35
Claudine Moulin / Iryna Gurevych / Natalia Filatkina / Richard Eckart de Castilho: Analyzing formulaic patterns in historical corpora.....	51
Roland Mittmann: Automated quality control for the morphological annotation of the Old High German text corpus. Checking the manually adapted data using standardized inflectional forms.....	65
Timothy Blaine Price: Multi-faceted alignment. Toward automatic detection of textual similarity in Gospel-derived texts	77
Gaye Detmold / Helmut Weiß: Historical corpora and word formation. How to annotate a corpus to facilitate automatic analyses of noun-noun compounds.....	91
Augustin Speyer: Object order and the Thematic Hierarchy in older German	101
Marco Coniglio / Eva Schlachter: The properties of the Middle High German “Nachfeld”. Syntax, information structure, and linkage in discourse	125
Stefanie Dipper / Julia Krasselt / Simone Schultz-Balluff: Creating synopses of ‘parallel’ historical manuscripts and early prints. Alignment guidelines, evaluation, and applications.....	137
Svetlana Petrova / Amir Zeldes: How exceptional is CP recursion in Germanic OV languages? Corpus-based evidence from Middle Low German	151

Alexander Geyken / Thomas Gloning: A living text archive of 15 th -19 th -century German. Corpus strategies, technology, organization	165
Christian Thomas / Frank Wiegand: Making great work even better. Appraisal and digital curation of widely dispersed electronic textual resources (c. 15 th -19 th centuries) in CLARIN-D.....	181
Bryan Jurish / Henriette Ast: Using an alignment-based lexicon for canonicalization of historical text	197
Armin Hoenen / Franziska Mader: A new LMF schema application. An Austrian lexicon applied to the historical corpus of the writer Hugo von Hofmannsthal.....	209
Thomas Efer / Jens Blecher / Gerhard Heyer: Leipziger Rektoratsreden 1871-1933. Insights into six decades of scientific practice	229
Stefania Degaetano-Ortlieb / Ekaterina Lapshinova-Koltunski / Elke Teich / Hannah Kermes: Register contact: an exploration of recent linguistic trends in the scientific domain	241
Esther Rinke / Svetlana Petrova: The expression of thetic judgments in Older Germanic and Romance	255
Richard Ingham: Spoken and written register differentiation in pragmatic and semantic functions in two Anglo-Norman corpora.....	269
Ana Paula Banza / Irene Rodrigues / José Saias / Filomena Gonçalves: A historical linguistics corpus of Portuguese (16 th -19 th centuries)	281
Natália Resende: Testing the validity of translation universals for Brazilian Portuguese by employing comparable corpora and NLP techniques	291
Jost Gippert / Manana Tandashvili: Structuring a diachronic corpus. The Georgian National Corpus project.....	305
Marina Beridze / Liana Lortkipanidze / David Nadaraia: The Georgian Dialect Corpus: problems and prospects.....	323
Claudia Schneider: Integrating annotated ancient texts into databases. Technical remarks on a corpus of Indo-European languages tagged for information structure	335

Giuseppe Abrami / Michael Freiberg / Paul Warner: Managing and annotating historical multimodal corpora with the eHumanities desktop. An outline of the current state of the LOEWE project “Illustrations of Goethe’s Faust”	353
Manuel Raaf: A web-based application for editing manuscripts	365
Gerhard Heyer / Volker Boehlke: Text mining in the Humanities – A plea for research infrastructures.....	373

Preface

Historical Corpora – Challenges and Perspectives

The present volume contains most of the papers read at the international conference “Historical Corpora 2012”, which was hosted by the LOEWE Research Cluster “Digital Humanities” of the State of Hesse at the University of Frankfurt on December 6-8, 2012. All in all, the conference comprised 27 individual papers, selected out of 45 applications in a meticulous peer-reviewing process by an international board, plus five keynote speeches, three of which have been kindly provided for publication in the present volume. It goes without saying that nearly all of the materials have been duly elaborated in the meantime in order to bring this volume up-to-date.

Both in arranging the conference program, which can be accessed on www.dhhe.de/historical-corpora, and in preparing the present volume, it became clear that the very title “Historical Corpora” opens a huge range of possible interpretations and, accordingly, topics, thus making it difficult beforehand to find a consistent order for the individual contributions. This is true, first of all, of the notion of “historical” itself. In many of the papers, this was taken to refer to older stages of given languages, be they “ancient”, “old”, “medieval”, or just not contemporary. In other cases, it involved a perspective across different stages in the history of a language; for this perspective, which is mostly concerned with linguistic change in time, the term “diachronic” would be more appropriate in order to distinguish it from a consideration of individual stages in a language’s past, which may be as “synchronic” as a study of contemporary language use. As the papers collected in the present volume show, the difference between these principles has a big impact on the structuring of corpora, their contents and their sizes. It may suffice here simply to mention a few points.

- a) Contemporary corpora can be multimodal (comprising written, spoken, audiovisual, elicited and other types of data); the same may be true for historical corpora both under a synchronic and a diachronic perspective, but only if the time-depth in question does not exceed 120 years, given that the oldest recordings of spoken language hardly antedate the year 1900.
- b) Contemporary corpora can always be thematically determined, provided the language in question is really “alive”; the same may be true of historical corpora both in a synchronic and a diachronic perspective, but only to a

certain extent, given that the further we go back in history, the fewer genres, registers, and text types we can expect to be represented in what has come down to us from the history of a given language.

- c) For the same reasons, only contemporary corpora can be truly “balanced” in the sense that they cover all modes, registers, genres, etc. of a given language to an equal extent. The balancing of historical corpora is always restricted by the materials that have survived.
- d) Contemporary corpora may be kept linguistically homogeneous, excluding, for instance, certain dialectal, sociolectal, or other strata. For historical corpora, this may be attempted as well, but only if they are synchronic; diachronic corpora can never be linguistically homogeneous as they cover linguistic change *per definitionem*.
- e) Contemporary corpora may be kept orthographically homogeneous, depending on the consistency of the orthographical rules of a given language. For synchronic historical corpora, this may be attempted, too, but it will again depend on the language in question and the time depth envisaged (as orthography in the modern sense is a rather recent phenomenon); for diachronic corpora, this will mostly be impossible as orthography changes over time everywhere, partly in accordance with linguistic change and partly independently.
- f) Contemporary corpora are “open” in the sense of being freely extensible at any time. For all kinds of historical corpora, however, extensibility is limited to what has been preserved, and the further we go back, the less material we can expect to find. Synchronic historical corpora can even be complete in the sense that they cover all the (written!) materials of a certain stage of a given language; an example of this is the TITUS corpus of Old Persian, <http://titus.uni-frankfurt.de/texte/etcs/iran/airan/apers/apers.htm>, a language that was written in cuneiform script between ca. 550 and 330 B.C. On the other hand, diachronic corpora cannot be complete if the time range to be covered includes contemporary usage.

In addition to the variety of perspectives on “historical corpora” emerging from these preliminary considerations, the contributions to the present volume differ, of course, with regard to the languages under concern. It is true that German – in nearly all its historical facettes – is the most widely addressed among them; however, the range of vernaculars treated extends far beyond

that, across the Romance languages into the Caucasus and from the recent past down into antiquity. Differences also concern the linguistic interests prevailing in the papers, which may focus on syntactic, semantic, pragmatic, lexicological or other phenomena. Beyond that, the program of the conference proves that historical corpora, in all senses of the term, have raised interest meanwhile not only in linguistics but also in neighbouring disciplines such as literary studies, history, philosophy, or theology, and we are delighted that some of the contributions to the present volume reflect views from outside linguistics. And of course, there are also contributions that are practically language-independent, dealing with more general issues of the structure of historical corpora (and the infrastructure required by them).

The arrangement of the papers in this volume tries to take these aspects into account as far as possible. There being no intrinsic principle that would be superior to others, we decided to let ourselves be guided by reader-friendliness, which simply presupposes that papers with related content should be placed close to each other. It goes without saying that there are no value judgments implied in the arrangement of any of the papers.

There are many persons and institutions to whom the editors of the present volume wish to express their gratitude: first of all, the keynote speakers, Gerhard Heyer, Karin Donhauser, Gerhard Lauer, Martin Durrell, and Anthony Kroch, who raised general topics of major interest and thus provided true highlights in a conference program of exceptional breadth and quality; second, the participants who sent their papers in for evaluation in due time and delivered them in Frankfurt, in some cases enduring long and unpleasant journeys; third, the peer reviewers, Pietro Beltrami, Anne Bohnenkamp-Renken, Nils Diewald, Karin Donhauser, Martin Durrell, Gerhard Heyer, Anthony Kroch, Gerhard Lauer, Henning Lobin, Anke Lüdeling, Rosemarie Lühr, Giovanna Marotta, Alexander Mehler, Cecilia Poletto, Andrea Rapp, Henning Reetz, Manfred Sailer, Maik Stührenberg, Marc van Oostendorp, Ulli Waltinger, and Helmut Weiß, who agreed to read and evaluate the papers submitted alongside their many other duties; fourth, the members of the staff of the LOEWE research cluster and the students of the Institute of Empirical Linguistics at the University of Frankfurt who helped us arrange and maintain the conference; fifth, the editors of the series “Korpuslinguistik und Interdisziplinäre Perspektiven auf Sprache – Corpus Linguistics and Interdisciplinary Perspectives on language (CLIP)”, Harald Lungen, Marc Kupietz and Christian Mair, as well as the Gunter Narr Verlag, Tübingen, for kindly accepting the present volume for

publication; sixth, the contributors of the present volume, who were ready to work through their papers again for publication after presenting them at the conference; and seventh, the Hesse State Ministry of Higher Education, Research and the Arts, which enabled us to work intensively on historical corpora and to organise the conference by granting the funding for our LOEWE Research Cluster.

Frankfurt, January 2015

Jost Gippert

MARTIN DURRELL

‘Representativeness’, ‘Bad Data’, and legitimate expectations

What can an electronic historical corpus tell us that we didn’t actually know already (and how)?

Abstract

The availability of electronic corpora of historical stages of languages has been welcomed as possibly attenuating the inherent problem of diachronic linguistics, i.e. that we only have access to what has chanced to come down to us – the problem which was memorably named by Labov (1992) as one of “Bad Data”. However, such corpora can only give us access to an increased amount of historical material and this can essentially still only be a partial and possibly distorted picture of the actual language at a particular period of history. Corpora can be improved by taking a more representative sample of extant texts if these are available (as they are in significant number for periods after the invention of printing). But, as examples from the recently compiled GerManC corpus of seventeenth and eighteenth century German show, the evidence from such corpora can still fail to yield definitive answers to our questions about earlier stages of a language. The data still require expert interpretation, and it is important to be realistic about what can legitimately be expected from an electronic historical corpus.

1. Introduction

My primary aim in this paper is to take the opportunity of standing back and taking a look at what we expect from historical linguistic corpora, consider the possibilities they provide and re-assess their inherent limitations, in particular in the light of the kind of caveats which have been voiced eloquently over the years by Rissanen (1989; 2008). These observations will be chiefly based on our recent experience in Manchester over the past few years of compiling a historical corpus of Early Modern German, the GerManC corpus – and in my own case coming relatively new to the whole field of corpus structure, compilation and design.

2. The problem of ‘Bad Data’

The obvious starting point is to consider the data of historical linguistics. In effect, historical linguistics has always been based on a corpus, although we haven’t always used the term. We have quite simply a body of data which is,

first of all, inherently finite – quite obviously so in the case of older languages like Gothic or Runic Norse. In this respect it is like any body of historical data in that we are wholly dependent on what has chanced to come down to us and we have to make sense of it, interpret it and make inferences from it on the basis of explicit (and hopefully sound) theoretical principles. Lass (1997: 42) summarizes the problem succinctly as follows:

The past is not directly knowable or independently available to us as such. But it is knowable through inference, which depends on theoretically directed interpretation and evaluation of witnesses, and where necessary on the actual construction of missing witnesses, which then become part of the record.

As Labov (1992) points out there is no avoiding the fact that all historical linguists have a limited set of accidentally preserved “Bad Data”, and in the field of corpus linguistics this has been taken up and emphasized by Nevalainen (1999). We cannot control it, nor can we appeal to native speaker intuitions. Essentially, as Lass (1997: 24) says, we cannot reconstruct the past but only “encounter it only indirectly, through theoretical judgements about what we take to be its witnesses”, in other words we must simply make hypotheses on the basis of the imperfect data we have in the light of our knowledge about language in general and the particular language at that point in its history.

Now, on the face of it, electronic corpora seem to offer us a way out of this dilemma. In the case of languages like those of medieval and early modern Europe, which are better attested than, say, Gothic, they seem to offer the prospect of affording easy access to an unprecedented amount of data. Instead of spending weeks or months in libraries ploughing through texts hunting for examples of a particular form, construction or vocabulary item, it is all available with a few keystrokes in the comfort of the scholar’s own study. As Cantos (2012: 102) says:

[...] corpus linguistics can fruitfully contribute to overcome the obstacles of the bad data problem; by allowing researchers to process simultaneously almost all the texts that have survived from a given period, corpus linguistics partly solves the fragmentary nature of historical material, and ensures that early varieties can be reliably reconstructed, [...].

However, concealed within this apparently positive claim are a number of very indicative hedges. Apart from the fact that, even with our present technology, it hardly seems a realistic prospect to “process simultaneously almost all the

texts that have survived” from, say, seventeenth century German, to say that “corpus linguistics can fruitfully contribute to overcome the obstacles of the bad data problem” is possibly still some way from overcoming it, and if it “partly solves the fragmentary nature of historical material”, the solution can still only ever be partial. The crucial point is that what we have is still written language data which has been preserved by chance. We may be able to access more of it more quickly and more simply, but it still has all the inherent qualities which led Labov to refer to it as “Bad Data”. It might not necessarily provide better insights than we already have, or give us a much clearer picture of the language at the particular point in time we are investigating. However much data we have, in historical linguistics, as in any historical discipline, we can only ever be dealing with “Bad Data”. An apposite example here would be the recent account by Jones (2009) of the passive auxiliary in the older Germanic languages, notably Old High German. Using extant electronic corpora he was able to propose a convincing and more comprehensive analysis of the distinction in function of the two passive auxiliaries in terms of *Aktionsart* of a kind which had eluded earlier scholars. However, the methods, procedures and theoretical foundations of his account were very much those of traditional philology and historical linguistics (and crucially his expertise in Latin and Greek, as well as in older Germanic), and there is no inherent reason why earlier scholars should not have been able to arrive at the same analysis without the benefit of electronic corpora. Electronic procedures simplified the searches and comparison of the examples, but what was crucial was that Jones (2009) simply asked the right questions within an adequate theoretical framework. On the other hand, where the data are insufficient, we will still lack adequate evidential base for a convincing account. An obvious example would be the still intractable question of whether aspect was a fully functioning grammatical category in the Gothic (and Germanic) verb, similar to Slavonic, with its exponence in prefixation, in particular through the prefix *ga-* (cf. Leiss 1992: 54-71). With or without the benefit of electronic corpora, we can only ever put forward well-founded hypotheses to understand and explain the data we have and try to evaluate them comparatively on the basis of our linguistic expertise.

3. Representativeness in historical corpora

For more recent historical periods, especially after Gutenberg, like that which was the basis of the GerManC corpus of Early Modern German, the amount of available material naturally increases exponentially and it is probably unrealistic to suggest that all the available material could be digitized, and even if that were possible the corpus could then run the risk of becoming unmanageable or inherently skewed. This means that we have to address the familiar issues of size, balance and representativeness (cf. Hunston 2008: 160). A large corpus obviously seems desirable, but with that, two things must be borne in mind. First, as mentioned earlier, however much material is included, we are still only dealing with what happens to have come down to us by chance, and a large corpus cannot solve *per se* the fundamental problem of “Bad Data”. Secondly, any corpus is in essence an artefact and entails all the kind of provisos and caveats indicated by Rissanen (2008: 64-67). It is a subset of the language as it existed at a particular time and it cannot answer the kind of questions which we are able to put to living speakers. We must beware of confusing a corpus with “the language” and of assuming that it gives us some kind of access to the grammar of a native speaker. And in this context it is important always to remember that we are dealing with written data, and the relationship between the varieties used in speech and writing may be rather problematic, especially after the development of widespread literacy or a widespread literary culture and the incipient stages of linguistic standardization (cf. Hennig 2009). Nevertheless, it is by no means certain that the assertion by Hunston (2002: 23) that “a statement about evidence in a corpus is a statement about that corpus, not about the language or register of which the corpus is a sample” is wholly tenable, since ultimately we have no real choice in historical linguistics but to extrapolate knowledge about the development of the language from such samples. As Rissanen (2008: 64-67) says though, no corpus, and especially no historical corpus, can be truly representative in a strictly statistical sense. Similarly, Wegea (2013: 64) points out that we can never know precisely what the relationship is between the sample and the language as a whole, and he refers to Köhler (2005: 5) who puts this very clearly:

Keine Stichprobe kann repräsentative Sprachdaten in dem Sinne liefern, dass in dem in der Statistik üblichen Sinne gültige Schlussfolgerungen auf die Population, auf das „Sprachganze“, möglich wären. Kein Korpus ist groß genug, um die Diversität der Daten im Hinblick auf Parameter wie Medium, Thematik,

Stilebene, Genre, Textsorte, soziale, areale, dialektale Varietäten, gesprochene vs. geschriebene Texte etc. repräsentativ abzubilden. Versuche, das Problem durch Erweiterung der Stichprobe zu lösen, vergrößern nur die Diversität der Daten im Hinblick auf die bekannten (und möglicherweise noch unbekannte) Variabilitätsfaktoren und damit die Inhomogenität.

If Leech (1991: 27) says that a corpus is representative if “findings based on its contents can be generalized to a larger hypothetical corpus”, that ultimately begs the question of how we can ever be in a position to establish how that hypothetical generalization can be carried out. Nevertheless, as Leech has said more recently (2007: 143-144), the debate about balance, representativeness and comparability might lead people

[...] to reject these concepts as being ill defined, problematic and unattainable. My attitude is different from this. [...] these are important considerations, and even if we cannot achieve them 100 per cent, we should not abandon the attempt to define and achieve them. We should aim at a gradual approximation to these goals, as crucial desiderata of corpus design. It is best to recognise that these goals are not an all-or-nothing; there is a scale of representativity, of balancedness, of comparability. We should seek to define realistically attainable positions on these scales, rather than abandon them altogether.

Nevertheless, the question still remains of how criteria might be established to assist us in seeking to define these positions.

4. The GerManC corpus

For historical periods after the introduction of printing by the use of movable type in Europe, the structure and design of any corpus will ultimately be determined by underlying research questions, i.e. what does the researcher want to know about the particular language (or language variety) at that stage in its development. In the case of the GerManC corpus, the primary objective was to provide a research resource which could be exploited to trace the process of standardization in German between the (conventionalized) end of the Early New High German (ENHG) period in 1650 and the relatively final stages of the process of codification at the end of the 18th century. For the period up to 1650 the *Bonner Frühneuhochdeutschkorpus* (Bonn corpus of Early New High German) is available, but standardization was in that period still in the process of selection of variants, and codification had hardly begun. Thus, much more variation still existed in the mid-seventeenth century than, say, in English or

French, and, characteristically for the history of German, this variation had a marked regional dimension. Despite any caveats that one might have about representativeness, it was obviously desirable to have available an electronic corpus which would provide as broad and balanced a picture as possible of the language during this period. The selection of texts was thus modelled on the notion of representativeness developed by Biber for the ARCHER corpus of English (ARCHER = “A Representative Corpus of Historical English Registers”, cf. Biber/Finegan/Atkinson 1993). An additional and important reason for this decision was the fact that David Denison, a colleague in the Department of English Language and Linguistics at the University of Manchester, was co-ordinating the team developing further versions of ARCHER, and a number of postgraduate students had been investigating comparative developments in English and German, for example Auer (2009) and Storjohann (2003). They had used the ARCHER corpus as their resource for historical material in English, but were hampered by the lack of a comparable systematic data collection for German. Using ARCHER as a model we thus considered that at least a greater degree of representativeness could be achieved by including in the first place a wider span of registers. These could not be identical with those of ARCHER because of differences in the types of texts which have been preserved for German, but the following registers were found to provide sufficient material: newspapers, narrative prose (not only fiction), drama, legal texts, sermons, personal letters, scientific texts and texts on humanities-based topics. The time-span of 150 years was divided into three sub-periods of 50 years (following the model of the Bonn Early New High German corpus), and given the continued importance of regional variation in German, the German-speaking area was divided into five major regions: North, West Central, East Central, South-West (including Switzerland) and South-East (including Austria). This proved adequate to cover the level of variation still present in the language. For the completed corpus three 2,000-word text samples were selected for each subdivision in terms of register, sub-period and region, and the whole thus contains nearly a million words. This is a relatively small corpus, but it reflects what could practically be achieved given the time and resources available (cf. Durrell et al. 2011).

It became clear from the earliest results that many previous investigations of the development of the forms and structures of the language in this period had indeed not been fully representative, since they had rarely taken the range of user-based or usage-based variation into account, tending to concentrate on

the prestige literary variety and consider almost exclusively the developments in that register. This reflected on the one hand the ideology of standard (cf. Milroy/Milroy 1999), as this register was commonly equated with the language as such, but on the other, of course, it was precisely those texts which were most readily accessible in the days before digitization. However, taking evidence from a single register clearly runs the risk of presenting a skewed picture of developments in the language as a whole. It was not just Bad Data, but an artificially restricted set of Bad Data which excluded a lot of the material which has actually come down to us.

4.1 The order of finite and non-finite verbs in subordinate clauses

A characteristic example of the sort of limitations this meant for research into the development of German is provided by the issue of the relative order of finite and non-finite verbs in subordinate clauses. Although this has long been regarded as one of the most interesting issues in the syntax of German (and other Germanic languages) from a theoretical perspective, it is noteworthy that the recent study by Sapp (2011) only covers the period up to 1650 and from 1800, as corpus data were not available for the intervening period, and the most comprehensive older study (Härd 1981) concentrates exclusively on developments in literary texts, as does the more recent account of developments in the intervening century and a half by Takada (1994).

Taking subordinate clauses with two verbs, the order of verbs within these groups was still fairly free in ENHG and three possible sequences are all relatively common:

- (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***

We considered examples with a modal auxiliary rather than the periphrastic perfect tense, since, as will be shown later, the perfect auxiliary is often omitted in subordinate clauses at this time. The only acceptable sequence in modern standard German is (c), and according to Härd (1981) this was already established as the dominant norm by 1600. The study by Lühr (1985) also estab-

lished that Luther used this sequence in nearly 90% of possible instances (cf. Fleischer 2011: 166). However, preliminary data from the GerManC corpus, given in Table 1, show that the older sequences do persist after 1650, even though they are relatively infrequent, with the highest proportion being in northern texts where they account for some 14% of cases in the first period.

	1650-1700			1700-1750			1750-1800		
	a	b	c	a	b	c	a	b	c
North	260	26	16	319	9	8	215	1	
WCG	234	7	4	148	2	2	168		
ECG	321	8	10	258	8	6	159		
WUG	177	12	4	174	5	5	185		
EUG	172	13	4	156	10	3	174	1	3

Sequence (a): [...], *dass du es heute [...]* **sollst**

Sequence (b): [...], *dass du es heute [...]* **sollst** *machen*

Sequence (c): [...], *dass du es [...]* **sollst** *heute [...]* *machen*

Table 1: Sequence of finite and non-finite verbs in subordinate clauses (Two-part sequences)

The picture is similar with three-part groups, as shown in Table 2 on the basis of passive constructions with a modal auxiliary.

	1650-1700		1700-1750		1750-1800	
	a	b	a	b	a	b
North	15	15	16	15	67	2
WCG	16	10	15	14	65	6
ECG	27	10	33	15	35	
WUG	19	16	13	10	14	2
EUG	12	7	11	12	10	4

Sequence (a): [...], *dass es [...]* *gemacht werden* **soll**

Sequence (b): [...], *dass es [...]* **soll** *gemacht werden*

Table 2: Sequence of finite and non-finite verbs in subordinate clauses (Three-part sequences)

These sequences are naturally less frequent, and variation continues over a considerably longer period. As Fleischer (2011: 167) points out, though:

Die Datensituation in Bezug auf die historische Entwicklung ist [...] widersprüchlich. Nach Hård (1981: 89) geht im 17. Jahrhundert „das finite Hilfsverb den infiniten Konstituenten voran.“ Dagegen schließt Takada (1994: 215) aus einer Korpusanalyse von Texten des 17. Jahrhunderts, dass sich die Nachstellung des Finitums auf Kosten der Voranstellung ausbreitet. Je nach analysiertem Korpus kommt man also zu verschiedenen Schlüssen.

Takada (1994) and Hård (1981) both use relatively limited sets of material with no allowance for representativeness, and it is perhaps not altogether surprising that their findings show marked differences. Hård (1981), unlike Takada (1994), was not using an electronic corpus, and his material shows a quite dramatic change after 1700, with almost total dominance of final position after that date, as in the modern standard (cf. Hård 1981: 170). By contrast, our corpus, which unlike these earlier studies includes material from a range of registers, shows a rather different picture, with variation persisting much longer and the two sequences in three-part groups evenly balanced until 1750, with the exception of East Central German – significantly the region whose usage, especially in literary genres, had high prestige and tended to function as a model for the developing standard. Nevertheless, the varying findings demonstrate that the problem of Bad Data in relation to the diachronic development of this feature is probably insoluble, since it is unlikely to be possible to find enough instances of these relatively rare constructions to provide an absolutely definitive picture of the process by which the variant which has become the modern norm was finally selected.

4.2 Genre-related variation in the order of finite and non-finite verbs

In practice, the problem of inadequate data even occurs with the two-place constructions. If we separate out the figures by genre, we find that a strikingly high proportion – in fact a majority – of the attestations for sequences with the finite verb first are in dramas, especially in North German.

Nothing comparable has been noted in earlier studies, despite the fact that we are dealing with a literary genre. The fact that most Period 1 dramas are in verse may be an additional complicating factor. However, verse is the norm for dramas of this period, and although we were aware of the problems this might

entail, we felt that we could not represent the genre properly if verse dramas were excluded. In fact even if it had been felt that prose dramas were to be preferred, it could have been difficult to find sufficient for our samples.

	1650-1700			1700-1750			1750-1800		
	a	b	c	a	b	c	a	b	c
North	25	23	9	34	4	2	11	1	
WCG	1	3		15			31		
ECG	40	6	2	19	1		19		
WUG	6	10	1	17	4	2	35		
EUG	8	6		7			29	1	

Sequence (a): [...], *dass du es heute* [...] ***machen sollst***

Sequence (b): [...], *dass du es heute* [...] ***sollst machen***

Sequence (c): [...], *dass du es* [...] ***sollst heute*** [...] ***machen***

Table 3: Sequence of finite and non-finite verbs in subordinate clauses
(Two-part sequences in Drama texts)

Nevertheless, the high proportion of instances of the finite verb being placed first cannot simply be explained by the exigences of rhyme or metre. First of all, this order must clearly still be grammatical, since ungrammaticality is not acceptable even in verse. It is also notably predominant in North German, and to a lesser extent in West Upper German, although with such small figures one hesitates to draw any firm conclusions. Interestingly, Takada's (1994) data also show a relatively high proportion of sequences with the finite verb first from northern texts. He actually refers to these as *Niederdeutsch*, but it is not Low German, but High German written by North Germans. However, it is not impossible that the order is calqued on Low German dialect. Equally we could here have, in an orally-oriented genre such as the drama, a reflection of general spoken norms, with persistence of variation, such as Hennig (2009) found in her data from *NaheSprache*. Finally, it would seem significant that the proportions in East Central German are quite different, with second position clearly predominant even in drama written in verse. And this is the region with the most highly developed degree of literary culture and whose language is most prestigious and often dominant in the selection of the variant which is ultimately selected as standard.

It would seem to be the case here that the effect of acquiring additional data in an electronic corpus has actually been to raise more questions than it answers. The picture provided by traditional selection of texts, such as Hård (1981) undertook, was fairly straightforward, with a relatively early selection of the variant which was eventually codified for the written standard. What we have found in our corpus, which includes texts from a wider variety of genres, could be a more rounded picture of developments in the language as a whole in that it shows variation persisting much longer and differing according to genre and region. But the picture is clearly much more complex, and the reasons underlying the variation and its persistence are more difficult to explain, such that one can only draw very tentative conclusions which need to be corroborated with further evidence, possibly of another kind altogether. In practice, we appear simply to have acquired more Bad Data, if not even Worse Data.

4.3 The a-finite construction

The problem of representativeness and data has come to light again recently through research currently being undertaken in Manchester using the GerManC corpus on the so-called a-finite construction of older German – the ellipsis of the auxiliary (most commonly the perfect auxiliary) in subordinate clauses, as in the following extract from the “Extraordinari Europæische Zeitung” No. 77, published in Hanau in 1701:

Es hat sich auch dieser Prælat solcher Commission aquitiret, ist aber darinnen so glücklich nicht gewesen/ als er wohl gewünschet hätte/ weil der Hr. Cardinal Bedencken trägt ferner etwas an die Stände dieses Königreichs gelangen zu lassen/ ehe und bevor dieselbe ihres Sentements auff das letzte Königl. Patent und sein Schreiben **so er dabey an dieselbe abgelassen**/ entdeckt und kund gemacht haben werden.

This construction is of considerable interest for general syntactic theory, as Breitbarth (2005) shows in her study of the feature. It emerged in late Middle High German and became frequent in Early New High German. Breitbarth's study is based on five texts of roughly 9,000 words each from the *Bonner Frühneuhochdeutschkorpus* for each of the periods covered by that corpus, and her findings show a rapid decline in the occurrence of the construction in the eighteenth century.

Breitbarth (2005) points out the potential limitations of her data sources, but claims that her figures are broadly in line with those obtained in earlier studies by Admoni (1967) and Härd (1981). Effectively, even though she does acknowledge that her data do not prove anything for the language as a whole, and that they can only be taken for what they are, i.e. the output of individual speaker's grammars, she does claim that she has been able to show a general tendency in the language and that the construction becomes much less frequent after 1700 and has pretty well disappeared by 1800.

sub-period	CLAUSE TYPE					
	RELATIVE		ADVERBIAL		ARGUMENT	
	percent	number	percent	number	percent	number
1450-1500	2.6	229	4.6	245	1.2	215
1500-1550	16.8	255	19.7	257	9.5	235
1550-1600	48.2	434	54.0	420	26.4	179
1600-1650	66.9	565	68.9	478	52.7	237
1650-1700	60.8	392	65.7	488	44.9	176
1700-1800	17.9	163	6.6	145	25.2	76

Table 4: The a-finite construction (Data from Breitbarth 2005)

However, these are actually quite broad conclusions, as emphasized in the title of her thesis, but ultimately they rely on a rather small number of actual texts which may lack adequate representativeness. Recent work by Thomas (2012) on the basis of the GerManC corpus, on the other hand, has revealed a very different picture. Even though she has initially only investigated texts from a single genre (Humanities) in a single region (West Central German) over the period 1650-1800, she found that, far from declining in the eighteenth century, the incidence of the a-finite construction actually increased markedly after 1700 and still accounted for a majority of instances in the second half of the eighteenth century.

These data are naturally still only preliminary, but they correspond closely to initial observations by the inputters, including the present author, in the rest of the corpus, and they may well actually be representative of general written usage (and it is important to bear this latter point in mind, since it is questionable

whether the construction was ever current in speech). Even so, it is by no means out of the question that when all genres and regions have been investigated systematically, the results may be closer to those obtained by Breitbarth (2005).

Period	finite	a-finite	Total	percentage a-finite
1650-1700	44	38	82	46.34%
1700-1750	24	87	111	78.38%
1750-1800	15	18	33	54.55%
TOTAL	83	143	226	63.27%

Table 5: The a-finite construction (Data from *Humanities* texts in the GerManC corpus)

Even if that were to be the case, though, we see here very forcibly the validity of Rissanen's (1989; 2008) caveats mentioned earlier, in particular that a corpus cannot be equated with "the language". If we found the assertion by Hunston (2002: 23) that "a statement about evidence in a corpus is a statement about that corpus, not about the language or register of which the corpus is a sample" rather too limiting, it does still flag up the potential risks which must always be borne in mind of basing broad conclusions on data whose representativeness is by no means assured and which cannot be queried directly by reference to actual language users. What findings we obtain and what conclusions we might draw are only tentative hypotheses based on what is inherently Bad Data.

4.4 Difficulties with identifying and marking "words"

Despite these caveats, a clear advantage of electronic corpora is that sophisticated tools can be (and have been) developed which enormously facilitate linguistic analysis and data collection (cf. McEnery/Hardie 2012). Corpora can be tagged, annotated for morpho-syntactic categories and parsed so that instances of particular forms and constructions can be found more quickly (and more reliably) than when laborious searches needed to be made in the original documents, even if the actual procedures of linguistic analysis are still essentially the same.

However, the exploitation of such tools, too, may not be as straightforward as one might wish, and Denison (2013) has shown that there are inherent problems involved in annotation which he has characterized (Denison 2010) as “WYSIWYTCH”, i.e. “What You See Is What Your Theory Can Handle”. As Denison (2013: 17) says:

[...] for grammatical mark-up, with few exceptions a given scheme must privilege one particular analysis for each word, sentence or other unit of analysis. [...] Grammatical mark-up remains essentially a matter of synchronic analysis, and the guiding principle is to be as specific as possible; tagsets routinely deploy a much finer set of distinctions than traditional word classes.

Denison argues that these principles can be problematic since certain forms may not allow of unambiguous allocation to a particular tag, with a particular problem in English being the porousness of word-class boundaries, especially in a diachronic context (cf. also Denison in prep.). Similar problems were encountered in the process of annotating GerManC (cf. Scheible et al. 2011). One striking example involves one of the oldest and most fundamental theoretical problems in linguistics, i.e. the question of what constitutes a “word”. For instance, as Denison (2013: 25-27) points out, the two most recent authoritative grammars of English differ on the analysis of complex prepositions like *on behalf of*, with Quirk et al. (1985: 670-673) claiming that eight out of nine indicators support a complex preposition analysis, i.e. as a single word, whilst Huddleston/Pullum (2002: 620-622) find no syntactic grounds for recognising such strings as complex prepositions. And, as Denison (2013: 27) points out, the British National Corpus tags every occurrence of *on behalf of* in two different ways at different levels of XML mark-up, i.e.:

- (a) [_{PRP} on] [_{NN1} behalf] [_{PRF} of]
 (b) [_{PRP} on behalf of]

That is, in (a) the three words are tagged individually as PRP (preposition) + NN1 (singular common noun) + PRF (preposition *of*), whereas in (b) the whole string is treated as a “multiword token” and tagged as a single preposition.

Historical stages of German throw up numerous instances of such intractable problems which affect tokenization, normalization, lemmatization and tagging. For instance, pronoun cliticization is very prevalent, especially (perhaps unsurprisingly) in the verse dramas, for example, from J.R. Karsten’s “Christ-rühmendes Schau-Spiel” (Frankfurt/Main 1668):

*Au! au! der Arm! du Hund! hast du ihn uns verrenkkt/
So **wirstu** ohne Gnad an Galgen aufgehenkkt!*

and we felt there was no alternative but to solve this by tokenizing the two elements as distinct words. However, this is clearly not entirely satisfactory since it makes searches for individual cliticized forms less straightforward. An even more pervasive problem is one which besets modern German, i.e. whether “words” should be written separately or together, and variation in respect of this is rampant in the period before codification of the orthography. To take a frequent example, there is much variation, even within single texts, between writing the infinitive particle *zu* separately from the verb or prefixed to it, e.g. *zugewarten* or *zu warten*. We decided after considerable discussion that the particle and the verb had to be consistently tokenized separately, not least to avoid potential confusion with *zu* as a verb prefix, and the simplex verb was required as input to normalization and lemmatization.

In practice, though, it is more frequent in Early Modern German for what are now seen as compound words to be printed separately – or, if they were written together, with internal capitalization or hyphens. Eventually we felt there was little alternative to take the forms as we found them, i.e. assigning compound nouns written as a single word (or with a hyphen) to a normalized lemma corresponding to the modern standard form, i.e. with no internal capitals or hyphens. Thus *Südseeinsel* would be used as the normalized form of *SüdSeeInsel* or *Südsee-Insel* or any other variant on these. However, compounds written in the text as two (or more) words were tagged as individual words, so that *Süd See Insel* is tagged as three words. This seemed the only practical solution, although it is clearly less than wholly satisfactory. A similar problem arose with verb prefixes written separately from the verb, e.g. *wahr nahm*, which is still very frequent in seventeenth century texts, whereas in modern usage they would be written together. However, in practice this turned out to be a rather less serious issue, because the verb prefix *wahr* could still be tagged as such, i.e. PTKVZ using the Stuttgart-Tübingen tagset (STTS, cf. Schiller et al. 1999), a possibility provided by the fact that in modern German prefixes can be separated from the verb and may thus be allocated a distinct tag.

As we saw, Denison (2013: 27) showed how the British National Corpus attempts to solve this kind of problem with two separate tags at different levels of XML mark-up, but this of course makes considerable demands in terms of time and resources. However, Early Modern German presents a more complex

variant of this problem with some conjunctions, in that it is not unusual for conjunctions which in the modern language are clearly single words, like *obgleich*, to appear as separate words in texts of this period. It seems straightforward to tag these as separate words using STTS tags, i.e. *ob* KOUS [...] *gleich* ADV, following the model provided for in the STTS tagset guidelines for tagging two word conjunctions of modern German like *als ob*, i.e. KOKOM *als* KOUS *ob* – despite the fact that it is perhaps not entirely satisfactory, since these “words” are clearly operating as a single semantic or syntactic “multiword” unit. However, in older German the parts of such “multiword” conjunctions are frequently separated by anything up to four words in the subordinate clause, as in the following example from “Drey Bücher Der Magnetischen Arzney-Kunst” by Guillelmus Maxwelllus (Frankfurt 1687):

*Er purgieret allein unter sich/ man darff sich auch keiner Salivation befahren/ ob man sich ihme **gleich** bey erfordernder Noth etlich mal gebrauchet*

Clearly these can still be tagged in the same way, with *ob* identified as the conjunction proper KOUS, and *gleich* as an adverb ADV, but it would seem that in one plausible analysis we are still dealing with a multi-word token which requires some appropriate identification which we have not (yet) been able to assign satisfactorily, especially as that could be the most helpful to the corpus user attempting to trace the development of this conjunction – and this is a criterion which must always be borne in mind, since a corpus is in the first instance ultimately a resource for researchers.

However, such cases only serve to further illustrate the central issue being addressed here. The existence of such constructions has long been known in German historical linguistics, so what has the corpus told us that we didn't know already? Are the problems just outlined simply a product of the difficulty of devising optimally efficient tools by means of which we can access and analyse the large amount of Bad Data which an electronic corpus may provide us with? Are we just engrossing ourselves in the fascination of the complex technology and the challenge of compiling programs to solve problems to which we may already know the answer (cf. Wegera 2013: 58)? Naturally, there may be examples of this, but it is evident that good practice must be to remain aware of these dangers and to always remember that the compilation of a corpus and the challenges of designing tools are not ends in themselves.

5. Conclusions

On balance, though, the existence of electronic datasets has facilitated huge steps forward in understanding language history and language change. Not only have sophisticated tools made it possible to ask questions which simply could not be considered previously, for example the research into complex patterns of change in usage reported in Hilpert (2011), with motion charts of development, even if such do depend on very big datasets of a kind to which we can only have recourse for fairly recent periods, or the work on linguistic networks in Late Latin by Mehler et al. (2013). The simple fact of the increased accessibility to data by large numbers of scholars has been immensely beneficial. It is no longer the case, for instance, that doctoral students have to laboriously and time-consumingly compile datasets, and this process has to be repeated by every new researcher. In this way, to return to the example of *obgleich*, we may have been aware that the construction existed, but we can now have a much clearer picture of when it emerged or how frequent it was in comparison to the compound, and whether it was used more in one genre or one region than another.

Nevertheless, it is still vital to be clear what one can legitimately expect from such corpora. We still do not have access to the whole of the language, but only what has chanced to come down to us, and that this is written language which may have autonomous norms at some remove from those of the spoken language, not least because of the development of standardized prescriptions. The a-finite construction discussed earlier may be an example of the problems entailed by this latter issue, since it is perhaps doubtful whether this was ever current in spontaneous spoken production. Even though it is important, in dealing with a period with a relatively large amount of preserved material, to sample what we have as widely as possible taking the variables we are aware of into account, we can only make statements in relation to those parameters, effectively formulating hypotheses on the basis of the Bad Data which we still have access to, in the light of our own (possibly limited) competence as historical linguists, our overall knowledge of the diachrony of the language involved and the circumstances in which the texts were produced, insofar as these are known – and these can be very limited, as for example in the case of early newspapers (cf. Durrell/Ensslin/Bennett 2008).

An electronic corpus means, first and foremost, that we can store very large datasets and access and query them very quickly. But you do have to know what you are looking for; even with a large electronic dataset it is still the case that the real work starts when the counting stops. Any findings from a corpus need to be carefully investigated and elucidated in the light of what else we know about the language in question at the period in question, and, as Rissanen (1989: 16-17) says:

In the analysis, synthesis and conclusions, the machine does not replace the human brain. We will be able to ask the right questions, draw inferences and explain the phenomena revealed by our data only if we develop a good overall mastery of the ancient language form we are studying.

References

- Admoni, Wladimir G. (1967): Der Umfang und die Gestaltungsmittel des Satzes in der deutschen Literatursprache bis zum Ende des 18. Jahrhunderts. In: Beiträge zur Geschichte der deutschen Sprache und Literatur (Halle) 89: 144-199.
- Auer, Anita (2009): The subjunctive in the age of prescriptivism. English and German developments during the Eighteenth Century. Basingstoke: Palgrave Macmillan.
- Bennett, Paul/Durrell, Martin/Scheible, Silke/Whitt, Richard J. (eds.) (2013): New methods in historical corpus linguistics. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache / Corpus Linguistics and Interdisciplinary Perspectives on Language (CLIP) 3). Tübingen: Narr.
- Biber, Douglas/Finegan, Edward/Atkinson, Dwight (1993): ARCHER and its challenges. Compiling and exploring a representative corpus of historical English registers. In: Aarts, Jan/de Haan, Pieter/Oostdijk, Nelleke (eds.): English language corpora. Design, analysis and exploitation. Amsterdam: Rodopi, 1-13.
- Breitbarth, Anne (2005): Live fast, die young. The short life of Early Modern German auxiliary ellipsis. PhD dissertation, Tilburg University. <http://www.lotpublications.nl/publish/articles/001464/bookpart.pdf> (last accessed December 6, 2013).
- Cantos, Pascual (2012): Corpora for the study of linguistic variation and change. Types and computational applications. In: Hernández Campoy/Silvestre, Conde/Camilo, Juan (eds.) (2012): The handbook of historical sociolinguistics. Malden etc.: Wiley-Blackwell, 99-122.
- Denison, David (2010): Category change in English with and without structural change. In: Traugott, Elizabeth Closs/Trousdale, Graeme (eds.): Gradience, gradualness and grammaticalization. Amsterdam/Philadelphia: John Benjamins, 105-28.

- Denison, David (2013): Grammatical mark-up. Some more demarcation disputes. In: Bennett et al. (eds.), 17-35.
- Denison, David (in prep.): English word classes. (= Cambridge Studies in Linguistics). Cambridge: Cambridge University Press.
- Durrell, Martin/Ensslin, Astrid/Bennett, Paul (2008): Zeitungen und Sprachausgleich im 17. und 18. Jahrhundert. In: Zeitschrift für deutsche Philologie 127: 263-279.
- Durrell, Martin/Bennett, Paul/Scheible, Silke/Whitt, Richard J. (2011): Investigating diachronic grammatical variation in Early Modern German. Evidence from the GerManC corpus. In: Konopka, Marek/Kubczak, Jacqueline/Mair, Christian/Šticha, František/Waßner, Ulrich H. (eds.): Grammatik und Korpora 2009. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache / Corpus Linguistics and Interdisciplinary Perspectives on Language (CLIP) 1). Tübingen: Narr, 539-549.
- Fleischer, Jürg (2011): Historische Syntax des Deutschen. Eine Einführung. (= Narr Studienbücher). Tübingen: Narr.
- Hård, John Evert (1981): Studien zur Struktur mehrgliedriger deutscher Nebensatzprädikate. Diachronie und Synchronie. (= Göteborger Germanistische Forschungen 21). Göteborg: Acta Universitatis Gothoburgensis.
- Hennig, Mathilde (2009): Nähe und Distanzierung. Verschriftlichung und Reorganisation des Nähebereichs im Neuhochdeutschen. Kassel: Kassel University Press.
- Hilpert, Martin (2011): Dynamic visualizations of language change: Motion charts on the basis of bivariate and multivariate data from diachronic corpora. In: International Journal of Corpus Linguistics 16: 435-461.
- Huddleston, Rodney/Pullum, Geoffrey K. (2002): The Cambridge grammar of the English language. Cambridge: Cambridge University Press.
- Hunston, Susan (2002): Corpora in applied linguistics. Cambridge: Cambridge University Press.
- Hunston, Susan (2008): Collection strategies and design decisions. In: Lüdeling/Kytö (eds.), Vol. 1: 154-168.
- Jones, Howard (2009): *Aktionsart* in the Old High German passive with special reference to the *Tatian* and *Isidor* translations. (= Beiträge zur germanischen Sprachwissenschaft 20). Hamburg: Buske.
- Köhler, Reinhard (2005): Korpuslinguistik. Zu wissenschaftstheoretischen Grundlagen und methodologischen Perspektiven. In: LDV-Forum 20: 1-16.

- Labov, William (1992): Some principles of linguistic methodology. In: *Language in Society* 1: 97-120.
- Lass, Roger (1997): *Historical linguistics and language change*. Cambridge: Cambridge University Press.
- Leech, Geoffrey (1991): The state of the art in corpus linguistics. In: Aijmer, Karin/Altenberg, Bengt (eds.): *English corpus linguistics*. Studies in honour of Jan Svartvik. London: Longman, 8-30.
- Leech, Geoffrey (2007): New resources, or just better old ones? The Holy Grail of representativeness. In: Hundt, Marianne/Nesselhauf, Nadja/Biewer, Carolin (eds.): *Corpus linguistics and the web*. Amsterdam: Rodopi, 134-149.
- Leiss, Elisabeth (1992): *Die Verbalkategorien des Deutschen. Ein Beitrag zur Theorie der sprachlichen Kategorisierung*. (= *Studia Linguistica Germanica* 31). Berlin/New York: de Gruyter.
- Lüdeling, Anke/Kytö, Merja (eds.) (2008-2009): *Corpus linguistics. An international handbook*. 2 vols. Berlin/New York: de Gruyter.
- Lühr, Rosemarie (1985): Zur Syntax des Nebensatzes bei Luther. In: *Sprachwissenschaft* 10: 26-50.
- McEnery, Tony/Hardie, Andrew (2012): *Corpus linguistics. Method, theory and practice*. Cambridge: Cambridge University Press.
- Mehler, Alexander/Schwandt, Silke/Gleim, Rüdiger/Ernst, Alexandra (2013): Inducing linguistic networks from historical corpora. Towards a new method in historical semantics. In: Bennett et al. (eds.), 257-274.
- Milroy, James/Milroy, Lesley (1999): *Authority in language. Investigating language prescription and standardisation*. 3rd. ed. London: Routledge & Kegan Paul.
- Nevalainen, Terttu (1999): Making the best use of “bad” data. Evidence for socio-linguistic variation in Early Modern English. In: *Neuphilologische Mitteilungen* 100: 499-533.
- Quirk, Randolph/Greenbaum, Sidney/Leech, Geoffrey/Svartvik, Jan (1985): *A comprehensive grammar of the English language*. London/New York: Longman.
- Rissanen, Matti (1989): Three problems connected with the use of diachronic corpora. In: *ICAME Journal* 13: 16-19.
- Rissanen, Matti (2008): Corpus linguistics and historical linguistics. In: Lüdeling/Kytö (eds.), 53-68.
- Sapp, Christopher D. (2011): *Verb order in subordinate clauses from Medieval to Modern German*. Amsterdam/Philadelphia: John Benjamins.

- Scheible, Silke/Whitt, Richard J./Durrell, Martin/Bennett, Paul (2011): Evaluating an 'off-the-shelf' POS-tagger on Early Modern German text. In: Proceedings of the ACL-HLT 2011 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2011). Portland, Oregon. <http://www.aclweb.org/anthology/W/W11/W11-1503.pdf> (last accessed December 6, 2013).
- Schiller, Anne/Teufel, Simone/Stöckert, Christine/Thielen, Christine (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical Report. Stuttgart: Institut für maschinelle Sprachverarbeitung.
- Storjohann, Petra (2003): A diachronic contrastive lexical field analysis of verbs of human locomotion in German and English. Frankfurt a.M. etc.: Peter Lang.
- Takada, Hiroyuki (1994): Zur Wortstellung des mehrgliedrigen Verbalkomplexes im Nebensatz im 17. Jahrhundert. Mit einer Beantwortung der Frage, warum die Wortstellung von Grimmelshausens 'Simplicissimus' geändert wurde. In: Zeitschrift für Germanistische Linguistik 22: 190-219.
- Thomas, Victoria (2012): Report of pilot search: auxiliary ellipsis in German embedded clauses 1650-1800. Unpublished working paper, School of Arts, Languages & Cultures, University of Manchester.
- Wegera, Klaus-Peter (2013): Language data exploitation. Design and analysis of historical language corpora. In: Bennett et al. (eds.), 55-73.

Das Referenzkorpus Altdeutsch

Das Konzept, die Realisierung und die neuen Möglichkeiten

1. Das DDD-Konzept für ein historisches Referenzkorpus des Deutschen

Das Referenzkorpus Altdeutsch ist kein singuläres, isoliertes Korpus einer einzelnen Sprachstufe des Deutschen, sondern es ist Teil eines viel umfassenderen Korpusunternehmens „Deutsch.Diachron.Digital“ (DDD), das sich zum Ziel gesetzt hat, für das Deutsche ein sprachstufenübergreifendes Referenzkorpus aufzubauen. Auf diese Weise soll der historischen Sprachwissenschaft des Deutschen ein innovatives Forschungstool an die Hand gegeben werden, das in elektronischer Form zuverlässige Textausgaben bereit stellt und umfangreich syn- und diachrone Recherchemöglichkeiten bietet, einschließlich der Möglichkeit quantitativer Untersuchungen. Den Anstoß zu diesem Projekt lieferte ein Treffen von an und mit Korpora arbeitenden germanistischen Sprach- und Literaturwissenschaftler/innen, die sich im Jahr 2002 darauf verständigt haben, den Aufbau eines solchen diachronen Korpus für das Deutsche in Kooperation anzugehen. In den folgenden Jahren wurde das DDD-Konzept für ein historisches Referenzkorpus des Deutschen unter maßgeblicher Beteiligung der Korpuslinguistin Anke Lüdeling (HU Berlin) und des Informatikers Ulf Leser (ebenfalls HU Berlin) im Detail ausgearbeitet. Es beinhaltet folgende Festlegungen:

(1) Zeitliche Abdeckung: In seiner Endausbaustufe soll das DDD-Korpus alle Sprachstufen des Deutschen erfassen und diachrone Recherchen zur Entwicklungsgeschichte des Deutschen ermöglichen von den Anfängen der Überlieferung bis an die Gegenwart (avisiert ist die Zeit um 1900). Der Schwerpunkt des DDD-Korpus liegt auf den frühen Sprachstufen des Deutschen bis ca. 1650, die Zeit nach 1650 wird nur in Ausschnitten erfasst.

(2) Räumliche Abdeckung: Das DDD-Korpus soll den gesamten deutschen Sprachraum erfassen. D.h., das Korpus ist nicht auf den hochdeutschen Sprachraum beschränkt, sondern es schließt auch die niederdeutschen Überlieferungen mit ein. Entwicklungsprozesse werden mithin über den ganzen heute vom modernen hochdeutschen Standard überdachten Sprachraum möglich sein.

(3) Textauswahl: Angestrebt wird ein im linguistischen Sinne ausgewogenes Korpus, austariert nach einem dreidimensionalen Raster von Zeit, Raum und Textart, das sich in dieser Form bereits bei anderen historischen Korpusprojekten (dem Bonner Frühneuhochdeutsch-Korpus und dem Korpus für die Neubearbeitung der mittelhochdeutschen Grammatik) bewährt hat. Dabei werden die drei Dimensionen in folgender Weise aufgeschlüsselt: Die zeitliche Gliederung erfolgt in Schritten von 50 Jahren. Die räumliche Gliederung orientiert sich an der üblichen Dialektgliederung des Deutschen. Darüber hinaus werden in der Dimension Textart Prosa und Vers voneinander unterschieden.

(4) Annotationstiefe: Das DDD-Korpus wird ein tiefenannotiertes Korpus sein. Es wird umfangreiche strukturelle (Satz, Zeile, Abschnitt, Seite usw.) und positionelle (Lemma, Wortart, Flexion usw.) Annotationen enthalten. Außerdem sind ausführliche Header-Informationen vorgesehen, die Auskunft über Herkunft von Texten geben und viele weitere Informationen enthalten, die für die sprachhistorische Bewertung der im Korpus enthaltenen Texte relevant sind.

(5) Flexibilität: Um das Korpus erweiterbar und maximal flexibel zu machen, wird eine Standoff-Architektur gewählt. D.h., die Annotationen werden nicht im originären Textfile, sondern in jeweils separaten Files abgespeichert. Damit wird sichergestellt, dass Nutzer beliebige zusätzliche Annotationsebenen einfügen und gegebenenfalls auch vorhandene Annotationsebenen für sich löschen oder überarbeiten können.

Das DDD-Konzept für ein historisches Referenzkorpus für das Deutsche wurde 2006 bei der Deutschen Forschungsgemeinschaft eingereicht und anschließend nach Gutachterhinweisen noch einmal überarbeitet. Entsprechend den Vorgaben der DFG wurden die DDD-Teilkorpora in der Folge separat und in zeitlicher Staffelung beantragt. 2009 begannen die Arbeiten an den Referenzkorpora Altdeutsch und Mittelhochdeutsch. 2012 startete das Referenzkorpus Frühneuhochdeutsch und 2013 das Referenzkorpus Mittelniederdeutsch. Die Arbeiten an den Referenzkorpora Altdeutsch und Mittelhochdeutsch sind mittlerweile abgeschlossen. Sie werden 2015 in einer initialen Version 1.0 über die ANNIS-Datenbank zugänglich gemacht, die ihrerseits im Rahmen des SFB 632 Informationsstruktur (Universität Potsdam/Humboldt-Universität Berlin) entwickelt und aufgebaut wurde.

2. Die Realisierung im Referenzkorpus Altdeutsch

Das Referenzkorpus Altdeutsch, das an den Standorten Berlin (Donhauser), Frankfurt (Gippert) und Jena (Lühr) erarbeitet wurde, ist mit ca. 1 Million Textwörtern (darunter ca. 300.000 Textwörter in Latein) das mit Abstand kleinste der DDD-Teilkorpora – gleichzeitig aber auch das einzige, das die Textüberlieferung einer ganzen Sprachperiode (hier von den Anfängen der Überlieferung bis Mitte des 11. Jahrhunderts) vollständig zur Verfügung stellt. D. h., das Korpus umfasst alle althochdeutschen und altsächsischen Textdenkmäler, die aus der Zeit überliefert sind. Dies ist einerseits natürlich ein Vorteil, andererseits aber auch ein Nachteil, denn die Textüberlieferung aus dieser Zeit ist im Vergleich zu späteren Sprachperioden recht klein und so ungleichmäßig verteilt, dass die Zellen des der DDD-Architektur zugrunde liegenden dreidimensionalen Rasters von Zeit, Raum und Textart bei weitem nicht alle gefüllt und nicht gleichmäßig mit Textvolumina belegt werden können. So ist z. B. mit dem umfangreichen Werk Notkers von St. Gallen fast die Hälfte der altdeutschen Textüberlieferung (ca. 340.000 Textwörter) auf nur einen Zeitraum-Slot, nämlich auf den Dialektraum des Alemannischen in der ersten Hälfte des 11. Jahrhunderts, konzentriert. Ein zweiter markanter Überlieferungsschwerpunkt findet sich im 9. Jahrhundert. Die aus diesem Jahrhundert überlieferten Texte (Otfrid, südrheinfränkisch; Tatian, ostfränkisch und Heliand, altsächsisch) sind ebenfalls vergleichsweise groß und beinhalten ihrerseits in etwa die Hälfte des verbleibenden Textvolumens. Das Referenzkorpus Altdeutsch ist mithin überlieferungsbedingt zwangsläufig nicht ausgewogen.

Die althochdeutschen und altsächsischen Texte, die über das Referenzkorpus Altdeutsch digital recherchierbar sind, werden in einer philologisch gesicherten und möglichst handschriftennahen Form bereitgestellt. Grundlage für die Textfassung ist die jeweils beste verfügbare Edition der jeweiligen Handschrift. Auf einen Rückgriff auf die Handschrift selbst, der sehr arbeitsaufwändig ist, aber im DDD-Konzept ursprünglich angedacht war, wurde im Referenzkorpus Altdeutsch nach reiflicher Überlegung aus Kostengründen verzichtet, weil die verfügbaren Editionen im Altdeutschen bereits relativ handschriftennah sind und – anders als dies im Mittelhochdeutschen der Fall ist – keine systematische Überformung in Richtung auf einen gedachten Standard vornehmen. Die Möglichkeit eines späteren Einbaus der Handschriftenversion ist im technischen Konzept der Referenzkorpus Altdeutsch aber bereits angelegt.

Der eigentliche Mehrwert, den das Referenzkorpus Altdeutsch für seinen Nutzer bietet, ergibt sich aus Art und Umfang der Annotationen, die mit den Texten verknüpft sind. Das Altdeutsch-Korpus richtet sich hier – ebenso wie die anderen DDD-Teilkorpora – nach den im DDD-Konzept erarbeiteten Vorgaben, die in folgender Weise umgesetzt wurden.

Header-Informationen

Die Metadaten, die im Altdeutsch-Korpus zu den einzelnen Texten verfügbar sind, umfassen fünf Gruppen von Informationen:

1. Informationen zum Text: Verzeichnet werden der Textname, das korpusinterne Kürzel sowie Angaben zur Textart (Prosa/Vers) und zum inhaltlichen Bereich (Textbereich), zu dem der Text gehört.
2. Informationen zum Überlieferungsträger, d.h. zu dem Codex, auf dem der Text verzeichnet ist: Hier geht es um Angaben zur Signatur des Codex, zum Aufbewahrungsort, zum Entstehungsort, der Entstehungszeit sowie zu Umfang und Gestaltung der Handschrift.
3. Informationen zur Niederschrift: Hier werden alle Angaben aufgeführt, die die konkrete Niederschrift des Textes betreffen, also Informationen zur Sprache (Althochdeutsch, Altsächsisch, Latein usw.), zum Dialekt- raum, zum Lateinbezug (mit oder ohne lateinische Vorlage) sowie erneut zum Entstehungsort und Entstehungszeit, die von Entstehungsort und -zeit des Überlieferungsträgers abweichen können, wenn der altdeutsche Text nachträglich auf einem bestehenden Codex verzeichnet wurde.
4. Informationen zur Vorlage: Für den Fall, dass ein Text eine Vorlage hat, werden bestimmte Basisinformationen zur Vorlage (Datierung, Lokalisierung, Autor, Sprache) ebenfalls mit angeführt.
5. Referenzen: Unter der Rubrik „Referenzen“ findet sich die Angabe der benutzten Textedition, dazu Verlinkungen zum Paderborner Repitorium (Handschriftenzensus) und – falls vorhanden – zum Online-Zugriff auf ein Handschriften-Faksimile.

Abbildung 1 enthält ein konkretes Beispiel der Header-Informationen, die im Korpus für den althochdeutschen Tatian verfügbar sind.

Text	Kürzel	T
	Name (weitere Namen)	(Ahd.) Tatian
	Textbereich	Religion
	Textart	Prosa
Überlieferungs- träger (HS)	Signatur	Cod. 56
	Überlieferungsträger	Handschrift
	Aufbewahrungsort	St. Gallen, Stiftsbibliothek
	alte Signatur	
	Entstehungszeit	9,1
	genauere Datierung	
	Entstehungsort	Fulda
	Schreiber/Drucker	
	Inhalt	Tatian-Bilingue
	Blattanzahl	171
	Format (Blattgröße in cm)	33,5 x 22,5
	Hände	6
	Schrift	karolingische Minuskel
	Spaltenzahl	2
	weitere HSS	
Besonderheiten	linke Spalte lat., rechte Spalte ahd. Text	
Text (Niederschrift)	Blattangabe	1-171
	Sprache 1	ahd.
	Sprache 2	lat.
	sprachl. Großraum 1	obd.
	sprachl. Großraum 2	
	Sprachlandschaft 1	ofrk.
	Sprachlandschaft 2	
	Schreibort	Fulda
	Entstehungszeit	9,1
	Schrift	karolingische Minuskel
	Hände	6
	Lateinbezug	Übersetzung
	Besonderheiten	
weitere Überlieferung	Ms. Jun. 13, Oxford, Bodleian Library (Tatian-Fragmente aus einer inzwischen verschollenen Hs, veröffentlicht ab 1597); Ms. Lat. 7641, Paris, Bibliothèque Nationale (einige aus den Schlusskapiteln des Tatian exzerpierte Wörter, Sätze und Phrasen)	
Text: Vorlage	Name	
	Datierung	um 170 n. Chr.
	Lokalisierung	Syrien oder Rom
	Autor	Tatian, syrischer Mönch
	Sprache	syrisch oder griechisch
Referenzen	Edition	Sievers, Eduard (Hg.), Tatian. Lateinisch und Altdeutsch mit ausführlichem Glossar. Paderborn, 1872.
	Handschriftencensus	http://www.handschriftencensus.de/16961
	online-Faksimile	http://www.e-codices.unifr.ch/de/preview/csg/0056

Abb. 1: Metadaten zum ahd. Tatian

Strukturelle Annotationen

Eine zweite Gruppe von Annotationen, die das Altdeutsch-Korpus in Übereinstimmung mit dem DDD-Konzept umsetzt, betrifft die Markierung von strukturellen Einheiten oberhalb der Basiskategorie 'Wort'.

Systematisch markiert werden dabei zum einen alle Ebenen der Textgliederung, die in den Handschriften selbst gekennzeichnet werden. Dabei geht es um die Untergliederung von Texten in Kapitel, Unterkapitel und Abschnitte ebenso wie um die Kennzeichnung von Manuskriptseiten und Manuskriptzeilen, die bekanntermaßen in einzelnen Texten von herausragender philologischer und auch linguistischer Bedeutung sind. So orientiert sich z. B. die althochdeutsche Übersetzung der lateinischen Evangelienharmonie weitgehend an den Zeilenvorgaben der lateinischen Vorlage. Die Übersetzung erfolgt in der Regel „interlinear“ – Verschiebungen über Zeilengrenzen hinweg sind vergleichsweise selten. Das Wissen um die Positionierung von Wörtern in Zeilen ist hier deshalb an vielen Stellen elementar für die Bewertung von linguistischen Phänomenen, insbesondere im Bereich der Wortstellung.

Ergänzend dazu werden Zeile und Seite der Edition indiziert. Dies hilft bei der Identifikation von Stellenangaben, wie sie in der einschlägigen sprachhistorischen Literatur verwendet werden.

Annotiert werden ferner die literarischen Einheiten Halbvers und Vers, die mit Metrik und Reim auf die sprachliche Organisation von Texten einwirken, vor allem aber auch die zentrale linguistische Einheit Satz, die hier im Sinne des englischen Begriffs „clause“ gebraucht wird.

Positionelle Annotationen

Der Schwerpunkt der Annotationen, die das Referenzkorpus Altdeutsch bereithält, liegt im Bereich der positionellen Annotationen, die auf die Einheit Wort bezogen sind. Hier geht die Annotation linguistisch in die Tiefe: Die Wörter werden lemmatisiert. Sie werden nach Wortarten klassifiziert und flexionsmorphologisch analysiert. Außerdem gibt es Angaben zur Wortbedeutung.

Das Referenzkorpus Altdeutsch greift dabei bewusst auf bereits vorhandenes Wissen zurück, das in den einschlägigen Wörterbüchern zum Altsächsischen und zum Althochdeutschen sowie zahlreichen Textwörterbüchern und Textglossaren niedergelegt ist. Diese Hilfsmittel wurden im Projekt digitalisiert.

Dann wurden relevante Informationen (Bedeutung, Wortart, Flexionsklassen usw.) ausgelesen und mit den zu bearbeitenden Texten verknüpft. Die auf diese Weise im Wesentlichen automatisch erzeugte Vorannotation wurde dann in einem zweiten Schritt von menschlichen Annotatoren ergänzt und korrigiert und abschließend – in einem dritten Schritt – einer semi-automatischen Konsistenzprüfung unterzogen.

Grundlage für die Wortart-Annotation im Referenzkorpus Altdeutsch ist ein spezielles Tagset (HiTS), das zusammen mit dem Referenzkorpus Mittelhochdeutsch erarbeitet wurde und auf dem in Stuttgart und Tübingen entwickelten Tagset für das moderne Deutsch (STTS) aufbaut. Dabei nimmt HiTS eine Reihe von Anpassungen vor, die strukturelle Unterschiede zwischen dem Neuhochdeutschen und dem Althochdeutschen, dem Altsächsischen oder dem Mittelhochdeutschen reflektieren. Beispiele dafür sind die Einführung des Tags PTKREL für die Relativpartikel *the*, die im Althochdeutschen belegt ist, oder aber die Möglichkeit, den Infinitiv mit Kasusmarkierungen auszuweisen. Es gibt aber auch tiefer gehende Eingriffe bzw. Umorganisationen. So verzichtet HiTS z. B. auf den Tag ART (Artikel), sieht dafür aber die Option vor, im Bereich der definiten und indefiniten Determinative (DD bzw. DI) artikelartige Verwendungen (DDA bzw. DIA) zu markieren. Dies trägt der Tatsache Rechnung, dass sich das Artikelsystem im Altdeutschen erst in Entwicklung befindet. Die Vergabe eines Tags ART würde vom Annotator also eine Entscheidung verlangen, die er oft nicht sinnvoll treffen kann. Die Kodierung als DDA oder DIA ist deutlich weniger voraussetzungsreich, weil keine Entweder-oder-Entscheidung vorzunehmen ist. Gleichzeitig wird so aber auch die Möglichkeit geschaffen, im Altdeutschen gezielt nach Vorformen des späteren Artikels zu suchen.

Eine Besonderheit der wortbezogenen Annotationen im Referenzkorpus Altdeutsch ist die „Doppelung“ der Annotation von Wortarten und Flexionsklassen, die sowohl mit Bezug auf das Lemma wie auch mit Bezug auf den konkreten Beleg indiziert werden. Im Regelfall sind diese Indizierungen identisch. Allerdings bietet diese Aufspaltung der Wortarten- und Flexionsklassen-Annotation hervorragende Möglichkeiten, etablierte Zuordnungen zu überprüfen und Sprachwandelprozesse wie Flexionsklassenübertritte oder Kategorienwechsel nachzuvollziehen. Auch können Schwankungen im Gebrauch von Wörtern und Wortformen auf diese Weise identifiziert und quantifiziert werden.

Die strukturelle Einheit „clause“ ist ebenfalls mit positionellen Annotationen unterlegt. Indiziert werden zum einen bestimmte formale und funktionale Merkmale von „Sätzen“:

Handelt es sich um Sätze mit finitem Verb (CF), um infinitivische (CI) oder partizipiale (CP) Strukturen oder um einen elliptischen Satz (CE)?

Ist ein Satz eingeleitet (C__I) oder uneingeleitet (C__U)?

Um was für einen funktionalen Typ von Satz handelt es sich? Unterschieden werden deklarative Hauptsätze (C__M), Frage- und Imperativsätze (C__Int, C__Imp), Subjekt- und Objektsätze (C__S, C__O), Adverbialsätze (C__Adv) und natürlich auch Relativsätze (C__Rel). Bei Adverbialsätzen ist zudem die Möglichkeit einer semantischen Spezifikation (Temp, Loc, Caus, usw.) vorgesehen.

Außerdem ist es möglich, eine Kennzeichnung vorzunehmen, wenn Sätze unterbrochen und nach einer Unterbrechung weitergeführt werden: C_S, C_C.

Dieses Vorgehen erhebt keinen theoretischen Anspruch. Es ist ausschließlich praktisch motiviert. Es möchte dem Nutzer des Korpus sinnvolle und wichtige Rechercheoptionen erschließen, ohne den Annotator zu überfordern oder ihm zeitaufwändige Klärungen aufzuerlegen. In diesem Sinne hat sich unsere Strategie, das Korpus in gewissem Umfang auch mit syntaktischen Informationen anzureichern, in der Arbeit bewährt.

Abbildung 2 zeigt ein Annotationsbeispiel im Format des Partitureditors ELAN, den wir im Projekt benutzt haben:

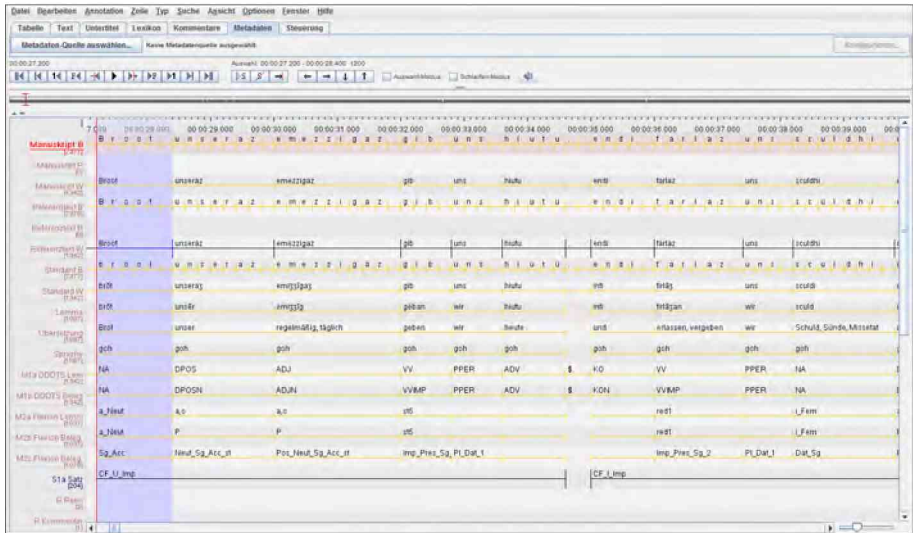


Abb. 2: Annotationsbeispiel – Strukturelle und positionelle Annotationen

Zugang

Der Zugang zum Referenzkorpus Altdeutsch erfolgt über die Suchoberfläche der ANNIS-Datenbank, die unter der Adresse www.deutschdiachrondigital.de erreichbar ist und die mit Blick auf die DDD-Korpora mit zusätzlichen Suchoptionen ausgestattet wurde. Insbesondere wurde ergänzend zu den bestehenden Möglichkeiten, Abfragen in einer formalen Abfragesprache zu formulieren, eine Art „Komfortzugang“ eingerichtet, der es erlaubt, auch ohne Kenntnis dieser Abfragesprache im Korpus zu recherchieren und dabei auch komplexe Anfragen vorzunehmen. So kann über den Komfortzugang direkt auf eine Liste der Annotationsebenen des Korpus und der dort vergebenen Werte zugegriffen und damit direkt Suchwerte ausgewählt werden. Dabei lassen sich mehrere Suchwerte miteinander verbinden, sowohl im Sinne eines ‘und’ wie auch im Sinne eines ‘oder’. Zudem können Kontexte für den Suchwert definiert werden. In einem weiteren Schritt kann eine solche Suche dann mit Metadaten verknüpft werden, z.B. mit Festlegungen in Hinblick auf einen bestimmten Zeitraum, bestimmte Dialekträume oder bestimmte inhaltliche Gruppen von Texten. Dieses Abfragesystem ist derzeit in Erprobung und wird bis zum Veröffentlichungstermin noch weiter verfeinert.

3. Die neuen Möglichkeiten

Wie die Korpusbeschreibung im vergangenen Abschnitt bereits deutlich macht, erhalten wir mit dem Referenzkorpus Altdeutsch eine qualitativ hochwertige und leistungsfähige Ressource, die es möglich macht, schnell und effizient in der kompletten altdeutschen Überlieferung zu recherchieren. Das ist definitiv eine neue Option, die in dieser Form noch keiner Forschergeneration zu Verfügung stand und geradezu dazu einlädt, unsere Wissensbestände über diese frühe Phase der Entwicklungsgeschichte des Deutschen einer systematischen Überprüfung zu unterziehen.

Wir haben projektbegleitend in der letzten Phase unseres Vorhabens eine Reihe von Probeuntersuchungen vorgenommen, die dieses Potential sichtbar machen, auch wenn die konkreten Ergebnisse dieser Pilotstudien sicher noch zu korrigieren sind, da zum Abfragezeitpunkt noch nicht das ganze Korpus zur Verfügung stand und auch noch nicht alle Korrekturvorgänge durchgeführt worden waren. Ein gutes Beispiel für eine solche 'Probebohrung' ist eine kleine Studie, die Sonja Linde 2012 zu den für das Althochdeutsche angesetzten Deklinationsklassen des Substantivs durchgeführt hat. Dabei geht sie u.a. der Frage nach, wie gut der Ansatz einer jeweils eigenen Deklinationsklasse von *ja*- und *jo*-Stämmen im Althochdeutschen mit Daten hinterlegt ist. Die folgenden drei Diagramme (Abb. 3-5) zeigen das Ergebnis der Korpusabfrage (Version 0, Stand Nov. 2012), die ins Verhältnis setzt diejenigen Flexionsformen bei den zu den *ja*- bzw. *jo*-Stämmen gerechneten Substantiven (Maskulina, Neutra und Feminina), die noch einen Reflex des *-j* aufweisen, zu denen, bei denen dies nicht mehr der Fall ist.

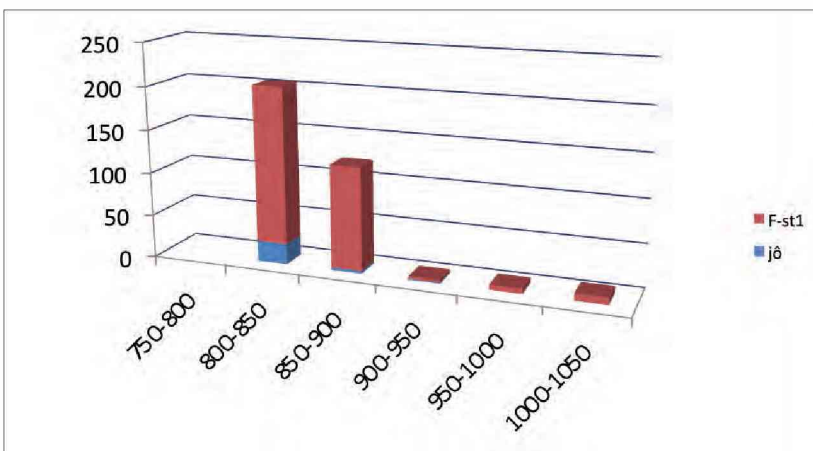


Abb. 3: *jo*-Feminina

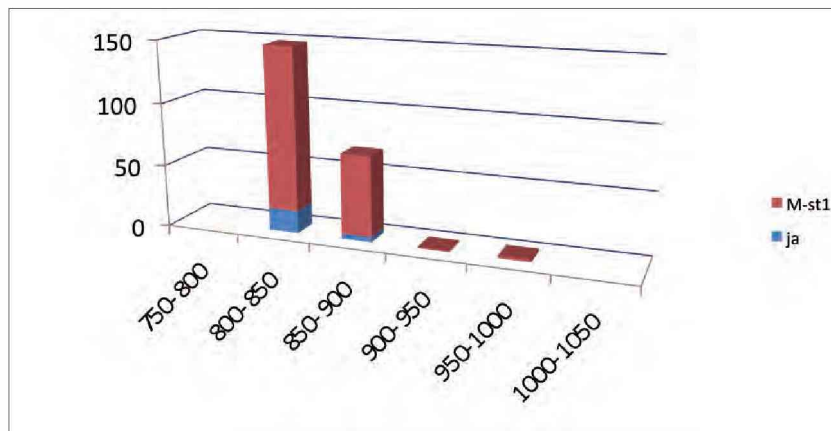


Abb. 4: ja-Maskulina

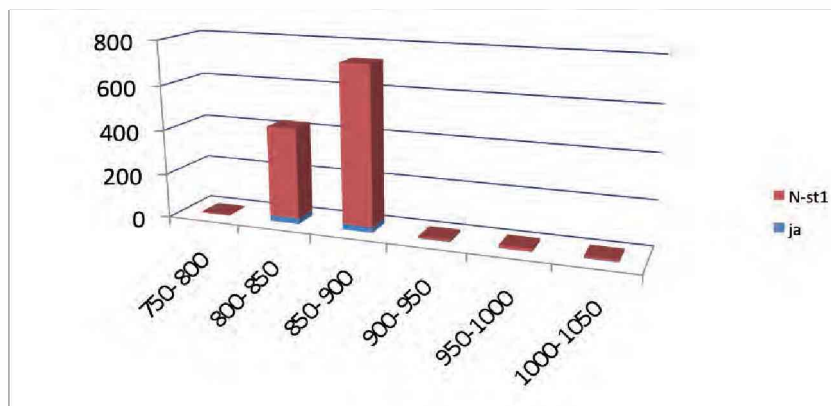


Abb. 5: ja-Neutra

Der Befund ist eindeutig und in allen drei Fällen der gleiche: *j*-haltige Formen (in den Diagrammen blau markiert) sind bei den einschlägigen Substantiven nur peripher nachzuweisen. Die Nachweise sind auf die erste Hälfte des 9. Jahrhunderts konzentriert. Nach Ende des 9. Jahrhunderts sind keine Vorkommen mehr belegt. Wenn sich diese Ergebnisse – was absehbar ist – bei der Nachrecherche in einer Korpusversion 1.0 bestätigen lassen, würde der Ansatz dieser Deklinationsklassen für das Althochdeutsche nachhaltig in Frage gestellt. In jedem Fall ist er stark historisierend. Wenn überhaupt, dann beschreibt er einen sprachlichen Zustand, der zeitlich deutlich vor dem Einsetzen der Überlieferung liegt.

Ergebnisse wie diese, die wir bei unseren ‘Probebohrungen’ zusammengetragen haben, werfen für uns letztlich aber eine viel grundsätzlichere Frage auf, nämlich, wie wir die hochauflösenden Daten, die sich aus dem Korpus entnehmen lassen, benutzen können, um ein möglichst differenziertes Bild von der Entwicklungsgeschichte des Deutschen in dieser frühen Phase zu zeichnen, das ohne die Projektion einer ‘Sprachstufe’ Althochdeutsch auskommt und ohne die Projektion einer Sprachgrenze, die das Altsächsische vom Althochdeutschen abtrennt.

Die Idee, die wir derzeit verfolgen, ist ein konsequent variationistischer Ansatz, der vom einzelnen Text ausgeht und kartographiert, wie bestimmte sprachliche Merkmale in Zeit und Raum über den altdeutschen Überlieferungsraum verteilt sind. Zu diesem Zweck experimentieren wir mit abstrakten chronographischen Karten, auf denen die einzelnen Überlieferungen in einem Raum-Zeit-Raster verankert sind. Abbildung 6 zeigt ein Beispiel für eine solche Karte, in diesem Fall für das Vorkommen von V1-Stellungen. Dabei wird auch der Umfang einer Überlieferung abgebildet – durch die Größe des Kreises, der erscheint, wenn ein Phänomen in dem jeweiligen Text belegt ist. Überlieferungen, die keinen Beleg für das untersuchte Phänomen aufweisen, werden durch Kreuze markiert. Die Frequenz der Vorkommen im einzelnen Text ist hier über Farbabstufungen zum Ausdruck gebracht.

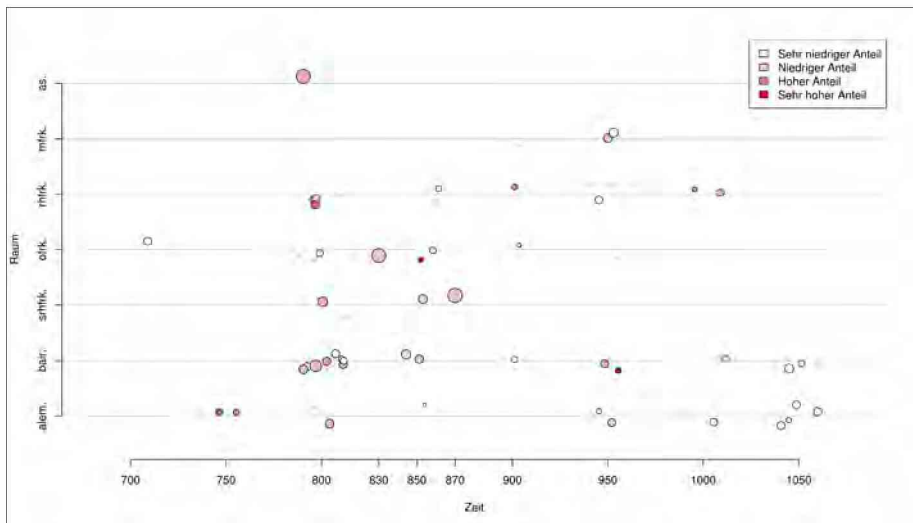


Abb. 6: Chronographische Karte: Verteilung von V1-Stellungen

Dieses Verfahren, an dessen Verfeinerung wir im Moment arbeiten, hat sich aus unserer Sicht grundsätzlich bewährt. So lässt sich relativ rasch und gut erkennen, wie ein Phänomen in Raum und Zeit verteilt ist – z.B. im Fall der V1-Sätze die raumunabhängige Verbreitung ebenso wie die tendenzielle Abnahme der Belege gegen Ende der altdeutschen Zeit.

In einem nächsten Schritt wollen wir die Informationen, die diese Karten enthalten, zusammenfügen und die Abstände zwischen einzelnen Textüberlieferungen berechnen. Auf diese Weise lässt sich ein – zunächst wenig voraussetzungshaltiges – Bild davon gewinnen, wie die altdeutsche “Überlieferungswolke” im Grundsatz strukturiert ist. Wie eng gruppiert oder wie weit von einander entfernt sind eigentlich diese Texte? Unterscheiden sich die altsächsischen Textdenkmäler wirklich so stark von den Texten althochdeutscher Dialekte, dass der Ansatz zweier Sprachen gerechtfertigt erscheint? Ist es primär die Dimension Zeit, die Differenzen zwischen Texten evoziert, oder doch eher die Kategorie Raum? Oder gibt es Hinweise auf weitere Dimensionen (z.B. Register), die wir in Ansatz zu bringen haben? Dies alles sind Fragen, die sich mit den neuen Möglichkeiten des Referenzkorpus Altdeutsch und perspektivisch mit den anderen DDD-Teilkorpora auch über weitere Überlieferungsstrecken des Deutschen adressieren lassen.

Literatur

1. Wörterbücher

Heffner, Roe-Merill Secrist (1961): A Word-Index to the Texts of Steinmeyer. Die kleineren althochdeutschen Sprachdenkmäler. Madison: The University of Wisconsin Press.

Hench, George Allison (1890): The Monsee Fragments. Straßburg: Trübner.

Hench, George Allison (1893): Der althochdeutsche Isidor. Straßburg: Trübner.

Kelle, Johann (1881): Glossar der Sprache Otfrids. Regensburg: Manz.

Sehrt, Edward (1955): Notker-Wortschatz. Halle: Niemeyer.

Sehrt, Edward (1966): Vollständiges Wörterbuch zum Heliand und zur altsächsischen Genesis. Göttingen: Vandenhoeck & Ruprecht.

Sievers, Eduard (1874): Die Murbacher Hymnen. Halle: Buchhandlung des Waisenhauses.

Sievers, Eduard (1892): Tatian. Lateinisch und althochdeutsch mit ausführlichem Glossar. 2. Aufl. Paderborn: Schöningh.

Splett, Jochen (1993): Althochdeutsches Wörterbuch. Berlin: de Gruyter.

Wadstein, Elis (1899): Kleinere altsächsische Sprachdenkmäler: mit Anmerkungen und Glossar. Norden und Leipzig: Dieder. Soltau's Verlag.

2. Sekundärliteratur¹

Dipper, Stefanie/Faulstich, Lukas/Leser, Ulf/Lüdeling, Anke (2004): Challenges in modelling a richly annotated diachronic corpus of German. In: Proceedings of the Workshop on XML-based richly annotated corpora, Lisbon, Portugal, May 2004. www2.informatik.hu-berlin.de/Forschung/Lehre/wbi/publications/2004/xbrac04_final.pdf.

Dipper, Stefanie/Donhauser, Karin/Klein, Thomas/Linde, Sonja/Müller, Stefan/Wege-
ra, Klaus-Peter (2013): HiTS: ein Tagset für historische Sprachstufen des Deut-
schen. In: Zinsmeister, Heike/Heid, Ulrich/Beck, Kathrin (Hg.): Das Stuttgart-
Tübingen Tagset – Stand und Perspektiven. JLCL 28(1): 85-137. [www.jlcl.org/
2013_Heft1/H2013-1.pdf](http://www.jlcl.org/2013_Heft1/H2013-1.pdf).

Faulstich, Lukas C./Leser, Ulf/Lüdeling, Anke (2005): Storing and querying historical texts in a relational database. Technical Report 176 des Instituts für Informatik der Humboldt-Universität zu Berlin, Januar 2005. <http://edoc.hu-berlin.de/series/informatik-berichte/176/PDF/176.pdf>.

Hoenen, Armin/Jügel, Thomas (Hg.) (2012): Altüberlieferte Sprachen als Gegenstand der Texttechnologie / Ancient languages as the object of text technology. JLCL 27(2). http://www.jlcl.org/2012_Heft2/H2012-2.pdf.

Linde, Sonja (2012): Manuelle Abgleichung bei automatisierter Vorannotation: Das Tagging grammatischer Kategorien im Referenzkorpus Altdeutsch. In: JLCL 27(2): 53-64. www.jlcl.org/2012_Heft2/4Linde.pdf.

Linde, Sonja/Mittmann, Roland (2013): Old German reference corpus. Digitizing the knowledge of the 19th century. Automated pre-annotation using digitized historical glossaries. In: Bennett, Paul/Durrell, Martin/Scheible, Silke/Whitt, Richard J. (Hg.): New Methods in Historical Corpora (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache / Corpus Linguistics and Interdisciplinary Perspectives on Language (CLIP) 3). Tübingen: Narr, 235-246.

Lüdeling, Anke (2007): Überlegungen zum Design und zur Architektur von diachronen Korpora. In: Sprache und Datenverarbeitung, Sonderheft Diachrone Corpora, historische Syntax und Texttechnologie 31(1-2) (Gastherausgeber: Hans Christian Schmitz und Jost Gippert), 7-14.

Lüdeling, Anke/Poschenrieder, Thorwald/Faulstich, Lukas C. (2005): DeutschDiachronDigital – Ein diachrones Korpus des Deutschen. In: Braungart, Georg/Eibl, Karl/Jannidis, Fotis (Hg.): Jahrbuch für Computerphilologie (6) 2004. Paderborn: mentis, 119-136. <http://computerphilologie.tu-darmstadt.de/jg04/luedeling/ddd.html>.

¹ Alle Internetadressen zuletzt aufgerufen am 28. Januar 2015.

- Mittmann, Roland (2012): Digitalisierung historischer Glossare zur automatisierten Vorannotation von Textkorpora am Beispiel des Altdeutschen. In: In: JLCL 27(2): 39-52. www.jlcl.org/2012_Heft2/3Mittmann.pdf.
- Mittmann, Roland (2013): Old German and Old Lithuanian: the creation of two deeply-annotated historical text corpora. In: Коллектив авторов (Ответственные редакторы: Захаров, Виктор П. / Митрофанова, Ольга А. / Хохлова, Мария В.) [Autorenkollektiv (Verantwortliche Redakteure: Zakharov, Victor P./Mitrofanova, Olga A./Khokhlova, Maria V.)]: Труды международной научной конференции «Корпусная лингвистика – 2013» / Proceedings of the international conference «Corpus linguistics – 2013». Санкт-Петербург: Санкт-Петербургский государственный университет, Филологический факультет / St. Petersburg: St. Petersburg State University, Philological Faculty, 103-111.
- Mittmann, Roland (2015, in diesem Band): Automated quality control for the morphological annotation of the Old High German text corpus. Checking the manually adapted data using standardized inflectional forms. In: Gippert, Jost / Gehrke, Ralf (Hg.): New methods in historical corpora. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache / Corpus Linguistics and Interdisciplinary Perspectives on Language (CLIP) 5). Tübingen: Narr, 65-76.
- Zeldes, Amir/Ritz, Julia/Lüdeling, Anke/Chiarcos, Christian (2009): ANNIS: A search tool for multi-layer annotated corpora. In: Proceedings of Corpus Linguistics 2009, July 20-23, Liverpool, UK. <http://edoc.hu-berlin.de/oa/conferences/reS4Xo05sncZc/PDF/29i8VT1zYfT3M.pdf>.

Analyzing formulaic patterns in historical corpora

Abstract

This paper aims to point out a linguistic phenomenon that due to the current stage of research can be analysed only insufficiently with the help of an electronic text corpus. In this way, the paper adds a new aspect to the discussion about historical corpora by tackling the question of how they should be designed in order to be useful for linguistic research on so-called *formulaic patterns*. The novelty of the question becomes apparent considering the fact that at present such historical corpora do not exist. In section 1, we define the term *formulaic pattern* because a clear understanding of this phenomenon is a prerequisite condition for collaborative research of it by historians of language and corpus and computer linguists. Section 2 gives a brief outline of the state of the art in the field of modern formulaic language within the framework of corpus and computer linguistics. Section 3 shows that some well known problems in this area are exacerbated when applied to historical texts. Section 4 presents a possible solution that has been implemented by the HiFoS Research Group at the University of Trier (Germany). Joint research efforts planned with UKP Lab at the TU Darmstadt (section 5) demonstrate that the restrictions posed by historical formulaic patterns are challenges to be overcome, rather than insurmountable obstacles.

1. Formulaic patterns at the crossroads of (historical) linguistics, corpus and computer linguistics: Looking for a common ground

Formulaic patterns is not yet a well-established linguistic term. The linguistic phenomena it is applied to have been studied mostly within the framework of phraseological research. *Phrasemes*, *set phrases* or in German *Phraseologismen* are understood here as expressions that comprise a minimum of two words (constituents) and a maximum of a sentence¹. They are syntactically more or less frozen and can (but need not) be semantically ambiguous or idiomatic (Burger 2010: 11-15). For different phrasemes, these semantic and syntactic features are characteristic to a different extent. Therefore, linguists distinguish various types of phrasemes: idioms (*to spill the beans*, *to kick the bucket*), binomials (*in black and white*), collocations (*to give a paper*, *boiling water*), proverbs (*Clothes make the man*), routine formulae (*Ladies and Gentlemen!*, *The floor is*

¹ From a historical point of view, these criteria can only serve as a starting point when identifying formulaic patterns. Cf. Filatkina et al. (2009) and section 3 in this paper for more details.

yours!), quotations (*To Be or not to Be – That is the question!*) and so on, to name just a few types. Phrasemes are conventionalised expressions reproduced by speakers in their particular structure and meaning. Furthermore, some types of phrasemes are strongly tied to a particular cultural background and transmit cultural elements through their image components (Dobrovolskij/Piirainen 2005). Since its establishment in the 1940s and its development into an independent, international branch of linguistics in the 1970s, phraseological studies have proven that phrasemes are a universal phenomenon typical for all modern languages but strongly dependant on the communicative and cultural conventions of a given language. Until very recently, phraseological research has mostly addressed modern languages spoken in Europe;² its main focus has been semantics and pragmatics.

A significant shift towards the investigation of the structure of phrasemes, their high potential for variation and occasional modifications occurred in the 1990s, partly driven by the advent of corpus and computer linguistics (Fellbaum (ed.) 2006, Heid 2008). As computer linguistics uses many methods that have evolved from corpus linguistics, we will only use the term computer linguistics for the rest of the paper and implicitly refer to corpus linguistics as well. What is described here with the terms *collocations* (Fellbaum 2007; Evert 2005, 2008) or *multi-word-units* (Tschichold 2000, Sag et al. 2001) differs from the above mentioned linguistic definitions. Collocations and multi-word-units are understood as statistically significant co-occurrences of mostly two lexical items. According to this definition, the example *New York* is considered to be a typical collocation in corpus linguistics whereas its consideration as a phraseme is questionable for a phraseologist.³ Despite these differences, corpus linguistics was one of the first disciplines to prove the high significance of syntactically more or less frozen constructions for human communication.⁴ Sinclair's idiom principle (1987), Hoey's lexical priming model (2005), or Wray's concept of formulaic language (2008) build on the basic idea that in a natural communication process the words of a language exist not in isolation but in a syntagmatic interplay with each other. As a result of this interplay, words evoke meanings attested to them by the members of their respective language com-

² Cf. a different approach in Piirainen (2012).

³ For more discussion concerning the definition problem cf. recently Sirajzade (2012).

⁴ Jackendoff (1995: 156) observes for English: "The lexicon of a language available to speakers in everyday situations contains at least as many multi-word expressions as single words." Cf. also Sag et al. (2001), Fellbaum (2007).

munity. While phraseological studies open their boundaries to the methods of computer linguistics, computer linguists start addressing the complex semantic nature of phrasemes in text corpora (e.g., disambiguation of literal and non-literal meanings; Li/Sporleder 2010, Li/Roth/Sporleder 2010) – the question that used to be central for classical phraseological research (as mentioned above).

As phraseological studies, computer linguistic research was mostly restricted to modern languages. However, the extensive research on historical German texts carried out in the HiFoS Research Group at the University of Trier (Germany) offers evidence that enables a sounder evaluation of some of the criteria used in determining formulaic diction (cf. section 3). Therefore, in the HiFoS-project, we speak about historical *formulaic patterns* rather than *historical phraseology* or *multi-word-units*. This term is more general than *phraseology* and enables scholars with different scientific interests to include even those patterns in their analysis that have a very low degree of syntactical stability and semantic idiomaticity. Formulaic patterns are not necessarily restricted by the length of two words on the one hand and by sentence boundaries on the other hand. These patterns can rarely occur in historical texts. Furthermore, in the HiFoS-project, an expression is considered to be formulaic if, in addition to its concise syntactic structure, its pragmatic functions are evident and central to a given text (Filatkina 2009a, Filatkina/Gottwald/Hanauska 2009).

2. Formulaic patterns in modern languages as “a pain in the neck” for corpus and computer linguistics

Despite the long tradition of computer linguistic research on formulaic patterns in modern languages, they are still considered to be “a pain in the neck” (Sag et al. 2001) from the technical point of view.

First of all, the question of what corpus size is sufficient for research on formulaic patterns remains unanswered. Scholars have already pointed out that it should be substantially larger than corpora used for research on other linguistic phenomena. According to Geyken (2004), even a 100 million words corpus might not be sufficient for the purpose of detecting formulaic patterns due to the discrepancy in the type-token-frequency: Even though different types of formulaic patterns are highly constitutive for many texts (cf. section 1), the token frequency of each type might be very low (cf. also Claridge 2008). Furthermore, the common practice in corpus linguistics to compile corpora from

sporadically chosen longer text excerpts does not appear to be helpful for research on formulaic patterns as the occurrence of formulaic patterns within a text excerpt can never be predicted in advance. One possible solution is to restrict the choice of texts and focus on certain genres. In reality, this approach turns out to be difficult as well because at the current stage of phraseological research little knowledge about the formulaic character of single text genres is available.

Secondly, no annotation standard or formal categorization scheme has so far been developed for formulaic patterns. One reason for this might be the difference in research perspectives between phraseological and computer linguistic research on formulaic patterns: while the former have, until very recently, been mostly semantically oriented, the latter can proceed to semantic analysis only after morphology and syntax have been reliably analyzed.

Thirdly, the (semi-)automatic identification and retrieval of formulaic patterns continues to be a challenging task. The existing approaches can be roughly grouped in 1) association measures tools, 2) formal preference tools, and 3) distributional semantics tools. At present, the tools of the first two groups are most common. They are statistical in nature and operate on the basis of statistically significant co-occurrences in shallowly annotated corpora. Association measure tools have proven to be particularly useful in the identification of set phrases with a frozen syntactic structure and of those consisting of not more than two lexical items. Formal preferences include morphological and syntactic restrictions as well as idiosyncratic features in the structure of formulaic patterns, but these are less efficient for the identification of syntactically absolutely regular patterns. Absolutely stable patterns, units limited to two lexical items or expressions with morphological and syntactic irregularities are far from being the majority of formulaic patterns. In addition to corpus linguistics, Natural Language Processing (NLP) has also addressed the question of identifying multi-word expressions. Here, the distinction of literal and non-literal usages was of particular interest. For distinguishing non-literal idioms, approaches looking at lexical cohesion (global lexical context, local lexical context, discourse context) and, to a lesser extent, at some syntactic features show good results (Li/Sporleder 2010), particularly as they allow generalization across idioms and address the problem of their low token frequency. At a more abstract level, some similarities can be found between identifying formulaic patterns and the task of relation extraction in Text Mining (Heyer/Quasthoff/Witting 2008). Answering the question how well different text re-

use techniques work in order to link the similar but still variant paraphrased passages (Büchler et al. 2011), related versions of the same text (e.g., different versions of the Bible) were required and analysed so far. As variants of formulaic patterns occur not only in related texts, these approaches need to be extended to texts of completely different origins. It is important to note that all tools and techniques mentioned above were developed and implemented on modern and/or normalized texts. Therefore, the development of efficient identification and retrieval tools for historical formulaic patterns remains a future research task so far.

3. Historical formulaic patterns of German and text corpora

A corpus-based study of historical formulaic patterns in German turns out to be an even more challenging task at present. Until very recently, no philologically reliable corpus of historical German texts existed that would allow for a systematic and consistent investigation into the dynamics of formulaic patterns from the beginning of the written tradition until the Early New High German period. Recently started projects such as *Referenzkorpus Altdeutsch*⁵ at the University of Frankfurt, Humboldt University of Berlin and University of Jena, *Das annotierte Referenzkorpus Mittelhochdeutsch (1050-1350)*⁶ at the University of Bonn (Dipper 2010), and *Referenzkorpus Frühneuhochdeutsch* at the Martin-Luther-University Halle-Wittenberg⁷ aim to fill this gap but, to the best of our knowledge, they have yet to be investigated from the perspective of formulaic patterns. The major difference between these corpora and the previously existing ones is the fact that they include full texts and not text excerpts, as well as texts of different genres and different authors. These are essential requirements for contemporary corpus-driven and corpus-based studies on historical formulaic patterns, which so far have been limited to just a few sources (cf. the overview in Filatkina 2009b; Filatkina/Gottwald/Hanauska 2009). In contrast to previous text collections, these recent corpora are based on diplomatic transcriptions of original texts. This makes them particularly suitable for research into morphological, syntactical and lexical variation of formulaic patterns, which is hardly possible on the basis of normalized text editions.

⁵ www.deutschdiachrondigital.de/.

⁶ www.linguistics.ruhr-uni-bochum.de/~dipper/project_ddd.html.

⁷ www.germanistik.uni-halle.de/forschung/altgermanistik/referenzkorpus_fruerneuhochdeutsch/.

The results of the HiFoS Research Group show that variation is the most characteristic feature of formulaic patterns in historical German texts (Filatkina 2009a, 2012). In order for language historians to carry out extensive research into variation models and their dynamics, this fact must be taken into consideration when answering the question about the depth of corpus annotation. It also sheds a new light on the process of (semi-)automatic identification of formulaic patterns. These questions have not yet been addressed by any of the existing historical corpus projects. Bennett et al. (2010: 65) claim that the identification of formulaic patterns will be the aim at a later stage of the GerManC-project; to our knowledge, no results are available to date.

The decision as to what a formulaic pattern is in a historical text is not trivial even from the point of view of historical linguistics. In a modern language, one can statistically measure the degree of the formulaic diction with the help of sufficiently large text corpora (cf. section 2), questionnaires or interviews. However, these methods of contemporary empirical linguistics are not possible when working historically. Consequently, a researcher is restricted to original texts and singular findings in texts that happened to be handed down from earlier times and whose number is incomplete. One of the more widely accepted criteria for identifying a formulaic pattern is its repetitious occurrence. It would seem a truism that this phenomenon can and indeed must be documented in order to employ the criterion. Thus, for the reasons given above it cannot be put at the centre of linguistic analysis of the historical data. Furthermore, the identification criteria that were established within the framework of phraseological research (polylexicity, syntactic stability, idiomaticity, cf. section 1) often do not apply to historical data where, e.g., polylexicity confronts the lack of orthographic norms or the problem of word/sentence boundaries and idiomaticity – the difficulties of hermeneutic interpretation of meaning caused by culture and time distances between present day and historical data. For any given formulaic pattern, it will be necessary to take into account relevant factors which apply to the transmission, geographical space, date of the text and evidence gathered from other languages as well as the cultural-historical role of the expression, including not only verbal media but also its visualisations. Corpus and computational studies in the field of formulaic patterns should take these circumstances as a starting point in order to facilitate linguistic research and to take it to a new level. More linguistic knowledge about historical formulaic patterns is required in order to support corpus compilation and the development of annotation tools and standards or, as Rayson et al. (2010: 2) put it:

[...] it has become increasingly obvious that in order to develop more efficient algorithms, we need deeper understanding of the structural and semantic properties of MWE's [NF: multi-word-expressions], such as morpho-syntactic patterns, semantic compositionality, semantic behaviour in different contexts, cross-lingual transformation of MWE properties etc.

4. HiFoS Research Group

One substantial step towards the collection of such knowledge was the foundation of the Research Group “Historical formulaic language and traditions of communication” or, in German, “Historische Formelhafte Sprache und Traditionen des Formulierens (HiFoS)” at the University of Trier.⁸

What historical research on formulaic patterns has so far been lacking is a systematic investigation of diachrony and synchrony in original historical texts, an investigation into the stability and variation of formulaic patterns in these texts, and the dependency of their usage upon the text genre as well as the intertextual specifics of their distribution. These are exactly the goals that HiFoS aims to achieve in order to create a strong disciplinary basis for further research, in particular for collaborative research with other philologies and scholarly disciplines. HiFoS investigates the historical development, stability and variation of different types of formulaic patterns in different German texts over the time period from ca. 700 to ca. 1700 with a strong focus on the oldest – Old High German – texts (ca. 700 to ca. 1050). Several theoretical and methodological principles were established by the HiFoS Group in order to answer the questions above:

- 1) In contrast to some other historical projects, HiFoS collects its data from original manuscripts rather than normalized text editions, because normalized forms of formulaic patterns might have never existed in original manuscripts.
- 2) Due to the current stage of historical corpora compilation and with regard to specific requirements posed by formulaic patterns (see sections 2 and 3) as well as the primary goals of the HiFoS Research Group, great effort was put into manual or rather intellectual extraction, documentation and detailed annotation of such single findings in different texts.

⁸ The project is financed by the Sofja Kovalevskaja Award 2006 of the Alexander von Humboldt Foundation. For further information cf. www.hifos.uni-trier.de as well as the publications of the group members listed on the project website.

- 3) As scholarly research generally has little information about what types of formulaic patterns can be found in which historical texts and why, the HiFoS project aims to cover German texts of different genres (poetry and fiction, historical legal texts, non-fictional sources such as travel reports, religious and pre-scientific texts). The data gained from the fiction texts can then be compared with those formulaic patterns and metalinguistic knowledge about them, which is subject of proverbial collections, dictionaries and grammar books.⁹
- 4) Formulaic patterns found manually in original text documents are collected in a database and encoded according to the standards of Unicode and the Text Encoding Initiative (TEI). After the context of the expression in question is noted, its type, different aspects of its morphology, syntax, semantics, pragmatics, and (if possible) cultural historical background, transmission lines and the interdependency of a given formulaic pattern on earlier similar expressions in other languages, particularly Latin and Greek,¹⁰ are analyzed.¹¹
- 5) In order to gain a more or less complete picture of the historical usage and dynamics of a formulaic pattern, the HiFoS database provides the possibility of grouping single findings in a so called *Formulierungstradition*. The collection of variants for one particular formulaic pattern is only one intended way of grouping data. Similar groupings are envisioned for formulaic patterns from specific texts, of specific authors, semantic or pragmatic equivalents and so on.

⁹ With this regard, the knowledge gathered within the DoLPh (“Dynamics of Luxembourgish Phraseology”; cf. <http://infolux.uni.lu/phraseologie/>) and OldPhras projects (www.oldphras.net) becomes particularly beneficial for HiFoS.

¹⁰ By doing so, we do not want to draw an immediate conclusion that a given formulaic expression is a loan pattern as this would require additional systematic research. Such joint research, extensive data exchange and close collaboration were established in 2010 between HiFoS, DoLPh, Aliento (“Analyse Linguistique, Interculturelle d’annoncés sapientiels et Transmission Orient-occident Occident-orient”; www.aliesto.eu), CASG (“Corpus der arabischen und syrischen Gnomologien”; <http://casg.orientphil.uni-halle.de/>) and SAWS (“Sharing Ancient Wisdoms”; www.kcl.ac.uk/schools/humanities/depts/bmgs/research-section/saw/). At present, the joint research interface and first publications are in preparation.

¹¹ For the complete description of all areas cf. Filatkina (2009b).

The compilation of extensively annotated data in form of a *Belegkorporus* is one of the major goals of the HiFoS project. But, at the same time, the database is also a research platform that has been already used in several smaller research projects and will be made publically available online by the end of the project. The database is crosslinked with an international bibliography about the historical German formulaic language and with the images of original manuscripts available online. This is particularly helpful in situations of a very complex contextualisation of a formulaic pattern, where a broad context needs to be considered. We are currently working on linking the database with similar databases for other languages as well as with the database of historical pieces of art.¹²

At present, all Old High German texts from the period of time 700 to 1050 have been analyzed. The HiFoS data corpus consists of ca. 30,250 fully annotated single formulaic patterns. Among them, ca. 9,494 entries come from Old High German texts (ca. 700-1050), ca. 11,644 from Middle High German (ca. 1050-1350) and ca. 8,973 from Early Modern High German texts (ca. 1350-1650).

5. Ubiquitous Knowledge Processing Laboratory (UKP Lab)

Despite the many differences between historical linguistics and corpus/computer linguistics in handling formulaic patterns, both disciplines show a growing interest in this phenomenon and a stronger awareness of the bilateral benefit for each other. More interdisciplinary projects need to be put forth in order to give research on formulaic patterns the solid frame that they merit with regard to the constitutive role they play in human communication, both in the past and today. The cooperation planned between the HiFoS Research Group and the UKP Lab at the Technical University Darmstadt is a good start on the way of improving the bad reputation of formulaic patterns as “a pain in the neck” of corpus and computer linguistics.

The texts from the various stages of the German language need to be normalized to a common language level in order to be prepared for automatic processing. Such normalization should support the application of tools that are primarily trained on modern language, and thus may constitute a modernization as described by Bollmann et al. (2012). That is, words that are no longer part of the modern language vocabulary are substituted by modern words of a similar meaning.

¹² Cf. footnotes 9 and 10.

The Darmstadt Knowledge Processing Repository (DKPro)¹³ maintained by the UKP Lab already covers many NLP components for modern languages. This includes components to handle and keep track of the modification of textual data as it is done when normalizing the texts (Eckart de Castilho/Gurevych 2009). Thus, any analysis results produced by applying the NLP components to the modernized forms can be mapped back to the original historic form. Concrete support for state-of-the-art normalization of historical texts is planned to be added to DKPro as part of the cooperation.

To aid in the identification of formulaic patterns, further components shall be added to DKPro which use the aforementioned approaches of association measures, formal preferences, and distributional semantics.

We understand that the nature of historical texts and of the task at hand requires that any results of automatic processing need to be inspectable and correctable by domain experts and such corrections need to be fed back to the automatic analysis system in order improve subsequent analysis runs. To facilitate this, a task-oriented user interface needs to be implemented which conveniently exposes the functionality of the automatic analysis without bothering the user with technical details. An example how we realized such a user interface for the task of finding uncommon and ambiguous grammatical structures in large corpora has been shown in Eckart de Castilho/Bartsch/Gurevych (2012).

Acknowledgements

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant no. I/82806, by the Hessian research excellence program “Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz” (LOEWE) as part of the research cluster “Digital Humanities”, by the German Federal Ministry of Education and Research (BMBF) under the promotional reference 01UG1110D (DARIAH-DE), and the Alexander von Humboldt Foundation in the framework of the Sofja Kovalevskaja Award 2006 (Research Group HiFoS).

¹³ www.ukp.tu-darmstadt.de/research/current-projects/dkpro/.

References¹⁴

- Bennett, Paul/Durrell, Martin/Scheible, Silke/Witt, Richard J. (2010): Annotating a historical corpus of German: A case study. In: Proceedings of the Conference “Language Resource and Language Technology Standards – State of the Art, Emerging Needs, and Future Developments (LREC10-W4)”, 18th of May 2010: 64-68. www.ims.uni-stuttgart.de/institut/mitarbeiter/scheible/publications/lrec2010.pdf.
- Bollmann, Marcel/Dipper, Stefanie/Krasselt, Julia/Petran, Florian (2012): Manual and semi-automatic normalization of historical spelling – case studies from Early New High German. In: Proceedings of the First International Workshop on Language Technology for Historical Text(s) (LThist2012), KONVENS, Vienna. www.linguistics.rub.de/~petran/papers/konvens2012.pdf.
- Büchler, Marco/Crane, Gregory/Mueller, Martin/Burns, Philip/Heyer, Gerhard (2011): One step closer to paraphrase detection on historical texts: About quality of text reuse techniques and the ability to learn paradigmatic relations. In: Thiruvathukal, George K./Jones, Steven E. (eds.): *Journal of the Chicago Colloquium on Digital Humanities and Computer Science 11*. Chicago, IL: University of Chicago.
- Burger, Harald (2010): *Phraseologie. Eine Einführung am Beispiel des Deutschen*. 4th ed. Berlin: Erich Schmidt.
- Claridge, Claudia (2008): Historical corpora. In: Lüdeling/Kytö (eds.), Vol. 1: 242-259.
- Dipper, Stefanie (2010): POS-Tagging of historical language data: First experiments. In: Proceedings of the 10th Conference on Natural Language Processing (KONVENS-10), Saarbrücken.
- Dobrovolskij, Dmitrij/Piirainen, Elisabeth (2005): *Figurative language. Cross-cultural and cross-linguistic perspectives*. Amsterdam/Philadelphia: Elsevier.
- Eckart de Castilho, Richard/Gurevych, Iryna (2009): DKPro-UGD: A flexible data-cleansing approach to processing user-generated discourse. In: Online-proceedings of the first French-speaking meeting around the framework Apache UIMA at 10th Libre Software Meeting (LSM/RMLL), Nantes, France. <http://e.nicolas.hernandez.free.fr/pub/rec/09/RMLL-cfp-en.html>.
- Eckart de Castilho, Richard/Bartsch, Sabine/Gurevych, Iryna (2012): CSniper (2012) – Annotation-by-query for non-canonical constructions in large corpora. In: *Association for Computational Linguistics: Proceedings of the 50th Meeting of the Association for Computational Linguistics (ACL) 2012 (Demo section)*: 85-90. www.aclweb.org/anthology/P12-3015.
- Evert, Stefan (2005): *The statistics of word cooccurrences. Word pairs and collocations*. Ph.D. thesis, University of Stuttgart. <http://elib.uni-stuttgart.de/opus/volltexte/2005/2371>.

¹⁴ All URLs have been checked and found valid as of late January 2015.

- Evert, Stefan (2008): Corpora and collocations. In: Lüdeling/Kytö (eds.), Vol. 2, 1212-1249.
- Fellbaum, Christiane (2007): Idioms and collocations. Corpus based linguistic and lexicographic studies. London: Continuum International Publisher.
- Fellbaum, Christiane (ed.) (2006): Corpus-based studies of German idioms and light verbs. Special Issue 19-4 of the International Journal of Lexicography.
- Filatkina, Natalia (2009a): Historical phraseology of German: regional and global. In: Korhonen, Jarmo/Mieder, Wolfgang/Piirainen, Elisabeth/Pinel, Rosa (eds.): Phraseologie global – areal – regional. Akten der Konferenz EUROPHRAS 2008 vom 13.-16.8.2008 in Helsinki. Tübingen: Niemeyer, 143-151.
- Filatkina, Natalia (2009b): Historische formelhafte Sprache als „harte Nuss“ der Korpus- und Computerlinguistik. Ihre Annotation und Analyse im HiFoS-Projekt. In: Linguistik online 39(3): 75-95.
- Filatkina, Natalia (2012): *Wan wer beschreibt der welte stat / der muoß wol sagen wie es gat*. Manifestation, functions, and dynamics of formulaic patterns in Thomas Murner's "Schelmzunft" revisited. In: Filatkina, Natalia/Kleine-Engel, Ane/Dräger, Marcel/Burger, Harald (eds.): Aspekte der historischen Phraseologie und Phraseographie. Heidelberg: Winter, 21-41.
- Filatkina, Natalia/Gottwald, Johannes/Hanauska, Monika (2009): Formelhafte Sprache im schulischen Unterricht im Frühen Mittelalter: Am Beispiel der so genannten „Sprichwörter“ in den Schriften Notkers des Deutschen von St. Gallen. In: Sprachwissenschaft 34: 341-397.
- Geyken, Alexander (2004): What is the optimal corpus size for the study of idioms? Lecture presented at the annual meeting of the DGfS, Mainz.
- Heid, Ulrich (2008): Computational phraseology. An overview. In: Granger, Sylviane/Meunier, Fanny (eds.): Phraseology. An interdisciplinary perspective. Amsterdam: John Benjamins, 337-360.
- Heyer, Gerhard/Quasthoff, Uwe/Witting, Thomas (2008): Text Mining: Wissensrohstoff Text. Konzepte, Algorithmen, Ergebnisse. Bochum: W3L GmbH.
- Hoey, Michael (2005): Lexical priming. A new theory of words and language. London/New York: Routledge Chapman & Hall.
- Jackendoff, Ray (1995): The boundaries of the lexicon. In: Everaert, Michael et al. (eds.): Idioms: Structural and psychological perspectives. New Jersey: Lawrence Erlbaum, 133-165.
- Li, Linlin/Sporleder, Caroline (2010): Using Gaussian mixture models to detect figurative language in context. In: Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2010). Short Papers, June 1-6, 2010, Los Angeles. www.coli.uni-saarland.de/~csporled/papers/naacl10.pdf.

- Li, Linlin/Roth, Benjamin/Sporleder, Caroline (2010): Topic models for word sense disambiguation and token-based idiom detection. In: Proceedings of the 48th Annual meeting of the Association for Computational Linguistics (ACL 2010), July 11-16, 2010, Uppsala, Sweden. www.coli.uni-saarland.de/~csporled/papers/acl10.pdf.
- Lüdeling, Anke/Kytö, Merja (eds.) (2008): *Corpus linguistics. An international handbook*. Two volumes. Berlin/New York: de Gruyter.
- Piirainen, Elisabeth (2012): *Widespread idioms in Europe and beyond. Toward a lexicon of common figurative units*. New York: Peter Lang.
- Rayson, Paul/Piao, Scott/Sharoff, Serge/Evert, Stefan/Moirón, Begoña Villada (2010). Multiword expressions: hard going or plain sailing? In: *Language Resources and Evaluation*, 44(1): 1-5.
- Sag, Ivan A./Baldwin, Timothy/Bond, Francis/Copestake, Ann/Flickinger, Dan (2001): Multiword expressions: A pain in the neck for NLP. In: *LinGO 2001-2003*. <http://lingo.stanford.edu/pubs/WP-2001-03.pdf>.
- Sinclair, John (1987): Collocation: A progress report. In: Steele, Ross/Threadgold, Terry (eds.): *Language topics: Essays in honour of Michael Halliday*. Amsterdam: John Benjamins, 319-331.
- Sirajzade, Joshgun (2012): *Das luxemburgischsprachige Œuvre von Michael Rodange (1827-1876). Editionsphilologische und korpuslinguistische Analyse*. PhD Thesis, University of Trier.
- Tschichold, Cornelia (2000): *Multi-word units in Natural Language Processing*. Hildesheim: Olms.
- Wray, Alison (2008): *Formulaic language: Pushing the boundaries*. Oxford: Oxford University Press.

Automated quality control for the morphological annotation of the Old High German text corpus

Checking the manually adapted data using standardized inflectional forms

Abstract

The project *Referenzkorpus Altdeutsch* ('Old German Reference Corpus') aims to establish a deeply-annotated text corpus of all extant Old German texts. As the automated part-of-speech and morphological pre-annotation is amended by hand, a quality control system for the results seems a desirable objective. To this end, standardized inflectional forms, generated using the morphological information, are compared with the attested word forms. Their creation is described by way of example for the Old High German part of the corpus. As is shown, in a few cases, some features of the attested word forms are also required in order to determine as exactly as possible the shape of the inflected lemma form to be created.

1. Introduction

When a morphologically annotated text corpus has been created without entirely automating the part-of-speech and inflectional annotation, an automated verification of these attributes appears to be a helpful approach for quality control. If the word tokens are already attributed to their lemmata, every lemma can be provided with the expected suffixes and endings (or also have its word stem altered according to the grammatical rules), and the result can be compared with the actually attested form. The choice of a consistent *sprachform* for the whole corpus would not only facilitate the creation of these standardized inflectional forms, but also facilitate search queries for specific inflectional forms of lemmata: an idealized Old High German word form *brungut* 'you (pl.) brought' – irrespective of its actual attested shape – is easier to key in than a lemma *bringan* plus the information that all forms of the second person indicative plural inflected according to the strong class IIIa are searched for. Moreover, for some kinds of corpus analyses, being able to neglect phonological and morphological variation will help to focus on the effective research question.

In the project *Referenzkorpus Altdeutsch* ('Old German Reference Corpus'), this course of action is applied. The project aims to produce a deeply-annotated corpus of all preserved texts from the oldest stages of German – Old High German (OHG) and Old Saxon (OS) –, which date from ca. 750 to 1050 CE. Comprising a total of about 650,000 word tokens, the corpus covers interlinear translations of Latin texts as well as free translations, adaptations and mixed German-Latin texts. These are complemented by a few texts composed entirely in an Old German language. The largest coherent sub-corpora are the OHG works of Notker and Otfrid of Weissenburg, an OHG translation of the Gospel harmony of Tatian, and the versified OS Gospel harmony known as the *Heliand*.

Since a consistent *sprachform* for OHG and OS could only be conceived in reconstructed Proto-West Germanic, and would thus differ too much from the attested word forms, the establishment of two different standards for the two languages was chosen. Below, the approach for OHG will be described.

2. The definition of a consistent *sprachform*

Whilst some OHG dictionaries use no consistent lemmatization, Splett (1993) does so for his *Althochdeutsches Wörterbuch*. As OHG is a "Sprachstufe, die keine allgemein gültige Leitvarietät besitzt",¹ Splett (1993: XXVIII) chooses "die Idealform des Ostfränkischen, das der Tatian überliefert"² (ibid.). It is therefore necessary to have a grammar that displays the inflectional endings in a language stage close to the one of Splett's lemmata. Fortunately, Braune (2004: 6) gives a description similar to Splett concerning his grammar: "In diesem Buch wird [...] die ostfrk. [ostfränkische] Sprache des ahd. [althochdeutschen] Tatian (2. Viertel 9. Jh.) zugrunde gelegt."³ – Thus, these two works seem very appropriate to our task.

Notwithstanding the above, a dictionary cannot cover all grapho-phonematic particularities: in several declensional classes, the ending of the dative plural is *-um* (or *-om*) in the oldest OHG texts and appears as *-un* or *-on* in the Tatian (cf. Braune 2004: 185). Sievers (1892: LXVIII) shows that the preference for either form depends mainly on the individual scribe. In total, he counts *-un*

¹ I.e., a "language stage that possesses no universally valid ideal form".

² I.e., "the ideal form of the East Franconian [language] that the Tatian passes down".

³ "In this book, the East Franconian language of the Old High German Tatian (2nd quarter of the 9th century) is taken as a basis."

119 times and *-on* 113 times. Even the adverbialized dative plural forms in the Tatian have a similar distribution (ibid.). Due to the lack of more criteria, <u> as the older one of the two vowels (cf. Braune 2004: 62) was eventually chosen. In contrast to this, the merger of final *m* and *n* into *n* in inflexional endings – almost completed in the Tatian (cf. Sievers 1892: XXVIII.) – was disregarded, as it could mark a semantic difference in some cases: e.g. *nāmum* ‘we took’ vs. *nāmun* ‘they took; later also: we took’ (cf. Braune 2004: 271).⁴

3. Description of the data at hand

The generation of the standardized inflexional forms starts from ELAN files, each containing a part of an OHG subcorpus. ELAN⁵ is software for the manual adaptation of multilevel annotations developed by the Max Planck Institute for Psycholinguistics at Nijmegen (the Netherlands). One example of an annotated word token in ELAN is shown in Figure 1.

Reference Text	altero
Standard Text	altero
Lemma	alt
Translation	alt, ausgewachsen, bejahrt
M1a PoS Lemma	ADJ
M1b PoS Record	ADJN*
M2a Inflection Lemma	a,o
M2b Inflection Record 1	n
M2c Inflection Record 2	Comp_Masc_Sg_Nom_wk

*) ADJN = Adjektiv, nachgestellt (adjective, postpositive)

Figure 1: Sample word annotation from the Tatian in ELAN format – before adapting the standardized inflexional forms (simplified representation)

The reference text comes from the printed edition of the text. Before the creation of the standardized text, the second tier contains only padding, i.e. copies of the reference text. The lemma and its (Modern German) translation are given as listed by Splett (1993). If several possible translations are given, only the appropriate one is adopted. For the remaining five tiers, a specially developed tagset is used, and their information comes from a range of glossaries, each of which covers a section of the corpus.⁶ They contain the part-of-speech

⁴ Splett (1993) apparently made the same decision, documented by expressions like *io dēm wîlôm* ‘on the spot’ from *wîla* ‘(a) while’ (p. 1124).

⁵ www.lat-mpi.eu/tools/elan. All URLs have been checked and found valid as of late January 2015.

⁶ Heffner (1961), Hench (1890, 1893), Kelle (1881), Sehrt (1955) and Sievers (1874, 1892).

types of lemma (M1a) and record (M1b), the lemma-dependent inflectional information pertaining to the lemma (M2a), the corresponding information of the individual record (M2b), and its record-specific inflectional information (M2c). The files have been automatically created using the OHG corpus on the TITUS website.⁷

The standardized inflectional forms are created from the standardized lemma, plus the inflectional information. The glossaries have been digitized and linked to the texts to save extensive manual work. However, their information is not always exhaustive and properly transferable. Because of this, the morphological information gained from the glossaries has to be disambiguated, corrected, and completed by hand before the standardized inflectional forms can be generated.⁸

4. Preparation of the data

To begin with the creation of the standardized inflectional forms, the files are again edited automatically. For every word token, its variant as occurring in the printed edition and its standardized lemma are extracted, as well as the morphological data with the exception of the lemma-dependent inflectional information pertaining to the lemma (M2a). The gender information concerning nouns, a part of M2b, is extracted to make M2b represent the inflectional class only.

In order to facilitate the replacements, two digraphs are substituted: <qu> is shortened to <q> to avoid <u> being treated as a vowel, and <ck> is replaced by <kk>, as doubled consonants are often simplified (cf. Section 5.1). When all replacement rules have been executed, these substitutions are undone. So, to the lemma *bok* ‘buck’, a genitive *bockes* will be created.

⁷ <http://titus.uni-frankfurt.de/texte/texte2.htm#ahd>.

⁸ A description of the digitization of the dictionaries can be found in Mittmann (2013). The extraction and processing of the glossary data, the adjustment of their lemmata to Splett (1993), the combination of the data with the TITUS text files, and the subsequent manual annotation of the ELAN files is treated of in detail in Linde/Mittmann (2013).

5. Generation of the standardized inflectional forms

The generation of the standardized inflectional forms is performed according to part-of-speech categories. There are two phenomena to be observed that cannot be attributed to a single category: consonantal simplification and the emergence of secondary vowels.

5.1 Consonantal simplification and secondary vowels

In the standard used, double consonants are simplified before consonants and at the end of a word. If, however, a word ends in a single consonant, it is not possible to determine whether this consonant is doubled in inflectional forms. Since Splett gives no information on that, a list of these words was needed.

The lemmata to be considered all feature a vowel or diphthong preceding the final consonant,⁹ and the consonant is only simplified in final or pre-consonantal position. Within the automatically created list of these words, the test is then performed by hand, and the final list is set up: recorded inflectional forms and derived words that have a vowel following the original word's final consonant are checked as to whether the consonant is doubled or remains simple.

In some rare cases, several lemmata have an identical form but differ with regard to the doubling. If there is a categorial difference between them, this can be used: the <r> of *far* is only doubled if this is a masculine noun ('bull'), but not if it is neuter ('harbour', 'beacon') or an adjective ('going'). Otherwise, the decision is made according to the spelling in the printed edition. This is, however, considered as a makeshift solution, as it compromises the ambition to create the standardized inflectional forms only from the standard lemma and the morphological information of the record.

The emergence of so-called secondary vowels, as in *zeihhan* 'sign, token', developed from Proto-Germanic **taiknan* with regular apocope of the original ending *-an*,¹⁰ leads to stem changes within the affected paradigms. However, it is often analogically transferred to other places in the paradigm: for the genitive plural, the Tatian has *zeihno*, *zeihhano* and *zeichano* (cf. Sievers 1892: 510). To

⁹ Although double consonants were simplified after long vowels or diphthongs in the course of the OHG era, Splett (1993) still gives their doubled form – as with *lūttar* 'limpid, noble' (p. 577) –, so this case cannot be excluded.

¹⁰ The secondary vowels arose before a resonant where the West Germanic resonant became syllabic due to loss of a vowel (cf. Braune 2004: 68).

avoid unnecessarily complicated paradigms for the inflected lemmata, the presence or absence of secondary vowels is assumed in accordance with the lemma.

5.2 Nouns (substantives)

With regard to the generation of the standardized inflectional forms, a first section converts the nouns, including the cardinal *dūsunt* ‘thousand’ and the pronouns *man* ‘one’ and *wiht* ‘something’, which inflect equally. The information on number and case is extracted from M2c. The addition of endings is performed depending on the inflectional class, and in some cases also on the gender. Some classes also trigger umlaut from <a> to <e>.¹¹ Here, the last <a> of the word, followed only by consonants (plus another vowel, if existing) before the end of the word, is converted, so that a nominative plural *gesti* is created from the masculine *i*-stem *gast* ‘guest’. Where it applies, consonantal doubling is performed, changing *bal* ‘ball’ into a genitive *balles*, but converting *wal* ‘whale’ into *wales*.

Since the rules needed for the genitive singular are very similar to those for the dative, they are treated together – and subsequently, only two rules are needed to transform any genitive into a dative form: final *-s* is removed, and final *-a* is replaced by *-u*. The accusative and instrumental cases are again created from the nominative. In the plural, a couple of rules can be collectively applied to all case forms; e.g. the addition of the *-ir*-suffix to the *z*-stems. For pluralia tantum, a correct inflection is also taken care of. Proper nouns of Latin origin ending in *-us* may keep or lose this ending when inflected: *Petrus* ‘Peter’ has both *Petre* and *Petruse* as its dative (cf. Sievers 1892: 298). Here, the attested records have to serve as the decisive aspect.

5.3 Adjectives, participles, determiners, pronouns, and numerals

A second section converts the remaining adjectives, participles, determiners, pronouns, and numerals. Grade, gender, number and inflectional type are extracted from M2c. The creation of the base forms of the participles is treated together with the verbs, but their inflection together with that of the adjectives.

¹¹ Henceforth, the term “umlaut” will be used only to describe an umlaut from <a> to <e>. Non-umlauted /e/ is written as <ē>.

Before the inflectional endings are attached, consonantal doubling is performed where required, and the comparative and superlative suffixes are added to the stems. In the case of lemmata with irregular comparison, the stems are replaced. Since there is no definite rule to determine whether *-ōr-* or *-ir-* (the latter also triggering umlaut) are added as the comparative suffix – and *-ōst-* or *-ist-* for the superlative, respectively –, this has once more to be decided according to the word form in the text, depending on whether the token contains *-or-/ -ōr-* or *-ost-/ -ōst-*. Thus, in Figure 1, *eltiro* is generated from *alt* ‘old’ and inserted as the standardized inflectional form, as the record *altero* does not show <ōr> or <or>. This criterion has been chosen because /ō/ usually retains its quality in middle syllables, whilst short /i/ is not uncommonly rendered as <e> (cf. Braune 2004: 64-66). Some stems had to be adjusted to the attachment of inflectional endings; e.g., *gēlo* ‘yellow’ has a stem *gēlaw-*. Consonantal doubling also occurs: *smal* ‘small’ remains *smal-*, but *snēl* ‘quick’ becomes *snēll-*.

De-adjectival adverbs are formed by adding the ending *-o* to the stem; *-i*-stem adjectives lose this vowel and their umlaut (*herti* ‘hard’ becomes *harto*). In the comparative and superlative, there is no ending at all. Apart from one adverbial positive that deviates from the corresponding adjective (*wola* ‘well’ from *guot* ‘good’), comparative and superlative forms are created similarly to those of the adjectives. One exception is that all adverbial comparatives feature *-ōr-* (cf. Braune 2004: 232).

Various pronouns and some of the cardinal numbers inflect divergently and are dealt with separately. The other cardinals do not inflect at all. *ein* ‘one, a’, *beide* ‘both’ and the ordinals behave like adjectives. The possessive pronouns *unsēr* ‘our’ and *iuwēr* ‘your’ feature both long and short stems – *unsēr-/iuwēr-* and *uns-/iuw-* – and have to be treated according to the attested word forms.

5.4 Verbs

For the verbs, from M2c, mood, tense, number and person are extracted. As the imperative only exists in the present and has forms differing from the indicative only in the singular, the two moods are treated together, and only in the singular are their forms determined separately. All modifications are carried out on the highest possible level, so the umlaut-like vowel changes of strong verbs from <io> to <iu> and from <ë> to <i>, which hold for the whole singular of the indicative and the imperative, need to be declared only once.

In the present tense, the first person indicative singular shows exemplarily that individual replacement rules for preterit-presents and other irregular verbs are required. The rules for weak and strong verbs are rather straightforward. The second and third person forms exhibit umlaut (except for weak verbs inflecting according to class II or III) as well as simplification of stem-final double consonants with some verbs featuring an infinitive ending in *-en*. For most consonants, the application of this rule depends on the cluster's history (cf. Braune 2004: 295f.), and as the evidence is again numerous for both cases, the word form from the printed edition has to be checked for a doubled consonant. So, *stellen* 'put' results in a form *stellis*, but *zellen* 'count, tell' gives *zelis*. Here, this approach is especially risky, as consonantal doubling or simplification is not rare in OHG inflectional morphology.

The imperative shows the same rule for the consonantal simplification of verbs in *-en* before the ending *-i*, but furthermore, verbs in *-an*, featuring a zero ending, exhibit a general simplification of final double consonants. In the plural and the whole subjunctive, fewer rules are needed, as only the personal endings deviate from the lemma.

In the past tense, weak verbs and preterit-presents have an *-(i)t*-suffix, and weak class Ia verbs feature consonantal simplification and absence of umlaut: *brennen* 'burn' becomes *branta*. The stem changes in the past tense require various replacement rules: not only do the stem vowels of all strong verbs change, but alternations also affect the consonants of some of these verbs. Apart from consonantal doubling and simplification, the most frequent reason for this is Verner's Law,¹² as shown by the example of the past tense form *wārun*, belonging to *wēsan* 'be'.

For most inflectional classes, Verner's Law does not affect the first and third person indicative singular. Nevertheless, it is more straightforward to apply the same rules to all other cases than to apply them to the whole past and undo them in these two cases – although the opposite would also be possible,¹³ given that all cases of two lemmata, one of which is affected by Verner's Law and has

¹² Verner's Law describes a peculiar consonantal mutation of Proto-Germanic: voiceless fricatives were voiced unless being word-initial or immediately preceded by the Indo-European word accent (cf. Verner 1877: 114). In OHG, the voiced fricatives have evolved to plosives, and /z/ has become /r/.

¹³ In stem classes where the mentioned two forms are unaffected by Verner's law, they also have a deviating stem vowel.

the same past tense stem as the other one, are ruled out in OHG.¹⁴ These cases constitute some of the instances of the elimination of Verner's Law within verbal paradigms in OHG, raising the question of whether it is more useful for possible queries in the corpus to keep the standard word forms close to the degree of perpetuation of Verner's Law observable in the Tatian; or to consider its execution in its original extent as the standard, so that all deviations from this could be more easily examined.

Most past participles have a prefix *gi-*, but there is no definite rule to discern whether this applies; so again, the word form appearing in the printed edition must dictate. The prefix may appear with <k>, <c> or <ch> instead of <g> and with <e> or <a> for <i>; this has to be considered as well. As some verbs have another prefix before the *gi-*, the correct position for the possible insertion has to be determined. First, it is checked whether the prefix appears within the word and the lemma lacks the digraph <gi>. If not, no insertion is performed. But if so, as for *nidargiuualzten* (cf. Sievers 1892: 486) from *nidarwelzen* 'bow down' (cf. Splett 1993: 1091), it is checked whether the grapheme of the word form directly after the prefix *gi-* is the same as the grapheme of the lemma at the same position, but without the prefix *gi-*. If this is the case, (-)*gi-* is added; otherwise – that is, if the initial grapheme of the verbal stem is different or the other prefix differs in length – the same check is performed again one by one with a list of initial graphemes or digraphs that often differ between record and lemma. Here, it is checked whether the altered grapheme has the same position in the record (after the *gi-* prefix) as the original one in the lemma without the *gi-* prefix. So from the lemma *nidarwelzen*, the M2b information *wk1a*, the M2c information *P_Pos_Masc_Pl_Dat_st*¹⁵ and the record *nidargiuualzten*, a standard word form *nidargiwalztēm* is created. Needless to say, if the record, but not the lemma, begins with the prefix *gi-*, it is added in any case. If this still does not help, another test ascertains whether the initial grapheme is shifted within the record to either direction (for instance, if the prefix *aba-* 'off' appears as *ab-*), and the variants of initial graphemes are tried out also. For the verbs that contain <gi> and are thus not covered by the test, it is performed again completely, taking this aspect into account. In the event that no possibility is found of placing the *gi-*, it is omitted and left to the manual annotation.

¹⁴ *fliohan* 'flee' should have *'flug-*, but this is normalized to *fluh-*, since *fliogan* 'fly' also has *flug-* (cf. Braune 2004: 279).

¹⁵ P: pronominal inflection.

Apart from the prefix, the past participles behave similarly to the past tense stems, and inflect like adjectives. Those of the weak class Ia (and adjectives based on them) have an *-it*-suffix, but lose the <i> and undergo the same stem changes as in the past tense if they acquire an inflectional ending: *gibrennit* ‘burnt’ becomes *gibrant-*. The base form of the present participle is generated from the infinitive plus *-ti* (with umlaut). Substantivized infinitives inflect like *a*-stem nouns, with their final <n> being doubled.

6. Final adaptation and insertion of the standardized inflectional forms

For the remaining part-of-speech categories, the lemmata remain unaltered by inflection. In a following step, identical standardized inflectional forms of double lemmata are reduced: e.g., a weak class III preterit of *sagen*, *sagēn* ‘say’ is reduced from *sagēta*, *sagēta* to a single *sagēta*.

With separable preverbs, a contextual approach is needed: in the phrase *ni ges thū thanan ūz* ‘thou shalt not come out thence’ (Sievers 1892: 51), the second person present subjunctive *ges* is attributed to the lemma *ūzgān* ‘to go out, to end’, as the meaning of verbs with separable prefixes differs from those of their constituents. To make the program generate a standardized inflectional form *gās*, not *ūzgās*, comparable to *ges*, the prefix has to be deleted. As separable preverbs have their own part-of-speech tag, this is achieved by taking them as a basis, and searching in both directions until a verb that begins with them is found.

Once the standardized inflectional forms have been generated, they can be inserted into the text files. To do this, they are assembled in a list, as are all language tags within the text file, after having been extracted. A third list is made of the internal index of every standard word form, unless it contains a punctuation mark and therefore does not have to be replaced. Then, for every element in the index list, the subsequent dummy for the standardized inflectional form – if the corresponding element in the language list is *goh* (“German, Old High”, according to ISO 639-3¹⁶) – is replaced by the real one. The altered code is written into a new ELAN file.

¹⁶ cf. www.sil.org/iso639-3/codes.asp.

7. Facilitating the data checks

In all cases where (a) the spelling of the printed edition is also considered, (b) a separable preverb is identified, or (c) double lemmata are reduced, information on this and the affected word forms is stored in a log file during the program execution to enable a manual check of potentially incorrectly generated word forms. Another program then compares the attested word forms and the standardized inflectional form: if they deviate considerably from each other, a mistake in the part-of-speech or the morphological annotation is likely. To this end, the *Relative Levenshtein Distance* is calculated: the average length of both word tokens is divided by the *Levenshtein Distance*, i.e. the minimal number of insertion, deletion and replacement operations required to convert one character string into the other (cf. Levenshtein 1966). The deviating word token pairs are output into another log file, sorted according to the largest deviance.

8. Conclusion

The automated generation of standardized inflectional forms for the Old German text corpus allows for a verification of the part-of-speech and morphological annotation, and offers an opportunity to do research on the corpus without having to consider cross-text variation within phonology or morphology. The creation of the standardized inflectional forms on the basis of standardized lemmata and inflectional information proves to be a manageable task in most cases, although some of them need complex examinations to generate the appropriate standard word forms. Sometimes, even the record from the printed edition has to be taken into account, as the inflectional information does not cover some specific cases. Nonetheless, enhancing the data quality of the corpus and facilitating its utilization using a high degree of automation justifies the effort of having to scrutinize the Old German grammars in every detail.

References

- Braune, Wilhelm (2004): *Althochdeutsche Grammatik. Band I: Laut- und Formenlehre*. 15th edition, ed. Ingo Reifenstein. Tübingen: Niemeyer.
- Heffner, Roe-Merill Secrist (1961): *A word-index to the texts of Steinmeyer. Die kleineren althochdeutschen Sprachdenkmäler*. Madison: The University of Wisconsin Press.
- Hench, George Allison (1890): *The Monsee fragments*. Straßburg: Trübner.

- Hench, George Allison (1893): *Der althochdeutsche Isidor*. Straßburg: Trübner.
- Kelle, Johann (1881): *Glossar der Sprache Otfrids*. Regensburg: Manz.
- Levenshtein, Vladimir I. (1966): Binary codes capable of correcting deletions, insertions, and reversals. In: *Soviet Physics Doklady* 10(8): 707-710.
- Linde, Sonja/Mittmann, Roland (2013): Old German Reference Corpus. Digitizing the knowledge of the 19th century. Automated pre-annotation using digitized historical glossaries. In: Bennett, Paul/Durrell, Martin/Scheible, Silke/Whitt, Richard J. (eds.): *New Methods in Historical Corpora* (= *Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache / Corpus Linguistics and Interdisciplinary Perspectives on Language* (CLIP) 3). Tübingen: Narr, 235-246.
- Mittmann, Roland (2013): Digitalisierung historischer Glossare zur automatisierten Vorannotation von Textkorpora am Beispiel des Altdeutschen. In: *Journal for Language Technology and Computational Linguistics* 27: 39-52. www.jlcl.org/2012_Heft2/3Mittmann.pdf.
- Sehrt, Edward (1955): *Notker-Wortschatz*. Halle: Niemeyer.
- Sievers, Eduard (1874): *Die Murbacher Hymnen*. Halle: Buchhandlung des Waisenhauses.
- Sievers, Eduard (1892): *Tatian. Lateinisch und althochdeutsch mit ausführlichem Glossar*. 2nd edition. Paderborn: Schöningh.
- Splett, Jochen (1993): *Althochdeutsches Wörterbuch*. Berlin: de Gruyter.
- Verner, Karl (1877): Eine Ausnahme der ersten Lautverschiebung. In: *Zeitschrift für vergleichende Sprachforschung auf dem Gebiete der indogermanischen Sprachen* XXIII: 97-130.

TIMOTHY BLAINE PRICE

Multi-faceted alignment

Toward automatic detection of textual similarity in Gospel-derived texts

Abstract

Ancient Germanic Bible-derived texts stand in as test material for producing computational means for automatically determining where textual contamination and linguistic interference have influenced the translation process. This paper reports on the results of research efforts that produced a text corpus; a method for decomposing the texts involved into smaller, more directly comparable thematically-related chunks; a database of relationships between these chunks; and a user-interface allowing for searches based on various referential criteria. Finally, the state of the product at the end of the project is discussed, namely as it was handed over to another researcher who has extended it to automatically find semantic and syntactic similarities within comparable chunks.

1. Introduction

The research and output discussed in this paper took place at the Goethe Universität Frankfurt during a period of over two years, from June 2011 to September 2013, as part of the LOEWE research cluster “Digital Humanities”. Formally named ‘Historische Wechselbeziehungen altgermanischer Sprachen’, the primary goal of the research project was to investigate methods for aligning and comparing old Germanic texts that historically related yet vary due to translation and, in particular, to reconfiguration through artistic retelling, as is the case with the Old Saxon *Heliand*.

Aligning like parts of texts is a primary stumbling block to any textual comparison. For texts that differ minimally, basic comparison based on structural differences can be performed by various popular software options, e.g. Microsoft Word or Adobe Acrobat, which can easily show where sections of text between two files match and/or have been altered. However, the scope of the project discussed here was to expand such capabilities to allow for comparison between texts that are less superficially related (different time periods, distinctive styles, and variant languages).

1.1 The plagiarism model

Plagiarism detection software is able to return the probability that one (chunk of) text has been copied from another. Once a certain threshold of similarity in the two texts is passed, the overall resemblance is considered “not due to chance,” suggesting material was copied. By analysing word combinations, i.e. *n-grams*, the computer compares arbitrarily sized chunks of text between the two documents. One way to extend these capabilities is to add syntactic parsing capabilities and lexica, allowing for analysis based on synonymy.

Plagiarism software is not absolutely perfect. Beyond a threshold of complexity, the problem becomes mathematically taxing. Increasing the number of texts to be compared also does this. Other complications can easily develop simply by removing certain assumptions, e.g. that a text was written by a single author.

1.2 Direct borrowing vs. generic influence

Multi-authored texts occur readily: citation of external works technically creates a multi-authored work. In the best-case scenario, an author gives credit to the original author. Yet academics and recent politics have shown us that “borrowing” is common and not always cited responsibly.

Even when an author is honest, it may still be possible that his writing was inspired by others in ways of which he is not fully aware. Unless he directly says so, it is difficult to know whether an author has come into contact with the works of another. This is important for anyone interested in the history of ideas.

Copying is a multifaceted problem: there exists a spectrum ranging from *copying-and-pasting*, through *paraphrasing*, toward *idea theft*. Each step rightward along this spectrum involves linguistic alternations that become increasingly more difficult to track. Between paraphrasing and idea theft is the realm of *translation*, from which yet another axis branches off.

1.3 Translation as “plagiarism”

Determining similarity between two texts in the same language is one thing; doing it between different languages is quite another. Techniques that aid such comparisons include employing sophisticated lexica and pre-determined divisional similarities, the latter being the scope of this paper.

Note, two texts that differ only in language are not normally considered plagiarism *per se*, rather translation. Nevertheless, the two practices are similar. A plagiarist rewords another's writing to fit a situation making the ideas look like his own. A translator also rewords another person's writing into a new situation – a new cultural realm. The difference tends to be a moral one. Still, the new expression of translator and plagiarist contains both copied and novel material.

Normally, a translator strives not to add to or remove from the information stream of the original. In reality, languages differ not only in the word forms used, but also in grammatical structure and in cultural context. Thus, translations will differ from their original to a variable degree depending on these linguistic and cultural differences. It is difficult to develop computational means to handle such flexibility in translations while still finding similarity between texts. In other words, one might consider software that recognizes a translation and its original to be one step beyond plagiarism software. It is no longer looking for common structures based only on *n-grams*; it is also simultaneously measuring these possibilities against the options allowed in the target language. One now needs to be able to handle another axis of potentially infinite complexity.

1.4 Highly disruptive alterations

There are likely many linguistic axes that might break the plagiarism detection model. One is style, i.e., authorial creativity. Through 'translation' we understand the transfer of information between languages with minimal alteration – semantic integrity trumps syntagmatic freedom. Yet not all translated texts fit this description: the assumption of meaning-over-form is likely too restrictive to describe all of what translation entails. Consider musical translation, where lyrics are more variable than a set melody, e.g. syncopation. In this case, the musical structure outweighs information provided by text. Creativity is allowed in semantic translation – and also more variance from the original intentions. A similar phenomenon occurs in translation of non-prose, where a translator may be forced to sacrifice certain poetic features of the original in order to focus on just one, e.g. end-rhyme over semantics.

Automatic text comparison faces this conundrum: texts often display disruptive alterations among otherwise recognisable similarity. In texts where multiple changes to an original have been undertaken simultaneously, *n-gram* per-

mutations are no longer the only concern. Grammatical limitations in the target language combine with poetic patterns affecting both word order and word choice. It may be necessary to deal with each of these challenges in piecemeal fashion, e.g. by limiting the amount of material being compared. Evaluating one extensive text with another is probably too expensive. Thus, for our project, where text sizes reach tens of thousands of lines, I have cut each into more manageable chunks. Where these cuts are to be made depends greatly upon each text.

2. Textual resources

The translations and derivative Biblical texts employed here are not without controversy. On the positive side, this reflects an interest within academia. Determining textual contamination within these texts is important toward answering questions regarding the texts' provenance.

2.1 Old Saxon *Heliand*

The Old Saxon (OS) *Heliand* was presumably written by an anonymous 9th-century author. Yet even this date is speculation assuming other resources must have been used by the author. Furthermore, material evidence from manuscripts gives only vague ideas about the circumstances of the *Heliand*'s creation, generally pointing to a date coinciding with attempts by neighboring tribes to Christianise the text's audience.

The *Heliand* is a re-working of the story of Jesus as presented traditionally. It is the only remaining full text in the now-lost language of an ancient European population, namely Old Saxon. It is not a translation, rather a re-composition of the original: 1) by translation, presumably from a Latin or Old High German (OHG) resource; 2) by change in style from prose to verse with grand elaborations; and 3) by reframing the story culturally through apparently deliberate omission of certain (likely culturally unacceptable) elements. Furthermore, there are some paleographic indications that *Heliand* was once sung. In other words, *Heliand* represents a text that contains multiple confounding alterations. Each of these alone challenges correct alignment with the original.

2.2 Tatian's *Diatessaron*

The *Diatessaron* is a harmony of the Gospels originally penned by the 2nd-century Syrian Tatian. A later translation into Latin (5th century) or its 9th-century translation into OHG is the proposed source behind the *Heliand*. Both Latin and OHG versions existed at the Benedictine scriptorium at Fulda in the early 9th century. Various theories (often circular in their logic) thus also place the *Heliand* there. Whatever the location of origin, the *Diatessaron* itself certainly represents the first steps of conversion from the original Gospels into OS, due to its similar re-organisation of the textual structure, wherein overlapping information from the four was reduced to a single text.

The stage of alteration from the Gospels to the *Diatessaron* is one that, though extensive, can quite easily be resolved using *n-gram* analysis, i.e. for determining which target material stems from which source material

The conversion of the *Diatessaron* to the *Heliand* is significantly more complicated, involving permutation, translation, and innovation. Thanks to his page-by-page linking of the *Heliand* with the OHG *Diatessaron*, Burkhard Taeger's (Behaghel/Taeger 1984) *Heliand* edition provides the reference notation that allows for smaller sections of both texts to be compared with one another relatively easily. This thus provides the bridge between the *Heliand* as target text and the Gospels as its source.

2.3 Otfrid's *Evangelienbuch*

Another old Germanic re-telling of the Gospels is the *Evangelienbuch*, penned in OHG by Otfrid of Weissenburg ca. 830. His presence at Fulda Monastery at that time is documented, which lends some credence to the theories noted above regarding the provenance of the OS *Heliand*.

Similar to the *Heliand*, Otfrid's work is a Gospel harmony retold in poetic verse. These facts have prompted theories that the *Heliand* and the *Evangelienbuch* were derived in succession from the *Diatessaron* at Fulda. However, the two texts differ; consequently, it is inconclusive as to whether the two are products of the same author.

These differences are evident primarily in style: the *Evangelienbuch* shows Latinate (or at least non-Continental Germanic) influence (end-rhyme, syllable counting), whereas the *Heliand* shows the traditional Germanic style (alliteration, loose metrics). Furthermore, the two linguistic varieties, OS and OHG,

though genetically related, were already at the time distinct languages, also suggesting that the texts are not simply the product of the same author.

Whatever relationships do exist between the two texts are best seen when compared via the *Diatessaron*. The resulting network of texts – the OS *Heliand*, the OHG *Evangelienbuch*, and the Latin *Diatessaron*, presents a challenge in alignment (due to more than just the differences in language might suggest), because Paul Piper’s 1884 reprint of Otfrid gives less-than-systematic chunks of the *Evangelienbuch* and their links with *Diatessaron*, meaning the full bridge between the interlinguistic texts is full of sizeable potholes. To close this gap, yet another version of the *Evangelienbuch*, this time in its Latin version, also needs to be taken into account.

2.4 Traditional Gospels

Eduard Sievers produced (1872) a side-by-side publication of the Latin and OHG *Diatessaron*. In it, he references where the Latin version corresponds to the Vulgate. With this, the complete bridge of relationships is built from the Gospels via the Latin and OHG *Diatessaron* and the OHG *Evangelienbuch* to the OS *Heliand*. It is significant to remember that, despite the linguistic varieties, only four texts are used.

Barring any significant differences resulting from the various languages of the texts, the resulting alignments via Sievers and Behaghel/Taeger create chunks of text that are manageable for comparison purposes: a handful of Gospel verses correspond to more or less individual *Diatessaron* sentences, as well as to ca. 30 lines of the *Heliand*. Section lengths from Otfrid vary widely, at most ca. 200 lines. Considering the full number of lines/verses of all texts involved reaches ca. 20,000 units, comparison groups averaging ca. 60 lines (from all texts) is an advancement on the necessity of comparing full texts with each other.

The Gospel translations relevant to the given project involve six languages (see “Digital Resources” below): Latin (Vulgate), Greek (*Novum Testamentum Graece*), Wulfila’s Gothic (5th century), Old English (OE: Wessex Gospels, ca. 990), and two Early New High German versions from Luther (*Septembertestament* [1521] and *Letzter Hand* [1545]). The Gothic and OE versions are mostly complete and stand in place of non-existent OS and OHG full Gospels. As genetically related Germanic languages, Gothic and OE are therefore the closest linguistic comparison between the derived texts (i.e., the *Heliand*, Tatian, Ot-

frid) and the full Gospels. The Vulgate provides the bridge between all the texts, since the Latin *Diatessaron* seems to correspond nearly word-for-word to it. The Gothic Gospels are generally thought to have been made from the Greek, which is therefore included. Any difference between the full Gospels in Latin, Greek, and Gothic reveals relationships that might have been passed down to texts derived from them. Gothic and OE provide a similar role for comparing the language of the derived Germanic texts.

The Luther translations are included because of rumors that Luther may have been influenced by the *Heliand*. Thus, Luther's disputed translations might owe some of their existence to this purported external influence. The variations in Luther thus provide another wrinkle of difficulty to test our measures, viz. by comparing his language a) to the Latin and Greek, b) to the *Heliand* (and potentially Otfrid, Tatian, Gothic, and OE), and c) to itself. This last point explains the inclusion of two Luther translations – his first and his last, which will aid in determining when his linguistic innovations came into being.

3. Alignment

A key element to the theory that the *Heliand*, the *Evangelienbuch*, and the OHG *Diatessaron* are related is a purported preference in all three for St. Matthew. A known resource at Fulda is the *Matthäuskommentar*, created by Hrabanus Maurus, alleged translator of the *Diatessaron* into OHG. Thus, one further goal is to test this theory: do the *Heliand* and the *Evangelienbuch* truly follow Matthew more commonly? Simply stated, the arguments made to support both of the texts having been penned in Fulda after the OHG *Diatessaron* was completed are cyclical in nature. By either proving or disproving the supposed preference for Matthew in these two derived texts, one will give substantial proof either towards the argument of their link to Fulda, or at least that this argument is unfounded (as such, other theories about the provenance of the *Heliand* and the *Evangelienbuch* might look more promising).

In order to test which Gospel any given section of the *Heliand* and/or the *Evangelienbuch* follows, one needs a summary of where the four Gospels overlap in theme. A table consisting of such relationships was produced first in the late 3rd century by Eusebius of Caesarea. His *Eusebian Tables* account for all parallel sections where any of the four Gospels correspond with all or any of the other Gospels, as well as where each Gospel contains unique storyline. Eusebius made use of divisions created by contemporary 3rd-century Christian philoso-

pher Ammonius of Alexandria, who developed the first means of subdividing the Gospels into verse-like chunks.¹

For the use of our project, the ten Eusebian Canon Tables and their subordinate Ammonian sections have been borrowed wholesale with one minor addition: an eleventh Canon to account for the Long Ending of Matthew, which developed into Christian canon after Eusebius' time.² The "chunk" divisions made to the *Heliand*, the *Evangelienbuch*, and the *Diatessaron* have all been matched via their aforementioned relationships to the full Gospels in the Vulgate (i.e., through the Latin *Diatessaron* analysis by Sievers). To account for material where the *Diatessaron* suggests Matthew as the source, one needs to also consider where these verses overlap with similar thematic material in the other Gospels. This is done via the Eusebian Canon Tables. I have extended the search for overlap to include all four Gospels as possible sources. Thus, a matrix of relationships develops whereby the *Heliand* and the *Evangelienbuch* are compared to thematically similar sections of the *Diatessaron* in both OHG and Latin, which is then compared to the Vulgate Latin equivalent as postulated by Sievers. Since the Gospel chapter-and-verse divisions hold true regardless of language, the other five versions of the Gospels can be referenced easily. From here, a failsafe to prove or disprove Sievers' reading of the Latin *Diatessaron* is provided by taking the Ammonian section to which his suggested Gospel reference belongs, and comparing this Ammonian section with those from the other Gospels (and, apparently in some cases, within the same Gospel) as provided by the Eusebian Canon Tables.

All this produces thematically reduced chunks of each text, presented side-by-side with all the possible source material from the Gospels to which one needs to turn to find the actual source material. Note, this result only produces the correct aligned materials. A future stage of the project will seek to develop the computational algorithms to analyse the linguistic material in all its variant languages to determine relationships and/or similarity.

¹ The now-traditional chapter and verse divisions of the New Testament are Renaissance and Reformation-era inventions by Stephen Langton (chapters, 13th century; Fenlon 1910, "Hebrew Bible") and Robert Estienne (verses, 16th century; Miller/Huber 2004: 173).

² As this material was not considered canon during the 2nd century, it was not included by Eusebius in his tables. Nevertheless, as its inclusion into the Canon was accomplished anciently, this material in St. Matthew has had historical impact on other Gospel translations and derivatives. It is thus necessary to include it somehow in the system for research purposes. I have done so by simply assigning it as the eleventh Eusebian Canon Table, in continuation of his systematic.

Despite not yet providing this analysis of similarity, the present stage of the research does provide a necessary tool. Though quite simple in appearance, the relationships produced by this process have successfully linked disparate texts from different time periods and in distinctive styles and variant languages by discovering where these texts parallel one another thematically. The results are 4,560 inter-text groups made up of the related chunks between all the originals and languages involved. Each of these 4,560 groups easily fits on a single webpage (sometimes with minimal scrolling) that any researcher will be able to access via the Internet. Simple naked-eye comparisons of the sub-texts can be performed now, where in the past simply finding where one line from the *Heliand* relates to which verse of, e.g., St. Mark, would have required many hours of tedium.

4. Representation

Access to the text groups is provided via a webpage,³ shown here in Figure 1. This page is divided into two horizontal areas, each with four columns. In the top area are the three derived texts (the *Heliand*, Otfrid's *Evangelienbuch*, Tatian's *Diatessaron*) and their "Gospel Equivalent", i.e. the verse(s) from which the *Diatessaron* section is purportedly derived. The bottom half of the page delivers the contents of the Ammonian section to which the "Gospel Equivalent" belongs, as well as the text of any parallel Ammonian sections from the other Gospel books. This bottom grouping is ordered according to western tradition (Matthew, Mark, Luke, John). Wherever a text section from any source is too long for the page space provided, the column is scrollable:

Figure 1 reflects a search performed by inputting the group number 680 (of the 4,560 total inter-text groups) in a search field on a previous page. A similar search can be performed on structural divisions of any of the texts involved. Thus, searching for *Heliand* fitt⁴ 34, line 2,885 will similarly bring up all the text subsections to which it has been related in the background database. Besides allowing for searches on book, chapter, and verse of the Gospels (lower half), one can also search according to Ammonian section numbers. The related parallel material of the other Gospels and the derived alignments will be called up automatically.

³ <http://titus.uni-frankfurt.de/database/diatessaron>.

⁴ The term historically used in reference to major subdivisions of the *Heliand*.

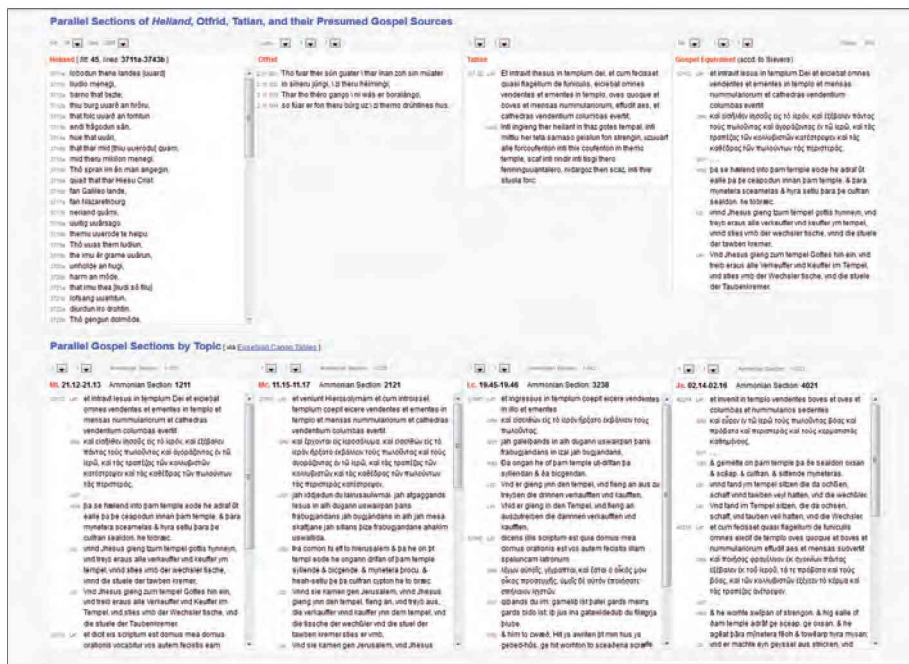


Figure 1: Screenshot showing parallel texts belonging to group 680

The HTML code draws its information from a centralized database with a main table containing information about the 4,560 inter-text groups. These in turn reference the major subdivisions of each text by their various systems: the *Heliland* by Taeger’s page numbering,⁵ Otrfrid by Piper’s analysis, Tatian via Sievers’ reprint, and the “Gospel Equivalents” by Sievers’ alignment of the Latin Tatian version back to the Vulgate. The Gospel sections are referred to via their associated Ammonian Sections.

Beyond this, the database contains tables with information for the various divisional structures used for each text. Through a Structured Query Language (SQL) search of either these tables directly, or via the main inter-text table, a third layer of tables is queried from which the text is pulled. In the case of the *Heliland* and Otrfrid, this third layer contains a single column containing the

⁵ I.e., not by *fitts*, since the device used to compare between the *Heliland* and the *Diatessaron* provided by Taeger occurs at the top of each page of his *Heliland* edition. As any given *fitt* can vary in the number of pages it covers in his edition, it has been more practical to use Taeger’s pages ad hoc as the unit of subdivision, though the text presented also indicates the given *fitt* and line number traditionally assigned to the *Heliland*.

text, while the other texts contain multiple columns separating the languages in which the text occurs.

5. Post-project directions

Due to the database structure chosen, each text can be accessed individually, and each language version of the multilingual texts can be accessed or hidden independently. Certain features were not implemented due to time constraints. These include: a set-up page in which the user can limit which of the Gospel languages he wishes to view. Also not realized was the ability to upload one's own Gospel translations and/or Gospel-derived texts, e.g. the Arabic *Injil*, which could theoretically still be made possible, as such translations/texts can be subdivided into comparably similar chunks as has been done for the *Heliand*, the *Evangelienbuch*, and the *Diatessaron*.

At the end of the research period, the results described in this paper were handed over for further development to Prof. Christian Chiarcos, also of the Frankfurt "Digital Humanities" group. Prof. Chiarcos has continued to build on these results by developing the means to automatically compare the linguistic material of the inter-text groups, in part by taking advantage of comparative semantic information delivered by another sub-project, namely 'Historische Sprachdatenbank (Simplex)', for which I was also responsible. For this sub-project, I collated digitized etymological and translation dictionaries of the languages dealt with above.

Using the resulting network of semantic and formal relationships, Prof. Chiarcos has succeeded in being able to automatically highlight the semantic relationships between the texts under concern here, as delivered by the inter-text subdivision. Thus, for example, by clicking on "intravit" of the Latin Tatian section shown in Figure 1, related words ("ingieng" in OHG Tatian, "quam" in *Heliand* line 3734a, "geng" in Otfrid, "eode" in OE [WSX],⁶ etc.) appears highlighted, directing the user's eye to more specifically, semantically related areas between the texts than what the current divisions provide.

Also under development by Prof. Chiarcos is the ability to compare multi-word units and *n-grams*, where such exist. A goal of this step is to reveal similar syntactic strings with semantic relations, e.g. "intravit Ihesus in templum dei" (Latin Tatian), "ingieng ther heilant in thaz gotes tempal" (OHG Tatian

⁶ I.e., Old English (Wessex).

117: 2), “Thô he an thene uúih innen, / geng an that godes hús” (*Heliand* 3733b-3734a.), “fúar er [...] \ zi themo drúhtines hus” (Otfrid Ev. 2 11:4), “se hælend into þam temple eode” (OE [WSX] Matthew 21:12), as well as the similar text from the other Gospels, e.g. “vnnð Jhesus gieng ynn den tempel” (L22 Mark 11:15), “ah galeipands in alh” (GOT Luke 19:45), “et invenit in templo” (LAT John 2:14), even where this differs from the “direct” translation by Luther (L45): “Vnd fand im Tempel sitzen”.

Digital resources⁷

Aland, Barbara/Aland, Kurt (2001): *Novum Testamentum Graece*. 27. Aufl. Stuttgart: Dt. Bibelges. Center for Computer Analysis of Texts (CCAT) database corrected and amplified by James Tauber. Accessed using E-Sword: www.e-sword.net/.

Behaghel, Otto/Taeger, Burkhard (1984): *Heliand und Genesis*. Tübingen: Max Niemeyer. Accessed at TITUS: <http://titus.uni-frankfurt.de/texte/etcs/germ/asachs/heliand/helia.htm>.

Luther, Martin (1522): *Biblia: die ganze Heilige Schriftt. Deudsch*. Wittenberg. Accessed at Wikisource: <http://de.wikisource.org/wiki/Lutherbibel>.

Luther, Martin (1546): *Biblia: Die gantze Heilige Schriftt Deudsch*. Wittenberg: Hans Lufft. Accessed at Wikisource: <http://de.wikisource.org/wiki/Lutherbibel>.

Piper, Paul (1884): *Otfrids Evangelienbuch: mit Einleitung, erklärenden Anmerkungen und ausführlichem Glossar / 2 Glossar und Abriss der Grammatik*. Paderborn: Schöningh. Accessed at TITUS: <http://titus.uni-frankfurt.de/texte/etcs/germ/ahd/otfrid/otfri.htm>.

Sievers, Eduard (1872): *Tatian. Lateinisch und altdeutsch / mit ausführlichem Glossar*. Paderborn: Schöningh. Accessed at TITUS: <http://titus.uni-frankfurt.de/texte/etcs/germ/ahd/tatian/tatia.htm>.

St. Jerome (405): *Latin Vulgate with Deuterocanon using Gallican Psalter*. Accessed using E-Sword: www.e-sword.net/.

Streiber, Wilhelm (1919): *Die Gotische Bibel: Der gotische Text und seine griechische Vorlage. Mit Einleitung, Lesarten und Quellennachweisen sowie den kleineren Denkmälern als Anhang*. Heidelberg: Carl Winter. Accessed at the Wulfila Project: www.wulfila.be.

Wessex Gospels (ca. 990): *Part of the York-Toronto-Helsinki Parsed Corpus of Old English prose*. Accessed at the Oxford Text Archive: <http://ota.ahds.ac.uk>.

⁷ All URLs have been checked and found valid as of late January 2015.

References

- Fenlon, John Francis (1910): *The Catholic Encyclopedia*. Vol. 7. New York: Robert Appleton Company.
- Miller, Stephen M./Huber, Robert V. (2004): *The Bible: A History: The Making and Impact of the Bible*. Intercourse, PA: Good Books.

Historical corpora and word formation

How to annotate a corpus to facilitate automatic analyses of noun-noun compounds

Abstract

In this paper we present some preliminary considerations concerning the possibility of automatic parsing an annotated corpus for N-N compounds. This should in principle be possible at least for relational and stereotype compounds, if the lemmatization of the corpus connects the lemmata with lexical entries as described in Höhle (1982). These lexical entries then supply the necessary information about the argument structure of a relational noun or about the stereotypical purpose associated with the noun's referent which can be used to establish a relation between the first and the head constituent of the compound.

1. Introduction

In our paper we present the outline of a research project on noun-noun (N-N) compounds in Old High German (OHG). The main focus of the paper lies on topics of corpus annotation with respect to word formation – a topic hardly ever addressed before to our knowledge.

The project is mainly concerned with three issues:

- 1) Structural aspects
- 2) Interpretation of N-N compounds
- 3) Corpus-linguistic aspects

In order to achieve our goals, we proceed in two steps:

First we build a data base which contains the N-N compounds listed in the OHG dictionary of Jochen Splett (1993). In a second step we try to investigate the question of which kind of corpus annotation is necessary to facilitate a quasi-automatic analysis of N-N compounds. The data base should deliver the empirical basis for the investigation of the structural aspects and the internal semantics of N-N compounds. It will contain all the relevant information for

¹ We want to thank two anonymous reviewers for helpful comments.

this purpose such as, e.g., the internal structure or the type of compound. To prepare a corpus annotation we try to investigate which and how much information must be given in the annotation to enable more or less automatic semantic analyses of N-N compounds in annotated corpora.

2. Classification of N-N compounds

In our project we make use of an approved classification of N-N compounds (cf. Olsen 2000, Meibauer et al. 2002) to analyze the OHG N-N-compounds listed in the OHG dictionary of Splett. We compile a data-base which contains all N-N-compounds and classifies each of these compounds according to this classification, as far as this is possible. We expect that this classification will be sufficient to analyze a great deal of the OHG compounds but have to be refined in some way (for example with respect to coordinative compounds, cf. Çinkılıç/Weiß 2012). We will only consider non-lexicalized forms, whose meaning is built up in a compositional way.

In this paper, we just give a short impression of the classification we are using (for further details on this topic, cf. Çinkılıç/Weiß 2013). Compounds could be subdivided into the following subclasses (of which only determinative compounds will be considered in the remainder of the paper):

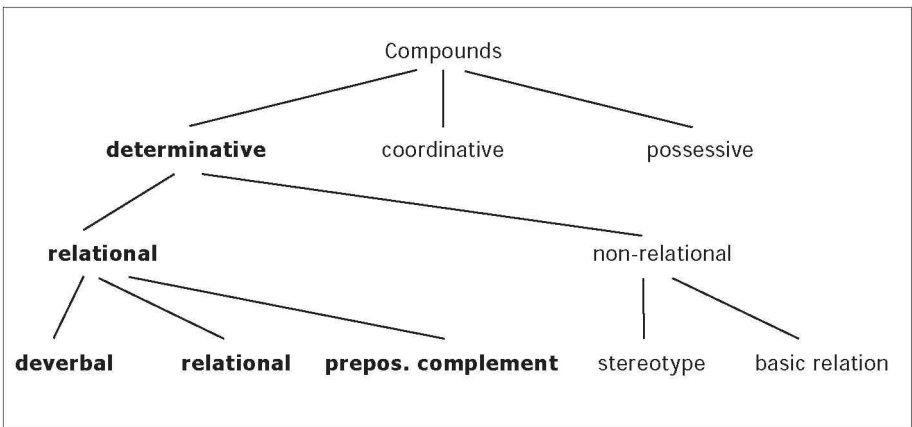


Figure 1: Subclasses of compounds

2.1 Relational compounds

Relational compounds are defined as compounds where the head is a relational noun and the first constituent is its argument. There are three subtypes:

Deverbal head

The first type includes relational compounds with a deverbal head, i.e. the noun is derived from a verb and the first constituent qualifies as an internal argument.

Examples:²

Hūs-eigo ‘house lord’ < *eigan* ‘to possess’

Wind-fanga ‘porch’ < *fāhan* ‘to grip, catch’

Obaz-traga ‘fruit carrier’ < *tragan* ‘to carry’

Noun with prepositional complement as head

The head of the second type is a noun which can take a prepositional complement. Note that the relevant examples *minna* and *lust* are not derived from a verb.

Examples:

minna ‘love of/for’

heim-minna = ‘love of home, patriotism’

māg-minna = ‘love of relatives’

lust ‘desire for’

minna-lust = ‘desire for love’

weralt-lust = ‘desire for the world’

Relational noun as head

The third type includes compounds with an inherently relational noun as head.

Examples:

sun ‘son’

basun-sun = ‘son of the aunt’

brouder-/swester-sun = ‘nephew’

huorūn-sun = ‘sun of a whore’

² All examples in this section are taken from Çinkılıç/Weiβ (2013).

A noun like *son* is inherently relational because it denotes a relation between two individuals (A is the son of B). The first constituent of the compound can function as member of such a relation, so a *basun-sun* is the son of the aunt and *swester-sun* denotes the son of the sister.

2.2 Non-relational compounds

Non-relational compounds are ones where the relation between first and head constituent which underlies the interpretation is not evident by the argument structure or the lexical meaning of the head, but comes from elsewhere. There are two types of non-relational compounds:

Stereotype

Because of our world knowledge we have built some stereotypes in our minds. Such stereotypes comprise, for example, the knowledge that artifacts are made for certain purposes, and this knowledge can be the basis of the interpretation, cf.:

teig-/wazzartroc ‘dough/water trough’

It is a stereotypical aspect of the lexical meaning of ‘trough’ that it is made to contain something – and this can be named as the first constituent.

bouhscrîni ‘bookcase’

A cupboard (= *scrîni*) is a piece of furniture used for storage, in this case for the storage of books.

Basic relations

The interpretational relations between the first and the head constituent can be basic relations which “are part of the mental algebra that processes meanings combined with other meanings” (Olsen 2000: 909). We use basic relations everywhere to classify things according to their shape, consistence, composition, purpose, or their tempo-spatial localization – and such basic relations can be inferred as holding between the constituents of an N-N compound. So a *Holzhaus* is a ‘house made of wood’ or the *Mittagessen* ‘the meal consumed at noon’.

As for OHG, there are numerous examples in Splett's dictionary for most of the basic relations, especially for LOC, TEMP and CONST (cf. Çinkiliç/Weiß 2013 for a more explicit description of the basic relations):

LOC	<i>hals-adra</i> 'carotid (artery)' – <i>herz-âdra</i> 'aorta'
TEMP	<i>âband-muos</i> 'dinner' – <i>âband-sterno</i> 'evening star'
PURP	<i>ambaht-hûs</i> 'workshop'
CONST	<i>agat-stein</i> 'agate stone' – <i>stein-ofen</i> 'stone oven' – <i>fig-flado</i> 'cake made with figs'
INSTR	<i>scif-wig</i> 'battle with ships' – <i>scilt-spil</i> 'fight with shields'
CAUSE	<i>wurm-âz</i> 'damage caused by worms'
THEME	<i>êwabuoh</i> 'law book'
SIM	<i>goldamaro</i> 'yellowhammer'
PART	<i>swert-helza</i> 'hilt of a sword' – <i>huot-snuora</i> 'hat string'

3. Corpus annotation

Apart from the data base and the qualitative (and quantitative) analysis of OHG N-N compounds, a further focus of the project lies on topics of corpus annotation. We try to elaborate which and how much information must be given in the annotation to enable more or less automatic semantic analyses of N-N compounds in annotated corpora. Questions to address here are, e.g., how to integrate context information or information from the lexical entry of the head constituent regarding its argument structure, its relational nature, etc. The challenge is, among others, to make visible information concerning the internal relation of the constituents (i.e. information below the actual 'word boundaries') and to guarantee access to it. To develop a special design for digital corpus studies of word formation is thus a further main object of investigation. In the following, we will present some preliminary considerations concerning this task.

As is known (cf. Linde 2011), annotating linguistic corpora involves morphological tagging concerning various kinds of information, e.g., parts of speech (POS) tagging, grammatical feature encoding, morphological segmentation, and morpheme type encoding. With this information at hand, it is already possible to automatically detect N-N compounds in an annotated text. If information about the type of the morphemes a complex word consists of is sup-

plied, it is possible to automatically identify compounds – as well as any other kind of complex word.

However, this information does not suffice to identify the ‘semantic’ type of a N-N compound. As described in section 2, the types of N-N compounds are distinguished according to the semantic relation between the head and the first constituent. The first type is established by relational compounds where the head is a relational noun and the first constituent represents an argument of the head. As mentioned above, the head can be a deverbal noun, a noun with a prepositional complement, or an inherently relational noun. So we need access to two types of information: (i) information whether the head noun is of such a relational kind, and (ii) information whether the first constituent can serve as an argument of the head noun.

At least the first kind of information can easily be retrieved if the lemmatization of the corpus connects the lemmata with lexical entries as described in Höhle (1982). Such lexical entries store information concerning the phonology, morphology, syntax, semantics, and pragmatics. For our purpose, it is crucial that lexical entries contain syntactic and semantic information of the necessary kind.

We will present a concrete example for how this automatic analysis could probably work (cf. Çinkiliç/Weiß 2013). The OHG lexicon of Splett (1993: vol 2, 1008) lists several compounds with *tragâri* ‘carrier’ as their head noun, e.g.:

<i>lioht-tragâri</i>	‘candle carrier’
<i>spera-tragâri</i>	‘spear carrier’
<i>stank-tragâri</i>	‘scent carrier’
<i>swert-tragâri</i>	‘sword carrier’
<i>wazzar-tragâri</i>	‘water carrier’
<i>wolla-tragâri</i>	‘wool carrier’

The head noun *tragâri* is a deverbal noun: it consists of the verbal root *trag* ‘to carry’ and a nominal suffix, *-âri*, which derives nouns from verbs:

[*trag*]_v + [*âri*]_{Nsuf}

Now, it is crucial that both morphemes of the head noun are listed separately in the lexicon, i.e., that the derivational suffixes also get lexical entries. The lexical entries of *-âri* and *trag-* would look like this:

Lexeme	-âri
PHON	/a:ri/
MORPH	Masculin
SYN	N _{af} [V _]
SEM	AGENT or INSTRUMENT which executes the V-action
PRAG	

Table 1: Lexical entry of -âri

Lexeme	trag-
PHON	/trag/
MORPH	Strong inflection
SYN	V [NP _{nom} 1, NP _{akk} 2]
SEM	Activity TRAG (x1, x2) x1: AGENS, x2: THEME
PRAG	neutral register

Table 2: Lexical entry of trag-

From the lexical entries we can see that *tragâri* is a deverbal noun. In addition to that, the lexical entry of the verb *tragan* provides the necessary information about its argument structure: it takes two arguments and one of them is a theme argument. So compounds with the head *-tragâri* could in principle be relational compounds – if the first constituent is a possible theme of *tragan*. However, the final decision would then require to look at the lexical entry of the first constituent – in our examples the lexical entries of ‘candle’, ‘scent’, ‘sword’, ‘water’, and ‘wool’ – to see whether they qualify as possible themes for the verb *tragan* or not.

However, at the moment it is hard to imagine how the information whether the first constituent is a possible argument of the relational head noun or not can be specified in the lexical entry. In addition, there are some further principal problems. One problem is that words can be used non-literarily; e.g., one cannot only carry material things like the ones in the examples mentioned above but also immaterial ones (cf. *lugi-tragâri* lit. ‘lie carrier’, i.e. ‘someone who spreads lies’). Another problem arises from the general possibility to in-

interpret relational compounds as non-relational ones; so a *Steinträger*, lit. ‘stone-carrier’, can denote a person who carries stones (as a profession), or it denotes a pillar made of stone. The second meaning is based on the basic relation CONST (‘X consists of Y’). Which interpretation is meant can only be decided on the basis of information coming from the co(n)text. Whether these problems can ever be solved in a satisfactory way remains open for further investigation.

Nonetheless the way sketched above is a conceivable way, at least to decide whether a given compound could be a relational compound or not.³ In contrast to this, compounds with a basic relation would be harder – if not impossible – to analyze automatically, since basic relations are independent of the lexical meaning of words (Olsen 2000: 909), so access to lexical entries would not help. Here, one may use some kind of exclusion rule like: if the head is not a relational noun, infer some basic relation for the interpretation of the relation between the first constituent and the head noun.

However, stereotype compounds may be treated in a similar way as relational compounds, because one can still use lexical information in a broader sense to interpret their semantics. Stereotypes arise from opinions about typical properties associated with the referents of the words (Olsen 2000: 910). This kind of ‘wor(l)d knowledge’ may be part of the description of the meaning of a word in the mental lexicon or at least connected with it (i.e., be part of a neural network representing the word meaning). However that may be realized in our brains need not concern us here. The important thing is that it is possible to integrate such information into lexical entries.

In section 2.2 we already mentioned examples of stereotype compounds from OHG: *teig-/wazzartroc* ‘dough/water trough’, *bouhscrîni* ‘bookcase’. A *troc* as in *teig-/wazzartroc* for instance is a trough where you can put something into – like dough or water. Troughs are artifacts and we know (whether a priori or from experience does not matter in our context) that artifacts are constructed to fulfill a certain purpose. This is one of the main differences to natural objects like trees or cats. It is obvious to think that the purpose an artifact is made for is part of the lexical meaning of the word denoting it. If this is the case, the lexical entry of *troc* ‘dough’ will contain a description of its possible purposes,

³ The other two subtypes of relational compounds can be treated in the same way. As shown in section 2.1, the meaning of an inherently relational noun like *son* contains per definitionem a relation, because it denotes a male human offspring of *somebody*. The same holds for nouns with a prepositional complement.

and access to this information can be used to establish a relation between the head and the first constituent. So we can apply the same procedure for stereotype and relational compounds.

4. Conclusions

In this paper, we have mainly presented some preliminary considerations concerning the possibility of automatic parsing an annotated corpus for N-N compounds. This should in principle be possible at least for relational and stereotype compounds, if the lemmatization of the corpus connects the lemmata with lexical entries as described in Höhle (1982). These lexical entries then supply the necessary information about the argument structure of a relational noun or about the stereotypical purpose associated with the noun's referent, which can be used to establish a relation between the first and the head constituent of the compound.

References

- Çinkılıç, Gaye/Weiß, Helmut (2012): Kopulativkomposita. In: *Linguistische Berichte* 232: 417-435.
- Çinkılıç, Gaye/Weiß, Helmut (2013): Historical word formation in German. On the interpretation of N-N compounds. In: Fisiak, Jacek/Bator, Magdalena (eds.): *Historical English word formation and semantics*. Frankfurt a.M./New York: Peter Lang, 211-227.
- Höhle, Tilman N. (1982): Über Komposition und Derivation: zur Konstituentenstruktur von Wortbildungsprodukten im Deutschen. In: *Zeitschrift für Sprachwissenschaft* 1: 76-112.
- Linde, Sonja (2011): Referenzkorpus Altdeutsch. Kurzbeschreibung. www.deutschdiachrondigital.de/data/home/manual/dateien/Manual.pdf (last accessed: January 29, 2015).
- Meibauer, Jörg/Demske, Ulrike/Geilfuß-Wolfgang, Jochen/Pafel, Jürgen/Ramers, Karl-Heinz/Rothweiler, Monika/Steinbach, Markus (2002): *Einführung in die germanistische Linguistik*. Stuttgart/Weimar: Metzler.
- Olsen, Susan (2000): Composition. In: Booij, Geert/Lehmann, Christian/Mugdan, Joachim (eds.): *Morphology. A Handbook of Inflection and Word Formation*. Berlin/New York: de Gruyter, 897-916.
- Splett, Jochen (1993): *Althochdeutsches Wörterbuch. Analyse der Wortfamilienstrukturen des Althochdeutschen, zugleich Grundlegung einer zukünftigen Strukturgeschichte des deutschen Wortschatzes*. 2 vols. Berlin/New York: de Gruyter.

Object order and the Thematic Hierarchy in older German¹

Abstract

The relative order of dative and accusative objects in older German is less free than it is today. The reason for this could be that speakers of the direct predecessor of Old High German organized the referents according to the Thematic Hierarchy. If one applies a Case Hierarchy Nom>Acc>Dat to this, the order Nom – Dat – Acc falls out. It becomes apparent that the status of the Thematic Hierarchy is not a factor governing underlying word order, but a factor inducing scrambling. Arguments from binding theory, whose validity is discussed, indicate that the underlying order is ‘accusative before dative’.

1. Introduction

This paper is concerned with the relative position of accusative (direct) and dative (indirect) object full noun phrases in the history of German. The central hypothesis, which was brought forward in Speyer (2011), is that German underwent a period in which the relative order was relatively rigid – this period includes Old and Middle High German (OHG; MHG) and the earlier parts of Early New High German (ENHG), while only from the 16th century onward we notice a considerable variability. In Speyer (2013) it was also shown that the new ‘freedom’ in positioning is mostly a phenomenon of written German, whereas in spoken German (and in written texts that are highly influenced by the spoken register, voluntarily or not) even today the order ‘dative object before accusative object’ (Dat > Acc) is almost the norm, far outnumbering the order ‘accusative object before dative object’ (Acc > Dat).

The main goal of this paper is to present a theory why the order of objects in older stages of German was relatively rigid. This can be explained by the concept of Thematic Hierarchy in the tradition of Dowty (1991), as further devel-

¹ This paper elaborates on material that I have presented at the Historical Corpora conference in Frankfurt, December 2012. Parts of this material were presented also at talks at the universities of Cologne, Göttingen, and Saarbrücken between March 2011 and July 2012 as well as the Linguistic Evidence 2012 conference at Tübingen in February 2012. My thank goes to the audiences at this talks and other colleagues, especially Jost Gippert, Roland Hinterhölzl, Michael Job, Jürgen Pafel, Ingo Reich as well as an anonymous reviewer, for their comments and suggestions. All remaining errors are my responsibility.

oped by Primus (2002a; 2002b; 2004; 2011). In addition to that, this paper follows a secondary goal, closely related to its main goal, viz. to address the question whether the Thematic Hierarchy determines the base order or is rather a factor for short scrambling. There is evidence in favor of the second option, namely the binding argument known from, e.g., Müller (1999). This argument was attacked, e.g., by Rothmayr (2006), so in order to use it it is necessary first to discuss whether Rothmayr's criticism is valid (which we will see is not the case).

The paper is organized as follows: In section 2, after mentioning some factors that influence the object order in Modern German (ModG), the central evidence that suggests that the object order was relatively rigid is reviewed. Section 3 offers an explanation in terms of thematic roles and the Thematic Hierarchy as known from Dowty and Primus. Section 4 addresses the question what the status of the factor is and what the consequences for the underlying structure are. The discussion is briefly summarized in section 5.

2. Evidence for the lacking freedom of relative object order

Part of the evidence was already presented in Speyer (2011 and 2013). I summarize here the pieces of evidence relevant for the present purposes and add some new evidence in Tables 1 and 2.

Let us say first a word about word order in ModG.² The influential factors that are under discussion are numerous.³ Alongside factors that are grammatical in the stricter sense (such as: the subject usually stands before the objects, the dative object tends to stand before the accusative object, pronouns are put before lexical noun phrases, and so on) there are factors of a more cognitive-semantic nature (such as: agent first, animated referents before inanimate referents, etc.) and pragmatic / information-structural factors (topic before comment, given information before new information, etc.). In an earlier paper (Speyer 2011) I concentrated on the interaction between the factor that an unmarked order arises if the dative object stands before the accusative object (referred to as *Object Condition*) with what I call *Animacy Condition* (animated referents are mentioned before inanimate referents) and the tendency

² I use the traditional term word order here, although it is clear that the discussion mostly centers on the relative order of constituents.

³ On factors cf., e.g., Lenerz (1977); Lötscher (1981); Höhle (1982); Zubin/Köpcke (1985); Reis (1987); Fortmann/Frey (1997); Hoberg (1997); Musan (2002); Primus (2004, 2011).

that given information is placed before new information, dubbed here *Givenness Condition*. In the present paper I keep to this selection.⁴

None of these factors dominates the other in ModG. In short, we observe an acceptable constituent order if at least one of those conditions is fulfilled. Nevertheless, the Object Condition (or, rather, grammatical factors in general) plays a slightly more central role in that a Dat>Acc order is unmarked in any case, no matter whether the Animacy or the Givenness Condition are obeyed or not, whereas sentences with the order Acc>Dat are unmarked only if the Givenness Condition is fulfilled (Lenerz 1977). The same goes for the Animacy Condition: There is a set of systematic exceptions from the Object Condition, in verbs such as *aussetzen* ‘subject s.o. to s.th.’, or *ausliefern* ‘sur-render s.o. to s.th.’. Here, Acc>Dat is the unmarked order (1).

- (1a) ... *weil der Sportlehrer die Schüler*
 because the PE-teacher [the pupils]_{ACC}
 dieser unmenschlichen Gefahr aussetzte.
 [this inhuman danger]_{DAT} subjected
- (1b) ?*... *weil der Sportlehrer dieser unmenschlichen Gefahr*
 because the PE-teacher [this inhuman danger]_{DAT}
 die Schüler aussetzte.
 [the pupils]_{ACC} subjected
 ‘because the PE-teacher subjected the pupils to this inhuman
 danger.’

As typically the accusative object denotes a person that is subjected or surrendered to something abstract or at least non-personal, the unmarked word order results from the Animacy Condition in these cases.

So the picture is quite complex in Modern German. A closer look on the historical stages of German might help to evaluate the factors. The null hypothesis would be that German behaves diachronically like English or the Romance languages in that it drifts from a relatively ‘free’ word order to a word

⁴ A factor which is necessary to include in future work is definiteness. A central question in this context is whether the definiteness effect (i.e. that definite NPs stand before indefinite NPs; cf. also Lenerz 2002) is due to definiteness in a semantic sense or to the formal marking of definiteness. In the latter case we should expect the definiteness effect to play a role only after the definite article had developed. Presently the investigation of this question is under progress.

order mostly governed by grammatical factors. Note that ‘free word order’ does not mean that anything goes or that German is non-configurational or the like,⁵ but simply that there is a big impact of non-grammatical factors on word order so that it is not possible to predict the word order in a given sentence by grammatical factors alone (which in English or French would be possible as a rule). So in OHG or other historical stages, we would expect that the word order in general and the object order in particular were mostly governed by factors such as the Givenness Condition and the Animacy Condition, and less so by the Object Condition.

A closer look at the OHG data suggests that this is not true (Speyer 2011). While the nature of the OHG data is such that it does not allow direct statements about this question,⁶ we can at least gather indirect evidence: In the *Evangelienbuch* by Otfrid (mid 9th century AD), all clauses containing a full NP accusative and dative object show the order Dat>Acc (2).

- (2) *bráht er therera worolti diuri árunti*
 brought he [the world]_{DAT} [precious message]_{ACC}
 ‘He brought a precious message to the world’
 (Otfrid Ev. 1,5,4)

This is the more surprising because one should expect more license in poetic texts. The fact that Otfrid does not make use of this license suggests that for him the production of a word order Acc>Dat is highly marked, at least too marked to be used just for making verses rhyme or the like, and perhaps even not possible.

The evidence from the translation texts points in the same direction. Of course we do find quite often cases in which the translator simply copied the Latin constituent order. But every now and then the translator deviates from the original. So we find cases in which the Latin original has Acc>Dat, but the translator renders it as Dat>Acc in German (3).

⁵ Cf. the discussion in Webelhuth (1992: 40ff.)

⁶ There are no sources of the nature necessary for such investigations, i.e., large original prose texts. All there is in terms of large texts, is an original poetic text (Otfrid von Weissenburg’s *Evangelienbuch*, a Gospel harmony in metrical form), and some translations from Latin, the *Tatian* (a Gospel harmony), a fragmentary translation of Isidor of Seville’s *Contra Iudaeos* (all 1st half of the 9th century) and the translations of several texts by Notker Labeo (around 1000 AD), among them a commentary to the psalms and Boethius’ *Consolatio Philosophiae*.

- (3) *Der allen mēnniscon ēzen gibit*
 who [all humans]_{.DAT} food_{.ACC} gives
 ‘who gives food to all people’
 (Notker Ps. 134 (505,17))⁸

Latin original:

qui dat escam omni carni
 who gives food_{.ACC} [all flesh]_{.DAT}

Interestingly, there are very few examples in which a Latin order Dat>Acc is translated as Acc>Dat in German. We should expect this to occur frequently, however, if Acc>Dat was an acceptable option in German. Under this view, the deviation from the Latin Acc>Dat order to German Dat>Acc actually gives a hint that the translator here ‘corrected’ according to his ‘*Sprachgefühl*’.⁷ So this evidence suggests strongly that in OHG the ‘normal’ word order was Dat>Acc, whereas Acc>Dat was either no option at all or very highly marked.

This impression is confirmed by looking at MHG and ENHG evidence (see Speyer 2011, 2013). In a selected part of Berthold von Regensburg’s *Sermons*,⁸ 94 of 96 clauses containing a full NP dative and accusative object show the order Dat>Acc (~ 2%). We get similar numbers for the early periods of ENHG. In fact, only after c.1500 the order Acc>Dat occurs with some frequency, and there it is due to the Givenness Condition which did not play a role for word order before.

So it looks as if the object order in OHG, MHG and early ENHG is governed mainly, if not exclusively, by the Object Condition.

⁷ We know that neither the translator(s) of Tatian nor Notker are consistent in correcting the text into native-like German. Sometimes they simply are not able to do it; in Tatian, for instance, the German line has to render the verbal material of the Latin line, and since the lines are relatively short, a deviation from the Latin text with respect to word order of full NPs was impossible in most cases. Sometimes they simply do not care about doing it, which has to do with the character of the translation which was more of a ‘crutch’ than a readable text on its own. This is true with other phenomena as well, e.g. the use of subject pronouns, which were almost obligatory in OHG but almost always dropped in Latin. Nevertheless, the translator of the Tatian drops it quite often, whereas Otfrid (remember, an original text) drops it seldom.

⁸ Sermons 1-15 in the edition by Pfeiffer; the sample contains roughly 80,000 words.

For MHG and ENHG, where we have direct evidence, the interaction between the Object Condition and the Givenness Condition is given in Table 1, the interaction between the Object Condition and the Animacy Condition in Table 2.⁹

	IO > DO				DO > IO			
	g>n	n>g	g>g	n>n	g>n	n>g	g>g	n>n
Berthold	29	13	30	16	2	–	–	–
Lancelot	9	2	16	4	1	2	3	1
ENHG sermons	1	–	4	2	1	–	2	2
ENHG narrative	16	–	9	4	1	3	3	1
total	55	15	59	26	5	5	8	4

Table 1: Given (g) and New (n) information in MHG and ENHG prose texts

	IO > DO				DO > IO			
	a>i	i>a	a>a	i>i	a>i	i>a	a>a	i>i
Berthold	69	–	17	8	–	2	–	–
Lancelot	27	–	7	–	–	3	4	–
ENHG sermons	4	–	2	1	–	3	1	1
ENHG narrative	24	–	6	–	–	7	1	–
total	124	–	32	9	–	15	6	1

Table 2: Animated (a) and inanimate (i) referents in MHG and ENHG prose texts

Table 1 illustrates that there is no interaction between the Object Condition and the Givenness Condition: given>new and new>given is attested with both orders, Acc>Dat and Dat>Acc. The picture changes dramatically if we look at Table 2. Here we note striking gaps: there are no examples of Dat>Acc that violate the Animacy Condition. At the same time, there are no examples of

⁹ The underlying texts are: Berthold: *sermons* 1-15; Lancelot: *Prosa-Lancelot*, part I, pp. 1-231; ENHG sermons: *Altdeutsche Predigten*; Johannes Tauler: *Predigt de Nativitate*; ENHG narrative: *Buch der Altväter*; Rulman Merswin: *Buch von den zwei Mannen*; Hans Mair: *Troja*; Helene Kottanerin: *Denkwürdigkeiten*. All ENHG texts are taken from the Bonner Frühneuhochdeutschkorpus (<http://korpora.org/Fnhd/>). The MHG texts have been searched through manually, mostly as a test whether the method to search for sample verbs offers satisfactory results. It does: The output of the exhaustive search in the MHG texts was almost completely dependent on verbs that stand on the sample verb list used for OHG and ENHG.

Acc>Dat that obey the Animacy Condition. All examples of Acc>Dat either are neutral with respect to animacy, or they violate the Animacy Condition.

By consequence, there must be a strong connection between the Animacy Condition and the Object Condition. In other words, the unmarked Dat>Acc order could be the result of the Animacy Condition as well as the Object Condition, simply because both conditions lead to the same output.

There are no parsed corpora available for either Old High German or any other historical stage of German (with a few exceptions in ENHG), so another method for which the available resources sufficed was used. There are corpora for each period of older German that are equipped with a word search engine. For OHG, the most relevant texts are available in the TITUS database (Gippert/Martinez/Korn (eds.) 1987ff.), for ENHG, there are some texts in the 'Bonner Frühneuhochdeutschkorpus' (Besch et al. (eds.) 1972-1985). The method here was to search for verbs with dative and accusative object. For OHG, a list of verbs could be composed using Greule (1999), for ENHG, a list of verbs was compiled for Modern German and applied to ENHG. The hits of searches for these verbs were manually filtered such that at the end only hits remained in which both dative and accusative objects are realized as a full noun phrases and in which neither of them stood in the pre-field (i.e., before the finite part of the verb form) or unambiguously in the after-field (i.e., after the infinite part of the verb form). In the case of MHG, the search had to be done manually as at least most of the Berthold text is not included in the Middle High German database (Springeth/Schmidt/Pütz (eds.) 1992).

3. The profiling of the Thematic Hierarchy

As mentioned before, there is one detail in the historical data that suggests an interaction between two factors. In Table 2, there is a conspicuous distribution pattern visible: While quite a number of Dat>Acc cases have also an order 'animated before non-animated referent', and several Dat>Acc cases are independent of the animacy of the referents, there is not one example in which Dat>Acc conforms to an order 'non-animated before animated referent'. This suggests a strong interaction. But what kind of an interaction is it?

Primus (2004, 2011) suggests that in many languages, word order is influenced by the Thematic Hierarchy. Following Dowty (1991), thematic roles are not conceptualized as discrete entities, i.e., participants with a well-defined set of properties, but rather as points on a scale. Depending on what agent-like

of patient-like properties the participants have, they can be ordered on the scale, the high-point of which is the maximum of agent-like properties (*Proto-Agent*), and the low-point, the maximum of patient-like properties (*Proto-Patient*). Table 3 gives a list of typical agentive / patient-like properties, adapted from Primus (2002a, b). The variable *e* denotes an event or state denoted by the verb, *x* denotes the role-bearer, *y* denotes some other participant in the verbal event. In a normal three-place predicate such as *give* or *show*, the agent (subject) would be high on the scale, the recipient (indirect object) would be somewhere in the middle, as it comprises agent-like properties such as physical activity (A5) with patient-like properties such as that s/he is not the participant that initiates the verbal action, but is causally affected (P3), while the patient (in *give*) or stimulus (in *show*) is lowest on the scale.

	Agent-like properties		Patient-like properties
A1	<i>x</i> is volitionally involved in <i>e</i>	P1	<i>x</i> is controlled by <i>y</i>
A2	<i>x</i> shows sentience / perception of <i>e</i>	P2	<i>x</i> is the target of sentience by <i>y</i>
A3	<i>x</i> causes <i>e</i>	P3	<i>x</i> is causally affected by <i>y</i>
A4	<i>x</i> causes change of state in <i>y</i>	P4	<i>x</i> undergoes a change of state
A5	<i>x</i> shows physical activity	P5	<i>x</i> is physically manipulated by <i>y</i>
A6	<i>x</i> undergoes movement (relative to the position of <i>y</i>)	P6	<i>x</i> is moved by <i>y</i>
A7	<i>x</i> exists independently of <i>e</i>	P7	<i>x</i> is dependent on <i>y</i> or <i>e</i>
A8	<i>x</i> possesses <i>y</i>	P8	<i>x</i> is possessed by <i>y</i>

Table 3: Typical agent-like and patient-like properties

Linking works in Primus' (2011) system such that a Case Hierarchy Nom>Acc>Dat is evoked in accusative languages like German. The participant highest on the Thematic Hierarchy links with the highest case in the Case Hierarchy. The second highest case, the accusative, does not link with the next highest element on the Thematic Hierarchy, but with the lowest element in the Thematic Hierarchy, while participants in the middle link with the dative as the lowest case in the Case Hierarchy (Fig. 1).

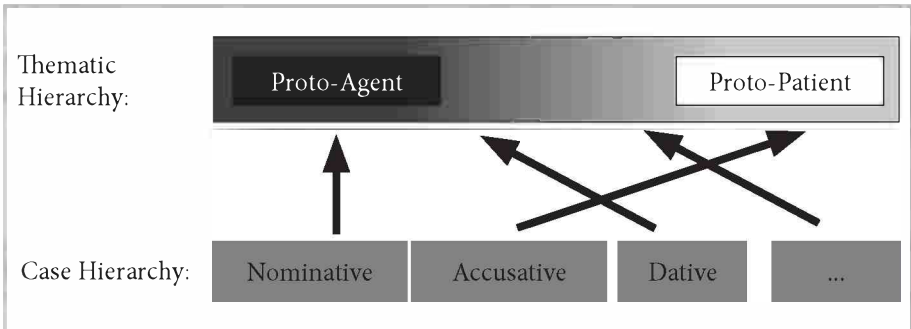


Figure 1: Linking in Primus' (2011) system

If in a language the word order is governed by the Thematic Hierarchy, and if the case linking works in the way Primus (2011) envisages, the order Nom>Dat>Acc in trivalent verbs comes for free as it is a direct result of the linking mechanism.

Similarly, the Animacy Condition falls out for free. Animacy is not a sufficient condition of agenthood. It is, strictly speaking, not even a necessary condition, but it is a typical property of agenthood nevertheless as it is a necessary condition of agent-like properties such as control of situation or physical activity. This is not to say that patients are typically inanimate, but animacy is no precondition for any prototypical patient-like property. So, if the participants in a verbal event are ordered according to the Thematic Hierarchy, chances are high that in the end animated participants are positioned before inanimate participants in the sentence.

If both the Object Condition and the Animacy Condition fall out from an ordering following the Thematic Hierarchy, we may conclude that the object order in OHG, MHG and early ENHG in reality did not follow either the Object Condition or the Animacy Condition, but instead was governed by the Thematic Hierarchy, which in turn was closely connected to the Case Hierarchy and followed the linking mechanism described by Primus (2011).

What happens is in later ENHG and up to ModG? Is the Thematic Hierarchy still a crucial factor governing word order? And is the linking to the Case Hierarchy still as strict? Relevant in this respect is the fact that from the 16th century onward, trivalent verbs like *aussetzen*, with an animate accusative and an inanimate dative object are attested (Speyer 2011). But note again that animacy or, rather, the lack thereof, is not a crucial condition of patienthood.

The referent of the accusative object shows rather typical patient-like properties (following the list of Primus 2002a and 2002b, see Table 3): it is controlled by another participant (P1), it is subjected to a situation that came about by another participant (P3), P4 and P5 can also apply. The referent of the dative object, i.e., the entity the referent of the accusative object is subjected to, on the other hand, shows neither agent-like nor patient-like properties. Thus, it is actually higher on the Thematic Hierarchy than the accusative. So the linking is as we would expect it from the Thematic Hierarchy: the dative is attached to the less patient-like object, the accusative to the more patient-like object. The word order, on the other hand, does not follow the Thematic Hierarchy in such cases, but rather the Animacy Condition which was originally a mere derivative side-effect of the ordering following the Thematic Hierarchy. So it looks as if German underwent a notable change: originally, that is, at some time before the OHG attestation sets in, the Thematic Hierarchy was the main if not the only factor governing object order.¹⁰ Descriptive generalizations following from the Thematic Hierarchy (which we can make but which language learners can make, too) are the Object Condition and the Animacy Condition. As all three conditioning factors lead roughly to the same output, how should a language learner decide which one is the decisive factor? The fact that verbs like *aussetzen* came about at some time suggests that at least from this time on, the Animacy Condition is independent from the Thematic Hierarchy and that language users view the Animacy Condition as the crucial factor, while they may view the Thematic Hierarchy as a subordinate factor or no factor at all. We may assume that something similar happened to the Object Condition at some stage. So the modern stage of affairs came about in which there is a multitude of factors that interact somehow.

¹⁰ I can speak confidently only of High German. It is possible that the changes described here in fact concern only High German. Preliminary searches on Old Saxon texts show that here the order of dative and accusative is actually less rigid than in OHG, which may suggest several things: 1) the Thematic Hierarchy and/or the Animacy Condition did not play such a crucial role for word order here, freeing the serialization for other, e.g. information-structural factors (cf. Petrova / Zeldes, this volume, on Middle Low German and Old Frisian constituent order); 2) the lexico-semantic structure of the verbs might have been somewhat different from OHG. Note that in Modern English, for instance, a verb such as *to give* can be represented in two ways (cf. Rappaport/Levin 1988): A: [[give]] = λzλyλx. [x cause [y to come into state [y has z]]] (yielding sentences like: *Mary gave John a book*), B: [[give]] = λyλzλx. [x cause [z to come into state [z is at y]]] (yielding sentences like *Mary gave a book to John*). This possibility is not (any more) possible with several trivalent verbs in German (among which is *geben* 'to give'). So there might be a connection. We are currently working on this problem.

4. On the structure of trivalent verbs in German

It is not immediately clear what the status of the word order Dat>Acc is in structural terms. The first question is whether both orders are base-generated, or whether one order is derived from the other. In the following I want to argue that the base order is Acc>Dat (the marked option), from which the Dat>Acc order is derived.

One may ask why the unmarked order Dat>Acc is not granted the status of a base-generated order. The reason is the well-known binding effect which clearly suggests that the accusative object c-commands (and thus binds) the dative object (cf., e.g., Müller 1999). This effect could not be derived if one assumed base-generation of the dative higher than the accusative. In the following I briefly illustrate the argument, mostly paraphrasing Müller (1999).

Crucial are sentence pairs like (4). Here, the accusative and the dative object refer to the same entity, that is to say, distributively to identical members of the same set. We know that in such cases the reference to one of the instances is done by an anaphor (reflexive or reciprocal pronoun in non-generative terminology). This is what sentence pair (4) shows. Principle A of Binding Theory (Chomsky 1981) requires that the anaphor has to be bound by the coreferent antecedent. The antecedent binds the anaphor by c-commanding it.

- (4a) *Ich sehe, dass Jörg [die Gäste]_i einander_i vorstellt.*
 I see that Jörg_{NOM} [the guests]_{ACC} each-other introduces
 'I see that Jörg introduces the guests to each other'

- (4b) **Ich sehe, dass Jörg [den Gästen]_i einander_i vorstellt.*
 I see that Jörg_{NOM} [the guests]_{DAT} each-other introduces
 'I see that Jörg introduces the guests to each other'

The crucial difference between (4a) and (4b) is that in (4a) the accusative object is realized by an R-expression, while the dative object is realized by an anaphor, whereas in (4b) it is the other way round. This has nothing to do with surface word order, as the sentence pair (5a, b), in which the anaphor precedes the R-expression, shows the same grammaticality contrast as (4a, b).

- (5a) *Ich sehe, dass Jörg einander_i [die Gäste]_i vorstellt.*
 I see that Jörg_{NOM} each-other [the guests]_{ACC} introduces
 'I see that Jörg introduces the guests to each other'

- (5b) * *Ich sehe, dass Jörg einander_i [den Gästen]_i vorstellt.*
 I see that Jörg_{NOM} each-other [the guests]_{DAT} introduces
 ‘I see that Jörg introduces the guests to each other’

As (4a) is grammatical, we may assume that Binding Theory is obeyed in this case, which requires a configuration in which the accusative object c-commands the dative object. This means that the accusative object has to be generated higher in the tree, resulting in a structure like (6a), which would result in a constituent order {NP-nom, NP-acc, NP-dat, V} if flattened out without further movement.

- (6a) [_{VP} NP-nom [_V [_{VP} NP-acc [_V NP-dat t₁]] V₁]]
 (6b) * [_{VP} NP-nom [_V [_{VP} NP-dat [_V NP-acc t₁]] V₁]]

As (4b) shows, the R-expression cannot be realized by a dative if the anaphor is realized by an accusative. This indicates that in German no grammatical configuration exists in which the dative object c-commands the accusative object (6b). The only way to deal with this grammaticality contrast is to assume that (6b) cannot be base-generated by the syntactic system of German (as opposed to other Germanic languages such as English). As the flattened out version of (6b) would be {NP-nom, NP-dat, NP-acc, V}, it is clear that this word order must be derived somehow and cannot reflect the base-generated word order. This comes surprising at first glance, since this serialization corresponds to the unmarked order with most verbs, but it is not unusual to have unmarked word orders that are structurally complex in that they obligatorily involve movement. The well-known clustering of pronouns to the left edge of the middle-field (Wackernagel-position) is clearly a case in which the pronominal dative and/or accusative object are moved over the non-pronominal subject (7a, b). This is most evident in cases like (7c) where the accusative reflexive is bound by the subject.

- (7a) ... *weil ihr Markus das Buch geliehen hat.*
 because her_{DAT} Markus_{NOM} [the book]_{ACC} lent has
 ‘because Markus lent her the book’
- (7b) ... *weil ihn ihr Markus vorgestellt hat.*
 because him_{ACC} her_{DAT} Markus_{NOM} introduced has
 ‘because Markus introduced him to her’

- (7c) ... weil sich_i ihr Markus_i vorgestellt hat.
 because himself_{.ACC} her_{.DAT} Markus_{.NOM} introduced has
 ‘because Markus introduced himself to her’

Another piece of evidence that Acc>Dat is the base order is that it is the unmarked order in the pronominal complex, as multiple movement should be structure-preserving. So nothing hinders us to assume that the base-generated order is Acc>Dat, as is suggested by the historical data.

This argument hinges crucially on the reliability of Binding Theory in such contexts. Rothmayr (2006) tried to show that this is not a case in which Binding Theory can apply. Her main arguments are the following:

- α) Dative plurals are no real plurals.
- β) Reciprocals are semantically complex and thus need not be subject to Binding Theory.
- γ) Picture nouns offer direct counterevidence against the binding argument.

Ad α): This argument rests partly on problematic premises, on which I will not elaborate. One piece of evidence is that dative objects fail to agree with the predicate in passivization, contrary to accusative objects (8; c, e adapted from Rothmayr 2006: 207).

- (8a) *Laura unterstützt die Kinder.*
 Laura supports [the children]_{.ACC}
 ‘Laura supports the children.’
- (8b) *Die Kinder werden unterstützt.*
 [the children]_{.NOM} become_{.PL} supported
 ‘The children are supported.’
- (8c) *Irmi hilft den Kindern.*
 Irmi helps[the children]_{.DAT}
 ‘Irmi helps the children.’
- (8d) * *Den Kindern werden geholfen.*
 [the children]_{.DAT} become_{.PL} helped
 ‘The children are helped.’

- (8e) *Den Kindern wird geholfen.*
 [the children]_{DAT} become_{SG} helped
 ‘The children are helped.’

Note, however, that we would not expect agreement in the first place as the dative object fails to be promoted to subject position and to receive nominative case, contrary to the accusative object in (8a, b). Nominative case seems to be a prerequisite for agreement. If we believe in an independent IP-architecture also in German, this would follow easily from movement of the subject to SpecIP (or the highest I-projection such as TP – the internal structure of IP does not play a role for the argument) where it both agrees with the verb by c-commanding it and receives nominative case. Datives never agree with the verb – but crucially accusatives do not either (9b). So this argument is not relevant, as we have no asymmetry between dative and accusative.

- (9a) *Mir/ Uns (ist/*bin/*sind) kalt.*
 Me_{DAT} Us_{DAT} is am are cold
 ‘I/we am/are cold’
- (9b) *Mich/ Uns (friert/ *friert/ *frieren).*
 Me_{ACC} Us_{ACC} freeze_{3.SG} freeze_{1.SG} freeze_{1.PL}
 ‘I/we am/are freezing’

The question is rather, why datives cannot be promoted to subject position. If we assume, contra Meinunger (2007), that dative is an inherent case in two-place predicates such as *helfen* ‘help’, whereas accusative is a structural case, the dative object, already being equipped with case, could not be moved to a position where it receives case, such as the subject position.¹¹

Ad β): This argument rests on the premise that the binding effect is to be seen only with reciprocals. The reason why the usual examples (such as (4)) exhibit reciprocals is simply that it is much easier to come up with examples in which a pluralic referent is used for direct and indirect object. That does not mean that it is impossible to find examples in which the common referent of the direct and indirect object is a singularic entity which is referred to by an ordinary reflexive. If we take a psychotherapy-context, for instance, and assume a case in which a patient is traumatized, part of the therapy is to make the patient go through the traumatic experience again mentally. In such a case it is

¹¹ Following a model that allows for IP in German, such as Suchsland (1988) or Sabel (2000).

possible to say that the patient is subjected to himself. The German version of this sentence is (10a), which sounds normal (if you accept the scenario) and in which the accusative binds the dative, whereas (10b) with the dative binding the accusative is deviant.¹² Note that it is not a matter of the unmarked word order acc>dat with verbs like *aussetzen* ‘subject s.o. to s.th.’. To demonstrate this, let us stay in the same context: The psychotherapist succeeds with his treatment, the patient is healed, and a friend of the patient congratulates the psychotherapist in saying that she gave the patient back to himself. The German wording (10c) involves again an accusative binding a dative, whereas the opposite case (10d) sounds ungrammatical. The unmarked word order with words like (*wieder*)*geben* ‘give (back)’ is not acc>dat.

- (10a) *Die Psychotherapeutin hat den Patienten_i sich_i ausgesetzt.*
 the psycho-therapist has [the patient]_{ACC} himself subjected
 ‘The psycho-therapist subjected the patient to himself’
- (10b) * *Die Psychotherapeutin hat dem Patienten_i sich_i ausgesetzt.*
 the psycho-therapist has [the patient]_{DAT} himself subjected
 ‘The psycho-therapist subjected the patient to himself’
- (10c) *Die Psychotherapeutin hat den Patienten_i sich_i wiedergegeben.*
 the psycho-therapist has [the patient]_{ACC} himself given-back
 ‘The psycho-therapist gave the patient back to himself’
- (10d) * *Die Psychotherapeutin hat dem Patienten_i sich_i wiedergegeben.*
 the psycho-therapist has [the patient]_{DAT} himself given-back
 ‘The psycho-therapist gave the patient back to himself’

So we receive the same contrast as in (4) with respect to binding in cases that do not involve reciprocals but ordinary reflexives. Consequently, the binding facts are real and cannot be dismissed with reference to the weirdness of reciprocals or problems specific to pluralic entities.

¹² The more natural way of expressing this would be with ‘sich selbst’, but without *selbst* it is marginally possible as well. The grammaticality contrast is the same with *selbst*.

Ad γ): This is Rothmayr's strongest argument. However, it is well-known that binding in picture nouns is a general problem of binding theory rather than a special problem of binding between direct and indirect object (see, e.g., Pollard/Sag 1992).¹³

That the problem is on more general lines can be demonstrated easily. In Rothmayr's example (31 on page 210), here slightly adapted in order to avoid the reciprocal (11), it seems indeed as if the dative object binds a reflexive pronoun inside the accusative object.

- (11) *weil Irmi [dem Kind]_i [ein Foto von sich_i] zeigte.*
 because Irmi [the child]_{DAT} [a photograph of h.-self]_{ACC} showed
 'because Irmi showed the child_i a picture of h.-self_i'

The problem is that we get similar effects also in configurations like (12), where an accusative object binds a reflexive inside the subject. Here it is obvious that the subject is generated higher than the object, yet we obtain the same binding effect.

- (12) *Es ist klar, dass [ein Bild von sich_i in der Zeitung]*
 It is obvious that [a picture of himself in the paper]_{NOM}
auch den Jörg_i beeindruckt.
 even [the Jörg]_{ACC} impresses
 'It is obvious that a picture of himself in the newspaper impresses even Jörg.'

Thus it is evident that the binding peculiarities of picture nouns do not offer evidence for structural dominance.

In sum, Rothmayr's arguments against the validity of the binding argument are in itself questionable so that, in my opinion, nothing speaks against assuming Acc>Dat as the base structure. This fits nicely with the historical data where it can be argued on independent grounds that Acc>Dat behaves as if it was the base-generated order, whereas Dat>Acc can be derived by other factors.

¹³ There have been several attempts to reconcile picture noun reflexives with Binding Theory (e.g. Pollard/Sag 1992, Reinhart/Reuland 1993) as well as psycholinguistic research on the matter (e.g. Runner/Sussman/Tanenhaus 2005), but a discussion would lead too far apart from the purposes of this paper.

So the order Dat>Acc must involve some sort of movement operation. It is clear that, if it is a sort of scrambling, it must be VP-internal scrambling. This can be demonstrated easily with examples that involve an adverbial that marks the boundary of VP, such as *gerne* ‘willingly’ (cf. Jackendoff 1972: 59ff.; Frey/Pittner 1998). If we assume that the subject is moved to SpecIP also in German (e.g. Sabel 2000), it is outside the VP; consequently the normal position of adverbs like *gerne* must be after the subject. This is shown to be correct in (13), where (13a) gives the version with Acc>Dat-order, while (13b) gives the (more natural sounding) Dat>Acc-version. The positioning of the adverb in front of the subject is definitely less acceptable, and it is not a felicitous answer to a question involving wide focus (13c), whereas (13b) would be a felicitous answer to the question in (13c).

(13a) ^(?) *dass* [_{IP} Jörg₂ [_{VP} *gerne* [_{VP} *t*₂ *sein neues Auto*
that Jörg_{NOM} willingly [his new car]_{ACC}
*seiner Freundin t*₁]] *zeigt*]₁
[his girl-friend]_{DAT} shows
‘that Jörg likes to show his new car to his girl-friend’

(13b) *dass* [_{IP} Jörg₂ [_{VP} *gerne* [_{VP} *t*₂ *seiner Freundin*₃
that Jörg_{NOM} willingly [his girl-friend]_{DAT}
*sein neues Auto t*₃ *t*₁]] *zeigt*]₁
[his new car]_{ACC} shows
‘that Jörg likes to show his new car to his girl-friend’

(13c) (Was siehst du?)
What see you
#/* *Ich sehe, dass gerne Jörg seiner Freundin*
I see that willingly Jörg_{NOM} his girl-friend_{DAT}
sein neues Auto zeigt.
his new car_{ACC} shows
‘What do you see? – I see that Jörg likes to show his new car to his girl-friend.’

We may assume that the adverb is adjoined outside of VP. This is demonstrated in (14). If the VP is moved to the prefield, adverbials like *gerne* must be stranded (14b), indicating that they are not part of the maximal projection of the moved verb *zeigen* ‘show’.

- (14a) ^(?) *Sein neues Auto der Laura zeigen wird Jörg gerne.*
 [his new car]_{.ACC} [the Laura]_{.DAT} show will Jörg willingly
 ‘Jörg will like to show Laura his new car.’
- (14b) *Der Laura sein neues Auto zeigen wird Jörg gerne.*
 [the Laura]_{.DAT} [his new car]_{.ACC} show will Jörg willingly
 ‘Jörg will like to show Laura his new car.’
- (14c) * *Gerne der Laura sein neues Auto zeigen wird Jörg.*
 willingly [the Laura]_{.DAT} [his new car]_{.ACC} show will Jörg
 ‘Jörg will like to show Laura his new car.’

Note that even under fronting of the VP the markedness (in Höhle’s sense) of the Acc>Dat-option remains, in the sense that Acc>Dat requires contrastive focus (14a, b). This indicates that the generation of the Dat>Acc-order does not involve adjunction outside VP, as then remnant movement of the VP should only be possible in the base-generated Acc>Dat-option. If an outside-adjoined adverbial like *gerne* must be stranded under VP-fronting, an outside-adjoined element like the dative object should have to be stranded as well if the order Dat>Acc were the result of adjunction of the dative outside the VP.¹⁴

It is not possible to find conclusive evidence that the dative object really has moved in cases of stylistically unmarked word order. A classical argument would derive from the Freezing Principle (Wexler/Culicover 1980, Müller 1998) which basically says that a moved constituent becomes an island, i.e., it is not possible to move material out of an already moved constituent. This test is not applicable to gather evidence for movement of the dative indirect object, as dative NPs functioning as indirect objects are islands anyway, no matter where they stand (15).¹⁵

¹⁴ Note that we are not talking here about scrambling to a position between subject and VP-boundary adverbials of the type *dass Hans das Buch gerne gelesen hat*. Chocano (2007: 14) suggests that the analysis of such scrambling with scrambling to a position higher than the subject is essentially the same, which seems to be correct. This does not entail that VP-internal scrambling in the sense used here can be subsumed under this account, too.

¹⁵ The effect is visible also in cases in which the dative is the only object: Whereas (i) is marginally acceptable, (ii) is definitely not. The sentence (iii), with the same meaning as (ii) but an accusative instead of the dative object, is acceptable.

(i) ? [Von wem]_i hat Peter [der Frau t_i] geholfen?
 of whom has Peter [the wife]_{.DAT} helped
 ‘The wife of whom did Peter help?’

(15a) * [Über wen]₂ verdankte Max [einer Studie t₂]₁
 on whom owes Max_{NOM} [a study]_{DAT}
 sein hohes Ansehen t₁?
 [his high reputation]_{ACC}
 ‘A study on whom helped Max gain his good reputation?’

(15b) * [Über wen]₁ verdankte Max sein hohes Ansehen
 on whom owes Max_{NOM} [his high reputation]_{ACC}
 [einer Studie t₁]?
 [a study]_{DAT}
 ‘A study on whom helped Max gain his good reputation?’

It can, however, be used as further evidence that the order Acc>Dat cannot be the result of a movement operation. As direct objects are not islands *per se*, we should get a Freezing effect in Acc>Dat orders if they were derived from Dat>Acc orders by movement of the accusative object. However, as (16) shows, there is no noticeable grammaticality difference between the Dat>Acc and the Acc>Dat version (apart from the general markedness of Acc>Dat orders).¹⁶

(16a) [Über wen]₁ hat der Max der Laura [ein Buch t₁]
 on whom has [the Max]_{NOM} [the Laura]_{DAT} [a book]_{ACC}
 geschenkt?
 given
 ‘On whom did Max give a book to Laura as a present?’

(16b) [Über wen]₁ hat der Max [ein Buch t₁] der Laura
 on whom has [the Max]_{NOM} [a book]_{ACC} [the Laura]_{DAT}
 geschenkt?
 given
 ‘On whom did Max give a book to Laura as a present?’

(ii) * [Gegen was]₁ hat Peter [einer Petition t₁] geholfen?
 against what has Peter [a petition]_{DAT} helped
 ‘Against what did Peter support a petition?’

(iii) [Gegen was]₁ hat Peter eine Petition t₁] unterstützt?
 against what has Peter [a petition]_{ACC} supported
 ‘Against what did Peter support a petition?’

¹⁶ Cf. Chocano (2007: 86ff.) on the general problems, with evidence from Freezing.

Still, one could say that the effect shown above (4, 5) that the dative noun phrase cannot bind an accusative anaphor – which was taken here as evidence that the accusative is necessarily structurally higher than the dative – might hinge on the particular verb *vorstellen* ‘to introduce’. But if other verbs are substituted for it, and the reciprocal changed to a simple reflexive, it is not as easy any more to come up with a naturally sounding scenario. However, if one succeeds, one sees that the same grammaticality judgments prevail (17, 18).

(17a) *Der Vater hat [die Eheleute]_i {sich/ einander}_i versprochen.*
 the father has [the spouses]_{.ACC} themselves/each other promised

(17b) **Der Vater hat [den Eheleuten]_i {sich/ einander}_i versprochen.*
 the father has [the spouses]_{.DAT} themselves/each other promised
 ‘The father has promised the spouses to each other’

(18a) *Jörg_j hat [den Mann]_i sich_{i,j} im Spiegel gezeigt.*
 Jörg has [the man]_{.ACC} himself in-the mirror shown

(18b) *Jörg_j hat [dem Mann]_i sich_{i,j} im Spiegel gezeigt.*
 Jörg has [the man]_{.DAT} himself in-the mirror shown
 ‘Jörg showed the man to himself in the mirror.’

5. Concluding remarks

The older stages of German (that is, High German until about 1500) show a relatively rigid object order with the dative before the accusative object. This effect could be traced back to the Thematic Hierarchy: German surface word order originally followed closely the Thematic Hierarchy, and as the case linking is directly dependent on the Thematic Hierarchy, we get the impression of a fixed argument order ‘nominative > dative > accusative’. We also get the impression of a serialization according to animacy, simply because animacy is a typical property of Proto-Agents; consequently, the arguments more to the ‘left’ (the agentive pole of the scale) are typically animated whereas the ones on the ‘right side’ (the patient-like pole of the scale) are not necessarily animated. This situation can easily lead to a re-interpretation of the factor governing serialization, in that the Thematic Hierarchy is not recognized any more as the crucial factor by language learners and instead, derivative factors such as case and animacy are interpreted as the decisive factors. This obvi-

ously happened at some stage in the history of German, at latest at the point at which accusative-dative verbs such as *aussetzen* ‘subject s.o. to s.th.’ begin to be used. Grimm’s dictionary gives no dates of first attestation of the accusative-dative verbs before around 1500 (see Speyer 2011). Note that this is the same time in which information structural factors begin to play a role for serialization, so we can be fairly certain to point the change away from the Thematic Hierarchy as the decisive factor for serialization around 1500. The change concerned only the surface word order. The underlying word order, at least in Modern German, is accusative before dative (this is suggested by Binding facts, the lack of Freezing effects) and probably was so throughout the history of the language (this is suggested by the distribution patterns of the data). One might ask why the underlying order never changed, even in the period in which the surface order was strictly ‘dative before accusative’, at least for lexical noun phrases, and thus the evidence for the language learner apparently overwhelming against Acc>Dat as the underlying order. The answer is probably that with pronoun NPs, the order typically is Acc>Dat, and since pronominal reference is much more common, especially in spoken language (which is the relevant register here), the evidence for Dat>Acc was in fact not that overwhelming but represented a minority pattern. Note that the ordering according to the Thematic Hierarchy concerns only lexical NPs, not pronouns. So the language learner, confronted with paradoxical evidence (pronouns Acc>Dat, lexical NPs Dat>Acc), can identify a non-grammatical factor for Dat>Acc, but not so for Acc>Dat; (s)he will settle for Acc>Dat as the underlying order from which Dat>Acc is derived if lexical NPs are involved.

References

Sources and aids¹⁷

- Berthold von Regensburg. Vollständige Ausgabe seiner Predigten mit Anmerkungen von Franz Pfeiffer. Mit einem Vorwort von Kurt Ruh. Bd. 1. Berlin: de Gruyter 1965 [Reprint; original: Wien: Braumüller 1862].
- Besch, Werner/Lenders, Winfried/Moser, Hugo/Stopp, Hugo (eds.) (1972-1985): Das Bonner Frühneuhochdeutschkorpus. URL: <http://korpora.org/Fnhhd/>.
- Gippert, Jost/Martinez, Javier/Korn, Agnes (eds.) (1987ff.): Thesaurus indogermanischer Text- und Sprachmaterialien (TITUS). URL: <http://titus.uni-frankfurt.de/indexe.htm>.

¹⁷ All URLs have been checked and found valid as of late January 2015.

- Greule, Albrecht (1999): Syntaktisches Verbwörterbuch zu den althochdeutschen Texten des 9. Jahrhunderts. Frankfurt a.M.: Peter Lang.
- Lancelot. Nach der Heidelberger Pergamenthandschrift Pal. germ. 147 hg. v. Reinhold Kluge. Bd. I. Berlin: Akademie Verlag 1948.
- Springeth, Margarete/Schmidt, Klaus M./Pütz, Horst P. (eds.) (1992): Mittelhochdeutsche Begriffsdatenbank. URL: www.mhdbdb.sbg.ac.at:8000/.

Research literature

- Chocano, Gema (2007): Narrow syntax and phonological form. Amsterdam/Philadelphia: John Benjamins.
- Chomsky, Noam (1981): Lectures on government and binding. Dordrecht: Foris.
- Dowty, David R. (1991): Thematic proto-roles and argument selection. In: *Language* 67: 547-619.
- Fortmann, Christian/ Frey, Werner (1997): Konzeptuelle Struktur und Grundabfolge der Argumente. In: d'Avis, Franz-Josef/Lutz, Uli (eds.): *Zur Satzstruktur im Deutschen*. (= Arbeitspapiere des SFB340 90). Stuttgart/Tübingen: Univ. Stuttgart/Univ. Tübingen, 143-170.
- Frey, Werner/Pittner, Karin (1998): Zur Positionierung der Adverbiale im deutschen Mittelfeld. In: *Linguistische Berichte* 176: 489-534.
- Hoberg, Ursula (1997): Die Linearstruktur des Satzes. In: Zifonun, Gisela/Hoffmann, Ludger/Strecker, Bruno: *Grammatik der deutschen Sprache*. Band 2. (= Schriften des Instituts für Deutsche Sprache 7.2) Berlin: de Gruyter, 1496-1680.
- Höhle, Tilman (1982): Explikationen für 'normale Betonung' und 'normale Wortstellung'. In: Abraham, Werner (ed.): *Satzglieder im Deutschen*. Tübingen: Narr, 75-153.
- Jackendoff, Ray S. (1972): *Semantic interpretation in generative grammar*. Cambridge, MA: MIT Press.
- Lenerz, Jürgen (1977): *Zur Abfolge nominaler Satzglieder im Deutschen*. Tübingen: Narr.
- Lenerz, Jürgen (2002): Scrambling and reference in German. In: Abraham, Werner/Zwart, C. Jan-Wouter (eds.): *Issues in formal German(ic) typology*. Amsterdam/Philadelphia: John Benjamins, 179-192.
- Lötscher, Andreas (1981): Abfolgeregeln für Ergänzungen im Mittelfeld. In: *Deutsche Sprache* 9: 44-60.
- Meinunger, André (2007): Der Dativ im Deutschen – eine Verständnishilfe für das Phänomen der gespaltenen Ergativität. In: *Linguistische Berichte* 209: 3-31.

- Müller, Gereon (1998): Incomplete category fronting. A derivational approach to remnant movement in German. Dordrecht: Kluwer.
- Müller, Gereon (1999): Optimality, markedness, and word order in German. In: *Linguistics* 37: 777-818.
- Musan, Renate (2002): Informationsstrukturelle Dimensionen im Deutschen. Zur Variation der Wortstellung im Mittelfeld. In: *Zeitschrift für germanistische Linguistik* 30: 198-221.
- Pollard, Carl/Sag, Ivan A. (1992): Anaphors in English and the scope of binding theory. In: *Linguistic Inquiry* 23: 261-303.
- Primus, Beatrice (2002a): Protorollen und Verbtyp: Kasusvariation bei psychischen Verben. In: Hummel, Martin/Kailuweit, Rolf (eds.): *Semantische Rollen*. Tübingen: Narr, 377-401.
- Primus, Beatrice (2002b): Proto-roles and case selection in optimality theory. In: *Theorie des Lexikons. Arbeiten des SFB 282*, 122: 1-39.
- Primus, Beatrice (2004): Division of labour: The role-semantic function of basic order and case. In: Willems, Dominique et al. (eds.): *Contrastive analysis in language: Identifying linguistic units of comparison*. Palgrave: Macmillan, 89-136.
- Primus, Beatrice (2011): Animacy, generalized semantic roles, and different object marking. In: Lamers, Monique/de Swart, Peter (eds.): *Case, word order, and prominence. Interacting cues in language production and comprehension*. Dordrecht: Springer, 65-90.
- Rappaport, Malka/Levin, Beth (1988): What to do with theta-roles. In: Wilkins, Wendy (ed.): *Thematic relations. (= Syntax and Semantics 21)*. New York: Academic Press, 7-36.
- Reinhart, Tanya/Reuland, Eric (1993): Reflexivity. In: *Linguistic Inquiry* 24: 657-720.
- Reis, Marga (1987): Die Stellung der Verbargumente im Deutschen – Stilübungen zum Grammatik:Pragmatik-Verhältnis. In: *Lunder germanistische Forschungen* 55: 139-177.
- Rothmayr, Antonia (2006): The binding paradox in German double object constructions. In: *Linguistische Berichte* 206: 195-215.
- Runner, Jeffrey T./Sussman, Rachel S./Tanenhaus, Michael K. (2005): Reflexives and pronouns in picture noun phrases: using eye movements as a source of linguistic evidence. In: Kepsers, Stephan/Reis, Marga (eds.): *Linguistic evidence. Empirical, theoretical and computational perspectives*. Berlin/New York: Mouton de Gruyter, 393-411.
- Sabel, Joachim (2000): Das Verbstellungsproblem im Deutschen: Synchronie und Diachronie. In: *Deutsche Sprache* 28: 74-99.

- Speyer, Augustin (2011): Die Freiheit der Mittelfeldabfolge im Deutschen – ein modernes Phänomen. In: Beiträge zur Geschichte der deutschen Sprache und Literatur 133: 14-31.
- Speyer, Augustin (2013): Performative Mündlichkeitsnähe als Faktor für die Objektstellung im Mittel- und Frühneuhochdeutschen. In: Beiträge zur Geschichte der deutschen Sprache und Literatur 135(3): 1-36.
- Suchsland, Peter (1988): Zur Interaktion von Morphologie und Syntax. In: Deutsch als Fremdsprache 25: 321-327.
- Webelhuth, Gert (1992): Principles and parameters of syntactic saturation. Oxford/New York: Oxford University Press.
- Wexler, Kenneth/Culicover, Peter W. (1980): Formal principles of language acquisition. Cambridge, MA: MIT Press.
- Zubin, David A./Köpcke, Klaus-M. (1985): Cognitive constraints on the order of subject and object in German. In: Studies in Language 9: 77-107.

The properties of the Middle High German “*Nachfeld*”

Syntax, information structure, and linkage in discourse

Abstract

There are several studies on the functions of the *Nachfeld* in Old High German and Present Day German, but no surveys have yet been conducted on Middle High German (and Early New High German). This paper presents some preliminary results from the Middle High German period, but is part of a larger corpus study also covering Early New High German texts. Based on an annotated corpus of Middle High German texts, it will be shown that syntactic, information and discourse-structural factors have a different incidence on the filling of the *Nachfeld*. It will be argued that the *Nachfeld* in Middle High German is characterized by conservative features but already displays the typical traits that can be observed in its modern use. Thus, on the one hand, focus still plays a fundamental role in occupying the *Nachfeld*, as was observed for Old High German. On the other hand, the Middle High German *Nachfeld* is already used as a means to establish, change or reactivate a discourse topic, as is often claimed for Present Day German.

1. Introduction¹

The object of our investigation is the German *Nachfeld* (‘final field’, henceforth NF) from a diachronic perspective. According to the topological *Feldertheorie* (‘field theory’), the NF is the position after the second element of the so-called *Satzklammer* (‘verbal’ or ‘sentence bracket’), which consists of the non-finite verbal elements. This syntactic space is particularly interesting because modern German is notoriously a verb-final language. Therefore, no constituents are expected to occur in the postverbal position. However, many exceptions to the verb-final restriction can be observed in both spoken and written language, as in the following example:

¹ This paper is part of the research conducted by the project B4 “The role of information structure in language change” (SFB 632 “Information structure: The linguistic means for structuring utterances, sentences and texts”). Special thanks go to Anke Gehrlein and Oxana Rasskazova for collecting and annotating the data, and to Andrew Murphy for reading the final proof. We would like to point out that this paper presents the provisional data that were available at the time of writing. A more recent survey based on a complete and larger corpus of texts is forthcoming.

- (1) *Diese Antwort muss sich einfügen in die künftige Architektur*
 This answer must fit in the future architecture
Europas. (H.-D. Genscher 06.040.1090, cited in Vinckel 2006: 2f.)
 of Europe.

‘This answer must fit in the future architecture of Europe.’

Obviously, factors such as the length of the postponed constituent and its informational value play a role in this process, which results in a clarification of sentence structure (Engel 1977, Hoberg 1981) and/or a better distribution of information (Zifonun et al. 1997: 1669). Thus, the speaker can use this position to deliver chunks of information at a later stage as “afterthoughts” (cf. Engel 1977, Zifonun et al. 1997) or to employ it in order to “highlight” the relevant information. Hoberg (1981) especially focuses on the rhematic character of the NF in the written standard variety. The same observation holds for different oral varieties. In this regard, Filpus (1994: 259) reports that rhematic information is much more often postponed (49%) than thematic information (11%).

At the same time, the strategy of postponing constituents in the NF seems to correlate with certain genres. It occurs less in spontaneous conversation than in planned speech, such as scientific lectures (Filpus 1994: 138) or political addresses (Vinckel 2006). In these genres, it is used for the focussing of the postverbal constituents, which are taken as the thematic part of the following sentence. Thus, the NF also seems to fulfill an important role for the thematic progression of the text (ibid., 208). More recently, Vinckel-Roisin (2011) shows that also in journalistic texts, filling the NF serves to mark a referent as the topic of the subsequent sentence. She argues that the NF position is relevant to either establish, change or reactivate a discourse topic to which the following assertions are anchored. This property is related to the position of the sentence within the article or within the section. For example, sentences with a filled NF at the beginning of an article or of a section mainly serve to introduce new discourse topics (cited in Vinckel-Roisin 2011: 390f.):

- (2) [...] *Man mag schon fast Mitleid haben mit Adolf Sauerland. Er ist ein Bild des Jammers. Der Oberbürgermeister der Stadt Duisburg muss sich am Tag der Trauerfeier verstecken. Er hat Angst vor seinen Bürgern, [...]* (sueddeutsche.de, July 30, 2010)

‘[...] You could almost feel sorry **for Adolf Sauerland**. He is a real pathetic sight. The Lord Mayor of Duisburg had to hide himself on the day of the memorial service. He was scared of his own citizens, [...]’

It is important to note that immediately after the occurrence in the NF, the referential expression *Adolf Sauerland* is referred to by the subject pronoun of the following sentence. Apparently, it is the proximity between antecedent and anaphora that is responsible for guaranteeing the coherence. Equally, most of the other subsequent sentences start with a subject referring to the same discourse referent. Thus, the NF can be considered a marked position in order to establish the discourse topic of the subsequent section.

2. The *Nachfeld* from a diachronic perspective

Occupying the NF in modern German is subject to a number of syntactic, information-structural, and discourse-related restrictions (Altmann 1981, Averintseva-Klisch 2009). However, an investigation into earlier stages of the language reveals that the kinds of restrictions imposed on constituents in the NF sometimes seems to be partially different from modern ones. In particular, earlier stages of the language clearly show that the NF was more open to host extraposed constituents. From a diachronic point of view, the question arises as to how the properties and functions of constituents that could be placed in the NF have changed.

It is well known that in New High German (NHG), mainly PPs and the second elements of comparative constructions are subject to extraposition (Zifonun et al. 1997: 1651). The situation is entirely different when we look at the first written records in Old High German (OHG). In the postverbal position, we find syntactic constituents such as objects, predicates and even pronouns which cannot be extraposed in NHG (ibid., 1651). This evidence is used as an argument in order to claim that the grammar of OHG still shows variation between OV and VO structure (Hinterhölzl 2004, Petrova 2009, Schlachter 2009, Haider 2010). At the same time, this variation is not random, but rather conditioned by information-structural factors. Thus, the VO order serves to mark postverbal constituents as focussed, as is often assumed for cases such as the following (Hinterhölzl 2004: 154, Petrova/Hinterhölzl 2010: 206):

- (3) *Inti thi thár habetun diuual* (T 59, 1)
 and those PRT had devil
 ‘And those who were possessed by the devil’
et qui demonia habebant

The question how this postverbal focus position was lost still remains unanswered. Until now, we only have an idea of the quantitative development of the NF in general: In Early OHG, the NF is filled by some constituent in about 40% to 50% of all subordinate clauses (Weiß to appear, Petrova 2009: 253). In Late OHG texts, the percentage decreases to 30%. Interestingly, this percentage grows again to 40% in Early Middle High German (MHG), but decreases to 35% in Late MHG (Prell 2003: 247) and drops to 20% during the Early New High German (ENHG) period (Schildt 1976). Even if we disregard the question of the underlying grammar, we need to know in more detail what kind of syntactic constituents occur in the NF.

Moreover, the function of the NF for building text coherence still requires further investigation. Until now, there have been hardly any studies on how, or indeed if, constituents in the NF are resumed in the following discourse. The very few investigations (Margetts 1969, Sigroth-Nellessen 1979) all lack the use of modern linguistic concepts.

The goal of our survey is to investigate the change in syntactic functions, information-structural properties, and discourse-related functions of the NF during the MHG period. The issues we want to investigate are the following:

- What kind of change in the grammatical functions of NF constituents can we observe? For instance, what are the conditions for extraposing accusative objects? When do extraposed objects disappear?
- Does information structure still have such a dominant influence on syntax as is assumed for OHG? Which information-structural categories are relevant for the diachronic development of the NF?
- Which are the discourse functions of extraposed constituents? Are they used to establish, change or reactivate a discourse topic to which the following assertions are anchored (cf. Vinckel-Roisin 2011)? What kinds of expressions are used for the anaphoric resumption of NF constituents?

3. Corpus and methodology

Our investigation mainly deals with MHG, the preliminary results of which are presented below. This period is particularly interesting because we can observe and describe the change of certain OHG patterns (for instance the development of the OV pattern) and compare the results of this investigation with what is claimed for NHG. Furthermore, for these periods there are no specific studies covering the three aspects of our investigation, namely syntax, information structure, and linkage in discourse.

For the present study, we created a corpus (with EXMARaLDA² Partitur-Editor 1.5) that, at the time of writing, includes texts covering MHG for which no annotated corpora are yet available.

One problem underlying all the aforementioned studies is that they are based on different text types. Therefore, for our investigation, we decided to keep the genre constant and to limit our corpus to sermons, speeches, and other text types that, beyond their appellative function, display a more general argumentative function. This has the clear advantage that they can be compared with NHG speeches (Vinckel 2006) and also with the OHG treatise Isidor (Schlachter 2009, 2010), which also show argumentative functions.

Moreover, the dialect of the texts plays an important role. Since there are not enough documents covering all dialect areas, the selection of texts was restricted to the Upper German language area.

The corpus used consists of the following MHG texts, which were chosen because each represents more or less one century. We excerpted passages of a total of around 12,000 words for each text.

1. *Speculum Ecclesiae* (1170-1200) [Speccl];
2. *Schwarzwälder Predigten* (end of the 13th century, according to Schiewers 2008: 6) [SchwP];
3. *Nikolaus von Strassburg* (first half of the 14th century) [NikP].

For our corpus, we decided to annotate the NF of main clauses only. The reason for this is that we still do not know much about the information structural properties of main clauses in older stages of the language. Previous work on OHG mainly focuses on subordinate clauses since only these are usually un-

² Available online: www.exmaralda.org. Cf. Schmidt/Wörner (2009).

ambiguous with respect to the presence of a full sentence bracket. In contrast, OHG main sentences may be ambiguous structures, given that they often display only a main verb in second position and thus provide no hints about the presence of NF material. However, since the number of sentences with periphrastic verbal forms increases in MHG, we obtain more relevant examples with full bracket structures.

In the following, we want to show in more detail how we treated sentences with a NF in our corpus with respect to the three different aspects: syntax, information structure, and discourse functions. Let us consider the following sentence and its annotation in Table 1:

- (4) *Der lerer der sol siniu schâf och retten vor dem beren.*
 the teacher this shall his sheep also save from the bear
 (SchwP 10, 9)
 ‘The teacher, he shall also save his sheep from the bear.’

	1841	1842	1843	1844	1845	1846	1847	1848	1849	1850	1851	1852	1853
[TEXT]	/be/	.	Der	lerer	d'	Sol	Siniu	schâf	och	rette-	vor	de-	bere-
[POS]	\$	DDA	NA	DDS	VMFIN	DPOS	NA	KON	VVINF	APPR	DDA	NA	\$
[CAT]			NP		NP	VP			VP		PP		
[SATZSTATUS]			MAINDECL										
[GF]			VOVF		SUBJ	VFIN	DO			INF	PP-OBJ		
[SATZKLAMMER]					LI				RE				
[VERLINKUNG]										UNLINK			
[WIEDERAUFNAHME]												KOREF1	
[GEGEBENHEIT]												GIV-INACTIVE	
[FOKUS]												NF	
[EDITION]		010,09; -lôben											
[HANDSCHRIFT]									007v,10				
[KOMMENTAR]		left dislocation											

Table 1: An example for the annotation of a sentence in EXMARaLDA

As shown in Table 1, the most important annotation layers are the following:

- Tier 5: Grammatical function [GF], where the syntactic function of each constituent in the sentence is annotated.
- Tier 7: [VERLINKUNG] (‘linking’). In combination with tier 5, this layer allows us to determine whether the extraposed XP forms a constituent of its

own (thus being unlinked) or if it is linked to some constituent in the middle field as part of a coordination or modification relation.

- Tier 8: [WIEDERAUFNAHME] (‘resumption’), where we annotate if the NF constituent is referred to in the following three sentences. If this is the case, the constituent in the NF is marked as KOREF 1 (‘coref(erential) 1’), while the anaphoric expressions in the subsequent sentences are marked as KOREF 2.
- Tier 9 and tier 10: [GEGEBENHEIT] (‘givenness’) and [FOKUS] (‘focus’), respectively, where we annotate the information status of the constituent in the NF and the possible presence of focus.

4. Preliminary results

First of all, our investigation tries to assess and quantify the changes in the syntactic categories of extraposed elements in the three MHG texts. Our results show a strong increase in the use of PPs in the Schwarzwälder Predigten and in Nikolaus. Another fundamental aspect is the clear decline in the use of subjects in NF position.

GF	Speculum	SchwPredigten	Nikolaus
subjects	21	9	0
objects	20	10	21
PPs	37	64	55
others	22	17	24
total	100	100	100

Table 2: Grammatical functions of NF constituents (percentages)

With respect to extraposed objects, we do not observe a quantitative change, but rather a qualitative one. In earlier MHG texts, such as *Speculum Ecclesiae*, their extraposition is still amply used even when the elements in the NF are not linked to the following discourse:

- (5) *Do wolte er in aller erste enbieten sine kvnft.* (Speccl 6, 9)
 then wanted he at first announce his coming
 ‘Then, he wanted to first announce his arrival.’

In later MHG texts (such as *Nikolaus*), objects are still attested in NF positions. But this is often due to their specific properties (such as heaviness) and links to the following discourse. Thus, for example, objects occur in the NF when they are modified by a following relative clause.

A second important issue of our survey is the determination of the information-structural status of NF constituents. Consider example (4) again and its annotation in Table 1. The postponed PP *vor dem beren* is annotated as GIV-INACTIVE. This means that the referent is already present in the preceding context, but is not referred to in the last sentence. It is explicitly reactivated in this utterance.

What we are interested in is the cognitive status of the referential expression used in the NF (cf. Gundel/Hedberg/Zacharski 1993). Table 3 provides some partial results of our investigation with respect to the information-structural status of postponed elements in the three MHG texts in our corpus:³

	Speculum	SchwPredigten	Nikolaus
GIVEN	16	33	17
ACCESSIBLE	46	30	44
NEW	38	37	39
total	100	100	100

Table 3: Information-structural statuses of NF constituents (percentages)

Table 3 shows no clear change in the information-structural status of NF constituents. Even though the number of postponed given and accessible referents changes in a significant way in the second text, *Nikolaus* displays values that are similar to those observed in the first text, *Speculum Ecclesiae*. Thus, it is not possible to show any substantial change in the information-structural function of the NF for MHG at the moment.

Besides their cognitive accessibility, the issue whether postponed constituents can be focussed is also addressed in our research. Focus is thus annotated independently from the cognitive status of a certain referential expression. So, for example, a focussed constituent in the NF is not necessarily new information. Our preliminary data show no clear change in the extraposition of focussed constituents, which are nonetheless robustly attested through the whole MHG period.

³ A finer classification is used in our corpus based on the proposal by Götze et al. (2007).

The last aspect we want to investigate is discourse structure, especially the functions of the NF for text-building strategies. In particular, the establishing of the discourse topics by means of NF postposition and referential choice are an important part of the research. The following issues are being investigated:

- the resumption of the NF constituent in the discourse. The function of extraposition is relevant for the discourse structure, since this strategy is used to either establish, change or reactivate a discourse topic;
- the lexical means of resumption, in particular definite expressions and pronouns.

Therefore, all expressions referring to identical objects have been annotated in order to investigate possible cases of linkage in discourse. To illustrate this, consider example (4) again, repeated here with some accompanying context:

- (6) *Der lerer der sol siniu schâf och retten vor dem beren. das ist vor dem libe. wan als der ber dem honege nach gat. vnd als er ez gerne izzet. als gat der lip der welte suezechait. vnd der welte geluste och nach.* (SchwP 10, 9)

‘The teacher, he shall also save his sheep from the bear, that is from the body. Because, as the bear seeks honey and he gladly eats it, the body also seeks the sweetness and lust of the world.’

The reactivation of the referent in the discourse (‘the bear’) is used as a strategy to introduce the topic of the following discourse. The referent is resumed by means of different (pronominal and non-pronominal) anaphora – *der ber* (‘the bear’) and *er* (‘he’) – according to the varying cognitive status of the referential object in the discourse.

In conclusion, we found that the change with respect to the grammatical function of NF constituents is characterized by means of a decrease in postverbal subjects, an increase in PPs and a qualitative change within the category of objects. No clear results can be noticed with respect to their information status, but we hope that a more fine grained annotation of these texts (and of ENHG ones, which are still in the process of being annotated at the time of writing) will allow us to gain further insights into the information status properties of the NF. Finally, discourse linkage seems to play an important role for postponing constituents, but at this point of our study we are still assessing the data collected. The final goal of our study is to provide insights into the specific functions of the topological NF in MHG and – at a future stage – ENHG by

taking into account syntactic, information- and discourse-structural aspects. Given that a specific investigation of these two historical epochs is still lacking, our research aims at providing the missing link in the diachronic development of the German NF, from its origins to the present day.

Primary texts

- [NikP] Pfeiffer, Franz (1845): *Deutsche Mystiker des 14. Jahrhunderts. Erster Band: Hermann von Fritzlar, Nicolaus von Strassburg, David von Augsburg*. Leipzig: G. J. Göschen'sche Verlagshandlung.
- [SchwP] Grieshaber, Franz Karl (1978): *Deutsche Predigten des 13. Jahrhunderts. Zwei Teile in einem Band. Nachdruck der Ausgabe 1844-1846*. Hildesheim/New York: Georg Olms.
- [Speccl] Mellbourn, Gert (1944): *Speculum ecclesiae. Eine frühmittelhochdeutsche Predigtsammlung (Cgm 39). Mit sprachlicher Einleitung*. Lund/Kopenhagen: Gleerup/Munksgard.
- [T] Masser, Achim (1994): *Die lateinisch-althochdeutsche Tatianbilingue*. Stiftsbibliothek St. Gallen Cod. 56. Göttingen: Vandenhoeck & Ruprecht.

References

- Altmann, Hans (1981): *Formen der „Herausstellung“ im Deutschen. Rechtsversetzung, Linksversetzung, Freies Thema und verwandte Konstruktionen*. Tübingen: Niemeyer.
- Averintseva-Klich, Maria (2009): *Rechte Satzperipherie im Diskurs. Die NP-Rechtsversetzung im Deutschen*. Tübingen: Stauffenburg.
- Engel, Ulrich (1977): *Syntax der deutschen Gegenwartssprache*. Berlin: Erich Schmidt Verlag.
- Filpus, Raja (1994): *Die Ausklammerung in der gesprochenen deutschen Sprache der Gegenwart*. Tampere: Universität Tampere.
- Götze, Michael/Weskott, Thomas/Endriss, Cornelia/Fiedler, Ines/Hinterwimmer, Stefan/Petrova, Svetlana/Schwarz, Anne/Skopeteas, Stavros/Stoel, Ruben (2007): *Information structure*. In: Dipper, Stefanie/Goetze, Michael/Skopeteas, Stavros (eds.): *Information structure in cross-linguistic corpora: annotation guidelines for phonology, morphology, syntax, semantics and information structure (= Interdisciplinary Studies on Information Structure 7)*. Potsdam: Universitätsverlag Potsdam, 147-187.
- Gundel, Jeanette K./Hedberg, Nancy/Zacharski, Ron (1993): *Cognitive status and the form of referring expressions in discourse*. In: *Language* 69: 274-307.

- Haider, Hubert (2010): Wie wurde Deutsch OV? Zur diachronen Dynamik eines Strukturparameters der germanischen Sprachen. In: Ziegler/Braun (eds.), 11-32.
- Hinterhölzl, Roland (2004): Language change vs. grammar change: what diachronic data reveal about the distinction between core grammar and periphery. In: Fuß, Eric/Trips, Carola (eds.): Diachronic clues to synchronic grammar. Amsterdam: John Benjamins, 131-160.
- Hinterhölzl, Roland/Petrova, Svetlana (eds.) (2009): Information structure and language change: new approaches to word order variation in Germanic. Berlin: Mouton de Gruyter.
- Hoberg, Ursula (1981): Die Wortstellung in der geschriebenen deutschen Gegenwartssprache. (= Heutiges Deutsch 1/10). München: Hueber.
- Margetts, John (1969): Die Satzstruktur bei Meister Eckhart. Stuttgart: Kohlhammer.
- Petrova, Svetlana (2009): Information structure and word order variation in the Old High German Tatian. In: Hinterhölzl/Petrova (eds.), 251-279.
- Petrova, Svetlana/Hinterhölzl, Roland (2010): Evidence for two types of focus position in Old High German. In: Ferraresi, Gisella/Lühr, Rosemarie (eds.): Diachronic studies on information structure. Language acquisition and change. Berlin: Mouton de Gruyter, 189-217.
- Prell, Heinz-Peter (2003): Typologische Aspekte der mittelhochdeutschen Prosasyntax. Der Elementarsatz und die Nominalphrase. In: Lobenstein-Reichmann, Anja/Reichmann, Oskar (eds.): Neue historische Grammatiken. Zum Stand der Grammatikschreibung historischer Sprachstufen des Deutschen und anderer Sprachen. Tübingen: Niemeyer, 241-256.
- Schiewers, Regina D. (2008): Die deutsche Predigt um 1200. Berlin/New York: de Gruyter.
- Schildt, Joachim (1976): Abriss der Geschichte der deutschen Sprache. Zum Verhältnis von Gesellschaft- und Sprachgeschichte. Berlin: Akademischer Verlag.
- Schlachter, Eva (2009): Word order variation and information structure in Old High German: An analysis of subordinate *dhazs*-clauses in Isidor. In: Hinterhölzl/Petrova (eds.), 223-250.
- Schlachter, Eva (2010): Zum Verhältnis von Stil und Syntax. Die Verbfrüherstellung in Zitat- und Traktatsyntax des althochdeutschen Isidor. In: Ziegler/Braun (eds.), 409-426.
- Schmidt, Thomas/Wörner, Kai (2009): EXMARaLDA – Creating, analysing and sharing spoken language corpora for pragmatic research. In: Pragmatics 19/4: 565-582.
- Sigroth-Nellessen, Gabriele von (1979): Versuch einer exakten Stiluntersuchung für Meister Eckhart, Heinrich Seuse und Johannes Tauler. München: Fink.

- Vinckel, Hélène (2006): Die diskursstrategische Bedeutung des Nachfelds im Deutschen: Eine Untersuchung anhand politischer Reden der Gegenwartssprache. Frankfurt a.M.: Universitätsverlag.
- Vinckel-Roisin, Hélène (2011): Wortstellungsvariation und Salienz von Diskursreferenten: Die Besetzung des Nachfeldes in deutschen Preetexten als kohärenzstiftendes Mittel. In: Zeitschrift für germanistische Linguistik 39/3: 377-404.
- Weiß, Helmut (to appear): Die rechte Peripherie im Althochdeutschen. Zur Verbstellung in *dass*-Sätzen. Ms. University of Jena.
- Ziegler, Arne/Braun, Christian (eds.) (2010): Historische Textgrammatik und Historische Syntax des Deutschen. Traditionen, Innovationen, Perspektiven. Berlin/New York: de Gruyter.
- Zifonun, Gisela/Hoffmann, Ludger/Strecker, Bruno/Ballweg, Joachim/Brauße, Ursula/Breindl, Eva/Engel, Ulrich/Frosch, Helmut/Hoberg, Ursula/Vorderwülbecke, Klaus (1997): Grammatik der deutschen Sprache. (= Schriften des Instituts für Deutsche Sprache 7). Berlin/New York: de Gruyter.

Creating synopses of ‘parallel’ historical manuscripts and early prints

Alignment guidelines, evaluation, and applications

Abstract

In this paper we introduce the task of aligning parallel historical texts, to create synopses for comparing similarities and deviations between them. We present guidelines for manually annotating corresponding words and phrases. A test annotation reveals that there is considerable high inter-annotator agreement, ranging from $\kappa = 0.76$ to 0.98 , depending on the specific text. In an application scenario we show a typical use case for which token and phrase alignments are of value.

1. Introduction

In this paper we introduce methods for aligning parallel historical texts. Alignment refers to the task of linking corresponding elements (words, phrases, paragraphs, etc.) between related documents. Documents can be related because, e.g., they go back to the same source or one document is a (close or loose) copy of the other. As a result, the documents are similar to each other to different degrees and deviations can provide interesting clues for linguistic or historico-cultural investigations.

The goal of this paper is to present and evaluate guidelines we developed for aligning different versions of the medieval passion treatise *Interrogatio Sancti Anselmi de Passione Domini* (henceforth referred to as *Anselm*). In this text, St. Anselm of Canterbury asks Virgin Mary to reveal the passion of her son Jesus Christ, beginning with the Last Supper and ending with his resurrection. There are 70 vernacular manuscripts and prints, which date from the 14th-16th century and represent dialects of Early New High German (ENHG), Middle Low German and Middle Dutch.¹ The degree of similarity between

¹ The further textual evidence covers 162 Latin texts and one Middle English text. The text is transmitted in different versions: verse, short prose, and long prose. For the alignment task introduced in this paper, we only deal with prose versions. Currently, 64 texts are transcribed and available for analysis (51 manuscripts and 10 prints in German, two manuscripts and one print in Dutch. 50 texts are preserved completely, 14 texts are fragments).

individual *Anselm* texts varies:² there are pairs which are almost identical as well as pairs that differ to a high degree, e.g. in vocabulary, word order, or content. In our guidelines, we model these differences with different annotation layers, which range from token level alignments to phrasal alignments.

There is a variety of conceivable applications where the marking of concordant elements is of value. Alignments can be used for analyzing certain kinds of linguistic variance (e.g. in word order patterns). Furthermore, they can be of interest for philological research, e.g. in preparing printed or digital editions with a critical apparatus.

The paper is organized as follows. In Section 2, we address related work. Sections 3 and 4 present the annotation guidelines and the results from our test annotation. In Section 5 we present an exemplary application where the results of the annotation process can be used for answering philological questions.

2. Related work

Word alignment has a strong tradition in translation studies which focus on the detection of translation correspondences in multilingual environments. Guidelines for the annotation of corresponding tokens have been described, e.g., in the Blinker Annotation Project (Melamed 1998) or in Macken (2010a). Relating to this, techniques for the automatic detection of correspondences are investigated in the field of Machine Translation (cf. Och/Ney 2000). Compared to translation tasks of contemporary documents, our project has to cope with a constantly varying degree of similarity between the single texts, making the alignment procedure more complex.

There are several philological projects which deal with parallel historical texts: For instance, the University of Vienna provides an online synopsis of the *Nibelungenlied*³ and the University of Bern aims at digitizing the complete tradition of the epic *Parzival* by Wolfram von Eschenbach.⁴ Another project that will publish the complete tradition of the 12th century *Kaiserchronik* started recently at the University of Cambridge.⁵ All these projects share the

² With the term *text* we refer to different instances (manuscripts or prints) of one (virtual) base text, the medieval passion treatise *Anselm*.

³ <http://germanistik.univie.ac.at/index.php?id=14531>. All URLs have been checked and found valid as of late January 2015.

⁴ www.parzival.unibe.ch/home.html.

⁵ www.mml.cam.ac.uk/german/staff/kaischron.html.

commonality that they concentrate on bringing together parallel versions in printed or online editions; they do not deal with word alignment.

3. Alignment guidelines

The goal of the alignment task described here is to cover all extant texts of *Anselm* and to align corresponding elements. Thus, in contrast to many edition projects, there is no a priori defined “central” text; instead all texts are of equal importance. Depending on the particular research question, any *Anselm* text can be chosen as the “central” version and basis of comparison. The main focus is on intertextuality and we are interested in similarities and differences between entire documents.

In our guidelines for aligning parallel texts, we distinguish between four separate annotation layers: Cognates (Section 2.1), Synonyms (Section 2.2), Coreference (Section 2.3), and Complex (phrasal) equivalents (Section 2.4). The layers differ in the degree of similarity between the aligned tokens or phrases.⁶ Each layer is annotated in a separate pass, i.e. all corresponding cognates are marked first, afterwards all synonyms are marked and so forth. Two tokens (or phrases) are taken to correspond if they share the same context. Ex. (1) shows such a comparable context from two manuscripts Ba1 and D4, where it is possible to align single corresponding tokens.

(1)

Bamberg (Ba1)
<i>Ain hoher lerer hiesz anshelmus, der pat vnser frauen lange weill vnd zeit wainent vasten vnd peten, Das sy im zu erkennen geb, wie vnser herre gemartert wer word</i>
'A high teacher was called Anselm, he asked our lady for a long time, crying, starving, praying, to show him how our lord was tortured'
Dessau (D4)
<i>Sant anszhelmüs // bischoff hat gebetten lang zeit mit vasten / weinen vnnd betten / Maria die reinen Iuncfrowen vnd müter gots / das si Im wolt volkommenlich offenbaren / das leyden Ires lieben soenes cristi iesu</i>
'Saint Anselm, the bishop, has asked long time – while starving, crying and praying – Mary, the pure virgin and mother of god, to completely reveal the passion of her son Jesus Christ'

⁶ Melamed (1998) and Macken (2010a) make similar distinctions, which Macken calls “regular links” (our first two levels) and “fuzzy links” (our levels three and four).

3.1 Cognates

On the first level of alignment, all corresponding tokens in two texts that are cognates are aligned. Cognates are words with a common etymological origin.⁷ Cognates can belong to different word classes; e.g., nouns can be aligned with verbs as long as both have the same stem or root, see Example (2), where the noun *marter* ‘martyrdom’ is aligned with the participle *gemartert* ‘martyred’. Only pairs of the form *token:token* are aligned at this level.

(2) Alignment of cognates⁸

Bamberg (Ba1)						
[<i>Das</i>] _{L1.1}	[<i>sy</i>] _{L1.2}	[<i>im</i>] _{L1.3}	<i>zu</i>	<i>erkennen</i>	<i>geb</i>	<i>wie</i>
that	she	him	to	recognize	gives	how
<i>vnser</i>	<i>herre</i>	[<i>gemartert</i>] _{L1.4}	<i>wer</i>	<i>word</i>		
our	lord	<i>tortured</i>	has	been		
‘ <i>That she show him how our lord was tortured</i> ’						
Stuttgart (Stu1)						
[<i>daz</i>] _{L1.1}	[<i>sy</i>] _{L1.2}	[<i>im</i>] _{L1.3}	<i>kunt</i>	<i>taetty</i>	<i>irs</i>	<i>aingebornes</i>
that	she	him	tell	does	her	only
<i>kindes</i>	[<i>marter</i>] _{E1.4}					
child’s	martyrdom					
‘ <i>That she tell him about her only child’s martyrdom</i> ’						

⁷ This includes suppletive forms (e.g. as in the paradigm of German *sein* or the English equivalent *be*).

⁸ Examples are organized as follows: L = level; L1 = level 1; L1.1 = level 1, alignment pair 1. All tokens belonging to the same alignment pair have the same ID.

3.2 Synonyms

In the second annotation pass, *real* synonyms of the type *token:token* are aligned. Two tokens are taken to be synonyms if they are interchangeable without a resulting change in meaning, as in Example (3) where *getwagen* is aligned with *gewaeschen*, both meaning ‘washed’.⁹

(3) Alignment of synonyms

Bamberg (Ba1)						
<i>do</i> _{Ei.1}	<i>mein</i> _{L1.2}	<i>kindt</i> _{L1.3}	<i>ir</i>	<i>fuezz</i> _{L1.4}	[getwagen] _{L2.1}	<i>het</i> _{L1.5}
when	my	child	their	feet	washed	has
‘when my child has washed their feet’						
Stuttgart (Stu1)						
<i>do</i> _{L1.1}	<i>min</i> _{L1.2}	<i>kint</i> _{L1.3}	<i>inen</i>	<i>die</i>	<i>fuezz</i> _{L1.4}	
when	my	child	them	the	feet	
<i>het</i> _{L1.5}	[gewaeschen] _{L2.1}					
has	washed					
‘when my child has washed them their feet’						

3.3 Coreference

This level aligns phrases that are coreferent, i.e., a pro-form (pronoun or adverb) is used in one text and a corresponding full phrase (NP, PP, VP) in the other. In contrast to the two previous two layers, in this layer entire phrases can be aligned; see Ex. (4) where the full noun phrase *den knecht* ‘the servant’ is aligned with the pronoun *in* ‘him’.

⁹ In German, certain verbs, called prefix verbs, can occur discontinuously. The verb and its prefix count as one token in these cases, since they belong to the same lemma, listed in standard lexicons. For our purpose, we use the *Duden* as a standard reference (www.duden.de). For extinct word forms (e.g. *getwagen* ‘washed’ in text Ba1 in Ex. (2a)) we use a dictionary for historical German, in our case a dictionary for Middle High German (www.woerterbuchnetz.de/lexen).

(4) Alignment of coreferential phrases

Stuttgart (Stu1)					
<i>vnd</i> _{L1.1}	<i>macht</i> _{L1.2}	[<i>den</i>	<i>knecht</i>] _{L3.1}	<i>wider</i>	<i>gesundt</i> _{L1.3}
and	makes	the	servant	again	healthy
'and restored the servant's health again'					
Bamberg (Ba1)					
<i>vnd</i> _{L1.1}	<i>machtet</i> _{L1.2}	[<i>in</i>] _{L3.1}	<i>zehant</i>	<i>gesundt</i> _{L1.3}	
and	makes	him	immediately	healthy	
'and restored the servant's health'					

3.4 Complex (phrasal) equivalents

In a last annotation pass, all remaining non-aligned tokens need to be checked for correspondences. Only entire phrases (NPs, PPs, VPs) can be aligned on this level. In Ex. (5) the individual elements of the verbal phrase *zu erkennen geb* 'to show' have no cognates or synonyms in the parallel text. But the phrase as a whole does have an equivalent in the other text: *kunt taetty* 'tell'. Hence, both phrases are aligned.

(5) Alignment of complex (phrasal) equivalents

Bamberg (Ba1)						
<i>Das</i> _{L1.1}	<i>sy</i> _{L1.2}	<i>im</i> _{L1.3}	[<i>zu</i>	<i>erkennen</i>	<i>geb</i>] _{L4.1}	<i>wie</i>
that	she	him	to	recognize	gives	how
<i>vnser</i>	<i>herre</i>	<i>gemartert</i> _{E1.4}	<i>wer</i>	<i>word</i>		
our	lord	<i>tortured</i>	has	been		
'That she signifies him how our lord was tortured'						
Stuttgart (Stu1)						
<i>daz</i> _{L1.1}	<i>sy</i> _{L1.2}	<i>im</i> _{L1.3}	[<i>kunt</i>	<i>taetty</i>] _{L4.1}	<i>irs</i>	<i>aingebornes</i>
that	she	him	known	does	her	only
<i>kindes</i>	<i>marter</i> _{E1.4}					
child's	martyrdom					
'That she tells him about her only child's martyrdom'						

4. Annotation

To evaluate the guidelines, we performed a test annotation with two annotators. In this section, we describe the annotation scenario, present results from the annotation, and compute the inter-annotator agreement.

4.1 Test annotation

For evaluating the guidelines, we extracted three text fragments with comparable content, from two Anselm texts that are rather similar to each other (Ba1 and Ba2) and from two texts that are rather dissimilar (Ba1 and D4). The fragments were taken from the beginning of the texts and consist of roughly 500 tokens (Ba1: 570; Ba2: 561; D4: 529).

	Ba1 : Ba2			Ba1 : D4		
	1:1	1:n, n:1	n:m	1:1	1:n, n:1	n:m
1 (cognates)	543	–	–	170	–	–
2 (synonyms)	4	–	–	60	1	–
3 (coref.)	–	–	–	–	6	–
4 (phrases)	0	4	3	3	20	26
All	547	4	3	233	27	26

Table 1: Number of alignments at different layers with two similar (Ba1 :Ba2) and two dissimilar (Ba1 :D4) fragments

The fragments were annotated independently by two student annotators who were well acquainted with the Anselm texts in general, but not with the alignment task. They had a short training phase: in a first meeting, they were introduced to the guidelines, next they aligned two short training fragments (of less than 500 tokens), followed by a discussion phase.

We used the annotation tool MMAX2 (Müller/Strube 2006), which we adapted to this task: The fragments to be aligned are placed next to each other in one MMAX window. Alignment links are then inserted using MMAX’s facilities for coreference links. Words are displayed in different colors, depending on the type of alignment that they participate in.

4.2 Results from the test annotation

After the test annotation, the annotators produced an adjudicated gold standard. Table 1 shows the number and types of alignments in the gold corpus. The two similar texts show an extremely high number of correlation and most alignments link cognates, meaning that the fragments even share most of their vocabulary. The two dissimilar *Anselm* texts behave very differently, sharing fewer links in total and fewer cognates in particular. The vast majority of alignments in both texts are 1 : 1 alignments.

Figure 1 confirms these findings. It plots the positions of the aligned tokens of both fragments. Aligning a text with itself would result in a diagonal. The plot on the left, displaying the links between Ba1 and Ba2, indicates that the correlation between both fragments is almost perfect. The plot of the dissimilar texts, Ba1 and D4, still clearly approximates the diagonal, which mirrors the fact that both fragments have the same topic. At the same time, it shows considerable deviations and alignment gaps.

	Ba1 : Ba2		Ba1 : D4	
	WAA	kappa	WAA	kappa
1 (cognates)	.99	.83	.95	.84
2 (synonyms)	.99	.39	.95	.63
3 (coref.)	–	–	.99	.57
4 (phrases)	.98	.58	.85	.67
All	1.00	.98	.91	.76

Table 2: Inter-annotator agreement (Word Alignment Agreement score and kappa) for all layers and fragments

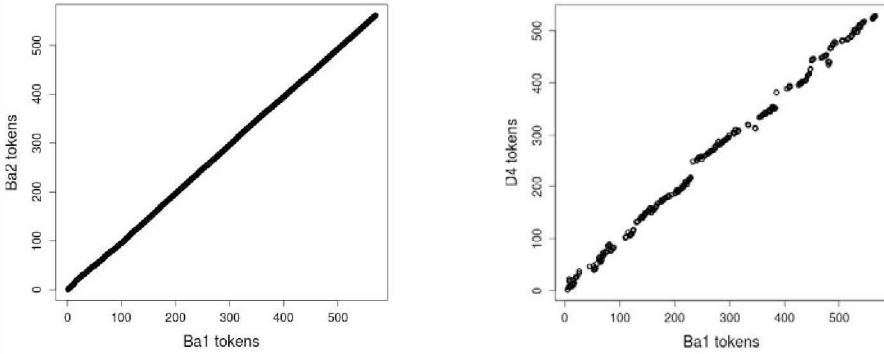


Figure 1: Plot of the aligned token positions in the similar (Ba1 :Ba2, left plot) and dissimilar (Ba1 :D4, right plot) fragments

4.3 Inter-annotator agreement

In similar alignment studies, Davis (2002), Daumé III/Marcu (2005) and Macken (2010b) evaluated manual word and phrase alignment by computing Word Alignment Agreement (WAA, Davis 2002) and the chance-corrected kappa score (Cohen 1960). WAA/kappa = 1 means perfect agreement, WAA = 0 means no agreement, while kappa = 0 means no agreement *above chance*. The task is considered a classification task: given a pair of <source-word, target-word>, one must decide whether the two words should be aligned or not.

Table 2 shows the results for the different layers and fragments. Observed agreement (WAA) is very high for almost all layers. The kappa score, which “punishes” highly skewed distributions, varies considerably. If we merge all layers into one, we see, however, that the annotators’ agreement on the alignments *as such* is nearly perfect (.98) or substantial (.76), while the decision about the type of the alignment (= its layer) is more controversial.

5. Case study

The following case study shows how the alignments can be used to answer philological questions. The treatise *Anselm* has been preserved both in verse and prose; the verse versions are a rather homogeneous corpus while the prose versions could be divided in two groups according to length, i.e. a short and a long form, depending on the word count. Our aim is to go beyond a quantitative classification of each *Anselm* text to a content-related grouping. For this

grouping exercise, it is important to find a way of combining keywords so that they form significant clusters.

Keywords are terms (cognates and synonyms) or phrases (complex phrasal equivalents) selected for their frequency and significance. The aim is to mark up a number of keywords in each of the *Anselm* texts as a profiling exercise to arrive at the ‘fingerprint’ of individual texts made up by their specific use of the keywords. Our assumption is that groups and sub-groups can be visualized by using alignments and measuring the distance between the annotated keywords.

In the following we present a case study taken from the opening sequence of the text. We annotated three keywords, namely all references to the persons ‘Anselm’, ‘Mary’ and ‘Jesus Christ’ (for single terms this has been tried manually, cf. Dipper/Schultz-Balluff 2013, Wegera 2014).

The corpus comprises 51 text instances which have preserved the opening sequence, i.e. 12 in verse, 20 in (long) prose, 15 in (short) prose, three Dutch copies, and one unclassified fragment. Seventeen texts start straight with the introduction, while the other 34 texts preface it by headings or preliminary remarks; most of them, however, were added at a later stage of transmission or edition, e.g. by another writer. It can therefore be assumed that the basic text had no definite title. This makes the specific forms of references to persons within these pre-text sequences especially important as identifying features of the different versions.

Table 3 shows the opening clause of texts representing the long prose version (PL), the short prose version (PS) and the verse version (V). The versions differ considerably with regard to references to persons (printed in boldface).

PL : B2, fol. 48r,5-17 (Ms. germ. qu. 2025, Staatsbibliothek zu Berlin, Preußischer Kulturbesitz)

Sante anhelm der bad *vnser liebe frauwe von hymelriche* alczü

lange zijt mit vil grofzer ynneger begerunge Mit falfen beden vnd mit wachen vnd mit andechtigem gebede vnd mit manichen heifzen drenen daz

fie yme künd wolde dün **vres evngbornes kindes** martele / wie ez von dem anbegynnen da erginge mit zü dem ende fynes lydens

<p>PS: M10, fol. 58v,8-59r,2 (CIm 14945, Bayerische Staatsbibliothek München)</p> <p>Ein hocher lerer hiez anthelmus Der pat <i>vnfer frawn</i> lange wainent vnd vaftent Daz fi im zerkennen gebe wie <i>vnfer h(er)re</i> gemartert wer</p>
<p>V: O, fol. 1r,1-14 (Cim. I.74, Landesbibliothek Oldenburg)</p> <p>ANcelmus was ein heilich man / De hadde langhe dar na ftan / Dat he gherne hedde weten / Wat <i>vnfe here</i> hedde be feten / Nv moghe gi horen wu he dede / he was ftede an finem bede / Beide nacht <i>vñ</i> dach / An finer venigen dat he lach / he fprak <i>maria bloygende rofa</i> / <i>Lylia vñ fittilofa</i> Goddes dure balf men fchirin / Lat mir hute dir werden fchin / Dattu mir moteft rede faghen / van finen iämerliken plaghen</p>

Table 3: Opening clause in the long prose version, the short prose version, and the verse version

In the fairly homogeneous verse version, there is very little variation in lexis and syntax. The first mentioning of ‘Anselm’, ‘Mary’, and ‘Jesus Christ’ is consistent: all texts give the attribute ‘holy man’ to Anselm, compare Mary to flowers such as rose, lily, and perennial, and address Christ as the ‘Lord’ (Table 4). The prose versions use two main clusters in reduced or expanded form around two cores. Beyond this, there is a larger number of texts which show singular combinations which can however be grouped again in clusters (Figure 2).

The possible relations of groups (core 1 and 2), sub-groups, and single combinations are shown in the following scheme: cluster 1 consists of the combination ‘St Anselm’ + ‘Our Lady from Heaven’ + ‘only child’; cluster 2 of ‘Honorable Teacher Anselm’ + ‘Our Lady’ + ‘Lord’.

The illustrations in Figure 2 and Figure 3 are meant as a spatial representation of the relation of different texts (referred to by their sigla, i.e. T, Hk etc.)¹⁰ and editions.

Keywords (person)	Forms of reference
Anselm	<i>Heiliger Mann Anselm</i> ‘holy man Anselm’
Mary	<i>Maria, Rose, Lilie und Zeitlose</i> ‘Mary, rose, lily, and perennial’
Jesus Christ	<i>Herr</i> ‘Lord’

Table 4: Keywords and forms of references in the verse version

¹⁰ For a complete list of all German and Dutch texts see www.rub.de/schultz-balluff/sanktanselmus.

Eight to nine texts can be linked to these two fixed clusters. These can be extended, reduced, and combined, i.e. ‘Our Lady from Heaven’ can be extended to ‘Our Dear Lady from Heaven’ or the two clusters can be combined to form ‘Teacher Anselm’ + ‘Our Dear Lady’ + ‘Dear Child’ (Fig. 2). We can assume that these clusters developed first and that the other combinations are variations of them.

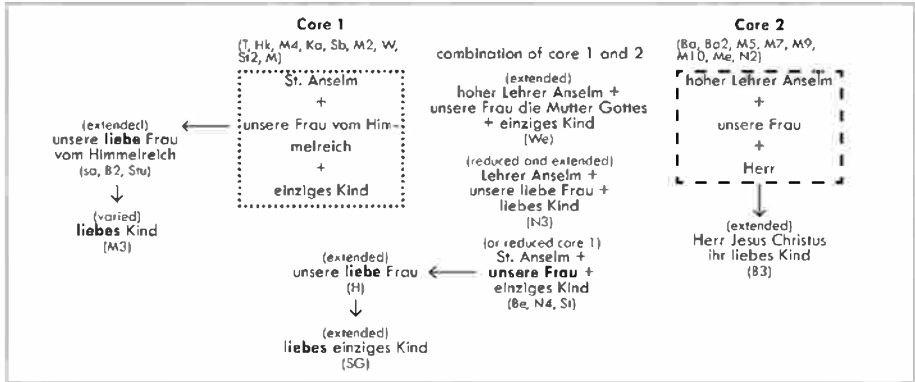


Figure 2: Main clusters in the prose version

The second scheme (Fig. 3) shows the correlation of disparate combinations and the identification of abstract cores. The ten Anselm texts show combinations of basic elements which suggest that they are also connected. It is possible to generate abstract clusters from the single elements which form virtual hubs, giving the ten text instances a clear position within the overall system.

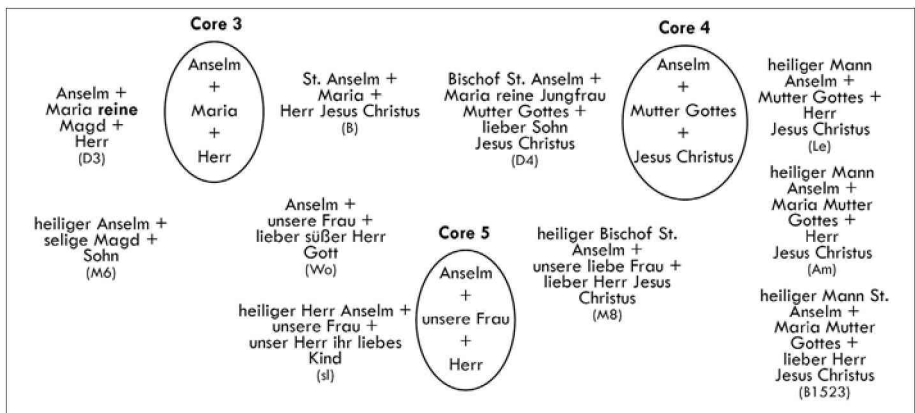


Figure 3: Sub-groups and single combinations in the prose version

Against this background, the preliminary distinction between long and short prose texts has to be reconsidered. With the help of keywords which will be applied to each text instance in its entirety, new constellations will become visible. This will allow us to take a fresh look at text production and changes during the transmission process. It will hopefully form a solid textual basis to build on for further insights into textual criticism, literary trends, and cultural change – ultimately resulting in a stable framework within which to place the shape-shifting sets of questions with which St Anselm confronts Mary.

6. Conclusion

In this paper we introduced a method for aligning parallel historical texts. The annotation guidelines focus on the problems aligning texts which differ in their degree of similarity. An evaluation annotation revealed considerable inter-annotator agreement, even when the two aligned texts were very dissimilar. Furthermore, we showed how the alignment can be of use for clustering parallel texts according to their use of specific vocabulary terms.

References

- Cohen, Jacob (1960): A coefficient of agreement for nominal scales. In: *Educational and Psychological Measurement* 20: 37-46.
- Daumé III, Hal/Marcu, Daniel (2005): Induction of word and phrase alignments for automatic document summarization. In: *Computational Linguistics* 31(4): 505-530.
- Davis, Paul C. (2002): *Stone soup translation: the linked automata model*. Ph.D. diss., Ohio State University.
- Dipper, Stefanie/Schultz-Balluff, Simone (2013): The Anselm corpus: methods and perspectives of a parallel aligned corpus. In: *Proceedings of the Workshop on computational historical linguistics at NODALIDA 2013*. (= NEALT Proceedings Series 18). Linköping: Linköping Electronic Conference Proceedings, 27-42. www.ep.liu.se/ecp/087/003/ecp1387003.pdf.
- Macken, Lieve (2010a): *Annotation guidelines for Dutch-English word alignment*. Version 1.0. Technical report, Language and Translation Technology Team, Faculty of Translation Studies. University College Ghent. www.lt3.ugent.be/media/uploads/publications/2010/Macken2010c.pdf.
- Macken, Lieve (2010b): *An Annotation Scheme and Gold Standard for Dutch-English Word Alignment*. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. www.lrec-conf.org/proceedings/lrec2010/pdf/100_Paper.pdf.

- Melamed, I. Dan (1998): Annotation style guide for the Blinker project, Version 1.0.4. (= IRCS Technical Report #98-06). Philadelphia: University of Pennsylvania.
- Müller, Christoph/Strube, Michael (2006): Multi-level annotation of linguistic data with MMAX2. In: Braun, Sabine/Kohn, Kurt/Mukherjee, Joybrato (eds.): *Corpus technology and language pedagogy. New resources, new tools, new methods.* Frankfurt a.M.: Peter Lang, 197-214.
- Och, Franz Josef/Ney, Hermann (2000): A comparison of alignment models for statistical machine translation. In: *Proceedings of the 18th International Conference on Computational Linguistics (COLING-ACL 2000).* Stroudsburg, PA: ACL, 1086-1090. <http://ucrel.lancs.ac.uk/acl/C/C00/C00-2163.pdf>.
- Wegera, Klaus-Peter (2014): *Interrogatio St. Anselmi de Passione Domini, deutsch. Überlieferung – Edition – Perspektiven der Auswertung.* Hrsg. von der Nordrhein-Westfälischen Akademie der Wissenschaften und Künste. Paderborn: Schöningh.

How exceptional is CP recursion in Germanic OV languages?

Corpus-based evidence from Middle Low German

Abstract

CP recursion, i.e. verb fronting to a lower CP in complement clauses selected by bridge verbs, is considered typical for the Germanic languages with VO basic order (Danish, Faroese, Norwegian and Swedish) but exceptional in those with underlying OV structure, with Frisian being the only modern OV language displaying this property. Historical evidence has been taken to weaken the exceptional status of CP-recursion in OV languages, but the data is sparse or provides no diagnostics allowing for a safe interpretation of the facts. In this paper, we want to present historical data from an OV language that helps us to overcome this empirical problem. Data search in a small syntactically annotated corpus from Middle Low German (MLG) produced examples for complement clauses including a complementizer, in which the verb is in second position, following a non-subject constituent. Additionally, the clauses contain typical middle field elements suggesting that the verb has left the VP and targeted the C-domain of the clause. Following these considerations, we argue in favour of an analysis of these orders as instances of CP recursion, which in turn suggests that it is not exceptional in OV languages. This pilot study is meant to emphasize both the relevance of MLG in research on comparative Germanic syntax as well as the role of historical corpora in solving open issues in theoretical linguistics.

1. Goals and issues

This paper presents the design and the results of a pilot study intended to illustrate how corpus-based investigations can be utilized in solving much debated issues in theoretical linguistics. We address the long-lasting discussion on the status of limited embedded V2 in Germanic languages whose underlying order is O(bject) – V(erb). It is well-known that in Danish, Faroese, Norwegian and Swedish, verb-second order (V2) in complement clauses in the presence of a complementizer is possible if the selecting element is a representative of the group of so-called “bridge verbs” like *to know* in (1). This property is called “limited embedded V2” (cf. Vikner 1995, Haider 2010: 47), in contrast to “generalized embedded V2” in Yiddish and Icelandic, where V2 occurs in all types of dependent clauses, including adverbial and relative ones. The same order is ungrammatical in modern German and Dutch, which only

allow for embedded V2 in the absence of complementizers (cf. Reis 1997 among others), as demonstrated by (2a) vs. (2b).

- | | | |
|------|---|-------------------------------------|
| (1) | <i>Vi ved at denne bog har Bo ikke læst</i>
We know that this book has Bo not read
'We know that Bo has not read this book' | Danish

(Haider 2010: 47) |
| (2a) | <i>*Maria sagt, dass Peter ist krank</i>
Mary says that Peter is ill | German |
| (2b) | <i>Maria sagt, Peter ist krank</i>
Mary says Peter is ill
'Mary says that Peter is ill' | German |

The variation in the position of the verb in dependent clauses in Germanic has been attributed to a fundamental difference in the grammars of the individual languages (cf. Vikner 1995). While generalized embedded V2 is derived by fronting of the verb to the head of a medial functional projection IP, which is a constant part of the grammar of the respective languages, limited embedded V2 is analysed as resulting from movement of V to the head of a lower CP licensed only in complement clauses of bridge verbs, assuming that the CP of such verbs is recursive (CP recursion).

Based on well-established typological properties of the Germanic languages that allow for limited embedded V2, it has been argued that CP recursion is a property typical for those with underlying VO grammar, with Frisian being the only modern OV language in which limited embedded V2 is attested (cf. Vikner 1995: 66; Haider 2010: 46). Several studies have addressed the assumed exceptionality of limited embedded V2 in Germanic languages with basic OV, presenting various kinds of potentially relevant data. However, the straightforward analysis of this data along the lines of CP recursion has revealed problems either on theoretical or on empirical grounds. Kemenade (1997) presents potential examples for CP recursion in Old English (OE), which according to her is underlyingly OV. Pintzuk (1993) shows that V2 in OE is not restricted to *that*-clauses of bridge verbs but also occurs in adverbial and relative clauses. Suchsland (2000) and Axel (2007) consider evidence from Old High German (OHG) but admit that conclusive examples are sparse. Modern German, finally, has been shown to display V2 in *dass*-clauses as well, see Freywald (2008), but also in complements to nouns, unlike the Scandinavian patterns. Until now, no attempt at showing that Germanic OV basic order is compatible with CP recursion has yielded successful results.

Petrova (2012, 2013), expanding on a brief mention in the descriptive literature (Rösler 1997: 190), has argued that MLG, the predecessor of the Germanic dialects spoken in Northern Germany, is the stage that provides the missing evidence concerning the compatibility of CP recursion with basic OV. In this study, we want show how properties diagnostic for limited embedded V2 can be isolated by searching an electronically available pilot corpus of MLG.

The outline of the paper is as follows. First, we will present the basic properties of MLG clause structure in order to show that it classifies as underlyingly OV and thus provides a suitable ground for testing the compatibility of CP recursion and basic OV. Second, we will describe the properties of the corpus compiled, and then outline the steps in the data search which yield the conclusive examples. Finally, we will summarize the basic arguments of the analysis, highlighting the importance of corpus-based work as well as the necessary refinements of the corpus for future studies.

2. Basic properties of Middle Low German clause syntax

MLG is a cover term applied to a number of regional varieties that remained unaffected by the Second (also called “High German”) Consonant Shift, a significant sound change that gradually took place in High German. It is documented in written sources dating back to the beginning of the 13th century and is well attested in texts of different genres up to the 17th century (Peters 2000). In sharp contrast with the spread and the amount of attestation available, MLG has remained unexplored in formal comparative studies on clause structure in Germanic, with the prominent exception of work on sentential negation (Breitbarth 2009).

Petrova (2013) demonstrates that MLG displays orders typical for Germanic OV. Note that in basic orders, objects precede their lexical heads, and non-finite verbs precede finite auxiliaries that select them as complements in compound tense forms, cf. (3):

- (3) *dat de lude de brut unde den brudegham to bedde*
 that the people the bride and the bridegroom to bed
gebrach hadden
 brought had
 ‘that the people had brought to bed the bride and the bridegroom’
 (LChr II, 103)

Additionally, as Petrova (2013) shows, MLG displays well-known diagnostic properties associated with basic OV in the research on Germanic (Haider 2000; Vikner 2001, among others) and employed at exemplifying the OV / VO contrast in Germanic, viz. empty objects (*e*) in VP coordination (4), scrambling (5a, b), preverbally adjacent particles and resultative predicates, (6a, b), and variable order in the verbal complex (7a, b, c):

- (4) *dat he [dat land]_i herede and e_i brande*
 that he DET land_i devastated and burned
 ‘that he should devastate the land and burn it’
 (SW 100, 4)
- (5a) *dat got [Moysese]_{DAT} [de ê]_{ACC} gaf*
 that God Moses the law gave
 ‘that God gave Mose the law’
 (SW 71, 41)
- (5b) *He gebot, dat men [de kindere]_{ACC} [den modern]_{DAT}*
 He asked that INDEF DET children DET mothers
wider gave
 back give-SBJ
 ‘He ordered to return the children to their mothers’
 (SW 118, 18)
- (6a) *de se den cristen **afghewunnen** hadden*
 REL they the Christians PRT-won had
 ‘which they had obtained from the Christians’
 (LChr I 71,4)
- (6b) *dat se sic ettelike **dot** staken*
 that they REFL lots-of dead slaughtered
 ‘that many people killed themselves’
 (SW 99, 10)
- (7a) *Do got der engele kore vullen₂ wolde₁*
 when God DET angels chorus fill-INF wanted
 ‘when God wished to complete the chorus of the angels’
 (SW 67, 6)

(7b) *do se weren₁ vordreven₂*
 when they were expelled
 ‘when they were expelled’
 (LChr I, 63)

(7c) *de sik hadde₁ wedder dat rike settet₂*
 REL REFL had against the kingdom set
 ‘who acted against the kingdom’
 (LChr I, 72)

Given this, MLG classifies as a language with basic OV and is a suitable basis for testing the compatibility of CP recursion and basic OV in Germanic.

3. Properties of the corpus

Project B4 of Collaborative Research Centre 632 on Information Structure at Humboldt-Universität zu Berlin has compiled a corpus containing parts of texts from the MLG period using the annotation tool EXMARaLDA (Schmidt/Wörner 2009). This tool granted us a multi-layer annotation architecture as well as the possibility to add tags and features according to the goal of the study (see Fig. 1). In order to exclude orders which are due to metrical demands

	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196
text	In	scole	gy	weten	dat	dat	hilghe	lant	is	dat	land	des	loftes	dat	god	van	hemmel	louet	hadde	Abraham
clause-status	MAIND					OOBJ								REL						
gf	ADV	VFIN	SUBJ	INF		SUBJ			VFIN	PREDN				DO	SUBJ			PARTPERF	VFIN	IO
syl_no																				
givenness																				
top-comm																				
aboutness																				
position																				
topic-marker																				
definiteness																				
loc-hg																				
loc-marker																				
context																				
comment																				
hbl	L.S. 02. 10			L.S. 02. 10																L.S. 02. 20

Figure 1: Multi-Level Annotation of *Ludolf von Sudheims Reise ins Heilige Land* using EXMARaLDA

or foreign influence, the corpus was restricted to original (non-translated) prose texts. Concentrating on one of these, the journey diary *Ludolf von Sudheims Reise ins Heilige Land* (“Ludolf of Sudheim’s Journey to the Holy Land”) which belongs to the early period of the attestation, we created a corpus comprising the first 20 pages of the available print edition (von Stapelmoor 1937), which were tagged for clause type and grammatical function. The corpus includes 6,690 tokens and is searchable via the ANNIS2 database (Zeldes et al. 2009) developed by Project D1 of the Collaborative Research Centre.¹

4. Steps in the data search

Note that limited embedded V2 correlates with the properties outlined in (8a-c):

- (8a) V2 occurs in a complement clause containing an overt complementizer *dat* ‘that’
- (8b) The instances are conclusive for genuine V2 because
 - i. V2 is not restricted to subject-verb orders but is also attested in complement clauses with a non-subject in first position
 - ii. V-to-C-movement takes place, i.e., the verb is above typical VP-elements like the sentential negator *nicht*
- (8c) Embedded V2 is triggered by the semantics of the selecting predicate, which is a representative of the class of bridge verbs listed in (9), cf. Vikner (1995).
- (9) *hint, indicate, answer, claim, report, emphasize, decide, learn, remember, find out, find, believe, hope, mean, say, see, feel, assume, know.*

In a first step, we performed a search for complement clauses introduced by *that* in which the finite verb is in second position. The corpus contains 20 complement clauses with *dat*, retrieved by the query in (Q1) and given in Appendix 1. In 7 of them, the verb is in second position, as found using the search in (Q2), with the relevant results in Appendix 2.

(Q1) `clause-status=/COBJ.* / & tok="dat" & #1_1_ #2`

(Q2) `clause-status=/COBJ.* / & tok="dat" & #1_1_ #2 & gf & gf="VFIN" & #2 . #3 & #3 . #4`

¹ www.sfb632.uni-potsdam.de/annis/.

As we expected, not all hits are conclusive for the aims of the study. First, in some of the examples, the order in the *that*-clause is subject-verb (10a). Second, the surface order in some of the examples is ambiguous because it can be derived both by leftward movement of the verb (10b) or by extraposition of the heavy XP to the right of the verb, which actually remains *in situ*, cf. (10c):

- (10a) *dat Rodijs is eyn alto schone land*
 that Rhodos is a very nice country
 ‘that Rhodos is a very nice country’
- (10b) [_{CP1} dat [_{CP2} Rodjjs, [is_i [_{VP} t_i en alto schone land t_i]]]]
verb fronting
- (10c) [_{CP} dat [_{VP} Rodjjs t_i is]] en alto schone land_i
extraposition

In order to be able to argue for limited verb second, we need to show that the verb is attested in second position after non-subjects, and preceding typical middle field elements such as the sentential negator *nicht*. The latter is evidence for verb movement out of the VP to the left periphery of the clause. We performed a query searching for these conditions in one and the same clause, cf. (Q3). But the search produced no examples in which all conditions in (8b i and ii) applied simultaneously. We thus separated the two conditions and selected V2-*dat*-clauses with non-subjects in first position (Q4) and V2-*dat*-clauses with the finite verb above the sentential negator *nicht* (Q5).

- (Q3) `clause-status=/COBJ.*/ & tok="dat" & #1_1_ #2 &
 gf!="SUBJ" & gf="VFIN" & gf="NEG" & #2 . #3 & #3
 . #4 & 4 . #5`
- (Q4) `clause-status=/COBJ.*/ & tok="dat" & #1_1_ #2 &
 gf!="SUBJ" & gf="VFIN" & #2 . #3 & #3 . #4`
- (Q5) `clause-status=/COBJ.*/ & tok="dat" & #1_1_ #2 &
 gf & gf="VFIN" & gf="NEG" & #2 . #3 & #3 . #4 &
 #4 . #5`

Both queries yielded positive results. The hits for (Q4) are listed in Appendix 3, a representative example is given in (11). Thus, we have evidence for *dat*-clauses in which the verb is in second position following non-subjects:

- (11) *Gy scholet ok weten dat in der stad wonet de keyser der*
 You should also know that in this town lives the king the-GEN
greken
 Greeks
 ‘You should also know that in this town lives the king of the Greeks’
 (LS 95, 10)

The refinement in (Q5) is intended to demonstrate that the superficial verb second order is not ambiguous but due to verb movement to the left. The relevant hit is given in (12). Although the number of hits is small, which may be due to the size of the corpus, we have evidence for leftward movement of the finite verb in a *dat*-clause in MLG.

- (12) *Nu scole gy weten, dat alle desse stede, die hir*
 Now should you know that all these places REL here
vor ghenomet zyn, sint nicht ene dachuart van Cipro
 earlier named are are not one day’s journey from Cyprus
 ‘Now you should know that all these places that are mentioned earlier, are less than a day’s journey away from Cyprus’
 (LS 113, 20)

Finally, we need to consider the properties of the selecting head. It is well-known that limited embedded V2 in Danish, etc. is only allowed under bridge verbs (cf. the list of verbs in (9) above), and only if these occur in affirmative, non-modalized contexts (de Haan/Weerman 1986). The latter has been related to the fact that V2 *dat*-clauses assert the existence of the embedded proposition as true, which is blocked if the matrix predicates is in the scope of a negative or modal operator.

To check this, we scanned the matrix structures in the list of hits produced by (Q2) in Appendix 2. We see that examples involve the verb *to know*, which is a well-known licenser of limited embedded V2 in the respective VO languages. But at the same time, we observe that the verb *know* in the examples is embedded under a modal verb *sc(h)olen* ‘should’, and this is a problem for the CP-recursion analysis. Note, however, that in the examples at hand, the modal verb is used to express a request, i.e., it appears in its root/deontic meaning. From the respective contexts we infer that the author has no doubt regarding the truth of the embedded proposition.

5. Conclusion

In this study, we showed how with the help of a corpus search, we can contribute to solving a long lasting debate in comparative Germanic syntax, namely the question regarding the status of CP recursion in Germanic OV languages. In searching an electronic corpus of a language stage ignored in previous research, we were able to provide conclusive evidence for limited embedded verb second in MLG, an OV stage of Germanic. This result clearly suggests that CP recursion is neither excluded nor exceptional in Germanic OV languages.

Note that the corpus available for search is a very small one. Petrova (2012: 165), on the basis of a manual search (in samples) from other texts, has found evidence suggesting that orders conclusive for limited embedded V2 are also found in *dat*-clauses selected by non-modalized, canonical affirmative predicates, cf. (13):

- (13) *De keiser der tatharen entschuldeghede sik dat*
 the Emperor of the Tatars apologized REFL that
uppe de tyd kunde he eme nichte helpen
 up to that time could he him NEG help-INF
 ‘The Emperor of the Tatars apologized that he had not been able to help him until then’
 (LChr I 147, 11)

The study highlights the relevance of the MLG attestation in comparative Germanic research and emphasizes the need for providing electronically searchable corpora of this language stage.

Appendix 1

1. antwerde den turken vnde bat [dat] se em gheuen vryst dre daghe dat se syk darvnder beredden
2. de vormunder synen vlyt daran [dat] he stedeliken werschopede vnde blide was vnde vrolich
3. stryde vnde bat de torken [dat] se ghinghen in syne dorntze de aldorghen slotaftich was vnde
4. de dar mede west weren [dat] se [so] vele gudes vnde roues ghenamen hadden dat se dat gud myt
5. baden vore laden vnde sprack [dat] de brodere gherne vrede vnde ghelofte maken wolden vnde holden myt den torken

6. Nu schole gy vort weten [dat] in Rodo is vele schones
hilghedomes dar is eyn cruce dat
7. vmmelang belegghen syn Me lyst [dat] Iaphet Noes sone de
erste was de den werder besettede vnde
8. wol ghehort van guden luden [dat] vnder des hemmels trone
nen schoner nen eddeler nen wunliker cleynat were dat god
deme mynschen to
9. deren Nu schole gy weten [dat] in Cipro synt de eddelsten
vnde de besten vnde de rykesten vorsten vnde de eddelsten
heren baronen ryddere vnde borgher de de werlt heft de
10. Dat heft me dar vore [dat] se sunte Lucas ghemalet hebbe
nach vnser vrouwen formen vnde de figure Dor werdicheyt
vnde ere des
11. Allexandria Nu schole gy weten [dat] alle desse stede de
hir vor ghenomet zyn sint nicht ene dachuart van Cipro Vnde
Allexandria is de erste
12. kan Nu scole gy weten [dat] dat hilghe lant is dat land des
loftes dat god van hemmel louet
13. rosten Gy scholet ok weten [dat] in der stad wonet de
keyser der greken [vnde patriarchen] dar se van holden alzo
14. hilghen lande wor me wyl [dat] nenes seghelendes not is
Ouerst eyn ander wech is
15. gy scholet dat vorwar weten [dat] in der stad to Venedyge
nerne eyn steynen pyler edder nerne eyn grot houwen sten is
he en sy van Troya
16. boke der martere sunte Aghaten [dat] de lude de in der
[stad] weren
17. droghen yeghen dat vur vnde [dat] vur vorghink dorch de
werdicheyt sunte Aghaten Van dem berghe Vilkanus Noch
18. vnde scholet dat vorware weten [dat] in der ieghennoten des
meres mer wen souen hondert werder sint mynner vnde groter
etlike myt luden bewonet etlike al woste Pathera was vor
iaren alto
19. werder Rodym Gy scholet weten [dat] Rodijs is eyn alto
schone land eft werder vnde vullen lustich vnde is sund
vnde was
20. Do de turken erst vornemen [dat] de brodere sunte Johannis
Rodem den werder syk vnderdanich ghemaket hadden do sammel-
den se syk myt

Appendix 2

1. stryde vnde bat de torken [dat] se ghinghen in syne dorntze de aldorghen slotaftich was vnde
2. Nu schole gy vort weten [dat] in Rodo is vele schones hilghedomes dar is eyn cruce dat
3. deren Nu schole gy weten [dat] in Cipro synt de eddelsten vnde de besten vnde de rykesten vorsten vnde de eddelsten heren baronen ryddere vnde borgher de de werlt heft de
4. Allexandria Nu schole gy weten [dat] alle desse stede de hir vor ghenomet zyn sint nicht ene dachuart van Cipro Vnde Allexandria is de erste
5. kan Nu scole gy weten [dat] dat hilghe lant is dat land des loftes dat god van hemmel louet
6. rosten Gy scholet ok weten [dat] in der stad wonet de keyser der greken [vnde patriarchen] dar se van holden alzo
7. werder Rodym Gy scholet weten [dat] Rodijs is eyn alto schone land eft werder vnde vullen lustich vnde is sund vnde was

Appendix 3

1. Nu schole gy vort weten [dat] in Rodo is vele schones hilghedomes dar is eyn cruce dat
2. deren Nu schole gy weten [dat] in Cipro synt de eddelsten vnde de besten vnde de rykesten vorsten vnde de eddelsten heren baronen ryddere vnde borgher de de werlt heft de
3. rosten Gy scholet ok weten [dat] in der stad wonet de keyser der greken [vnde patriarchen] dar se van holden alzo

Primary texts

- [LChr] Die lübeckischen Chroniken in niederdeutscher Sprache, ed. Ferdinand Heinrich Grautoff, 1829-1830. Hamburg: Perthes & Besser.
- [LS] Ludolfs von Sudheim Reise ins Heilige Land. Nach der Hamburger Handschrift, ed. Ivar Stapelmohr, 1937. Lund: C.W.K. Gleerup.
- [SW] Sächsische Weltchronik, ed. Ludwig Weiland, 1980. Unveränderter Nachdruck der Ausgabe 1877. München: Monumenta Germaniæ Historica.

References²

- Axel, Katrin (2007): *Studies in Old High German Syntax: left sentence periphery, verb placement and verb-second*. Amsterdam/Philadelphia: John Benjamins.
- Breitbarth, Anne (2009): A hybrid approach to Jespersen's Cycle in West Germanic. In: *Journal of Comparative Germanic Linguistics* 12: 81-114.
- Freywald, Ulrike (2008): Zur Syntax und Funktion von *dass*-Sätzen mit Verbzweitstellung. In: *Deutsche Sprache* 36: 246-285.
- Haan, Germen de/Weerman, Fred (1986): Finiteness and verb fronting in Frisian. In: Haider, Hubert/Prinzhorn, Martin (eds.): *Verb second phenomena in Germanic languages*. Dordrecht: Foris, 77-110.
- Haider, Hubert (2000): OV is more basic than VO. In: Svenonius, Peter: *The derivation of VO and OV*. Amsterdam: John Benjamins, 45-67.
- Haider, Hubert (2010): *The Syntax of German*. Cambridge: Cambridge University Press.
- Kemenade, Ans van (1997): V2 and embedded topicalization in Old and Middle English. In: Kemenade, Ans van/Vincent, Nigel (eds.): *Parameters of morphosyntactic change*. Cambridge: Cambridge University Press, 326-352.
- Peters, Robert (2000): Soziokulturelle Voraussetzungen und Sprachraum des Mittelniederdeutschen. In: Besch, Werner/Betten, Anne/Reichmann, Oskar/Sonderegger, Stefan (eds.): *Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung*. Vol. 2.2. Second edition. Berlin/New York: de Gruyter, 1409-1422.
- Petrova, Svetlana (2012): Multiple XP-fronting in Middle Low German root clauses. In: *Journal of Comparative Germanic Linguistics* 15(2): 157-188.
- Petrova, Svetlana (2013): *The Syntax of Middle Low German*. Habilitation thesis, Humboldt-Universität zu Berlin.
- Pintzuk, Susan (1993): Verb seconding in Old English: verb movement to Infl. In: *The Linguistic Review* 10: 5-35.
- Reis, Marga (1997): Zum syntaktischen Status unselbständiger Verbzweit-Sätze. In: Dürscheid, Christa/Ramers, Karl Heinz/Schwarz, Monika (eds.): *Sprache im Fokus. Festschrift für Heinz Vater zum 65. Geburtstag*. Tübingen: Niemeyer, 121-144.
- Rösler, Irmtraud (1997): *Satz – Text – Sprachhandeln. Syntaktische Normen der mittelniederdeutschen Sprache und ihre soziofunktionalen Determinanten*. Heidelberg: Winter.
- Schmidt, Thomas/Wörner, Kai (2009): EXMARaLDA – Creating, Analysing and Sharing Spoken Language Corpora for Pragmatic Research. In: *Pragmatics* 19(4): 565-582.

² All URLs have been checked and found valid as of late January 2015.

- Suchsland, Peter (2000): ... *ibu dū mi ēnan sagēs, ik mi dē ðdre uuēt*. Zur Syntax des Hildebrandliedes. Eine Fallstudie. In: Haustein, Jens/Meineke, Eckhard/Wolf, Norbert Richard (eds.): *Septuaginta quinque*. Festschrift für Heinz Mettke. Heidelberg: Winter, 355-379.
- Vikner, Sten (1995): *Verb movement and expletive subjects in the Germanic languages*. New York/Oxford: Oxford University Press.
- Vikner, Sten (2001): *Verb movement variation in Germanic and optimality theory*. Unpublished habilitation thesis, University of Tübingen.
- Zeldes, Armir/Ritz, Julia/Lüdeling, Anke/Chiarcos, Christian (2009): ANNIS: A search tool for multi-layer annotated corpora. In: Mahlberg, Michaela/González-Díaz, Victorina/Smith, Catherine (eds.): *Proceedings of the Corpus Linguistics Conference, CL2009*, University of Liverpool, UK, 20-23 July 2009. http://ucrel.lancs.ac.uk/publications/cl2009/358_FullPaper.doc.

A living text archive of 15th-19th-century German

Corpus strategies, technology, organization

Abstract

The corpus situation for the study of Early New High German (ENHG) and the older stages of New High German (oNHG) is far from satisfying. Indeed, there is no integrated and balanced corpus of a sufficient size that is publicly available. In this paper we give an outline of a living text archive of ENHG and oNHG in a distributed infrastructure that is supposed to fill this gap in the coming years. Such an archive is designed as a collaborative, sustainable and interoperable platform where historical texts are integrated together with relevant metadata and expert computational technology. We briefly mention examples of research questions and of corpus building scenarios and comment on the organizational aspects of such an archive including its potential as a basis for reference corpora for specific purposes.

1. Introduction

The notion of a corpus is intimately linked to three basic ideas: first, the idea of a corpus as a balanced or even representative sample of a language, a language stage, a variety of a language or a specific form of language use; second, the idea that a corpus is designed in order to fulfill a specific research question on an empirical basis; third, the idea that different research questions, purposes and layers of investigation (like semantics, morphology, textual organization) may require very different types of corpora in respect of size and composition. When Hoffmann (1998) gave his overview on historical corpora of German, there were not many electronic items available. Today, historical language corpora are usually digitized collections and historical corpus linguistics is closely connected with the use of computational tools.

However, the corpus situation for the study of Early New High German (roughly 15th-17th cc.; ENHG) and the older stages of New High German (roughly 17th-19th cc.; oNHG) is far from satisfying. There is no integrated and balanced corpus of a sufficient size that is publicly available.¹ Small or middle

¹ The Bonn Corpus of Early New High German (www.korpora.org/Fnhhd/) is balanced as to parameters of time, region and text types; however, it is relatively small and was built and encoded specifically for the analysis of grammatical aspects.

sized subcorpora of research projects are usually not made available to other researchers. Up to now, there is no platform and no established culture of publicly sharing ENHG and oNHG resources among historians of the German language. Moreover, the electronic data that are produced for the publication of printed editions are not normally further processed for corpus usage, and in many cases, digital rights are given away to publishing houses without negotiation of moving wall solutions for the later use of the text in core corpora of the German language. On the other hand, there *are* many electronic texts on the hard drives of individual scholars and out there on the web, but they come in a huge diversity of technical formats and transcription schemas, so that it is hardly possible to work with these resources in a systematic way.

In addition to ongoing corpus projects funded by the Deutsche Forschungsgemeinschaft (German Research Council),² a curation project within the CLARIN-D infrastructure project³ funded by the Bundesministerium für Bildung und Forschung (Federal Ministry for Education and Research) aimed to establish a distributed infrastructure for what we call a *living text archive* of ENHG and oNHG. Such an archive is designed as a collaborative, sustainable and interoperable platform where historical texts are integrated together with relevant metadata and expert computational technology. While the profile of such an archive will be ‘opportunistic’, it will be possible to choose specific subcorpora in a systematic way by using metadata on linguistic variation parameters. In a first round, an inventory of suitable legacy resources was compiled and, as a feasibility study, more than 78,000 pages were curated and integrated into the infrastructure, covering a wide range of text types, topics, dates of publication etc. This forms the basis for a long term curation initiative in a distributed platform that supports corpus based research on historical texts in general.

² Referenzkorpus Frühneuhochochdeutsch (Reference Corpus of Early New High German), see: <http://gepris.dfg.de/gepris/projekt/200609649> (10/05/2014); Deutsches Textarchiv (German Text Archive), see: www.deutschestextarchiv.de. All URLs quoted were last accessed on May 5, 2014.

³ On CLARIN-D see www.clarin-d.de. On the curation project see www.deutschestextarchiv.de/clarin_kupro and <http://de.clarin.eu/de/fachspezifische-arbeitsgruppen/f-ag-1-deutsche-philologie/kurationsprojekt-1>.

In this paper we give an outline of such a living text archive of ENHG and of oNHG in a distributed infrastructure (section 2), we briefly mention examples of research questions and of corpus building scenarios (section 3), and comment on the technical (section 4) and the organizational aspects (section 5) of such an archive including its potential as a basis for reference corpora for specific purposes. Technical aspects and aspects of workflow are described in more detail in the paper by Thomas and Wiegand (this volume).

2. The idea of a living text archive of Early New High German and older New High German

The basic idea of the CLARIN-D curation project (Kurationsprojekt 2012; Thomas/Wiegand, this volume) was to ‘feed’ the existing infrastructures at the Deutsches Textarchiv (DTA) of the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW)⁴ and at the Digital Library of the Herzog August Bibliothek (Wolfenbüttel)⁵ with historical texts from the 15th to the 19th centuries that are already available: (i) in collections like wikisource, (ii) in the legacies of research projects, (iii) from individual scholars and (iv) scattered across different places on the web. Such texts had first to be evaluated with respect to their quality. The ones to be integrated into the repository had to be enriched by metadata (e.g. on the transcription schema, on the degree of accuracy, and on linguistic variation parameters), and finally they had to be brought into a TEI compliant format. The texts are now available in a distributed infrastructure at the BBAW, the HAB and in the near future at the Institute for the German Language (IDS, Mannheim), both for federated search and for download under a creative commons license.

The aim of the curation project was threefold: (i) to produce an inventory of available electronic texts in our time frame; (ii) to provide an evaluation and integration infrastructure that can be used on a long term basis; (iii) to curate and to integrate 35,000 pages of historical corpus texts into the respective infrastructures in a ‘first round’. At the end of the project, more than 78,000 pages were integrated. The result of this integration of digital texts from the 15th to the 19th centuries is not yet a balanced reference corpus but rather an opportunistic repository. However, the use of metadata on linguistic variation parameters (e.g. date, place of publication, subject field, text type) will allow

⁴ www.deutschestextarchiv.de.

⁵ www.hab.de/de/home/bibliothek/digitale-bibliothek-wdb.html.

the building of specific subcorpora according to criteria of selection and their combination (see section 4.3, sampling procedure).

In the medium term the infrastructures both at the BBAW and the HAB will be used for ongoing work on the *living archive* of historical German from the 15th to the 19th centuries. Basically, work on the enhancement of the archive will consist in a number of different kinds of activities:

- adding new texts and new text groups according to criteria like time zone (decades), subject field (e.g. balneology), text type (e.g. newspaper reports), gender, the manuscript/print ratio, and others;
- adding new and/or more fine-grained metadata, e.g. with respect to subject fields or historical text types (cf. Budin/Kabas/Mörth 2012);
- improving aspects of balance/representativeness in a systematic way, e.g. in respect of the time zones (decades, centuries), the text types, the topics and subject fields, the gender ratio and other criteria of corpus balance;
- improving the quality of texts: while there are already numerous high quality texts in the Deutsches Textarchiv, we intend to include in the archive a number of working transcriptions as well. This means that together with image data of the underlying prints or manuscripts it will be possible to do serious work with the texts even if there may be a few remaining transcription errors that will be corrected in the course of the work.

Apart from these dynamic activities on core aspects of the texts and of the text collection as a whole we provide the opportunity of adding or connecting new (layers of) linguistic information and the results of other kinds of scientific investigation with the texts. This possibly includes introducing new (types of) metadata categories according to specific research questions and according to the results of specific research projects. E.g., cataloguing texts that belong to a certain historical controversy has not been one of the core aspects of corpus architecture; it is, however, an important aspect of Early Modern intellectual history. Therefore, such kinds of further information from specific research projects should be integrated as well.

3. Research questions and corpus building scenarios

Empirical work on specific research questions in the history of the German language must be based on suitable subcorpora. The design of specific subcorpora from a large textual archive for specific purposes requires both fine-grained metadata and knowledge about the history of the German language, its text types in a historical perspective, a broad overview on the historical development of the German language and sound knowledge of factors of language change and linguistic evolution. We will now illustrate this point with some examples.

- (i) *Word usage and historical semantics.* Assume that a historical lexicographer is working on an article, say on the history of the foreign word *Influenz* (Engl. *influence*) or on the adjective *billig* (Engl. *cheap*). The word *Influenz* is interesting for its early usage in a cosmological, astrological sense. The word *billig* is particularly interesting for its semantic development with its shifts from the senses ‘fair, proper’ to ‘cheap, not expensive’ and then to ‘of poor quality’. Here the interesting question is in which contexts these shifts took place and how common knowledge (a fair price is a low price; a cheap product is usually of poor quality) fostered specific understandings that eventually were generalized and conventionalized. In this situation a lexicographer might be interested to see all the quotations and in further steps to filter the quotations according to centuries or textual groups. In this use case, the production of KWIC concordances based on the whole archive, on temporal subcorpora (for *billig*) or on subcorpora for specific subject fields like alchemy, astronomy and cosmology (for *Influenz*) allows to describe the historical semantics of these words more precisely along important evolutionary parameters of word usage. The example shows that metadata (time, subject fields) are technical instruments that have to be ‘geared’ by knowledge or by hypotheses about linguistic dynamics.
- (ii) *Linguistic profiles of historical text types in a synchronic and diachronic perspective.* How did newspaper reports, culinary recipes, printed sermons, medical case studies, technical descriptions of machines, literary translations and many other text types look like around 1600, around 1700, around 1800, around 1900? How did they evolve over time? What are the (changing) principles of textual organization? Are there typical syntactic patterns? What are the basic aspects of organization of their lexi-

con? In order to treat these types of questions, the metadata of the living archive will enable users to choose subcorpora for specific text types. They allow for straightforward searches for those aspects that can be ‘translated’ to purely form-based queries (e.g., show all quotations for *hat man*⁶ in 17th-century newspapers in order to find a news-specific type of construction; or: show all word formations in *-ung* together with their spelling variants that come from texts marked for the subject field ‘politics’). For other questions, the annotation of subcorpora will be the way to go.

- (iii) *The search for and automatic retrieval of specific constructions and grammatical phenomena* is not a trivial matter. Depending on the complexity of the research question, word based searches (e.g., *werden* for the German *werden* passive) will be helpful. In other cases (e.g., the structure of the NP in different historical text types of German), suitable subcorpora will facilitate further research on algorithms for automatic retrieval and analysis.
- (iv) *The modelling of linguistic change and of the processes of linguistic dissemination and evolution* requires large historical corpora. Past experiences, e.g. in the history of modal verbs, show that word senses are often specific to certain domains and that the modelling of paths of change requires very large and richly diversified corpora.
- (v) *Other use cases* include research on historical terminology development, on the evolution of spelling variants, on the history of morphological forms and systems, on the language use of certain groups of persons (e.g. women), on the connection of language use and the history of ideas (e.g. pietism, rise and fall of phlogiston theory), the language of famous authors, and many others.

It is one of our aims in the future not only to contribute further texts together with their metadata to the distributed CLARIN-D infrastructure, but also to delineate a number of typical use cases that show how the texts, the metadata and forms of annotation can be put to use in order to work on specific research questions in the history of the German language.

⁶ A literal english translation für *hat man* is *has one/one has*. Formulae of the type *Aus Prag hat man, dass ...* were common in Early New High German newspapers to indicate that news were based on some kind of source.

4. Technical aspects

As stated above a distributed infrastructure for what we call a *living text archive* of ENHG and oNHG should be designed as a collaborative, interoperable and sustainable platform where historical texts are integrated together with relevant metadata and expert computational technology.

Since it is beyond the scope of this paper to discuss all technical aspects in detail, we will focus on those aspects that seem crucial for the collaborative aspect of the envisioned infrastructure and mention other aspects only briefly. In the following, we discuss three central technical components: an agreed standard for corpus encoding (4.1), a web-based infrastructure for ensuring quality assurance, for the integration of legacy data and for further annotation of data (4.2), and a sampling procedure that extracts a reference corpus from the entire document collection (4.3). Furthermore, historical texts pose specific problems to full text retrieval. This is due to the absence of consistent orthographic conventions in historical texts, which presents difficulties for any system requiring reference to a fixed lexicon. These issues as well as software to overcome these difficulties are addressed, e.g., in Gotscharek et al. (2009) and Jurish (2010).

4.1 An agreed standard for corpus encoding

The set of annotation schemes developed by the Text Encoding Initiative (TEI; Burnard/Bauman 2012) is more and more wide-spread in current corpus projects and is going to become a de facto standard of corpus encoding for historical texts. Large infrastructure projects such as CLARIN-D and DARIAH recommend this de-facto standard. In its current version, TEI-P5 is a very flexible scheme that is adoptable for a large variety of text types. Due to this flexibility, TEI-P5 compliant corpora are generally not interoperable per se.⁷ However, interoperability of corpora is one major backbone of a collaborative infrastructure on several levels. On the level of metadata, interoperability assures that texts can be uniformly processed and stored in central databases, thus providing a larger visibility of the corpus data. Standards such as OAI-PMH⁸ are available for metadata exchange and have been adopted by many infrastructure projects. For object data, a common encoding facilitates the

⁷ For a discussion of some of the problems arising from this fact cf. Unsworth (2011).

⁸ Open Archives Initiative Protocol for Metadata Harvesting, a standard for the interchange of metadata.

exploitation of the text for further computational processing such as text mining or the annotation of the text with syntactic information. It also facilitates the common indexing of texts: texts encoded in a common way are searchable via search engines or federated search interfaces without further conversion work. Last but not least, texts with a common metadata and object data encoding can directly be stored in the repository of a larger corpus infrastructure, thus making corpus data sustainable.

How can the interoperability of corpus data be assured? First of all, it is important to state that the willingness to share data, to provide it with a license that enables the reuse of corpus data, is a prerequisite to the technical notion of interoperability. Second, interoperability of corpus data is comparatively new to corpus encoding. Up to the end of the 1990s, corpus compilation on the basis of the TEI was mainly a project-specific activity. Corpus documents were validated against a project-specific document grammar, possibly private character encodings were used, and the documents were transformed into proprietary formats in order to be indexed for full text retrieval. In that era of project-specific encoding, exchange of documents across projects was no goal per se and, in general, character encoding problems as well as differences in the document type grammar (DTD) were obstacles to a broader exchange of data. With the advent of XML and Unicode, documents encoded according to the recommendations of the TEI became interchangeable, but the problem of a lack of interoperability persisted due to the complexity and flexibility of TEI-P5. More recently, several attempts were made to increase the interoperability among different document collections by creating common formats. Subsets of TEI-P5 were created in such a way that the number of elements was largely reduced with respect to the full set of elements of the TEI Guidelines.⁹ Also, the number of attributes and their corresponding values were restricted in order to obtain a better control of documents encoded in that format. Such formats – technically expressed as XML schemas – should allow for a basic structuring of all written texts and therefore serve as a starting point from which more detailed, possibly project-specific text structuring could start.

⁹ The TEI recommends the definition of a subset of TEI elements appropriate to the anticipated needs of the project rather than to base the annotation of a corpus on the whole TEI tagset (Burnard/Baumann 2012: ch. 15.5) and promotes formats like TEI Tite (Trolard 2011), TEI Lite (Burnard/Sperberg-McQueen 2012) or the Best Practices for TEI in Libraries (TEI SIG on Libraries 2011). Other formats, such as TEI Analytics (Unsworth 2011, Pytlik-Zillig 2009), IDS-XCES (Institute for the German Language, Mannheim) and Textgrid's Baseline Encoding for Text Data in TEI P5 (Textgrid 2007-2009) were created.

The base format of the DTA project (henceforth DTABf) is a TEI-P5 subset that ensures the interoperability of corpus texts (cf. Geyken et al. 2012b). DTABf draws on the experiences of the DTA where a selection of 1,300 important texts of different text types (fictional, functional, and scientific texts), originating from the 17th to the 19th century, is currently being digitized and annotated. Linguistic analyses are added to the digitized text sources in a stand-off format for further corpus research. The tagset of the DTABf is a strict subset to the TEI-P5 guidelines, i.e., no new elements or attributes were added to the TEI-P5 tagset. It consists of about 80 TEI-P5 elements needed for the basic formal and semantic structuring of the DTA reference corpus. The purpose of the DTABf is to provide a faithful page per page presentation of the entire works and to maintain coherence on the annotation level (i.e., similar structural phenomena should be annotated similarly).

The DTABf attempts to meet the criteria of interoperability mentioned by Unsworth (2011) in that it “focuses on non-controversial structural aspects of the text and on establishing a high quality transcription of that text”. Therefore, the goal of the DTABf is to provide as much expressiveness as necessary by being as precise as possible. For example, DTABf is restrictive not only considering the selection of TEI-elements but also with respect to attribute-value pairs, and allows only a limited set of values for a given attribute. Unlike initiatives such as TEI Analytics (as presented in Pytlik-Zillig 2009), the goal of DTABf is not to build a schema that validates as many cross-collections as possible but to convert resources from other corpora so as to keep the structural variation as small as possible.

The necessity of a common standardized format for the annotation of printed texts seems to be opposed to the fact that different projects usually have different needs as to how a corpus may be exploited. Therefore annotation practices vary according to the variable queries on a certain corpus. This problem may be addressed by defining different *levels of text annotation* that represent different text structuring depths. The TEI Recommendations for the Encoding of Large Corpora foresee four different levels of annotation defining required, recommended, optional, and proscribed elements.¹⁰

The DTABf consists of such annotation levels, which serve as classes subsuming and, by that, categorizing all available ‘base format’ elements:

¹⁰ Cf. www.teic.org/release/doc/tei-p5-doc/en/html/CC.html.

- Level 1/required: elements that are mandatory for the basic semantic structuring of corpus texts.
- Level 2/recommended: elements that are recommended for the semantic structuring of corpus texts. These elements are systematically used in the DTA-corpus.
- Level 3/optional: elements that need not be considered for text annotation. Level 3 elements are not (yet) part of the DTA guidelines and are therefore not used systematically in the texts of the DTA corpus. They are, however, compatible with the DTA schema.
- Level 4/proscribed: elements that were explicitly excluded from the DTA guidelines. They should be avoided in favour of the solutions offered in the DTA guidelines.

4.2 An infrastructure for the lifecycle of a digital text

The second step towards an interoperable corpus platform is, from a technical point of view, to provide software enabling an easy integration of legacy data into the DTABf as well as to correct the texts in case their quality does not meet the criteria established by the community.

DTABf is a flexible format in the sense that it consists of mandatory, recommended, and optional criteria. Hence, software that validates legacy data against DTABf can do this on those three levels. Thus, legacy texts can be integrated in the platform even though they are possibly not DTABf compliant on all levels. Currently, two generic conversion tools are provided in the CLARIN-D context: a generic web-based software (TEI-Integrator, Th. Eckart, Univ. Leipzig¹¹) that assists the user in the conversion and the upload process of legacy data, and a specific oXygen-Framework where any TEI-P5 compliant text can be validated against the DTABf and a GUI is used to assist the user in evaluating the amount of work needed to convert the legacy document into a valid DTABf document.¹²

Since it can be expected that legacy data from heterogeneous origins do not meet all the required criteria, a collaborative platform is needed where uploaded texts can be proofread and evaluated. Since generic collaborative proof-reading platforms did not exist for TEI-P5 texts, such a platform was

¹¹ <http://clarin.informatik.uni-leipzig.de/program>.

¹² <http://lecture2go.uni-hamburg.de/konferenzen/-/k/13952>.

implemented for the DTA (DTAQ).¹³ Apart from the proof-reading facilities where the user has the possibility to check the text for errors against the image page per page, the user is also provided with several presentation formats of the text: the original XML/TEI format, a formatted HTML presentation, a pure text format, and a normalized view of the text after being processed by a lemmatizer of historical German word forms (Jurish 2010).

The third issue of the corpus management infrastructure concerns the fact that an electronic document is always ‘living’, i.e., it is always subject to further correction and structural or semantic annotation. From the point of view of the ‘living archive’ it is important to note here that additional annotations are not only carried out for a specific research but can also be valuable for other researchers. Therefore, these additional annotations (as far as the DTABf is concerned) should be carried out ‘within’ the archive, thus enabling other researchers to benefit from previous annotations. In addition, such a platform should integrate computational tools such as Named-Entity-Recognizers, tools for statistic computations, or tools for the automatic analysis of citations. Such tools are currently in preparation in the CLARIN-D infrastructure and will be accessible to the DTAQ platform as web-services. Finally, the versioning of the documents as well as the use of persistent identifiers play a role here since stable references to the documents are a necessary requirement for any sustainable infrastructure. Solutions for these problems exist and are currently implemented on a larger scale for infrastructure projects such as CLARIN¹⁴.

4.3 Corpus chooser

The result of the previous steps is an interoperable distributed corpus repository annotated with more or less finegrained metadata information. This repository is not per se a reference corpus. As stated above, there may be more than one reference corpus at hand depending on the research purpose. In order to establish a subset of the repository that qualifies as a reference corpus for a given research task, sampling procedures are needed that exploit the metadata in such a way that an optimal subcorpus of the repository can be extracted. There are at least two possible ways to deal with these sampling procedures: either as a corpus bitmap where any user can choose the appropriate “virtual” subcorpus for her or his research (Kupietz et al. 2010), or as a

¹³ www.deutschestextarchiv.de/dtaq.

¹⁴ www.clarin.eu/files/pid-CLARIN-ShortGuide.pdf.

complex set of criteria chosen by the user beforehand, which is then computed by a sampling procedure trying to find an optimal subcorpus for these criteria (cf. Geyken 2007).

We illustrate this method with the example of building a maximally balanced corpus out of the entire text collection that is equally distributed over all the decades of the 15th to the 19th century. The balance of the selection follows a previously determined distribution of text types. In order to determine the maximally balanced corpus, we identify the decade with the least number of tokens. The sampling process then calculates the expectation value according to the text type distribution. For practical reasons, this algorithm cannot be applied strictly since the decades are not equal from the 15th to the 19th century; e.g., in the period of the Thirty Years War there was a much sparser text production than in the 1890s. Therefore we assume that the token number for some decades can deviate from the mean value. The deviation depends on the difference between the minimally balanced corpus and the corpus size to be attained. Moreover, the sampling procedure begins with a so-called 'initial corpus'. The initial corpus consists mainly of texts by major writers and scientists as well as texts that are considered to be of high interest. This guarantees that works by Goethe, Marx, Büchner, Boltzmann or Liebig will be present independently of the sampling procedure whereas, for example, serial corpus texts such as newspaper articles or exchange of letters are selected randomly. Of course, the initial corpus must not exceed the boundaries of the required text distribution, i.e., it will have to be ensured beforehand that no decade and no text class in the start corpus exceed the required token number. Further sampling strategies will be demonstrated in the future with respect of specific use cases.

5. Organizational aspects of collaboration on a living text archive

A living text archive basically needs (i) contributors who are willing to share, and to work on, textual resources, and (ii) an infrastructure with a core team whose members organize collaboration on a long term basis.

- (i) From a contributor's point of view, the main question is why she or he should share textual resources. At present, there are three partial answers to this question.

- We aim to create a system of ‘reputation’ for making resources available, including personal ascription in the metadata of work that has been donated to the archive and the possibility to produce contributor profiles. The success of such mechanisms depends on the importance that is granted to these contributions to infrastructures in comparison to traditional publications.¹⁵
- We are able to display the texts that have been given to the archive *locally* via a JSON interface. In this way texts can be consulted both in the large archive *and* on the websites of research teams, universities etc. in a professional way, together with expert corpus technology.
- There is also a moral aspect to this issue: If we as researchers want *huge* corpora, we all have to contribute by integrating into a common infrastructure the data many of us have produced. On the other hand, sharing one’s resources is not only altruistic but is also a way to make this work visible to others and to receive recognition for it. Thus, sharing produces a win/win-situation.

In addition, to share texts together with different annotations allows to combine research aspects beyond the interests of the original annotators. E.g., to project a syntactic annotation of text X by person A onto an annotation of textual structures of text X by person B allows to describe typical syntactic structures of textual elements in a given text of a certain text type. At present, this type of collaborative use of different annotations is rare in research on the history of German.

- (ii) A new culture of sharing and collaborating in a living text archive needs infrastructure centers with core teams whose members organize collaboration on a long term basis. Among the ongoing tasks of such a core team are: to catalogue newly available digital resources in a canonical way; to evaluate and to acquire new digital texts; to support and advice in conceptual, technical, and legal matters; to adjust the principles of work to new developments; to take care of and to advance corpus technology for his-

¹⁵ The German Council of Science and Humanities (Wissenschaftsrat) is encouraging the scientific community to count contributions to digital infrastructures as ordinary publications: “Auch um die Behebung des Reputationsdefizits ist der Wissenschaftsrat bemüht, wenn er anregt, Infrastrukturarbeit als eigenständige Forschungsleistungen nicht anders als Publikationen zu bewerten. Bei der Besetzung wichtiger Posten sollen technische und wissenschaftliche Qualifikation von gleichem Gewicht sein.” (www.faz.net/aktuell/feuilleton/forschung-und-lehre/digital-humanities-eine-empirische-wende-fuer-die-geisteswissenschaften-11830514.html).

torical texts; to provide best practices, use cases and strategies for the use of subcorpora for specific research questions; etc. We hope that the CLARIN-D curation project will demonstrate the fruitfulness of such a long term living archive of ENHG and oNHG.

References

- Budin, Gerhard/Kabas, Heinrich/Mörth, Karlheinz (2012): Towards finer granularity in metadata. Analysing the contents of digitised periodicals. In: *Journal of the Text Encoding Initiative* 2 (February 2012): 1-8. <http://jtei.revues.org/416>.
- Burnard, Lou/Bauman, Syd (2012): P5: Guidelines for electronic text encoding and interchange, Version 2.1.0, June 17th, 2012. www.tei-c.org/release/doc/tei-p5-doc/en/html/.
- Burnard, Lou/Sperberg-McQueen, Michael C. (2012): TEI Lite: Encoding for Interchange: an introduction to the TEI. Final revised edition for TEI P5, August 2012. www.tei-c.org/Guidelines/Customization/Lite/.
- Geyken, Alexander (2007): The DWDS corpus: A reference corpus for the German language of the 20th century. In: Fellbaum, Christiane (ed.): *Collocations and idioms: linguistic, lexicographic, and computational aspects*. London: Continuum, 23-41.
- Geyken, Alexander/Gloning, Thomas/Stäcker, Thomas (2012a): Compiling large historical reference corpora of German: quality assurance, interoperability and collaboration in the process of publication of digitized historical prints. Panel DH2012, Hamburg. [Abstract and video lecture].
- Geyken, Alexander/Haaf, Susanne/Wiegand, Frank (2012b): The DTA 'base format': A TEI-subset for the compilation of interoperable corpora. In: Jancsary, Jeremy (ed.): *11th Conference on Natural Language Processing (KONVENS): Empirical Methods in Natural Language Processing. Proceedings of the Conference on Natural Language Processing 2012*. (= *Schriftenreihe der Österreichischen Gesellschaft für Artificial Intelligence* 5). Wien: ÖGAI, 383-391. www.oegai.at/konvens2012/proceedings.pdf#page=383.
- Geyken, Alexander/Gloning, Thomas/Kupietz, Marc/Stäcker, Thomas/Thomas, Christian/Witt, Andreas (2012c): Integration und Aufwertung historischer Textressourcen des 15.-19. Jahrhunderts in einer nachhaltigen CLARIN-Infrastruktur, Vorhabensbeschreibung für ein Kurationsprojekt der F-AG 1 Deutsche Philologie. www.deutschestextarchiv.de/doku/clarin_kupro_index.
- Geyken, Alexander/Haaf, Susanne/Jurish, Bryan/Schulz, Matthias/Thomas, Christian/Wiegand, Frank (2009): TEI und Textkorpora: Fehlerklassifikation und Qualitätskontrolle vor, während und nach der Texterfassung im Deutschen Text-

- archiv. In: *Jahrbuch für Computerphilologie* 2009. <http://computerphilologie.tu-darmstadt.de/jg09/geykenetal.html>.
- Gotscharek, Annette/Neumann, Andreas/Reffle, Ulrich/Ringlstetter, Christoph/Schulz, Klaus (2009): Enabling information retrieval on historical document collections: the role of matching procedures and special lexica. In *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data*. New York: ACM, 69-76.
- Haaf, Susanne/Wiegand, Frank/Geyken, Alexander (2012): Measuring the correctness of double-keying: Error classification and quality control in a large corpus of TEI-annotated historical text. In: *Journal of the Text Encoding Initiative* 4. <http://jtei.revues.org/pdf/739>.
- Hoffmann, Walter (1998): Probleme der Korpusbildung in der Sprachgeschichtsschreibung und Dokumentation vorhandener Korpora. In: Besch, Werner/Betten, Anne/Reichmann, Oskar/Sonderegger, Stefan (eds.): *Sprachgeschichte*. Zweite, vollständig neu bearbeitete und erweiterte Auflage. Volume 1. Berlin/New York: de Gruyter, 875-889.
- Jurish, Bryan (2012): Finite-state canonicalization techniques for Historical German. PhD thesis, Universität Potsdam. http://opus.kobv.de/ubp/volltexte/2012/5578/pdf/jurish_diss.pdf.
- Jurish, Bryan (2010): More than words – using token context to improve canonicalization of Historical German. In: *Journal for Language Technology and Computational Linguistics (JLCL)* 25/1: 23-39.
- 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 2010)*. Paris: ELRA, 1848-1854.
- Pytlik-Zillig, Brian (2009): TEI Analytics: converting documents into a TEI format for cross-collection text analysis. In: *Literary and Linguistic Computing* 24(2): 187-192. doi:10.1093/lc/fqp005.
- Sinclair, John (2005): Corpus and text – basic principles. In: Wynne, Martin (ed.): *Developing linguistic corpora: a guide to good practice*. Oxford: Oxbow Books: 1-16.
- TextGrid (2007-2009): TextGrid's baseline encoding for text data in TEI P5. www.textgrid.de/fileadmin/TextGrid/reports/baseline-all-en.pdf.
- Unsworth, John (2011): Computational work with very large text collections. Interoperability, sustainability, and the TEI. In: *Journal of the Text Encoding Initiative* 1. <http://jtei.revues.org/pdf/215>.

Making great work even better¹

Appraisal and digital curation of widely dispersed electronic textual resources (c. 15th-19th centuries) in CLARIN-D

Abstract

Numerous high-quality primary textual resources – in the context of this paper, this means full-text transcriptions (and corresponding image scans) of German texts originating from the 15th to the 19th century – are scattered among the web or stored remotely on institutional or private servers. They are often filed on degrading recording media and are encoded in out-of-date or inflexible storage formats. Often, textual resources are accompanied by scarce, insufficient or inaccurate bibliographic information, which is only one further reason why valuable resources, even if available on the web, remain undiscovered. Additionally, idiosyncratic, project-specific markup conventions often hinder further usage and analysis of the data. Because of these and other problems, a great amount of the abovementioned transcriptions of historical sources can hardly be found, let alone accessed by third parties, and are of little use to the wider research community. This situation is unsatisfying from the perspective of a (corpus-)linguistic project like the one described here, but also from the perspective of any text-based research in the humanities and social sciences. The integration of as many of these ‘dispersed’ high-quality primary textual resources as possible into an encompassing repository like the sustainable, web and centres-based research infrastructure of CLARIN-D² is an important step and at least a necessary prerequisite to solve this problem. This paper summarizes the work of an 18-month project funded by the German Federal Ministry of Education and Research (BMBF) which dealt with the curation and integration of historical text resources of the 15th-19th century into the CLARIN-D infrastructure.

¹ This paper is a thoroughly revised version of the original full paper by the same title, handed in for the International Conference “Historical Corpora 2012”, December 6-9, 2012; Goethe University, Frankfurt am Main, Germany, and published in October 2012 on the edoc-server of the Berlin-Brandenburgische Akademie der Wissenschaften (BBAW), URN: urn:nbn:de:kobv:b4-opus-23081, URL: <http://edoc.bbaw.de/volltexte/2012/2308/> [last retrieved April 30, 2014, as for all URLs cited in this paper].

² CLARIN-D: Common Language Resources and Technology Infrastructure, <http://clarin-d.de/>. Funded by the Federal Ministry of Education and Research (BMBF), CLARIN-D is the German contribution to the EU-wide project CLARIN. It develops a web and centres-based research infrastructure, primarily for language-centred research in the social sciences and humanities. CLARIN-D aims at providing linguistic data, tools and services, and offers a federated content search and sophisticated retrieval facilities. Its service centres share their data and tools in an integrated, interoperable and scalable way, and will see to their long-term availability and archiving to ensure persistent public access.

1. The Mission: curating and integrating distributed text resources into a large text repository

The work described in this paper was carried out in the context of a joint ‘curation project’ (duration: September 2012 until February 2014) of the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW), the Justus-Liebig-Universität (JLU) Gießen, the Herzog August Bibliothek (HAB) Wolfenbüttel, and the Institut für Deutsche Sprache (IDS) in Mannheim as partner institutions in CLARIN-D.³ Digital curation in the context of the project described here entails the careful selection, refinement and analysis, archiving and ongoing maintenance of digital assets.⁴ The stated objective of this project was to process the equivalent of approx. 35,000 pages printed between the 15th and the 19th century from large text collections, digital libraries, ongoing and terminated research projects, scholarly editions, etc. When the project terminated in February 2014, more than 79,000 pages (encompassing approx. 21 million tokens) were integrated, thereby even doubling the targeted number of pages. The three major reasons for this over-achievement are worth mentioning:

³ Cf. the project’s web page “Integration und Aufwertung historischer Textressourcen des 15.-19. Jahrhunderts in einer nachhaltigen CLARIN-D-Infrastruktur”, Kurationsprojekt 1 der Facharbeitsgruppe 1 Deutsche Philologie, www.deutschestextarchiv.de/clarin_kupro. The project was counseled by the discipline-specific working group German Philology in CLARIN-D and coordinated at the BBAW. It was carried out by the CLARIN-D service centres at the BBAW and the IDS, and by the HAB. The basic ideas behind a cooperation like this and the more general aims and methods of corpus compilation are described in this volume in the contribution of Alexander Geyken and Thomas Gloning: “A living text archive of 15th-19th-century German. Corpus strategies, technology, organization.”

⁴ According to the Digital Curation Centre (DCC) (2007), “Digital curation is maintaining and adding value to a trusted body of digital research data for current and future use; it encompasses the active management of data throughout the research lifecycle [...], including the provision of access to data and data reuse. Meeting this obligation will be enabled by good data stewardship.” While ‘digital curation’ puts the emphasis on the cycle of creation, selection and preservation, ‘digital stewardship’ is used in a somewhat broader sense: it emphasises the activities of curation as crucial, but equally stresses the responsibility for ongoing, active work on preserved objects in the asset. Quite often, however, and also in DCC’s definition quoted above, the terms are used interchangeably or in the sense that one concept entails the other, cf. for example Whyte/Wilson (2010), Lee/Tibbo (2007) or Rusbridge et al. (2005: 2). For the purpose of this paper, the definition given above will suffice. For an overview of recent publications on this topic cf. Bailey, Jr. (2012).

1) The project could rely on the elaborated corpus building infrastructure and the well-documented workflow set up at the cooperating project Deutsches Textarchiv (DTA)⁵ at the BBAW.

2) A great amount of text (the equivalent of more than 36,000 pages) was integrated from projects associated with the HAB. Since they were already TEI⁶-encoded, these documents could easily be converted into the specific TEI-format of the DTA; after proofing some representative sample documents, the process of integration could be entirely automated for the rest of the HAB-corpus. For another large amount of text (approx. 20,000 pages) integrated from Wikisource, the manual effort could be reduced significantly with the help of a specialised web form.⁷ This script-based integration form parses the cumbersome 'MediaWiki'-syntax and transforms as many elements as possible into TEI-XML.

3) It has to be kept in mind that the curation of digital assets is an ongoing process that does not end with the integration. For some of the HAB texts, but also for texts from the Max Planck Institute for the History of Science (MPI-WG) and other collections, further work has to be done to improve the quality of text and metadata. In the course of the project, we decided to first of all convert and integrate these texts into the corpus infrastructure at the DTA, in order to then be able to use the quality assurance mechanisms provided by the DTA and thereby support the ongoing process of data curation.

The integrity and significance of the collections in general and of each single item in particular was evaluated thoroughly with respect to the project's qualitative criteria described below. The selected items were integrated into the partner's respective repositories and, from there, made available in the CLARIN-D framework under a Creative Commons license.⁸ In preparation for

⁵ Deutsches Textarchiv (DTA), www.deutschestextarchiv.de. The DTA is funded by the German Research Foundation (DFG). All DTA texts are available for download in different formats: in TEI-XML, HTML, in the Text Corpus Format (TCF) used by WebLicht services in CLARIN-D, and as plain text transcriptions. Metadata, available as TEI-Headers, formatted in Dublin Core (DC, cf. <http://dublincore.org/>) and according CLARIN's Component MetaData Infrastructure (CMDI, cf. www.clarin.eu/cmdi), can be harvested via an Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), available under www.deutschestextarchiv.de/api.

⁶ TEI: Text Encoding Initiative, www.tei-c.org. Cf. Guidelines for Electronic Text Encoding and Interchange www.tei-c.org/Guidelines/.

⁷ URL: www.deutschestextarchiv.de/dtae/submit/wikisource.

⁸ Note that CLARIN-D is only one example of a wide-span research infrastructure. By offering the

and all through the course of the project, large scale collections such as Wiki-source and Gutenberg.org as well as smaller, more specific sources were critically reviewed to identify text resources appropriate to serve as valuable extensions of a growing reference corpus for the historical German language. The selected items were aggregated and standardized with respect to their storage and annotation format; structural and bibliographic information was enhanced and corrected, if necessary.

To reach its aims, the curation project fortunately could make extensive use of the elaborated technical infrastructure and, not less important, the encompassing documentation of transcription- and format-specific guidelines developed by the DTA.

2. Exemplary workflow: the DTA and its enhancement module DTAE

The DTA project started in 2007 and is building a TEI-XML-annotated full-text corpus of German-language texts. More than 1,300 volumes printed between the 17th and the 19th century will be processed and published online until 2014/15. Scientific texts, as well as fiction, poetry, drama, essays and everyday literature combine to a comprehensive collection documenting the development of the modern German language. TEI-XML-annotated full-text transcriptions of the primary sources accompanied by detailed bibliographic metadata are made available for free download and are displayed on the internet alongside digital facsimiles. The transcriptions are true to the source, show a high level of accuracy and are annotated with structural information following the TEI P5-compliant DTA ‘base format’ (DTABf).⁹ The electronic full-texts are enriched with linguistic information in stand-off markup gained

data via OAI-PMH, the resources aggregated in the course of the project described here can be made available also within other national, European or international infrastructures such as DARIAH, Europeana, TextGrid, Project Bamboo, etc.

⁹ DTABf: Deutsches Textarchiv – Basisformat, www.deutschestextarchiv.de/doku/basisformat. The DTABf is a subset of TEI P5, containing about 100 elements and their possible attributes and values. It restricts the number of elements from the TEI Guidelines in order to reduce the application of inconsistent tagging for similar structural phenomena within the corpus. By this means, the DTABf aims at gaining coherence at the annotation level, given the heterogeneity of the DTA texts regarding time of origin (~1600-1900) and text type (e.g. fiction, functional texts, or scientific texts). Cf. Geyken/Haaf/Wiegand (2012). The DTABf is recommended as best practice for the structural encoding of historical printed texts within CLARIN-D. Cf. CLARIN-D AP 5 (2012): Part II, ch. 6, subsection “Text Corpora”.

through tokenization, lemmatization, and part-of-speech-analysis. Each text is analyzed with CAB, a set of rewrite rules for automated orthographic normalization of historical text material.¹⁰

The prospect of substantially more than 1,300 original texts from three centuries to be published until 2014/15 is promising for (computer-aided) research in linguistics, semantics, typology, and other areas. But still, certain discourses and genres, subject fields or domains are less well represented in the corpus than others, and the number of witnesses per decade may – for some purposes – seem relatively small. So, to enhance DTA's 'core collection', i.e. to substantially broaden the text base and to improve the balance of the corpus, the software module DTAE ("E" for Extensions) was developed. By the time this article was written (April 2014), there are 1,312 texts (core corpus plus curated extensions) dating from between the 16th and early 20th century online, comprising a total of more than 423,000 digitized pages with more than 691 million characters and roughly 100 million tokens. More than 500 additional volumes, mainly from the period between 1600 and 1780, are prepared to be published and will likewise be made freely available under a Creative Commons license. In the course of the curation project described here, DTAE was used as the platform for conversion and publication of high-quality resources from various contexts. The resources were integrated into DTA's extended corpus, and at the same time into the CLARIN-D research infrastructure, where tools for further analysis of the data are provided and their long-term preservation is taken care of.

DTAE provides routines and scripts for the conversion of metadata, text and images, as well as tools for the (semi-)automatic conversion from different source formats (HTML, doc, docx, plain text, PDF, ...) into the DTABf. Thereby, the DTAE infrastructure and tools facilitate the production of new high-quality transcriptions of primary sources in cooperation between the DTA and external researchers, as well as allowing for the integration and enhancement of existing resources. Both ways of corpus building have been followed successfully at the DTA: co-operative text production – for example, together with the Alexander-von-Humboldt-Forschungsstelle and the Marx-Engels-Gesamtausgabe (MEGA) project at the BBAW, as well as the Forschungsstelle für Personalschriften at the Philipps-Universität Marburg (Arbeitsstelle der Akademie der Wissenschaften und der Literatur, Mainz) – and corpus build-

¹⁰ Cf. Jurish (2010; 2011). CAB provides an automated normalization of the historical orthography in order to allow for lemma-based, spelling-tolerant corpus searches.

ing via integration and enhancement of existing resources – for example, from born-digital scholarly editions like those of the works of J. v. Sandrart, J. F. Blumenbach, and from the centenary-spanning ‘Polytechnisches Journal’ founded by J. G. Dingler.¹¹ The integration of these text resources is relatively straightforward, thanks to the TEI-compliant encoding provided by the projects mentioned. Therefore, instead of going into detail any further on this aspect, the remainder of this paper will deal with the much higher obstacles on the way to identify, enhance, refine and integrate text converted from various storage formats in the context of the curation project.

3. Digital curation: select, enhance and preserve distributed resources

3.1 Criteria: what to look for?

Among a great number of possible sources, appropriate items for the curation project were identified with the help of a set of criteria. Generally speaking (and deliberately in contrast to other corpus building approaches), the curation project put an emphasis on quality over quantity:¹² This meant rather ‘hand-picking’ than following a ‘DownThemAll!’ approach, where, as a prize for the greater amount of data to be gained in one single sweep, one has to put up with the downside, i.e. the minor quality a considerable number of single items in the collection – and, as a result, of the corpus as a whole – will display. To overcome problems of format obsolescence and inflexibility, conversion of the data into a consistent, standardised and flexible format such as (TEI-) XML was central. However, the amount of time and manual work this process requires differs strongly depending on the data base. With this in mind, it was decisive for the success of the curation project to have kept a sound balance between the effort it took to integrate the ‘chosen ones’ and the (anticipated) value they represent to the research community addressed in CLARIN-D, and to carefully have weighed the quality against the quantity of the aggregated resources.

¹¹ For further information on these editions, cf. the projects’ respective web sites: www.sandrart.net, www.blumenbach-online.de, and www.polytechnischesjournal.de.

¹² Nevertheless, selected ‘working transcriptions’ were also integrated and revised step by step to finally meet the curation project’s criteria. Especially in this respect, recommendations of CLARIN-D’s discipline-specific working groups were taken into consideration, and members of the community were encouraged to help improve the resources, e.g., by proofreading and correcting.

The criteria described in the following were defined in accordance with the general guidelines of the DTA.¹³ First of all, the digitized print sources should be first or early editions of the text represented. As a project with a strong orientation towards historical text/corpus linguistics and lexicography, the DTA offers text true to the primary source, without later ‘normalizations’ in spelling and other severe intrusions distorting the historical text. Any alterations, e.g. the replacement of certain letters like the long *s* (*f*) by the ‘modern’ round *s*, the dissolution of ligatures, the correction of printing errors, etc., should be documented and be done consistently. Line breaks, or at least page breaks found in the source document should be marked in the transcription.¹⁴ The transcription should prove high accuracy on the level of characters (preferably 99.5+ %) and, with respect to the annotation, should contain at least the most basic structural information (i.e. divisions/chapters, headers, paragraphs). Furthermore, the texts in question should be expressive witnesses of the development of the New High German Language, and/or relevant to a certain field of scientific or cultural history, and/or instances of a certain special discourse, documenting specific aspects of different kinds of language use, including everyday language. The transcribed text should contain or be accompanied by information about the method of data acquisition (uncorrected, ‘dirty’ OCR or OCR with proofing, single-handed transcription, double keying, ...), its creator and its editing status (completed, draft, working transcription, ...). The image scans should show a high resolution, preferably be full-colour master copies with ≥ 300 DPI in TIFF or JPEG2000 format. The metadata describing the source should be accurate and as detailed as possible (while it certainly still has to be complemented in the curation process, if only for the purpose of marking versions and stating editing responsibilities in the life cycle of the document).¹⁵ Certainly, legal aspects concerning text, metadata and images will have to be sorted out: each item, i.e. images, metadata and text should be available under a free license at least for reuse in a scientific context.

¹³ Cf. DTA-Leitlinien, www.deustchestextarchiv.de/doku/leitlinien.

¹⁴ The marking of page breaks is essential for the (automated) alignment of source images and transcribed text. It also allows for a rough, general comparison between source and derived text in order to evaluate the quality of the resource. In this sense and beyond that, it facilitates anticipative as well as retrospective quality assurance, e.g. proofreading. For a documentation of DTA’s profound experience with quality assurance in large text corpora cf. Geyken et al. (2012) and Haaf/Wiegand/Geyken (2013).

¹⁵ See DCC (2007) for an illustration of a Curation Lifecycle Model.

Although, at first glance, these criteria might seem to form quite a low threshold, in the course of the curation project they still helped to guarantee a high quality and the integrity of the acquired data and provided a good orientation to separate the wheat from the chaff.¹⁶ For example, most of the texts represented in the text collection of Gutenberg-DE¹⁷ – and, although with some notable exceptions, also those of zeno.org¹⁸ – did not meet the curation project's criteria in every respect. A great number of the transcriptions available there are based on philologically questionable editions, bristling with undocumented and, often enough, inconsistent alterations of the original text. In some cases, forewords, dedications, and other 'supplementary' parts printed in the primary source remained unconsidered altogether, and some transcriptions simply did not show the accuracy required. So, instead of the two large collections Gutenberg-DE and zeno.org, quantitatively smaller, but – considering the curation project's criteria – qualitatively better sources became the major points of interest for the project.

3.2 Sources: where to look?

3.2.1 Large text (and image) collections: Wikisource and Project Gutenberg

The German partition of Wikisource and German-language texts from the American Project Gutenberg (PG) proved to be the most fruitful sources for a considerable amount of documents fulfilling the criteria described above.¹⁹

¹⁶ As a welcome side effect, the criteria helped to narrow the focus of the project described here to a manageable amount of text resources.

¹⁷ Projekt Gutenberg-DE, <http://gutenberg.spiegel.de/>.

¹⁸ Zeno.org, www.zeno.org. In 2009, the whole collection was acquired by the research infrastructure project TextGrid, funded by the BMBF. The text files from zeno.org were converted into the TextGrid 'Baseline Encoding', a TEI-conformant basic encoding format used mainly to allow for project-specific as well as cross-text queries within the TextGrid Repository (Cf. TextGrid 2007-2009: 6). In this process, basic structural information was gained by automated analysis of the source markup. XML-IDs were added to each line of the transcription to allow for more exact referencing. Since July 2011, the data stock of the literature folder is available for download. The original transcriptions of historic works for zeno.org were almost exclusively derived from partly modernized editions from the 19th/20th century. During the transformation to TextGrid, they were not proofed against reliable scholarly editions or compared to the primary sources. Likewise, proofing and correction of the metadata is yet to be done, cf. www.textgrid.de/en/digitale-bibliothek/.

¹⁹ Of course, but with the reservations mentioned above in mind, selected, high-quality items from zeno.org and Gutenberg-DE meeting the project's criteria were also integrated. For example, the accurate transcription of Hans Stadens "Warhaftige Historia und beschreibung eyner Landschafft

The focus in the following passage is on Wikisource, which proved to be the richest source for appropriate texts. The quality of the single resources assembled in ‘opportunistic’ collections like Wikisource with its many individual contributors differs strongly, but nonetheless several high quality representations of historic documents could be discovered. The site offers accurate transcriptions of historic primary sources, often along with corresponding image scans in good quality. Unfortunately, the best items were somewhat hidden among the vast total number of objects. To make sure that its integration would be worth an effort, each possible candidate was evaluated following the criteria described in the previous section – a non-trivial task itself, given the amount of approx. 30,000 German-language texts (as of April 2014) in the German Wikisource.²⁰

The metadata describing the collected objects displayed on the website often proved to be not sufficient to serve as a basis for a systematic selection of single items. The navigational structure of the site is rather opaque, and the on-site retrieval facilities are quite basic. The options to browse and search the collection are rather limited and it is hard to get an overview.²¹ This holds true for Wikisource, but also for Project Gutenberg and other large scale collections under consideration. Therefore, the sites in focus had to be critically scoured manually pursuing different strategies. From Wikisource, the most prolific source among the large scale collections, 1,891 high-quality texts containing almost 20,000 pages were identified and integrated.²²

der Wilden / Nacketen / Grimmigen Menschfresser Leuthen [...]” (Marpurg [Marburg], 1557), www.deutschestextarchiv.de/staden_landschaftt_1557, was integrated from Gutenberg-DE. A number of works of female writers, a group notoriously under-represented in corpora of historical printed works, was drawn from zeno.org, cf. www.deutschestextarchiv.de/doku/clarin_kupro_liste?g=zeno. Further additions were derived from Sophie – A Digital Library of Works by German-Speaking Women (<http://sophie.byu.edu/>), for example Louise Aston’s “Aus dem Leben einer Frau” (Hamburg, 1847), www.deutschestextarchiv.de/aston_leben_1847.

²⁰ Cf. <http://de.wikisource.org>, Hauptseite > Wikisource Aktuell > Statistik.

²¹ Unfortunately, Wikisource offers no query or download API for ingesting the full descriptive metadata of the project’s resources, although its development obviously has been discussed for some time, cf. http://de.wikisource.org/wiki/Wikisource:Metadaten#Weitergabe_der_Metadaten and http://de.wikisource.org/wiki/Wikisource:Skriptorium/Archiv/2006/3#Professionalisierung_von_Wikisource.

²² Cf. www.deutschestextarchiv.de/doku/clarin_kupro_liste?g=wikisource.

3.2.2 Research projects and scholarly editions

As a second domain for historical text resources, research projects and scholarly editions were taken into account, as their data in general incorporate the expertise and scrutiny of acknowledged specialists. Without doubt, the fruits of their labour were of high interest for the purpose of this curation project,²³ but first of all, the data had to be retrieved and often enough legal issues had to be solved: sometimes, access was impossible even to the ‘raw’ data of the project, e.g. because of restrictive contracts with publishing houses.²⁴ Both tasks, retrieving the data and securing access to it, were even harder to accomplish in cases where the research project in question had already ended: staff members were off to other places, while the work done – especially the fundamental steps *before* the publication of the research outcomes – often was hardly documented. As one result of this, the project-specific transcription and the markup conventions applied had become hard to comprehend by others. They had to be reconstructed a) in order to be able to evaluate the resource in the first place and b), if the item was to be integrated, in order to perform a lossless conversion of the data into the DTABf.

If the data was available for integration, a further and no less severe problem concerned its storage format. Until recently, the majority of scholarly editions of historical text material were produced with the goal of a printed (or print-

²³ A very successful cooperation was established with the editors of the historical-critical edition of Karl Gutzkow’s works and letters, <http://projects.exeter.ac.uk/gutzkow/Gutzneu/>. The HTML-representation of Gutzkow’s primary works was converted into the DTABf, encompassing linguistic analysis and indexing for full-text queries (which the Gutzkow edition itself does not offer) and allowing for collaborative quality assurance in DTAQ (cf. below, ch. 3.3). By preserving all editorial and other comments originally present in HTML and by discussing and carefully documenting the steps of conversion, the integration of all texts from the Gutzkow edition into the DTA corpus remains reversible. This was a crucial point for the fruitful cooperation between the projects. For example, if transcription errors are corrected or the text base is changed for other reasons in the process of quality assurance in DTAQ, or if the transcription is annotated more deeply (for example by annotating named entities), the results of this work can be re-transformed from DTA’s site into the Gutzkow project.

²⁴ ‘Raw’ data in the context of this paper could mean an uncommented, but exact transcription of the primary source, which forms the basis of almost every scholarly edition of a text. These transcriptions would be of great value to other projects (not only the curation project described in this paper, but also for corpus projects like the DTA in general), which seldom seems to be considered while negotiating the terms of publication. Often, this ‘raw data’ is taken less care of in the process of critical editing and commenting, and therefore it even more likely becomes outdated and inaccessible by (storage) format evolution over time.

like) documentation of the work in mind.²⁵ Therefore, the text base was produced with the help of GUI-based text processors and other office tools. It was published and/or stored in formats such as MS doc or docx, Adobe InDesign, PDF or LaTeX. The most severe problems are the evolving obsolescence of certain (esp. proprietary) data formats (older versions of MS Word, WordStar, WordPerfect, etc.), and the fact that GUI-based text processors and their output formats tend to indistinguishably mix layout information with structural information. Therefore, the data demanded a notable amount of manual labour in reformatting to preserve the intellectual work explicitly and implicitly contained in the documents.²⁶

3.2.3 Special collections and single resources

Finally, and in addition to large scale collections and scholarly projects, smaller compilations of texts on a certain topic, representing a particular discourse or epoch were considered. Often built and run by enthusiastic private scholars or layman investing a lot of energy and spare time, these thematic collections may reveal astonishing discoveries. Single findings were integrated, fortunately always with the approval and sometimes also with the support of their producers.²⁷

²⁵ Of course, this is still a wide-spread conduct, while it would be of great benefit for the research community to produce and preserve data in exchangeable, well documented formats like (TEI-)XML from the beginning.

²⁶ Two examples to illustrate the outcome of this laborious, but worthwhile effort of data conversion are Theresia Lindnerin's "Koch Buch zum Gebrauch der Wohlgebohrenen Frau" (around 1780), which was transcribed, annotated and published as a PDF file under <http://geb.uni-giessen.de/geb/volltexte/2009/7361/> by Thomas Gloning from the University of Gießen, and is now available in TEI-XML under www.deutschestextarchiv.de/lindnerin_kochbuch_1780; and a transcription of "Petrus de Crescentiis zu teutsch mit Figuren. Speyer, ca. 1493"; www.deutschestextarchiv.de/crescentiis_figuren_1493, which Jakub Šimek had published as part of his University of Heidelberg's Magister thesis (cf. <http://crescenzi.dyskanti.com/>) and stored in LaTeX. We are grateful to both editors for offering their documents and their instructive comments on how to convert the valuable information.

²⁷ See, for example, Joseph Schauberg's three-volume "Vergleichendes Handbuch der Freimaurerei" (1861-63), www.deutschestextarchiv.de/schauberg_freimaurerei01_1861, [.../schauberg_freimaurerei02_1861](http://www.deutschestextarchiv.de/schauberg_freimaurerei02_1861) and [.../schauberg_freimaurerei03_1863](http://www.deutschestextarchiv.de/schauberg_freimaurerei03_1863), derived from the "Portal to the World of Freemasonry", www.internetloge.de. Altstuhlmeister Franz-L. Bruhns, webmaster and editor of this non-commercial web page, happily agreed to the re-use of the HTML-representation of the "Handbuch" (www.internetloge.de/symhandb/symb.htm) on the one condition that www.internetloge.de is appropriately credited as creator of the original transcription.

Now that the major sources of high-quality resources have been described and before the process of their integration in the course of the curation project is outlined, a word on appreciation and responsibilities is due. In order to establish a culture of shared access and usage, the importance of a reputation system must not be omitted. Therefore, and for each single item integrated, the appreciation of the work of others was made visible in the source documentation. Also, responsibilities in every stage of the text refinement were made transparent.

3.3 Integration: how to proceed?

Once a relevant resource meeting the named criteria was identified, the full-text transcription, image scans and metadata were acquired and integrated into DTA's enhancement module DTAE. In the course of this, the electronic documents were enriched with the acquired and enhanced bibliographic and structural information. In the next step, the bibliographic data and full-text transcription were converted into the DTABf. The text and metadata were then published alongside the corresponding image scans via the DTAE framework; it was also made available via the BBAW's CLARIN-D repository.²⁸ Each text was analyzed with CAB for automated orthographic normalization of the historical text: the great variance in spelling of terms is being mapped onto its modern form, thereby allowing for spelling-tolerant and complex queries in the growing text corpus. The linguistic analysis furthermore encompasses tokenization, lemmatization, and PoS-tagging in stand-off markup.

Each integrated text can now be displayed page-wise in an HTML representation automatically rendered from the underlying TEI-XML (Fig. 1), it can be searched and explored as a single resource, in the context of the different sub-corpora compiled by the DTA, in the context of the DTA 'core corpus' and in the greater context of all corpora available in CLARIN-D. The resource descriptions and bibliographic information are standardized conformant to authority formats (e.g. CMDI or DC) in order to be shared via OAI-PMH and to be integrated into CLARIN-D's service architecture.

²⁸ Cf. the repository at the CLARIN Service Center of Zentrum Sprache at the BBAW, <http://clarin.bbaw.de/>.

The screenshot displays the DTA interface for the Wikisource item 'Deß Weltberuffenen SIMPLICISSIMI Pralerey und Gepräng mit seinem Teutschen Michel' by Hans Jakob Christoffel von Grimmelshausen. The interface is organized into three main sections:

- Image:** A scanned image of the title page of the book, showing the title and author information in a historical font.
- Transcription:** A rendered HTML version of the title page text, which is clean and readable. The text includes: 'Deß Weltberuffenen SIMPLICISSIMI Pralerey und Gepräng mit seinem Teutschen Michel / Jedermänniglichen / wanns seyn kan / ohne Lachen zu lesen erlaubt Von Signeur Meßmahl. Gedruckt unter der Preß / in dem jetzigen Land / darinnen das selbige löbliche Geschick erstmahls erhanden worden / Als selbne Liebe Inwohner neben an Dertl Völkern anfragen / Den Jahren Vassers Hellwuch / In gleichem Jahr. ZV Zählten 1673.'
- Metadata and Information:** A sidebar on the right containing various fields such as 'Informationen zum Werk', 'Metadaten zum Werk', and 'Anmerkungen zur DTA-Ausgabe'. It also includes a search bar and navigation links.

Figure 1: Wikisource-item integrated into the DTA corpus: image, transcription in rendered HTML, metadata and further information on the transcription and annotation guidelines applied in the production of the resource. Grimmelshausen, Hans Jakob Christoffel von: Deß Weltberuffenen SIMPLICISSIMI Pralerey und Gepräng mit seinem Teutschen Michel. [Nürnberg], 1673, Title Page / image 7. In: Deutsches Textarchiv, www.deutschestextarchiv.de/grimmelshausen_michel_1673/7

In parallel, all items can be accessed via DTA's quality assurance platform DTAQ.²⁹ In DTAQ, texts may be proofread page by page in comparison to their source images (Fig. 2). This way, errors that may have occurred during the former transcription and annotation process, or that were overlooked or not taken care of during integration can be detected and corrected. While the transcription can best be inspected in the rendered HTML version, the underlying annotation can conveniently be checked in TEI-XML. The automated analysis of the full-text with CAB can be checked as well.

²⁹ Deutsches Textarchiv – Qualitätssicherung (DTAQ), www.deutschestextarchiv.de/dtaq. [Users must register and have their accounts activated by a DTA staff member.]

The screenshot displays the DTAQ (Deutsches Textarchiv - Qualitätssicherung) web interface. At the top, the logo 'DTAQ' is visible, along with navigation links for 'zuletzt gelesen', 'Hilfe', 'Zufallsseite', and user information for 'FrankWiegand'. The main content area shows a document titled 'storm_waldwinkel_1875 (Wikisource)'. The document text is rendered in HTML, with some words highlighted in red to indicate errors. A 'Ticket-Details #60529' overlay is open, showing the following information:

- angelegt von:** ChristianThomas, 2013-11-11T11:58
- Typ:** Transkriptionsfehler
- Zusammenfassung:** Wurstel
- Beschreibung:** (empty)
- Fundstelle:** Wurstf
- Typ:** Transkriptionsfehler
- Zusammenfassung:** Wurstel
- betrifft:** Seite 177, ganzes Buch
- Priorität:** normal
- Relevanz:** normal
- Actions:** als neu fassen, lösen als behoben, zuweisen an, annehmen

Buttons for 'Ticket ändern' and 'Abbrechen' are visible at the bottom of the ticket overlay. On the right side of the interface, there are sections for 'Buchdaten', 'DTA-Informationen', 'Korrekturstatus', and 'Tickets für diese Seite'.

Figure 2: Quality Assurance in DTAQ: image, transcription in rendered HTML, 'ticket' system to report findings, e.g. transcription errors, printing errors, and inconsistencies in annotation. Storm, Theodor: Waldwinkel, Pole Poppenspäler. Novellen. Braunschweig, 1875, p. 173 / image 177. In: Deutsches Textarchiv – Qualitätssicherung, www.deustchestextarchiv.de/dtaq/book/view/storm_waldwinkel_1875?p=177 [retrieved 2013-11-11; the transcription errors highlighted in the illustration have been corrected in the meantime]

4. Conclusion

In the course of the CLARIN-D curation project described here, the equivalent of more than 79,000 pages was integrated into a large corpus for the written German language between the 15th and the 19th centuries. From the large text repository consisting of the DTA's, HAB's, IDS' and many other partner's corpora available under the roof of CLARIN-D, more balanced reference corpora can now be derived. In this respect, the curation project helped to improve the situation for corpus-based research, particularly in historical linguistics, but also in the humanities in general. By applying a consistent,

interoperable encoding based on the recommendations of the TEI³⁰ and by integrating the resources into the CLARIN-D infrastructure, the data can now be explored in a broader context. Access to and sustainability of these resources were thereby improved substantially. A formerly dispersed, large variety of corpus texts can now be processed by the elaborated tool chain CLARIN-D offers. By establishing methods of interoperation, a system of quality assurance and credit, and a set of technical practices that allow the integration of resources of different origin, CLARIN-D contributes significantly to the scholarly community. The idea of curating and sharing corpus resources in a collaborative manner was put into practice with excellent results – which hopefully will encourage similar initiatives.

Affiliation

CLARIN-D curation project “Integration und Aufwertung historischer Textressourcen des 15.-19. Jahrhunderts in einer nachhaltigen CLARIN-Infrastruktur”, cf. www.deutschestextarchiv.de/clarin_kupro for an overview on the project and for a list of texts integrated.

References

- Bauman, Syd (2011): Interchange vs. interoperability. Presented at Balisage: The Markup Conference 2011, Montréal, Canada, August 2-5, 2011. In: Proceedings of Balisage: The Markup Conference 2011. (= Balisage Series on Markup Technologies 7). www.balisage.net/Proceedings/vol7/cover.html, doi:10.4242/BalisageVol7.Bauman01.
- Bailey, Jr., Charles W. (2012): Digital Curation Bibliography: Preservation and Stewardship of Scholarly Works. <http://digital-scholarship.org/dcpb/dcb.htm>.
- CLARIN-D AP 5 (2012): CLARIN-D User Guide. Version: 1.0.1, Publication date: 2012-12-19. www.clarin-d.de/en/language-resources/userguide.html.
- Digital Curation Centre (DCC) (2007): What is digital curation? www.dcc.ac.uk/digital-curation/what-digital-curation; DCC curation lifecycle model. www.dcc.ac.uk/resources/curation-lifecycle-model.
- Geyken, Alexander/Haaf, Susanne/Wiegand, Frank (2012): The DTA ‘base format’: A TEI-subset for the compilation of interoperable corpora. In: Jancsary, Jeremy (ed.): 11th Conference on Natural Language Processing (KONVENS): Empirical Methods in Natural Language Processing. Proceedings of the Conference on

³⁰ Cf. Bauman (2011) and Unsworth (2011).

- Natural Language Processing 2012. (= Schriftenreihe der Österreichischen Gesellschaft für Artificial Intelligence 5). Wien: ÖGAI, 383-391. www.oegai.at/konvens2012/proceedings.pdf#page=383.
- Geyken, Alexander/Haaf, Susanne/Jurish, Bryan/Schulz, Matthias/Thomas, Christian/Wiegand, Frank (2012): TEI und Textkorpora: Fehlerklassifikation und Qualitätskontrolle vor, während und nach der Texterfassung im Deutschen Textarchiv. In: Jahrbuch für Computerphilologie, online version. www.computerphilologie.de/jg09/geykenetal.html.
- Haaf, Susanne/Wiegand, Frank/Geyken, Alexander (2013): Measuring the correctness of double-keying: error classification and quality control in a large corpus of TEI-annotated historical text. In: *Journal of the Text Encoding Initiative* 4. <http://jtei.revues.org/739>, doi:10.4000/jtei.739.
- Jurish, Bryan (2010): More than words: using token context to improve canonicalization of Historical German. In: *Journal for Language Technology and Computational Linguistics (JLCL)* 25/1: 23-39. http://media.dwds.de/jlcl/2010_Heft1/bryan_jurish.pdf.
- Jurish, Bryan (2011): Finite-state canonicalization techniques for Historical German. PhD thesis, Universität Potsdam. <http://opus.kobv.de/ubp/volltexte/2012/5578/urn:nbn:de:kobv:517-opus-55789>.
- Lee, Christopher A./Tibbo, Helen R. (2007): Digital curation and trusted repositories: steps toward success. In: *Journal of Digital Information (JoDI)* 8(2): Digital Curation & Trusted Repositories. <http://journals.tdl.org/jodi/article/view/229/183>.
- Rusbridge, Chris/Burnhill, Peter/Ross, Seamus/Buneman, Peter/Giaretta, David/Lyon, Liz/Atkinson, Malcolm (2005): The Digital Curation Centre: a vision for digital curation. In: *Proceedings from the IEEE Conference Local to Global: Data Interoperability – Challenges and Technologies*. Forte Village Resort, Sardinia, Italy, 2005: 1-11. <http://eprints.erpanet.org/82/>.
- TextGrid (2007-2009): TextGrid's Baseline Encoding for Text Data in TEI P5. www.textgrid.de/fileadmin/TextGrid/reports/baseline-all-en.pdf.
- Unsworth, John (2011): Computational work with very large text collections. In: *Journal of the Text Encoding Initiative* 1. <http://jtei.revues.org/215>, doi:10.4000/jtei.215.
- Whyte, Angus/Wilson, Andrew (2010): How to Appraise and Select Research Data for Curation. (= DCC How-to Guides). Edinburgh: Digital Curation Centre. www.dcc.ac.uk/resources/how-guides.

Using an alignment-based lexicon for canonicalization of historical text

Abstract

This paper addresses issues in orthographic normalization of historical German text. We investigate the utility of a finite deterministic canonicalization lexicon semi-automatically constructed from a corpus of historical and contemporary editions of the same texts by comparing its performance on a simulated information retrieval task to that of a robust generative finite-state canonicalization architecture, as well as that of a hybrid method which uses a finite lexicon to augment a generative canonicalizer.

1. Introduction

Virtually all conventional text-based natural language processing techniques – from traditional information retrieval systems to full-fledged parsers – require reference to a fixed lexicon accessed by surface form, typically trained from or constructed for synchronic input text adhering strictly to contemporary orthographic conventions. Unorthodox input such as historical text which violates these conventions therefore presents difficulties for any such system due to lexical variants present in the input but missing from the application lexicon. *Canonicalization* approaches (Rayson/Archer/Smith 2005; Jurish 2012, Porta/Sancho/Gómez 2013) seek to address these issues by assigning an extant equivalent to each word of the input text and deferring application analysis to these canonical cognates.

Traditional approaches to the problems arising from an attempt to incorporate historical text into such a system rely on the use of additional specialized (often application-specific) lexical resources to explicitly encode known historical variants. The simplest form such lexical resources take is that of simple finite associative lists or “witnessed dictionaries” (Gotscharek et al. 2009b) mapping each known historical form w to a unique canonical cognate \tilde{w} . Since no finite lexicon can fully account for productive morphological processes like German nominal composition, and since manual construction of a high-coverage lexicon requires a great deal of time and effort, such resources are often considered inadequate for the general task of canonicalizing arbitrary input text (Kempken/Luther/Pilz 2006).

In this paper, we investigate the utility of a finite deterministic lexicon induced from a corpus of aligned historical and contemporary editions of the same texts (Jurish/Drotschmann/Ast 2013). We compare the canonicalization performance of the induced lexicon to that of the robust generative finite-state canonicalization architecture described in Jurish (2012), and to that of a hybrid method which uses a finite lexicon to augment a generative canonicalization architecture.

1.1 Related work

Rayson/Archer/Smith (2005) describe an automatic “variant detector” for canonicalization of historical English, reporting a substantial improvement in accuracy on a small test set compared to conventional spell-checkers. Inverse canonicalization approaches mapping modern query words to (potential) historical variants have been described by Gotscharek et al. (2009b) and Ernst-Gerlach/Fuhr (2007). Recent work on canonicalization for historical text has focused on the use of context for disambiguation of historical “false friends” (Jurish 2012, Reffle et al. 2009), the induction of rule-sets for mapping historical to modern forms (Baron/Rayson 2009, Bollmann/Petran/Dipper 2011), and the rigorous characterization of the mapping task (Jurish 2012, Porta/Sancho/Gómez 2013).

The use of specialized canonicalization lexica for historical document collections has been described by Hauser et al. (2007), who note in particular that explicit lexical mappings may succeed where pattern-based cognate approaches fail, as in the case of extinct historical word forms such as *marcken* (“to trade”). Gotscharek et al. (2009a) describe the manual construction of a canonicalization lexicon and its application in the context of information retrieval. Manually constructed corpora annotated with canonical cognates are described by Scheible et al. (2011) and Dipper and Schultz-Balluff (2013), while Jurish/Drotschmann/Ast (2013) present a semi-automatic bootstrapping procedure for canonicalization corpora using historical and modern editions of the same texts.

2. Materials

For the current investigations, we used the semi-automatic procedure described in Jurish/Drotschmann/Ast (2013) to bootstrap a canonicalized corpus of historical German in which each (historical) token w is explicitly associated with a (modern) canonical form \tilde{w} by aligning historical texts with contemporary editions of the same texts. The construction is based on the assumptions that the contemporary editions used in the construction adhere to

modern orthographic conventions and can therefore be interpreted as the desired canonical output for the respective historical texts on the one hand, and that a large portion of the canonicalization pairs can be expected to be identity pairs in which the historical form is in fact a valid modern form on the other. In attempt to minimize manual annotation effort while maximizing the accuracy of the relevance relation implied by the canonicalization pairs, the historical and contemporary editions were first automatically aligned, and subsequently subjected to a two-phase manual review process of the non-identity alignments. The procedure was applied to 126 volumes of historical German¹ originally published between 1780 and 1901 drawn from the *Deutsches Textarchiv*² and contemporary editions of the selected volumes provided by the online libraries Project Gutenberg³ and Zeno.⁴ The resulting corpus contained 5,642,813 tokens of 212,028 distinct (w, \tilde{w}) -pair types.

Subsequent experience with the resulting corpus indicated that the assumptions underlying the construction procedure were not in fact borne out by the editions used in our construction. In particular, the assumption that the contemporary editions themselves systematically adhere to contemporary orthographic conventions appears to have been unjustified in the current case. While some orthographic normalization – such as the conversion from historical *th* to contemporary *t* as in the mapping pair *Theil* \mapsto *Teil* (“part”) – was indeed undertaken by the editors of the contemporary editions, these texts still exhibit a substantial number of unnormalized historical spelling variants. Such unnormalized historical spellings lead to identity canonicalizations of the form (w, w) during the alignment phase of the corpus construction procedure, which were accepted into the output corpus without manual confirmation⁵. The presence of such unnormalized words in the contemporary editions thus leads to identity canonicalizations which are not in fact valid contemporary forms, and therefore do not accurately represent a ground-truth canonicalization, such as *andre* \mapsto *andre* \neq *andere* (“other”), *kömmt* \mapsto *kömmt* \neq *kommt* (“comes”), *nich* \mapsto *nich* \neq *nicht* (“not”), and *ward* \mapsto *ward* \neq *wurde* (“was”).

¹ About 80% of the source texts were *belles lettres*; the remainder were academic prose.

² www.deutschestextarchiv.de.

³ www.gutenberg.org.

⁴ www.zeno.org.

⁵ Nearly half of the output corpus types representing over 81% of tokens were identity pairs, and over 59% of types representing over 87% of tokens were identical modulo deterministic transliteration of extinct characters such as long “s” or superscript “e”.

In an attempt to ameliorate these shortcomings, the entire corpus was subjected to a document-level review phase. Five volumes (197,925 tokens) were dropped from the corpus due to pervasive orthographic violations in the contemporary editions used for alignment – typically, these were volumes of verse using non-standard capitalization conventions. Using page-wise diagnostic heuristics, a total of 204 pages in 41 volumes were manually selected and purged from the corpus, chiefly due to heavy use of pseudo-phonetic dialect or foreign-language material. For example, the entirety of the story *Von den Fischer und seine Fru* (“Of the fisherman and his wife”, written entirely in Low German) was purged from the Grimms’ fairy tales in this fashion.

Many alignment errors were found to result from irregular hyphenation, explicit elisions or genitive marking using apostrophes, and tokenization errors involving beginning-of-line quotes. In order to remove these errors from the corpus, all 9,250 types (16,300 tokens) containing an apostrophe, quotation mark, or mixture of alphabetic and non-alphabetic characters were flagged as invalid, effectively removing them from further consideration. Finally, the remaining corpus tokens were heuristically checked for consistency with an independently constructed canonicalization lexicon derived from an online error database, and target forms were checked against the TAGH morphology system for contemporary German (Geyken/Hanneforth 2006). Inconsistent pairs and unknown target forms were flagged as suspicious and subjected to an additional manual review phase. A total of 12,121 types (57,542 tokens) were flagged as suspicious in this manner, and at the time of writing (November, 2012), 55,059 tokens of 9,686 suspicious types had been manually checked and re-incorporated into the corpus. The final trimmed corpus used for the current experiments contained 5,444,888 tokens of 205,055 distinct pair-types. Of these, 4,916,639 tokens of 173,532 distinct types occurred in sentences containing no suspicious or purged material.

2.1 Test corpus

We used that subset of the corpus which had been subjected to the most thorough manual scrutiny⁶ as a ground-truth test corpus for evaluation. After applying the corpus trimming heuristics described above, the test corpus contained 378,300 tokens of 28,012 distinct pair types in 17,472 sentences. Of these, 319,866 tokens of 27,561 distinct pair types contained only alphabetic

⁶ The “prototype corpus” as described in Jurish/Drotschmann/Ast (2013).

characters and were thus considered “word-like”. Identity canonicalizations accounted for 250,382 word-like tokens (78%) of 15,454 distinct types (56%).

2.2 Training corpus

In order to achieve as accurate as possible a picture of the effectiveness of a corpus-induced canonicalization lexicon, all works by or about any author represented in the test set were excluded from the training set. The final training corpus used for the current experiments contained material from 45 distinct authors distributed over 101 volumes published between 1785 and 1901. After removing all sentences containing questionable material, the training corpus contained 4,180,924 tokens of 161,148 distinct pair types in 194,678 sentences. Of these, 3,511,679 tokens of 158,074 distinct pair types were “word-like”. Among word-like tokens, 2,737,398 (78%) of 79,882 distinct types (51%) were identity canonicalizations of the form (w, w) .

2.3 Canonicalization lexicon

The training corpus described above was used to bootstrap a finite canonicalization lexicon. Raw frequency counts $f(w, \tilde{w})$ for pairs of historical source word w and contemporary target word \tilde{w} were computed over the pruned training corpus. The finite corpus-based canonicalization lexicon was defined by the simple expedient of mapping each source type w represented in the training corpus to that target type $\text{CLEX}(w)$ with which it occurred most frequently:⁷

$$(1) \text{CLEX}(w) = \arg \max_{\tilde{w} \in \Sigma^*} f(w, \tilde{w})$$

Of course, such a deterministic type-wise mapping cannot account for any ambiguity whatsoever, but the frequency-maximization heuristic should act to ensure that the correct target form is returned for most input tokens of any known source type. Only 856 of the training corpus source types (<1%) had ambiguous canonicalizations modulo letter case, and only 1,626 training corpus tokens (<0.1%) would have been incorrectly canonicalized by the frequency maximization heuristic from equation (1). For the effectiveness of a corpus-trained lexicon on previously unseen text, unknown words – words present in the input for which no training data was available – have a far greater impact. In order to extend the finite function $\text{CLEX}(\cdot)$ to a total canonicalization func-

⁷ Σ is a finite character alphabet. In case multiple maximally frequent target forms were found, one was chosen randomly.

tion $\text{LEX}: \Sigma^* \rightarrow \Sigma^*$ which produces some output string for every possible input string, a fallback strategy was implemented which maps any unknown input word to itself:

$$(2) \text{LEX}(w) = \begin{cases} \text{CLEX}(w) & \text{if defined} \\ w & \text{otherwise} \end{cases}$$

2.4 HMM canonicalizer

The robust generative canonicalization architecture described in Jurish (2012: Ch. 4) employing a dynamic Hidden Markov Model to disambiguate type-conflation hypotheses was used here to provide a generic corpus-independent token-level canonicalization function.⁸ For the current experiments, an “intensional” phonetic equivalence cascade with an infinite weighted target lexicon derived from the TAGH morphology transducer (Geyken/Hanneforth 2006) was used in place of a finite target lexicon.

3. Method

We used the relevance relation derived from the test corpus to compare three different canonicalization techniques: the generic corpus-independent canonicalizer (HMM) from section 2.4, the corpus-based canonicalization function with identity fallback (LEX) from section 2.3, and a hybrid method (HMM+LEX) which canonicalizes known words according to the finite corpus canonicalization function $\text{CLEX}(\cdot)$, passing any unknown words to the generic HMM canonicalizer. More precisely, the hybrid method passed all sentences in their entirety through the HMM canonicalizer, but each token instantiating a known word type w was assigned a singleton set of canonicalization hypotheses containing only the unique lexicon entry $\text{CLEX}(w)$ for that type, effectively restricting the output of the model for known words while still allowing context-dependent disambiguation of unknown words, and even allowing the model to make direct use of the corpus-based canonicalizations in its computation of path probabilities.

⁸ A “hypothetical dictionary” in the terminology used by Gotscharek et al. (2009b).

3.1 Evaluation measures

The various canonicalization methods were evaluated using the ground-truth test corpus from Section 2.1 to simulate an information retrieval task. Formally, let $G = \langle g_1, \dots, g_N \rangle$ represent the test corpus,⁹ where each token g_i is a pair $\langle w_i, \tilde{w}_i \rangle$ such that \tilde{w}_i is the (modern) canonical cognate for the (historical) word w_i . Let $C = \{\text{HMM}, \text{LEX}, \text{HMM}+\text{LEX}\}$ be the finite set of canonicalizers under consideration. Then, for each test corpus token g_i and for each canonicalizer $c \in C$, let $\llbracket \tilde{w}_i \rrbracket_c$ represent the unique canonical form returned by the canonicalizer c for the token g_i . Let $Q = \bigcup_{i=1}^N \{\tilde{w}_i\}$ be the set of all canonical cognates represented in the corpus, and define for each canonicalizer $c \in C$ and query string $q \in Q$ the sets $\text{relevant}(q)$ and $\text{retrieved}_c(q) \subset \mathbb{N}$ of *relevant* and *retrieved* corpus tokens as:

$$(3) \text{ relevant}(q) = \{i \in \mathbb{N} : q = \tilde{w}_i\}$$

$$(4) \text{ retrieved}_c(q) = \{i \in \mathbb{N} : q = \llbracket \tilde{w}_i \rrbracket_c\}$$

Token-wise precision ($\text{pr}_{\text{tok},c}$) and recall ($\text{rc}_{\text{tok},c}$) for the canonicalizer c can then be defined as:

$$(5) \text{ pr}_{\text{tok},c} = \frac{|\bigcup_{q \in Q} \text{retrieved}_c(q) \cap \text{relevant}(q)|}{|\bigcup_{q \in Q} \text{retrieved}_c(q)|}$$

$$(6) \text{ rc}_{\text{tok},c} = \frac{|\bigcup_{q \in Q} \text{retrieved}_c(q) \cap \text{relevant}(q)|}{|\bigcup_{q \in Q} \text{relevant}(q)|}$$

Type-wise measures $\text{pr}_{\text{typ},c}$ and $\text{rc}_{\text{typ},c}$ are defined analogously, by mapping the token index sets of Equations (3) and (4) to corpus types before applying Equations (5) and (6). We use the unweighted harmonic precision-recall average F (van Rijsbergen 1979) as a composite measure for both type- and token-wise evaluation modes:

$$(7) F(\text{pr}, \text{rc}) = \frac{2 \cdot \text{pr} \cdot \text{rc}}{\text{pr} + \text{rc}}$$

⁹ More precisely, only the “word-like” tokens of the test corpus were considered for evaluation purposes, and differences in letter case were ignored.

4. Results and discussion

Type- and token-wise precision (pr), recall (rc), and harmonic precision-recall average F for the three canonicalization techniques with respect to the test corpus are given in Table 1. Immediately obvious from the observed data is that while both the HMM and corpus-based methods are quite effective in their own right ($F_{tok} > .99$ in both cases), the best performance across the board is achieved by the hybrid method HMM+LEX, as anticipated in light of the data from Gotscharek et al. (2009b). The data also show a clear discrepancy in type-wise recall between the LEX and HMM methods. This effect can be attributed to insufficient data in the training corpus: the finite corpus-based canonicalization lexicon itself provided canonicalizations for only 81.7% of test corpus types representing 97.3% of test corpus tokens;¹⁰ the remaining types were handled by the string identity fallback strategy for the LEX condition. Less than half (40.9%) of these “unknown” word types were correctly canonicalized by the fallback strategy, representing slightly more than half of the unknown tokens (51%). It is worth noting that the recall of the identity fallback strategy was substantially poorer on unknown words than test-corpus globally, where it achieved a type-wise recall of 55.7% and a token-wise recall of 78.5%. This implies that a disproportionately large number of the test corpus types not present in the training corpus were in fact non-trivial historical spelling variants, since valid contemporary forms would be canonicalized correctly by the identity fallback strategy.

Replacing the naïve identity fallback strategy with the HMM canonicalization architecture in the condition HMM+LEX resulted in correct canonicalization for 80.1% of the unknown types representing 77.8% of unknown tokens. This is relatively unsurprising, since the HMM canonicalizer is explicitly designed to deal with previously unseen input types, whereas the corpus-based canonicalization lexicon can only be hoped to correctly canonicalize those types for which training data was available with any reliability. The benefits of combining corpus-based and robust generative techniques were not all one-way however: the HMM canonicalizer also benefited from inclusion of the corpus-based exception lexicon. The hybrid method HMM+LEX incurred 18-31% fewer type-wise errors and 33-53% fewer token-wise errors than the HMM canonicalizer on its own, although these differences are of smaller absolute magnitude com-

¹⁰ The high coverage rate of the corpus-based lexicon is itself predicted by Heaps' Law (Heaps 1978), a correlate of the more widely known Zipf rank-frequency correlation (Zipf 1949) which states that there is a log-linear correlation between vocabulary size in types and corpus size in tokens.

pared to the effects on type-wise LEX recall. Differences in this region of the evaluation scale must be viewed with a modicum of skepticism for a test corpus of the current size, since the observed discrepancies result from differences in the canonicalizations of only 511 types (2,595 tokens). Nonetheless, we believe that given the quality of our test corpus, the observed recall improvements at least are robust enough to survive replication on a larger scale.

	Types			Tokens		
	pr_{typ}	rc_{typ}	F_{typ}	pr_{tok}	rc_{tok}	F_{tok}
LEX	.990	.878	.931	.998	.985	.992
HMM	.983	.936	.959	.996	.985	.985
HMM+LEX	.986	.957	.971	.998	.993	.995

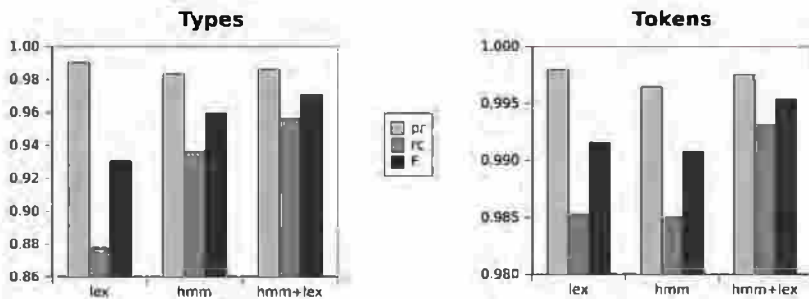


Table 1: Comparison of three canonicalization techniques: a generic Hidden Markov Model canonicalizer (HMM), a corpus-induced exception lexicon (LEX), and a generic canonicalizer supplemented by a corpus-induced lexicon (HMM+LEX). The maximum value in each column appears in boldface type.

5. Conclusion

We used a simulated information retrieval task over a semi-automatically constructed ground-truth corpus of historical German text to compare the performance of three different canonicalization techniques: a generic dynamic Hidden Markov Model disambiguation cascade, a static type-wise canonicalization lexicon trained from a canonicalized corpus, and a hybrid architecture which uses the generic method to canonicalize only those input words for which no training data was available. The observed results showed that while both the HMM and corpus-based techniques were quite effective on their own, the hybrid technique outperformed both of them in both type- and token-wise F .

The most drastic improvements were observed in type-wise recall for the hybrid method with respect to the corpus-based lexicon, assumedly due to data sparsity problems for the corpus-based method from which the HMM method does not suffer as acutely. Substantial improvements were observed in both precision and recall for the hybrid method with respect to the HMM canonicalizer as well, which suggests that these two methods complement one another if both a large canonicalized training corpus and a robust canonicalization cascade are available.

Acknowledgements

The work described here was funded by a *Deutsche Forschungsgemeinschaft* (DFG) grant to the project *Deutsches Textarchiv*. Additionally, the authors would like to thank Marko Drotschmann, Alexander Geyken, Susanne Haaf, Thomas Hanneforth, Lothar Lemnitzer, and Kai Zimmer for helpful feedback, questions, comments, and assistance with various stages of the work described here.

References¹¹

- Baron, Alistair/Rayson, Paul (2009): Automatic standardization of texts containing spelling variation, how much training data do you need? In: Mahlberg, Michaela/González-Díaz, Victorina/Smith, Catherine (eds.): Proceedings of the Corpus Linguistics Conference CL2009. University of Liverpool, UK, 20-23 July. http://uclrel.lancs.ac.uk/publications/cl2009/314_FullPaper.pdf.
- Bollmann, Marcel/Petran, Florian/Dipper, Stefanie (2011): Rule-based normalization of historical texts. In: Proceedings of Language Technologies for Digital Humanities and Cultural Heritage Workshop. Hissar, Bulgaria, 16 September, 34-42. www.linguistics.ruhr-uni-bochum.de/~dipper/papers/ranlp11.pdf.
- Dipper, Stefanie/Schultz-Balluff, Simone (2013): The *Anselm Corpus*: methods and perspectives of a parallel aligned corpus. In: Proceedings of the workshop on computational historical linguistics at NoDaLiDa 2013. (= NEALT Proceedings Series 18 / Linköping Electronic Conference Proceedings 87). Linköping: Linköping University Electronic Press, 27-42. www.ep.liu.se/ecp/087/003/ecp1387003.pdf.
- Ernst-Gerlach, Andrea/Fuhr, Norbert (2007): Retrieval in text collections with historic spelling using linguistic and spelling variants. In: Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '07). New York: ACM, 333-341. http://leland.is.inf.uni-due.de/bib/pdf/ir/Ernst_Fuhr_07.pdf.

¹¹ All URLs have been checked and found valid as of late January 2015.

- Geyken, Alexander/Hanneforth, Thomas (2006): TAGH: A complete morphology for German based on weighted finite state automata. In: *Finite State Methods and Natural Language Processing, 5th International Workshop, FSMNLP 2005, Revised Papers.* (= Lecture Notes in Computer Science 4002). Berlin: Springer, 55-66. http://dx.doi.org/10.1007/11780885_7.
- Gotscharek, Annette/Neumann, Andreas/Reffle, Ulrich/Ringlstetter, Christoph/Schulz, Klaus U. (2009a): Enabling information retrieval on historical document collections: the role of matching procedures and special lexica. In: *Proceedings of AND '09*. New York: ACM, 69-76. <http://doi.acm.org/10.1145/1568296.1568309>.
- Gotscharek, Annette/Reffle, Ulrich/Ringlstetter, Christoph/Schulz, Klaus U. (2009b): On lexical resources for digitization of historical documents. In: *Proceedings of DocEng '09*. New York: ACM, 193-200. <http://doi.acm.org/10.1145/1600193.1600236>.
- Hauser, Andreas/Heller, Markus/Leiss, Elisabeth/Schulz, Klaus U./Wanzeck, Christiane (2007): Information access to historical documents from the Early New High German period. In: *Proceedings of the IJCAI-07 Workshop on Analytics for Noisy Unstructured Text Data (AND-07)*. New York: ACM, 147-154. http://research.ihost.com/and2007/cd/Proceedings_files/p147.pdf.
- Heaps, Harold S. (1978): *Information Retrieval: Computational and Theoretical Aspects*. Orlando: Academic Press.
- Jurish, Bryan (2012): *Finite-state canonicalization techniques for Historical German*. PhD thesis, Universität Potsdam. http://opus.kobv.de/ubp/volltexte/2012/5578/pdf/jurish_diss.pdf.
- Jurish, Bryan/Drotschmann, Marko/Ast, Henriette (2013): Constructing a canonicalized corpus of historical German by text alignment. In: Bennett, Paul/Durrell, Martin/Scheible, Silke/Whitt, Richard Jason (eds.): *New Methods in Historical Corpora.* (= *Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache / Corpus Linguistics and Interdisciplinary Perspectives on Language (CLIP) 3*). Tübingen: Narr, 221-234.
- Kempken, Sebastian/Luther, Wolfram/Pilz, Thomas (2006): Comparison of distance measures for historical spelling variants. In: Bramer, Max (ed.): *Artificial Intelligence in Theory and Practice*. Boston: Springer, 295-304. http://dx.doi.org/10.1007/978-0-387-34747-9_31.
- Porta, Jordi/Sancho, Jose-Luis/Gómez, Javier (2013): Edit transducers for spelling variation in Old Spanish. In: *Proceedings of the workshop on computational historical linguistics at NoDaLiDa 2013.* (= NEALT Proceedings Series 18 / Linköping Electronic Conference Proceedings 87). Linköping: Linköping University Electronic Press, 70-79. www.ep.liu.se/ecp/087/006/ecp1387006.pdf.
- Rayson, Paul/Archer, Dawn/Smith, Nicholas (2005): VARD versus Word: A comparison of the UCREL variant detector and modern spell checkers on English histori-

- cal corpora. In: Proceedings of the Corpus Linguistics Conference 2005, Birmingham, UK, 14-17 July. http://eprints.lancs.ac.uk/12686/1/cl2005_varword.pdf.
- Reffle, Ulrich/Gotscharek, Annette/Ringlstetter, Christoph/Schulz, Klaus U. (2009): Successfully detecting and correcting false friends using channel profiles. In: *International Journal on Document Analysis and Recognition* 12(3): 165-174. <http://dx.doi.org/10.1007/s10032-009-0091-y>.
- Scheible, Silke/Whitt, Richard J./Durrell, Martin/Bennett, Paul (2011): A gold standard corpus of Early Modern German. In: Proceedings of the 5th Linguistic Annotation Workshop. Portland, Oregon: ACL, 124-128. www.aclweb.org/anthology/W11-0415.
- van Rijsbergen, Cornelis J. (1979): *Information Retrieval*. Newton, MA: Butterworth-Heinemann.
- Zipf, George K. (1949): *Human Behaviour and the Principle of Least-Effort*. Cambridge, MA: Addison-Wesley.

A new LMF schema application

An Austrian lexicon applied to the historical corpus of the writer Hugo von Hofmannsthal

Abstract

In this paper, the creation and representation of a digital dialect lexicon from existing internet sources and books is presented. The creation procedure can serve as a role model for similar projects on other dialects and points toward a new cost saving way to produce NLP resources by use of the internet in a manner similar to human-based computation. Dialect lexica can facilitate NLP and improve POS-tagging for German language resources in general. The representation standard used is LMF. It will be demonstrated how this lexicon can be used as a tool in literature science, linguistics, and computational linguistics. The critical edition of Hugo von Hofmannsthal's works is a well-suited corpus for the aforementioned research fields and the inspiration to build this tool.

1. Introduction

In NLP, lexical resources are employed for various tasks such as POS-tagging, stemming, information retrieval and so forth. In order to study corpora with dialectal components, the lexicon must contain the dialectal words. For the digital representation of lexica, various standards have emerged throughout the last decades. The ISO standard of LMF (“Lexical Mark-up Framework”; ISO 24613, 2005) is one of the most versatile platforms for lexicon representation, and it integrates several state-of-the-art features. In its beginning a UML specification, the website www.lexicalmarkupframework.org/ provides an XML DTD¹ with one base file and several extensions. For the linguistic features, LMF uses another ISO standard encapsulated into a *feat* tag, the Data Category Registry.²

In the next section we propose an LMF model as the data exchange format used in our system. In section 3 we present the application framework for which the lexicon has been designed, followed by a recapitulation of the pecu-

¹ www.tagmatica.fr/lmf/DTD_LMF_REV_16.dtd. All URLs have been checked and found valid as of late January 2015.

² www.isocat.org/.

liarities of Austrian German and the description of the reusable workflow of a dialect lexicon creation procedure. The corpus investigated in section 6 contains the works of Hugo von Hofmannsthal, an Austrian writer who lived between 1874 and 1929, thus subject to the effects of two orthographic conferences. It encompasses the critical edition volumes 6, 7, 16, 17, 19, 21, 22, 25-1, 27, 33 and 34 and includes, amongst others, dramas, poems, essays, and narratives. Besides the authorized texts of Hofmannsthal, his literary legacy contains a number of fragments that are published in the critical edition of the Freies Deutsches Hochstift in Frankfurt/Main. Although the critical edition in its apparatus contains a glossary in volume 34, the glossary is not comprehensive and we aim at providing a more comprehensive resource along with a generalizable workflow of its creation. Section 7 concludes with a brief summary and an outlook.

2. The LMF model

Various standards for the digital representation of lexica have emerged. Detailed descriptions of the developments can be found in Budin/Majewski/Mörth (2012). Francopoulo et al. (2006) give an insight into the emergence of the ISO standard of LMF, which integrated various former projects. LMF as an ISO standard was chosen as the primary representation format, but it is planned to offer TEI and RDF as alternative data exchange formats. Based on the two ISO standards of LMF and Data Category Registry (DCR), which is “maintained as a global resource by ISO TC37” (Francopoulo et al. 2006), a new representation model for lexica has been worked out, which is being applied to a number of historical corpora within the research cluster “Digital Humanities” of the LOEWE initiative of the state of Hesse.³ The following representation based on these standards was developed for integrative use in the module called *Lexicon Browser* within the humanities computing platform of the *eHumanities Desktop* (Mehler et al. 2009) described in section 3.

2.1 The header

The header specification follows the LMF proposal quite strictly and incorporates the structure in a minimal way. The language and its ISO code are represented.

³ www.digital-humanities-hessen.de/.

Listing 1: The LMF header for a Lexical Resource

```

<GlobalInformation>
  <feat att="languageCoding" val="ISO 639-3" />
</GlobalInformation>
<Lexicon>
  <feat att="language" val="deu" />
  * * *

```

2.2 Lexical entries

The representation of *Lexical Entries* can become complex. In the following example, a dialectal/historical spelling variant is encoded as a *LexicalEntry* featuring *FormRepresentations* of its *WordForms*:

Listing 2: An example word form

```

<LexicalEntry id="18339941">
  <Lemma>
    <feat att="type" val="dialectal" />
    <feat att="dialect" val="Austrian" />
    <feat att="label" val="Abschnittzeile" />
    <feat att="description" val="abgeschnittenes kleines Stück" />
    <feat att="part-of-speech" val="noun" />
    <feat att="gender" val="n" />
  </Lemma>
  <WordForm>
    <FormRepresentation>
      <feat att="id" val="18339942" />
      <feat att="label" val="Abschnittzeile" />
      <feat att="case" val="nominativeCase" />
      <feat att="number" val="sg" />
    </FormRepresentation>
    <FormRepresentation>
      <feat att="id" val="18339943" />
      <feat att="label" val="Abschnittzeile" />

```

```

        <feat att="case" val="nom inativeCase" />
        <feat att="number" val="sg" />
    </FormRepresentation>
</WordForm>
<WordForm>
    <FormRepresentation>
        <feat att="id" val="18339944" />
        <feat att="label" val="Abschn itze ls" />
        <feat att="case" val="gen itiveCase" />
        <feat att="number" val="sg" />
    </FormRepresentation>
</WordForm>

```

...

Each lexical entry must by definition have a *Lemma*, must carry the attributes of *type*, *label*, *description*, and *part-of-speech* plus additional features that are not subject to change within the inflectional paradigm of the present part of speech. Additional grammatical attributes are features of the corresponding *WordForm*. For a noun, e.g., gender is a feature of the *Lemma*, as it never changes regardless of case, while case itself is a feature of the *WordForm* and therefore not pertaining to the *Lemma*.⁴ These restrictions are language specific. An important feature of the *Lemma* is the *type*, which in our system can have the value *dialectal*; the dialect's name is specified in the next feature, *dialect*. The lemma's ID corresponds to the ID of the lexical entry as a whole, i.e. the lexicon is sorted by lemmata as the upmost hierarchical layer, word forms having their own IDs subsequent to their lemma's ID. If an entry combines different spellings, as is often true for historical or dialectal word forms, these are encapsulated in a *FormRepresentation*. All *FormRepresentations* constitute one word form.

It is exactly this representation that is used to display dialectal and variant spellings. *Synsets* are used to group synonymous semantics or senses for dialectal items and their standard counterpart if existent. Additional export formats include RDF and TEI.

⁴ If a word form happens to represent a different part of speech than its lemma and there is therefore a conflicting value in the parent-lemma's default set, the word form's feature overwrites the lemma's feature.

3. The application framework

Any lexicon which uses the above described LMF as an input/output format can be managed within the *eHumanities Desktop*. The *eHumanities Desktop* is a browser-based web-interface allowing users to share, organize, and analyse resources. Once uploaded into the *Lexicon Browser* via LMF, the user interface makes a lexicon browseable, performs search operations and obtains statistics connected with every single entry. The interface shows, e.g., word forms connected to a query and provides grammatical information pertaining to them in a human-readable way. Additionally, it displays the information graphically in a network. The lexicon can further be connected with a text. Frequency distributions and collocation statistics are available through another module within the *eHumanities Desktop* called *Historical Semantics Corpus Management* (see Jussen/Mehler/Ernst 2007 and Mehler et al. 2011). The user can annotate, re-annotate and perform online reindexations in order to keep the (statistical) information up to date. In Gleim/Mehler/Ernst (2012) the application framework of lexica and corpora management and its architecture are described in greater detail.

4. Austrian German

German is a so-called pluricentric language (Clyne 1992), i.e., there is more than one center from which standardisation processes spread, leading to a mosaic of different partly overlapping substandards, varieties, and dialects. Additionally, a plurality of countries with German as a national language exist. One of these countries is Austria. Austrian German has as many as three different neighbouring non-Germanic language families (Slavic, Finno-Ugric, and Romance) plus some additional sources for calques and loans like Yiddish or Rotwelsh (cf. Beyerl/Hirtner/Jatzek 2009; Wiesinger 1990). Research on Austrian began at least as early as 1774 under the empress Maria Theresia, under whom the abbot Johann Ignaz Felbiger created a schoolbook with first lists of Austrian terms (Back et al. 2009). The orthography of German in Austria was administered in Vienna, while throughout the 19th century Prussia and other German regions kept defining their own standards. In order to resolve differences within the German speaking countries, two orthographic conferences (1876 and 1901) were held.

Research encompassing the Austrian variety of German in former times led to the production of various printed lexica, the most important of which continues to be used in Austrian schools (Back et al. 2009). There is also an EU protocol of some 30 Austrian terms with their counterparts in Standard German (Markhardt 2005). On the internet, various sites with dialectal content can be found in guestbooks, forums and chats (Bashaikin 2005: 444), and a Wikipedia for the Bavarian dialect group exists.⁵ Still, to the best knowledge of the authors there is no digital annotated dialect lexicon or word list publicly available. The ICLTT⁶ offers the *Wörterbuch der bairischen Mundarten in Österreich* on a commercial basis.

Linguistically, Austrian dialects belong to the Bavarian dialect continuum. Different subvarieties are attested (see, for instance, Rowley 1990; Wiesinger 1990). This adds an element of complexity to the lexicon structure. Concerning the orthography, Auburger (2011) notes that there is no widely accepted standard yet for the written manifestation of the Bavarian dialects. In written language, the following non-phonological features are widely applied distinguishing the dialect from the standard (Wiesinger 1990):

- a) lexical features;⁷
- b) 2nd plural verb endings;⁸
- c) diminutives;⁹
- d) gender differences;¹⁰
- e) differences in the use of prepositions;¹¹

⁵ <http://bar.wikipedia.org/>.

⁶ Institut für Computerlinguistik und Texttechnologie, Austrian Academy of Sciences.

⁷ Certain words like *Schmäh* (“nonsense”) or terms from cuisine like *Zibebn* (“raisins”).

⁸ The second person plural is marked by an *s* distinguishing it from the first and third person plural in verbal inflection, while in Standard German and Dutch the second person plural forms are not marked in the same way, cf. *Ihr gebts des dem Hansl* as opposed to *Ihr gebt das dem Hansl/Hänslein/Hänschen* (“You will give this to Hans”).

⁹ *Hans*, a proper name, becomes *Hansl* with the reduced form of the diminutive suffix *-lein* as used in the standard language. The standard has another diminutive suffix *-chen*, which is not used in Bavarian. Note that for Bavarian, no umlaut takes place while the standard diminutive form of *Hans* would be either *Hänschen* or *Hänslein*.

¹⁰ E.g., *das Teller* (“the plate”, neuter) as opposed to *der Teller* (masculine).

¹¹ *Ich möcht beim Calafatti fahrn* (“I want to go to Mr. Calafatti”). Here the standard would use the preposition *zu* instead of *bei*, *bei* would even be ungrammatical.

f) formation of the perfect tense with the ‘to be’ auxiliary for all verbs of motion and posture.¹²

For additional features see Wiesinger (1990).

5. Lexicon creation

As Geyer points out, “Large-area-lexica like the lexicon of the Bavarian dialects in Austria (WBÖ) have a long collection period and a long publication phase.”¹³ The author refers to the time-range of the emergence of this print lexicon, which accounts for 107 years (1913-2020). The main sources are informants, i.e. laymen, who have in this case answered questionnaires. In our case, the advent of the digital age enabled us to significantly shorten this period. First, the digitized glossary of Austriacisms from the critical edition of von Hofmannsthal’s works was included in our lexicon, but additionally, the internet was the source for the majority of entries. Certain sites do provide lexica for Austrian. The major portion of these are created and maintained by laymen and some are cooperative sites, where each entry stems from a blogger who adds it to the compilation.

The procedure being applied is making use of the internet in much the same way as human-based computation.¹⁴ The people who created the lexicon sites for the public enabled the project to build a lexical resource without having to rely on the time consuming process of constructing questionnaires. This accelerates the production of an annotated NLP dialect lexicon significantly. The following sites were taken as a basis for the creation in March 2012:¹⁵

oesterreichisch.net,
 oewb.retti.info,
 ostarrichi.org,
 unsere-sprache.at,
 de.wikipedia.org/wiki/Liste_von_Austriazismen,

¹² Standard German mostly uses “to have”: *ihr seids da gesessn* (“you sat there”) vs. *ihr habt da gesessen*.

¹³ “Großlandschaftswörterbücher wie das Wörterbuch der bairischen Mundarten in Österreich (WBÖ) haben eine lange Sammelphase und eine lange Publikationsdauer“ (Geyer 2005: 195).

¹⁴ See for instance http://en.wikipedia.org/wiki/Human-based_computation.

¹⁵ All entries of pages other than www.openthesaurus.de/, the Wikipedia, the EU Protocol, and <http://oewb.retti.info> have been deleted from the resulting list as they are not freely available.

openthesaurus.de/synset/variation/at,
sistlau.blogspot.de,
german.about.com/od/vocabulary/a/Austrian.htm, and
das-oesterreichische-deutsch.at.

The EU Protocol 10¹⁶ issued when Austria entered the EU was also taken into account, as well as entries from a dialect guide published by Beyerl/Hirtner/Jatzek (2009). Having a wider range and a bigger variety of tools was one advantage of merging many sites, although one disadvantage of this approach was that it also required some additional unification and filtering efforts. After having identified the resources, the following steps were used in the lexicon creation process:

- a) downloading the entries from the internet
- b) unifying the format
- c) merging the entries
- d) resorting and deleting duplicate entries
- e) removing Standard German entries without the loss of “false friends”
- f) detecting and relating synonyms
- g) annotating POS-tags
- h) expanding the word forms
- i) testing the application of the lexicon.

In order to ensure correctness, a lot of manual labour was involved for control of the output at every step. The workflow is a general procedure that can be applied to any such task. Entries were first downloaded through a script in the Java programming language, followed by a unification of format. The chosen format is a table-layout, featuring columns and rows, where each row is a new word form while each column covers an information type. The first column contains the Austrian word; the second column bears additional grammatical information if present; the third column includes a verbal description of the meaning of the word; the fourth column features additional information; and the last column contains the sourcesite URL. Later on, two additional columns for POS-tags and lemmas were filled. The lexicon has been integrated with the resources of the POS-tagger freely available under <http://api.hucompute.org/>

¹⁶ http://ec.europa.eu/translation/german/guidelines/documents/austrian_expressions_de.pdf.

preprocessor. In the following subchapters, some of the steps will be recapitulated in detail for better understanding.

5.1 Removing Standard German entries without losing “false friends”

In order to avoid capturing Standard German words, which are already present in the non-dialectal lexicon, tokens had to be removed if they were congruent with the standard in both meaning and spelling. As has been shown several times, there is no objective criterion to draw a clear line between the notion of language and dialect (see, e.g., Thomason 2001: 2). For the authors of dialectal lexica, this is a challenge. Certain words will clearly be borderline cases. Some of the judgements might even depend on the lexicon author’s command of standard and dialect. For the sake of consistency and comprehensiveness, every word can be accepted for input, so as to certainly capture all the core items. However, for the reasons mentioned above,¹⁷ in order to be more restrictive, a check for entries overlapping with Standard German was performed. The list of the 100,000 most frequent German words as published by the IDS Mannheim¹⁸ has been intersected with the lexicon. The widely known items were removed. Many such words were colloquial or curse words, a category barely written and therefore more easily perceived as dialectal (considering that the formula “that which you see written is the standard, that which you hear is the dialect” is easy to understand and memorize and at the same time a sufficient explanation for the dichotomy of language and dialect). If a Standard German term had a completely different meaning in Austrian German, commonly labeled a “false friend”, it would be kept. A more subtle case for the decision on tokens were entries which have a Standard German equivalent with the exact same meaning, but which are affected by minor sound alternations. In principle, the aim of the lexicon was to capture those elements which are not at all present in dictionaries of the standard and which are not necessarily detected as variants by established measures like the Levenshtein distance (Levenshtein 1965). Hence, tokens with phonologically close cognates were only accepted if the number of phonological differences (or their degree) rendered the token incomprehensible when appearing in a standard context. For the decision, native (non-Austrian German) speaker intuition was used as the benchmark.

¹⁷ For instance, an Austrian term that is widely used in the standard and has no immediate alternative synonym there should not be included in the dialectal lexicon.

¹⁸ www1.ids-mannheim.de/kl/projekte/methoden/derewo.html.

5.2 Synonyms and variants

Another step in the creation of the lexicon was the treatment of synonyms and variants. Synonyms were treated as interconnected item sets with synonym relations to the ids of other tokens, and tokens with different senses were duplicated and displayed as non-connected. Table 1 and Table 2 display token relationships.

unique ID	token	translation	synonym set
9541	Hopertatsch	ungeschickter Mensch	[9541,9550,231601]
9550	Hoppadatschi	ungeschickte Person	[9541,9550,231601]
23160	Hirsch	ungeschickte Person	[9541,9550,231601]

Table 1: Synonyms and variants

unique ID	token	translation
9558	hoppertatschert	überheblich/ungeschickt
9559	hoppertatschig	überheblich/ungeschickt
9558	hoppertatschert	überheblich
9559	hoppertatschert	ungeschickt
9560	hoppertatschig	überheblich
9561	hoppertatschig	ungeschickt

Table 2: Separating homonyms

5.3 POS-tagging and lemmatising

In order for the lexicon to become a useful NLP resource, basic parts of speech (ADJ, NN, V, PART, ADV) were annotated. The major part was annotated automatically according to some rules, others after that by hand, at the same time checking for mistakes of the automatic assignment. Luckily, nowadays and in the critical edition German nouns are capitalised, so with very great certainty¹⁹ capitalised entries are annotated as nouns (rule1); verbs were more demanding, but as lexicon entries they often end in the infinitive ending *-en* (rule 2).

¹⁹ The beginning of a sentence is an exception. Each word regardless of part of speech is capitalized here.

The evaluation of the automatic assignment is given in Table 3. Finally, adjectives were annotated on the basis of suffixes such as *-lich* (rule 3).

As for the lemmatisation, lemmas were easily annotated as a lexicon entry is already a lemma. At the end of this step the lemma list (not yet a lexicon) contained 19,479 tokens: 12,192 nouns, 3,144 verbs, 1,389 adjectives, 388 adverbs. 2,061 entries contained at least one space character, thus using two words at minimum²⁰, which were categorized as multi-word units (MWU). Additionally there were 305 items which were either particles or suffixes or had the possibility to be interpreted as more than one part of speech. These were manually annotated.

word class	precision
(a) nouns	0.975
(b) verbs	0.96
(c) adjectives and other	0.65
mean ($\frac{a+b+c}{3}$)	0.93

Table 3: Automatic POS-Tagging

5.4 Expansion

For the ca. 20,000 obtained lemmas, an expansion scheme was set up according to the inflectional paradigms given in Wiesinger (1990). For nouns, verbs and adjectives, we produced ca. 112,000 word forms, each connected to their lemmas. This made the product a subset of a *full form lexicon*.²¹ An LMF version of the lexicon is available in the *eHumanities Desktop*.

6. Detection

The lemma list can serve as an input to detect patterns of dialect usage throughout a text. In order not to capture items present in the standard language, the lemma list was separated into lexical and semantic Austriacisms by again detecting the overlap with the big German lexicon used in the POS-tagger by

²⁰ Articles are not counted here; they had been separated beforehand.

²¹ *Full form lexicon* refers to a computational resource where each inflected word form is stored connected to its lemma. For a discussion see Carstensen et al. (2004: 519) on the German equivalent “Vollformenwörterbuch”.

Waltinger (2010). The lemma list was used for the detection of Austriacisms in a text. The detection was augmented by a simple matching of some idiosyncratic features of the Bavarian dialects not necessarily included in the model, such as:

- inflected auxiliaries in the second plural as listed by Wiesinger (1990) (*derfts, gehts, mögts, müssts, sollts, wollts* combined with *ihr*, as well as *habts, seids, tuts*);
- suffixes characterising Austrian diminutives through a regular expression, filtering false positives afterwards;
- the relative clause entry sequence *die wo* (Eroms 2005).

6.1 Application of the lexicon in linguistics

In linguistics, code-switching refers to various patterns of the use of different languages within the same discourse (Thomason 2001). This applies not only to languages but also to dialects (Niebaum/Macha 2006: 9), where the degree of dialect usage interwoven with the standard provides an additional layer of complexity. Within Hofmannsthal's oeuvre, dialectal elements appear only in a minority of his works. When they do, the dialect is graded, just like that of non-literary dialectal speech of competent speakers all around the world. Niebaum/Macha (2006) report on three typical types of dialect speakers identified by Lausberg (1993), namely code-switchers, code-mixers, and permanent dialect speakers. In Hofmannsthal's unfinished work *Wiener Pantomime* ("Viennese Pantomime"), a text which was never actually published and is only attested in fragments, not having undergone any revisionary or editorial processes, dialect is used in different ways, sometimes even within one discourse (dialectal material is underlined):

der römische Kaiser: Jetzt wieder schlafen gehen, Schmarr'n!
the Roman Emperor: Going to sleep again now, nonsense!

die Schäferin: Ich möcht beim Calafati fahr'n!
the Shepherdess: I want to go to the Calafatti!

der Herrnhuter: Was fallt Dir ein! Dir wer i's zeig'n!
the Herrnhuter: What are you thinking! I'll have the last laugh!

In the example, dialect has been applied in different degrees. The only word in the speech of the emperor which is clearly dialectal is his last word, *Schmarr'n* (“nonsense”), a lexical Austriacism. If we look at the personal pronouns and the verbal forms, the shepherdess and the Herrnhuter use different patterns. While the shepherdess uses dialectal verb-forms but the standard language's first person singular personal pronoun *ich*, the Herrnhuter uses both dialectal verbal inflection and the dialectal form of the first person singular pronoun (*i*). Thus they display different degrees of dialect application. Linguistic hypotheses that would have to be proven by looking at many more examples could be, for instance, whether:

- when dialectal pronouns are used, they are used for the entire class of pronouns and never only for one;
- when dialectal pronouns are used, dialectal verb-forms must be used as well, but never the other way round.

A digital text together with a detection tool would in this case facilitate the process of data acquisition.

An example of dialectal usage in spoken English at the intersection of linguistics and literature science stems from Gardner-Chloros (2009: 3) who finds a potential narrative use of varieties: “Sebba [a discourse participant] suggests that code-switching is used here to ‘animate’ the narrative by providing different ‘voices’ for the participants in the incident which is described.” The author refers to code switching between Creole and Standard English. This, however, is an example of unconscious usage. The next section will highlight some aspects of conscious usage of dialectal components for literary purposes.

6.2 Application of the lexicon in literature science

A quantitative analysis of the distribution of dialectal words can reveal in which genre of the critical Hofmannsthal edition Austriacisms accumulate. Austrian words are not restricted to the lexical level but include morphemes and inflectional elements as well. Figure 3 illustrates that besides the “Roman-Fragment”, it is especially Hofmannsthal's scenic works (Dramen, Libretti) that exhibit a remarkable cluster of dialectal terms. An ostensive example is “Der Schwierige”, because explicit Austrian terms as well as correspondent grammatical forms appear in large numbers in this play (Mauser 1982: 115). The critical edition offers a list of foreign terms and their meaning in the appendix

including English and French loanwords specific to Austria. In his article about “Der Schwierige”, Wolfgang Mauser ascribes the application of the existing Austriacisms to the accentuation of Austrian traditions (Mauser 1982: 115): the protagonists are members of the Austrian nobility and thus prefer an exalted lifestyle. The location of the story line – Austria – plus the clientele of the play are considered as the motivation for the application of the corresponding dialect. Apparently, the vernacular is used to refer to a sense of tradition and patriotism of the figures (Mauser 1982: 115). The used dialectal forms not only describe specific Austrian customs or local dishes which are unknown to foreigners but also comprise common terms such as *schurigeln* (“to bedevil”) or *tentieren* (“to intend”). Thus Austriacisms are not only utilized in cases when no Standard German term is available but are also applied instead of existing High German equivalents. Furthermore, Mauser claims that Hofmannsthal systematically exaggerates the application of Austrian terms and thus generates humour (Mauser 1982: 115). The conclusion for the usage of the Austrian dialect would thus be the ironicization of ancient Austrian traditions of the nobility by the author.

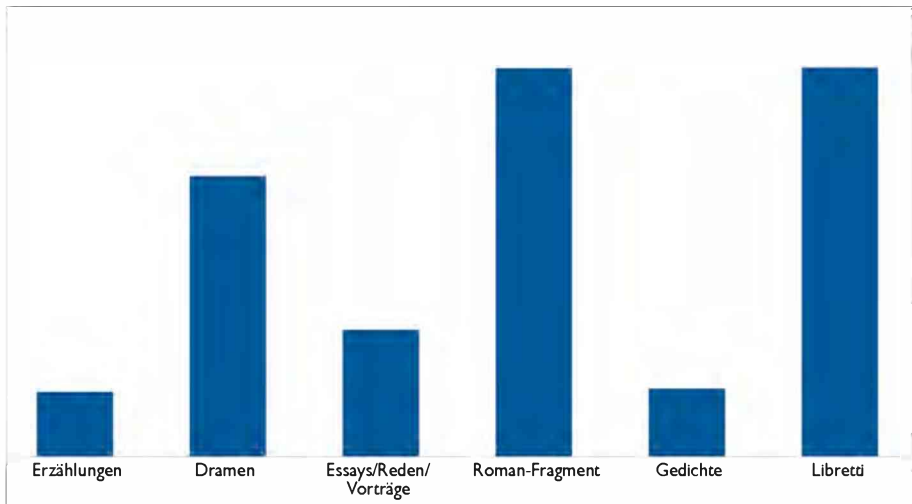


Figure 1: Austriacisms in different genres

Another starting-point regarding the analysis of the Austrian dialect as a stylistic device is the social context of literary characters. In 1919 Hofmannsthal himself wrote in the magazine “Die Theater- und Musikwoche” about his libretto “Die Frau ohne Schatten”:

Ich wollte das Ganze als Volksstück, mit bescheidener begleitender Musik, machen, zwei Welten gegeneinanderstehend, die Figuren der unteren Sphären im Dialekt. (Hofmannsthal 1998: 236).

The hierarchy of the two worlds here implicitly expressed as “high” versus “low” relates to a difference in the social status of the protagonists – a royal couple on the one side and a dyer and his wife on the other side. Whether the missing dialect in the speech of the sovereign couple can be ascribed to their exalted educational background must be left open.

Less obvious is the application of Austriacisms in Hofmannsthals opera “Der Rosenkavalier”. Although social hierarchies are illustrated via varying speech levels, the boundaries between dialectal forms and High German are vague here. The maid Mariandl speaks in the vernacular but her mistress and other aristocratic figures include dialectal terms and phrases in their talk, too (code-mixing). Hofmannsthal himself describes the setting of the opera – Vienna – as a city where social differences are mirrored in the manner of speaking:

[...] dieses Wien von 1740, eine ganze Stadt mit ihren Ständen, die sich gegeneinander abheben und miteinander mischen, mit ihrem Zeremoniell, ihrer sozialen Stufung, ihrer Sprechweise oder vielmehr ihren nach Ständen verschiedenen Sprechweisen [...]. (Hofmannsthal 1986: 549)

The Austrian historian Adam Wandruszka emphasises the dialectal dimension in the “Rosenkavalier”. He points out that this linguistic characteristic could be ascribed to a diary which was published a short time before Hofmannsthal began writing his libretto (Wandruszka 1967: 562). This diary was written by the controller of Maria Theresia’s household and describes the life at court in the contemporary Viennese dialect. Wandruszka hypothesises that Hofmannsthal was familiar with these texts (ibid.).

The unfinished piece “Wiener Pantomime” introduces a traditional Viennese character, “*den lieben August*” as its protagonist. His speech is distinctively Austrian on the lexical, grammatical and syntactic level. In contrast to August, other figures, for example the nymphs, are explicitly supposed to speak High German in the way “children recite wishes” (Hofmannsthal 2006: 139). Their speech is non-dialectal because they emerge from a mythological context. Besides, yet other characters expose a hybrid embodiment: even though the sovereign Ypsilanti is a Greek warrior for freedom (ibid., 682), his comments show a strong Austrian accent. Even the Roman emperor speaks in the Austrian dialect. The differentiation between the vernacular and Standard Ger-

man lies in the context of the figures, namely whether they have a mythological or historical background. Hence the dialectal speech in this context could possibly be connected to the condition of being human.

The dictionary for Austriacisms thus provides assistance for the analysis of literary texts in two ways. Besides the common function of a reference book to clarify terms, the dictionary can show via quantitative evaluation in which literary genre vernacular terms accumulate and help to analyse the respective motivation for the use of Austriacisms. Exemplarily it was ascertained above that Austrian dialectal forms occur mostly in the dramatic genre. The scenic character with direct speech can thus be evaluated as a criterion for the increased application of the dialect. Other possible indications are historical contexts and social hierarchies.

7. Conclusion

We presented an LMF model for the representation of a lexicon in the humanities computing environment *eHumanities Desktop*, discussed the peculiarities of dialectal lexica and the Austrian German dialect, and described a lexicon creation procedure which can serve as a role-model for other NLP dialect resources. This lexical resource is now part of the *eHumanities Desktop*. We showed how the resource can be used for philological or linguistic analyses by taking the example of the works in Austrian German by Hugo von Hofmannsthal.

The successful cooperation between specialists from the Humanities and computer science repeatedly involves decisions in NLP tasks where the corpus-suited development of methods and error rates must be counterweighed against the effort of manual labour, in trying to collaborate in the most cost-efficient way possible. With very large data, this might only be achievable by the application of software; for very small corpora, on the other hand, manual labour may be the quicker and more efficient way (consider the German saying *mit Kanonen auf Spatzen schießen* – “to shoot at sparrows with canons”). Historical corpora which are at the border between these two cases with respect to their size may be especially well-suited and fruitful for a digital humanities cooperation. The creation of this lexicon may serve as an example.

Acknowledgements

The LMF representation and the lexicon have been developed for the *eHumanities Desktop* in the lab for text-technology (computer science) at Goethe University Frankfurt in collaboration with the *Freies Deutsches Hochstift*. We would like to thank the federal state of Hesse's LOEWE program which is the financial source for the research cluster "Digital Humanities".²² Lastly, we would like to thank the authors of the websites from which we extracted our tokens.

References

- Auburger, Leopold (2011): Boarische Orthographie. Berlin: Pro Business.
- Back, Otto/Benedikt, Erich/Blüml, Karl/Ebner, Jakob/Hornung, Maria/Möcker, Hermann/Pohl, Heinz-Dieter/Tatzreiter, Herbert (2009): Österreichisches Wörterbuch. Wien: öbv.
- Bashaikin, Nikolaj (2005): Dialekt im Cyberspace. Überlegungen zu einigen sozio- und pragma-linguistischen Aspekten. In: Kraemer-Neubert (ed.), 439-450.
- Beyerl, Beppo/Hirtner, Klaus/Jatzek, Gerald (2009): Wienerisch – das andere Deutsch. Bielefeld: Reise Know-How Verlag.
- Budin, Gerhard/Majewski, Stefan/Mörth, Karlheinz (2012): Creating lexical resources in TEI P5. A schema for multi-purpose digital dictionaries. In: Journal of the Text Encoding Initiative 3: 2-18. <http://jtei.revues.org/522>.
- Carstensen, Kai-Uwe/Ebert, Christian/Endriss, Cornelia/Jekat, Susanne/Klabunde, Ralf/Langer, Hagen (2004): Computerlinguistik und Sprachtechnologie. München: Spektrum Verlag.
- Clyne, Michael (1992): Pluricentric languages. Differing norms in different nations. Berlin: Mouton de Gruyter.
- Eroms, Hans-Werner (2005): Relativsatzmarkierung im Bairischen. In: Kraemer-Neubert (ed.), 75-88.
- Francopoulo, Gil/George, Monte/Calzolari, Nicoletta/Monachini, Monica/Bel, Nuria/Pet, Mandy/Soria, Claudia (2006): Lexical markup framework. In: Proceedings of Language Resources and Evaluation Conference (LREC) 2006. Paris: ELRA: 233-236. www.lrec-conf.org/proceedings/lrec2006/pdf/577_pdf.pdf.
- Gardner-Chloros, Penelope (2009): Code-switching. Cambridge: Cambridge University Press.

²² www.digital-humanities-hessen.de.

- Geyer, Ingeborg (2005): Belegdarbietung in Grosslandschaftswörterbüchern im Spannungsfeld von Zeit und Raum am Beispiel des Wörterbuchs der bairischen Mundarten in Österreich (WBO). In: Kraemer-Neubert (ed.), 195-204.
- Gleim, Rüdiger/Mehler, Alexander/Ernst, Alexandra (2012): SOA Implementation of the eHumanities Desktop. In: Proceedings of the Workshop on Service-oriented Architectures (SOAs) for the Humanities: Solutions and Impacts, Hamburg: Digital Humanities, 24-29. www.clarin-d.de/images/workshops/proceedingssoasforthehumanities.pdf
- Hofmannsthal, Hugo von (1986): Der Rosenkavalier. Zum Geleit. In: Hoffmann, Dirk O./Schuh, Willi (eds.): Hugo von Hofmannsthal. Sämtliche Werke XXIII. Operndichtungen I. Frankfurt a. M.: S. Fischer Verlag, 5-101.
- Hofmannsthal, Hugo von (1998): Die Frau ohne Schatten. In: Koch, Hans-Albrecht (eds.): Hugo von Hofmannsthal. Sämtliche Werke XXV.1. Operndichtungen 3.1. Frankfurt a. M.: S. Fischer Verlag, 7-79.
- Hofmannsthal, Hugo von (2006): Wiener Pantomime. In: Schmid, Gisela Bärbel/Krabiel, Klaus-Dieter (eds.): Hugo von Hofmannsthal. Sämtliche Werke XXVII. Ballette, Pantomimen, Filmszenarien. Frankfurt a. M.: S. Fischer Verlag, 134-140.
- ISO 24613 (2005): Language resource management – Lexical markup framework. ISO Geneva.
- Jussen, Bernhard/Mehler, Alexander/Ernst, Alexandra (2007): A corpus management system for historical semantics. In: Sprache und Datenverarbeitung. International Journal for Language Data Processing 31: 81-89.
- Kraemer-Neubert, Sabine (ed.) (2005): Bayerische Dialektologie. Heidelberg: Winter.
- Lausberg, Helmut (1993): Situative und individuelle Sprachvariation im Rheinland. Variablenbezogene Untersuchung anhand von Tonbandaufnahmen aus Erfstadt-Erp. (= Rheinisches Archiv 130). Köln/Weimar/Wien: Böhlau.
- Levenshtein, Vladimir I. (1965): Binary codes capable of correcting deletions, insertions, and reversals. In: Doklady Akademii Nauk SSSR 163: 845-848.
- Markhardt, Heidemarie (2005): Das österreichische Deutsch im Rahmen der EU. Frankfurt a. M. etc.: Peter Lang.
- Mauser, Wolfram (1982): Österreich und das Österreichische in Hofmannsthals „der Schwierige“. In: Recherches germaniques 12: 109-130.
- Mehler, Alexander/Gleim, Rüdiger/Waltinger, Ulli/Ernst, Alexandra/Esch, Dietmar/Feith, Tobias (2009): eHumanities Desktop – eine webbasierte Arbeitsumgebung für die geisteswissenschaftliche Fachinformatik. In: Hoepfner, Wolfgang (ed.): GSCL-Symposium Sprachtechnologie und eHumanities. Duisburg-Essen University. Duisburg: Abteilung für Informatik und Angewandte Kognitionswissenschaft, 72-90. <http://duepublico.uni-duisburg-essen.de/go/technische-berichte>.

- Mehler, Alexander/Schwandt, Silke/Gleim, Rüdiger/Jussen, Bernhard (2011): Der eHumanities Desktop als Werkzeug in der historischen Semantik: Funktionsspektrum und Einsatzszenarien. In: *Journal for Language Technology and Computational Linguistics (JLCL)* 26: 97-117.
- Niebaum, Hermann/Macha, Jürgen (2006): *Einführung in die Dialektologie des Deutschen*. Tübingen: Niemeyer.
- Russ, Charles Victor Jolyon (ed.) (2006): *The dialects of modern German*. London: Routledge.
- Rowley, Anthony A. (1990): North Bavarian. In: Russ (ed.), 417-438.
- Thomason, Sarah G. (2001): *Language contact – An introduction*. Washington DC: Georgetown University Press.
- Waltinger, Ulli (2010): *On social semantics in information retrieval: from knowledge discovery to collective web intelligence in the social semantic web*. Saarbrücken: Südwestdeutscher Verlag.
- Wandruszka, Adam (1967): Das Zeit- und Sprachkostüm von Hofmannsthals „Rosenkavalier“. In: *Zeitschrift für deutsche Philologie* 86: 561-570.
- Wiesinger, Peter (1990): The Central and Southern Bavarian dialects in Bavaria and Austria. In: Russ (ed.), 438-519.

Leipziger Rektoratsreden 1871-1933

Insights into six decades of scientific practice

Abstract

The aim of this paper is to introduce university archives as valuable sources for document-centric historical research. That comprises the history of science as well as the history of society. With the example of Leipzig as a university city with an outstanding wealth of archived academic material we want to stress on the great and in many cases not yet digitally explored potential of such sources.

We then focus on a collection of annual administrative speeches called “Rektoratsreden” that span over 60 important years of Leipzig’s university life. We discuss some of the possibilities for content analysis using methods of Natural Language Processing (NLP). The focus lies on facilitating the access to larger corpora. We present a minimalist process chain for a distant-reading, explorative approach on the Rektoratsreden-corpus. For more general considerations we also highlight some of the digitization efforts that took place in Leipzig and reflect on how archive material as well as archival workflows can benefit from research infrastructure and vice versa.

1. Introduction

Historically-minded researchers of all fields should not miss the opportunity to visit their local university’s archive. With the chance to study – firsthand – the witnesses of a time long passed one is easily immersed in learning about past academical habits. While it then may seem as if time had stood still, by its very nature, time never stands still in an archive. New items keep on arriving constantly and new methods have to be applied to keep up with this dynamics. We can observe the advent of a digital age of archival work. When we want to highlight the special role of its archive for Leipzig’s university, we first have to briefly explain where it comes from.

The University of Leipzig originates from one of the latest medieval University foundations in Europe. In a tense time between Czech and German nation-building and early European ideas of church reformation, a massive dispute at the University of Prague led to the emigration of many scholars and successively to the founding of a new university in Leipzig.

From the very beginning – and rooted in the special founding circumstances – there was a strong urge to record and save written testimonials of all the relevant documents, certificates, matriculation lists, official seals and insignia. In that spirit, the university archive already got founded within the first statutes from 1409, with the headmaster of the university (the “Rector”, i.e. vice-chancellor) as the responsible holder.

In the course of the centuries, the University of Leipzig ranged among the six largest universities in Germany. In a flourishing industrial and cultural environment Leipzig eventually rose to be one of the World’s leading universities at about 1900. The turning point came with World War I, disconnecting Leipzig from its international peers. The ensuing economic difficulties in the Weimar Republic shrunk its capacities further. The following Nazi regime and the devastating World War II destroyed the intellectual and material foundations of science. In times of the GDR, strict ideological barriers were imposed upon the “Karl-Marx-Universität”, until the German Reunification finally brought the freedom for a re-orientation.

2. Inventory of the Leipzig University Archive

Today, after six centuries of German and European scientific development, there are about 140 million sheets of paper stored in the Leipzig University Archive, accounting for more than 7000 meters of shelf space. This vast amount of sources is searched by about 800 researchers each year, leading to a new publication at almost weekly frequency.

Each month about 1500 new (physical) files are stored in the archive. In the last years, about 50,000 digital files (summing up at ca. 800 GB) were collected – all have to be stored and conserved for at least 30 years. The inventory is managed with 1,2 million database entries. Around half of them describe digitized pages from university files. The early digitization efforts were cost intensive and of vague benefits.¹ While quality and accessibility of digitized documents have improved, the costs of producing them are still relatively high. Only about 5% of the stored university files are yet digitized and available online.

¹ At the end of the 1990s, some TIFF-Files of 50 MB size were hard to handle while today their resolution is considered too low for many purposes.

Not only from the perspective of corpus-based studies, many of the archival tasks resemble the requirements for general research infrastructure: Easy and reliable access, long-time conservation, complete metadata handling, and the like are common concerns in the e-humanities and beyond. The European infrastructure projects for arts, humanities and linguistics – foremost CLARIN² and DARIAH³ – aim at a level of service completeness that covers the complete resource lifecycle and therefore have to deal with archival aspects to a great extent. The NLP group in Leipzig is part of CLARIN-D, the German CLARIN initiative. CLARIN-D operates several data- and computing-centers, of which two (Garching and Jülich) are responsible for archiving. This decentralized system of distributed labour within a service-oriented environment is an important aspect of modern infrastructures of that scale. Today regional university archives are in the process of planning a cloud-based distributed infrastructure connecting Halle, Jena and Leipzig. It can only be to everyone's advantage to develop synergies, ensure a high level of compatibility and to actively exchange experiences between such initiatives. Consequent service-orientation, an open source policy and a lot of standards work is the technological and organizational key for sustainable and future-proof infrastructures.

Documents of contiguous temporal and contentual nature are often of special historical interest. Several corpora with diachronic features are available in Leipzig's university archive. Among the already digitized ones are, for example, university newspaper corpora from the GDR era, as the official university newspaper (1957-1991) or the science-related newspaper "Wissenschaftliche Zeitschrift" (1951-1991). In the following section we want to present a corpus that dates back earlier and covers the "Golden Age" of our university, when the university was still organized with an intact academic self-administration.

3. Rektoratsreden-Corpus

The yearly transfer of administrative power from the rector to an elected successor was a long-standing tradition in Leipzig. Within a solemn framework, usually held on reformation day, it formed the ritual highlight of the academic year. Main parts were the inaugural speech ("Antrittsrede") of the new rector and the annual report ("Jahresbericht") of the respectively replaced rector.

² www.clarin.eu. Unless stated otherwise, all URLs have been checked and found valid as of late January 2015.

³ <http://dariah.eu>.

The annual report contained important news, events, staff changes and the like while the new rector used his speech for science communication, presenting his respective field of study. The possible insights from both speeches differ but they complement each other in an interesting way.

The speeches were published in written form soon after the event. In the year of 2004, in the preparation phase for the university's 600th anniversary, 123 of those "Rektoratsreden"-speeches were compiled into a two-volume edition (Häuser (ed.) 2009). Based on the digitized texts from 2340 scanned images of the original prints,⁴ the edition was created in a manual process. A textual correction was performed, retaining the original grammar and spelling, but correcting obvious printer's or typographical errors as well as faulty OCR results.

The principle of the edition work was to reproduce the speeches unabridged in authentic length and depth. In order to keep a reasonable page count, no extensive commentary or inclusion of the speakers' biographic data was added. This process resulted in 1790 printed pages for the corpus consisting of ca. 700,000 tokens in about 5,1 MB of plain text. For accessing the vast amount of information stored in that two-volume edition, a comprehensive index with about 6500 entries for people, places and topics was created. The manually performed index creation accounted for more than two thirds of the whole project phase.

The edition covers the time from 1871, when the first speech was published in print, to the year 1933, when a shift from freely elected rectors to appointed rectors took place. Afterwards, the speeches were merely occasions where conformity with the Nazi propaganda was demonstrated, rather than performing the accustomed reflections on the inner structure, events and positions of the university.

4. Supplementary resources for corpus analysis

The documents collected at the archive do not only form diachronic corpora of prose but also hold large lists of structured or semi-structured data that can be used to analyze those corpora. While the digitization of (mostly) handwritten lists is a very time- and resource-consuming task, it can provide a valuable framework of data entries to which other documents can be linked.

⁴ Fraktur-typeset documents were copied manually, OCR was used for other typesets.

In Leipzig, several of these lists were made digitally accessible. There are digital editions of matriculation lists (containing names and, for entries after 1810, also a lot of bibliographical information), databases of employee lists and bursary data (“Quästur”) and other special lists⁵ available. These lists are not only a great entry point for researchers but also constitute an invaluable input of diachronic and entity-centric data that exactly matches the time and place of the corpora. The university archive is already working on a research-friendly GIS-application to hold about 280,000 person-related entries with geo-temporal anchors. An ongoing cooperation with the archive of the University of Prague can further extend the data volume. The collected data already powers several online research interfaces such as a statistical overview over the historical popularity of certain given names.⁶

Not only can digitized lists be used to retrieve names and name variants that were popular in the past. The corpora themselves give the chance for even further extraction and validation of names, for example by the extraction of context-based rules as presented in Schlaf/Remus (2012).

5. Setup for a first explorative analysis

The goal of the analysis prototype was to showcase how a minimalist NLP-based approach on the corpus can result in new means of browsing the texts in a somewhat more topic-oriented manner. Instead of an in-depth analysis, novel method development or a display of recent algorithms, we simply want to encourage experiments with standard tools and promising corpora as an interdisciplinary “brainstorming”.

The idea was to extract Named Entities from within the speeches and use them as anchor points for navigation, creating an alternative, largely automated way of indexing the resources. Without using any pre-existing indices and corresponding page references, we want to show how to build basic cross-document links across sources and entities.

For the task of Named Entity Extraction we assembled a simple process chain using the GATE framework and specifically the robust ANNIE workflow package (Cunningham et al. 2002). Basic look-up annotations were created using a

⁵ E.g., a list of detention cell (“Karzer”) punishments: www.archiv.uni-leipzig.de/digitale-archivalien/datenbanken/karzerstrafen/ (last accessed: May 2014).

⁶ www.archiv.uni-leipzig.de/vornamensuche/.

gazetteer that contained all given names and family names from the above mentioned databases as well as a short, manually compiled list of titles and abbreviations. Using GATE's annotation processing language "JAPE", we then created simple rules for a context-aware matching of the basic annotations to recognize the mentioning of people's full names. For simplicity we skipped more advanced methods of the ANNIE toolbox such as Co-Reference Resolution, as this would have required a POS tagging which is relatively hard to produce for such a special corpus with significant diachronic features. The JAPE-based rules followed an intuitive and rather conservative approach, focusing on precision rather than recall – although both values have not been scientifically evaluated.⁷ Possible entities of type "Person" for which only a family name could be retrieved (e.g. people being introduced with just "Professor" or "our dear colleague" instead of a given name) were disregarded in the later analyses.

For simplicity we decided not to extract other NE-types: The added complexity, e.g. resulting from the frequent occurrence of places in ambiguity with family names would be too much. As a further idea we initially planned to also define abstract concepts such as "war" and observe their context including a corresponding impact on used terminology. We skipped such analyses because of the serious manual effort needed to provide a fitting terminological or ontological framework with a respective linking to the involved terms, used from the 1870s to the 1930s. But such an approach of conceptual terminology extraction will certainly be addressed in further work.

In order to construct a "bigger picture" of the people mentioned in the corpus, a fitting representation of the extracted NE-annotations was sought. So the single speech documents were each regarded as an (addressable) unit, where the occurrence of people within that unit was recorded. The extracted links were then exported as a graph structure. We chose the GEXF format⁸ since it is well integrated with Gephi (Bastian/Heymann/Jacomy 2009), our graph-

⁷ The edition's index could not efficiently be used as gold standard because it mixes different NE-types. Instead, the rules have been iteratively altered to boost the perceived precision until moderately-sized random samples did not contain errors anymore (still finding reasonably many different full names). While this trial-and-error-principle inarguably lacks methodological elegance, it may be generally a good approach for inter-disciplinary work, allowing for a lot of communication on the processing quality in tight feedback-loops and keeping all parties close to the data as well as to the algorithm.

⁸ <http://gexf.net>.

analytics tool of choice, and because its “Dynamic Graph” feature allows for a later extension of our output with temporal constraints.

Sub-units of speeches like paragraphs could allow for better understanding of the “relatedness” between people, but unfortunately the corpus is too sparse on the temporal axis with only two documents per year.

6. Results

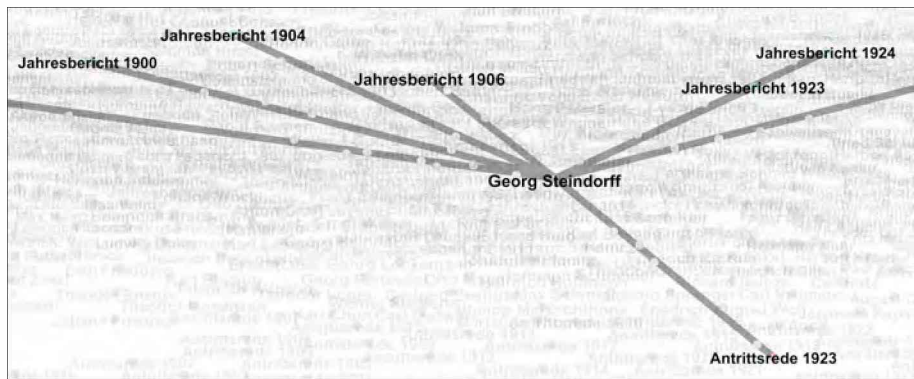


Figure 1: Occurrences of famous archaeologist Georg Steindorff (rector 1923/24), as highlighted in the Gephi tool. From the density of the greyed-out nodes in the background it can be seen that the graph is much sparser at the bottom where the non-administrative inaugural speeches (that naturally contain fewer references to people) are represented.

With our prototype we could extract 2391 named entities of type “Person” in a fully automated way. This includes a few name deviations denoting identical persons (e.g. “Dr. med. Ernst Leberecht Wagener” instead of the correct family name “Wagner”). Luckily their low Levenshtein-Distance makes it possible to automatically compile a small set of candidates for merging both name labels into a single representative for the entity that can be manually processed.

We created two complementary graphs out of the extracted data: First an “occurrence graph” with all the speeches and the recognized persons as nodes and second a “co-occurrence-graph” of persons only. The occurrence graph works just like an index. While an index entry is usually pointing to a page number, people entries are here pointing to speeches (Figure 1). Figure 2

shows the visual effect of filtering the occurrence graph so that only 680 nodes with a degree of two and greater remain (people with more than one speech mentioning them). Only over those nodes a navigation step from one speech to another can be performed. To comfortably get from one person to another via the speeches mentioning both, we constructed the co-occurrence graph that is connecting two people if they both appear in the same speech – the edge weight is the number of such speeches. Figure 3 shows the co-occurring nodes for Leipzig-born artist Max Klinger. While there may not be obvious semantically motivated connections between those people (such as Karl Marx co-occurring with Friedrich Engels) it nevertheless depicts some kind of temporal and topical similarity that can then be further investigated by reading the passages where both appear in the speeches.

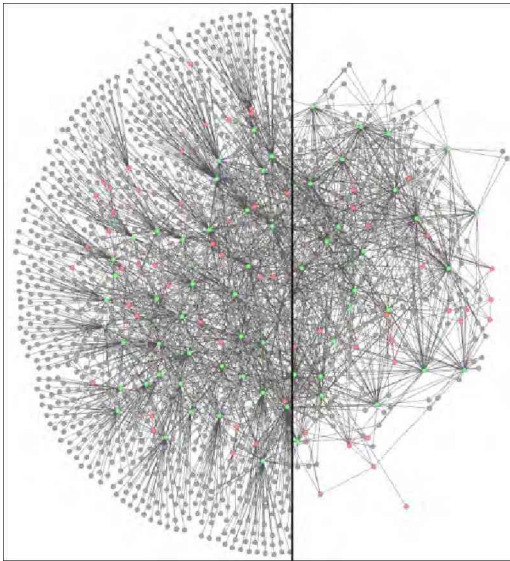


Figure 2: Force-directed layout of the occurrence-graph; left: full graph; right: filtered by node degree 2 and greater, revealing more of the structure

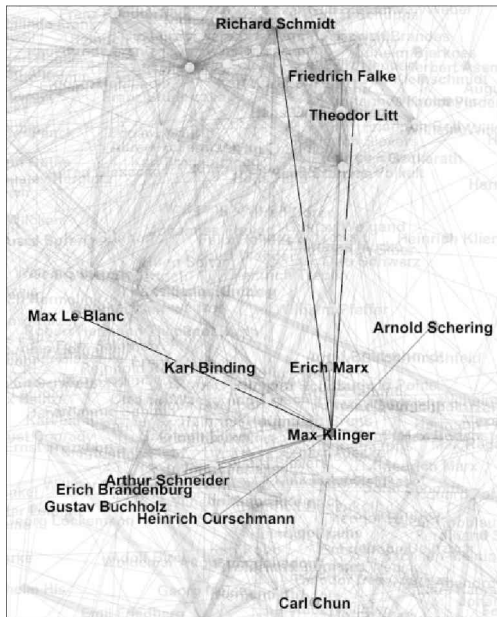


Figure 3: Co-occurrences for Max Klinger with edges weight of 2 or more (being mentioned together in at least 2 speeches)

We believe that NLP methods can be valuable tools to facilitate the creation of indices for people, places and topics for classical print publications. When extending the rule set to gain more recall, the lower precision can easily be compensated in later manual analyses in the edition process. But printed editions do not have to be the final result of edition work. Online editions can provide highly interlinked hyper-textual documents. Specific corpus browsers with an explorative focus could encourage a more “casual” browsing and research behavior, opening up the possible range of readers beyond the purely academic world.

A next step should be the integration with other resources such as linked data hubs like Freebase⁹ or the Personennormdaten (PND) identifiers from the German National Library. Using such unambiguous anchors for entities enables an inter-corpus linking. Identity-aware matching methods such as that described in Rizzo et al. (2012) can be employed. The narrow geographical focus of our corpus may allow for a simpler disambiguation in many cases by

⁹ www.freebase.com.

just finding “Leipzig” in full text descriptions or by finding a corresponding edge in the semantic (hyper-)graph towards the resource for the city of Leipzig. Further restrictions can be made e.g. by rejecting all people born after the date of the speech that mentioned their name.

7. Outlook and future possibilities

We showcased a merely quantitative, entity-based analysis of a diachronic scientific corpus. This allows for a rather distant view on “texts” and “protagonists” resembling the publication-author-model which is central to bibliometrics. This discipline employs various forms of network analysis with a focusing on temporal aspects and can therefore be seen as related work with many suitable approaches to diachronic document collections. It would be most interesting to extend bibliometrical methodology to work on NLP-processed documents and to combine the extracted information from within the texts with classical bibliographic information where possible. In a way this could possibly even constitute a next step towards a better understanding of Science Dynamics (Scharnhorst et al. 2012).

The inventory of science archives is applicative in multiple fields of e-humanities: Not only can historical studies and social sciences gain insights from the material – other disciplines may benefit too, like historical linguistics, extracting knowledge about academic language and its development. To meet all scientist’s quantitative requirements it is necessary to digitize more documents, which is costly and often connected with caveats such as insufficient OCR-quality, especially for older types and of course handwriting.

Working on the Rektoratsreden-corpus it became evident that the introduction of NLP methods in archival and editorial work can drastically reduce manual workload, especially in the indexing, linking, classifying, and general processing of digitized documents. In return, the historical data from the archives can enhance the applicability of NLP methods on historical texts. Within the University of Leipzig, a cooperation beyond the exchange of data and methods is planned. We want to encourage a fruitful collaboration of Computer Scientists and their respective University Archives throughout the country. Therefore we want to establish a close interexchange on the level of archival and research-infrastructure in Leipzig in order to prepare the way for other university locations. As mentioned earlier, archival infrastructure and research infrastructure share a large set of common concerns that should lead

to synergies – not necessarily by mixing data or workflows but by collaborations in standardizing and defining methods. Compatibility (or even better interoperability) between both systems is an important topic. For example, any archive infrastructure that used a service oriented approach and a common metadata categorization system like IsoCat (Broeder et al. 2010) could easily be extended to interact with the CLARIN infrastructure. Another aspect of digital archive work is the inclusion of the ever growing contemporary archive material that is now created mostly in digital form. Especially infrastructure projects for “digital born documents” are needed. A “private cloud” infrastructure based on open source technology and hosted across several regional university locations is already planned. An ongoing general goal of NLP is to find novel ways to transform documents into a suitable representation form for further processing in the emerging field of Visual Analytics (e.g. Keim et al. (eds.) 2010). New, scalable methods from there can produce alternative views on document collections, allowing for an explorative and quantitative research while retaining the ability to access every textual occurrence in place. After all, a better accessibility of university archive material could also lead to interest from researchers from various non-historical fields and contribute to fruitful reflections on the roots and development of science itself as well as its possible forms of institutionalization.

References

- Bastian, Mathieu/Heymann, Sebastien/Jacomy, Mathieu (2009): Gephi: an open source software for exploring and manipulating networks. In: International AAAI Conference on Weblogs and Social Media, May 17-20, 2009, San Jose, California, USA. <https://gephi.org/publications/gephi-bastian-feb09.pdf>.
- Broeder, Daan/Kemps-Snijders, Marc/Van Uytvanck, Dieter/Windhouwer, Menzo/Withers, Peter/Wittenburg, Peter/Claus Zinn (2010): A data category registry and component-based metadata framework. In: Proceedings of the 7th International Conference on Language Resources and Evaluation. Paris: ELRA, 43-47. www.lrec-conf.org/proceedings/lrec2010/pdf/163_Paper.pdf.
- Cunningham, Hamish/Maynard, Diana/Bontcheva, Kalina/Tablan, Valentin (2002): GATE: a framework and graphical development environment for robust NLP tools and applications. In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002. <https://gate.ac.uk/sale/acl02/acl-main.pdf>.

- Häuser, Franz (ed.) (2009): Die Leipziger Rektoratsreden 1871-1933. Herausgegeben zum 600-jährigen Gründungsjubiläum der Universität im Jahr 2009. Berlin: de Gruyter.
- Rizzo, Giuseppe/Troncy, Raphael/Hellmann, Sebastian/Brümmer, Martin (2012): NERD meets NIF: Lifting NLP extraction results to the linked data cloud. In: LDOW, 5th Workshop on Linked Data on the Web, Lyon, France, April 16, 2012. <http://events.linkedata.org/ldow2012/papers/ldow2012-paper-02.pdf>.
- Schlaf, Antje/Remus, Robert (2012): Learning categories and their instances by contextual features. In: Proceedings of the 8th International Conference on Language Resources and Evaluation. Paris: ELRA, 1235-1239. www.lrec-conf.org/proceedings/lrec2012/pdf/181_Paper.pdf.
- Keim, Daniel A./Kohlhammer, Jörn/Geoffrey, Ellis/Mansmann, Florian (eds.) (2010): Mastering the information age – solving problems with visual analytics. Goslar: Eurographics Association.
- Scharnhorst, Andrea/Börner, Katy/van den Besselaar, Peter (eds.) (2012): Models of science dynamics: encounters between complexity theory and information sciences. Berlin: Springer.

Register contact: an exploration of recent linguistic trends in the scientific domain

Abstract

In this paper, we present a study of selected lexico-grammatical features showing possible diachronic linguistic trends in specialized academic domains that have emerged by contact with computer science (computational linguistics, bioinformatics, digital construction, microelectronics). The data basis for this study is the English Scientific Text Corpus (SciTex) which covers a time range of roughly thirty years (70/80s to early 2000s) (Degaetano et al. 2013).¹ In particular, we are looking for trends of linguistic standardization and diversification across the disciplines under investigation. The theoretical framework we use as a basis is Systemic Functional Linguistics (SFL) and its specific incarnation of register theory (Halliday/Hasan 1985).

1. Introduction

Academic registers have been the subject of many studies ranging from the description of single registers (e.g., Halliday 1988, O'Halloran 2005) as well as differences between registers (e.g., Hyland 2007, Charles 2006) through studies from applied linguistics (for instance for educational purposes, e.g., Fang/Schleppegrell/Cox 2006) to analyses of academic language as such (e.g., Halliday 1993, Ventola 1996, Biber 2006). Compared to investigations of recent language change in English overall, diachronic studies related to registers are still fairly rare (see e.g., Halliday (1988), Banks (2008) in the area of academic registers). The motivation for our own work is to trace some of the diachronic trends that have been attested for English in general (e.g., colloquialization, standardization; Kytö/Romaine 2000, Mair 2006) in scientific English, as well as to discover other, possibly competing trends such as, for instance, densification and (registerial) diversification.

While linguistic standardization is a general, longer-term historical process that arises due to changing social, political and commercial needs, specialized discourse domains (such as the scientific domain) underlie additional, specific kinds of pressure such as, e.g., efficiency of communication, individual

¹ Our work has been funded by Deutsche Forschungsgemeinschaft (DFG) under grant TE-198/2 *Registers in Contact* (Regico).

tion or diversification. The specific situation we are interested in is the linguistic contact between two or more scientific registers due to joint research activities of two or more scientific disciplines (e.g., computer science and biology/genetics in the field of bioinformatics). Such contact will be reflected linguistically in the adoption of the linguistic conventions of the “seed registers” (in the present example: computer science, biology) on the part of the newly emerging discipline (in the present example: bioinformatics), but as the new field develops, it will create its own, individual “language”. In other words, some kind of linguistic standardization will set in within the new field over time and it will tend to be increasingly linguistically distinct from the “language” of other fields.

The remainder of this paper is organized as follows. In Section 2, we present the theoretical background, our data basis as well as a short overview of our methodology. Section 3 shows some examples of the kinds of analyses we conduct in the search of diachronically significant linguistic features. Section 4 concludes the paper with a summary and envoi.

2. Theoretical framework, data and methodology

Our theoretical basis is Systemic Functional Linguistics and its theory of register (Halliday/Hasan 1985), which says that registers are instantiated by particular sets of lexico-grammatical features. These features are associated with three contextual variables which characterize a register: *field* (what is being talked about), *tenor* (participants involved in the discourse and the relationship between them) and *mode* (the role of language in the interaction) (see Table 1). Particular settings of these variables will be associated with typical distributions/clusters of lexico-grammatical features, thus forming a register.

contextual variable	lexico-grammatical domains	feature examples
field	processes, participants, circumstantials	terminology, semantic roles
tenor	roles and attitudes of participants	stance expressions, modality
mode	thematic structure, information distribution	theme-rheme, given-new

Table 1: Contextual variables and lexico-grammatical features

In our approach, we analyze the distribution of lexico-grammatical features across registers and time. As a data basis we use the English Scientific Text corpus (SciTex; see Figure 1), a diachronic corpus which covers a time range of approximately thirty years (70/80s to early 2000s) (Teich/Fankhauser 2010, Degaetano et al. 2013). The corpus is annotated in terms of sentence boundaries, tokens, lemmas and parts-of-speech, and contains roughly 34 million tokens. For more information on our project and methodology see Degaetano et al. (2013).

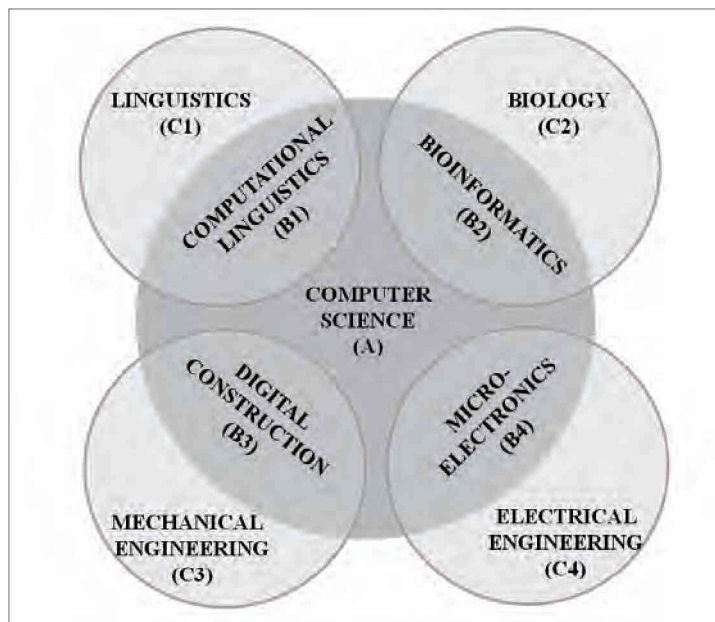


Figure 1: Scientific registers in the SciTex corpus

To explore possible linguistic trends reflecting, for instance, standardization, we look for differences in the use of lexico-grammatical features comparing the “contact registers” (i.e., the interdisciplinary fields; B subcorpus) and the corresponding “seed registers” (A and C subcorpora). Here, several comparative dimensions have to be taken into account: (a) comparative analyses of lexico-grammatical features in A-B-C triples of subcorpora (e.g., computer science, computational linguistics, linguistics), (b) a comparison of the contact registers as a whole (B subcorpus) with the seed registers (A and C subcorpora), (c) diachronic comparison of lexico-grammatical features across all registers.

For the present paper, we have selected two examples of features (see Section 3 below). We extract the features from SciTex with the Corpus Query Processor (CQP; CWB 2010, Evert 2005). CQP allows feature extraction in terms of regular expressions as well as annotation of features at multiple layers. To determine whether the differences obtained by the comparison of the extracted features are significant, we employ various univariate and multivariate methods, such as Pearson's chi-square test, correspondence analysis, clustering and automatic classification (cf. Teich/Fankhauser 2010, Degaetano/Lapshinova-Koltunski/Teich 2012).

3. Analyses of selected lexico-grammatical features

In this section, we show selected analyses of lexico-grammatical features potentially involved in the linguistic evolution of the contact registers under investigation (B subcorpora). The features are (1) the noun-*of*-noun feature (e.g., *date of issue*, *set of expressions*) reflecting one aspect of the field of discourse, and (2) modals and semi-modals of obligation (e.g., *need to*, *ought to*) reflecting one aspect of the tenor of discourse (cf. Table 1 above).

3.1 Discourse field: the noun-*of*-noun pattern

Features typically associated with the field of discourse lie in the area of experiential lexis. Given that nouns are the prototypical carriers of experiential meaning, we have selected a nominal group pattern which is constituted by a noun followed by a post-modifying *of*-phrase (see Examples 1-2), the most frequently occurring type of noun complement according to Biber/Johansson/Leech (1999). Also, this pattern has been shown to decrease over time between the 1960s and 1990s (cf. Leech 2012).

- (1) *In each iteration, it uses the current **set of rules** to assign labels to unlabeled data.*
- (2) *Machine learning algorithms are typically designed to optimize some objective function that represents a formal **measure of performance**.*

Considering the distribution of this pattern across the single registers of SciTex (see Figure 2), they either show a very similar frequency of noun-*of*-noun phrases across the two time slices (e.g., B1-computational linguistics) or exhibit a slight increase (e.g., C4-electrical engineering).

Taken on its own, the mere distribution of noun-*of*-noun phrases is thus not very revealing, except that the trend is the opposite to the one of English in general for most of the registers in SciTex (compare again with Leech 2012). In order to find evidence of standardization and/or diversification over time, we need to look at the lexical fillers of both the heads and the modifiers in the noun-*of*-noun phrases.

Considering the twenty most frequent combinations, some kind of trend emerges regarding the contact registers. Tables 2 and 3 display the two contact registers computational linguistics (B1) and bioinformatics (B2) across the two time slices for the twenty most frequent noun-*of*-noun phrases. While the modifiers exhibit a wide range of variation with a number of different functions (they can appear after quantifying collectives (e.g., *set of books*), species nouns (e.g., *kinds of questions*), etc. (see Biber/Johansson/Leech 1999 for a full list), the head noun *number* diachronically increases in frequency and combines with a wider range of *of*-phrases (B1: from 2 to 9; B2: from 6 to 8). The same holds for digital construction (B3: from 10 to 12) and microelectronics (B4: from 4 to 14), which are not shown here. In contrast, this tendency is not present in the seed registers (C1: from 2 to 4; C2: from 1 to 3; C3: from 5 to 3; C4: from 2 to 4) except in computer science (A: from 4 to 9).²

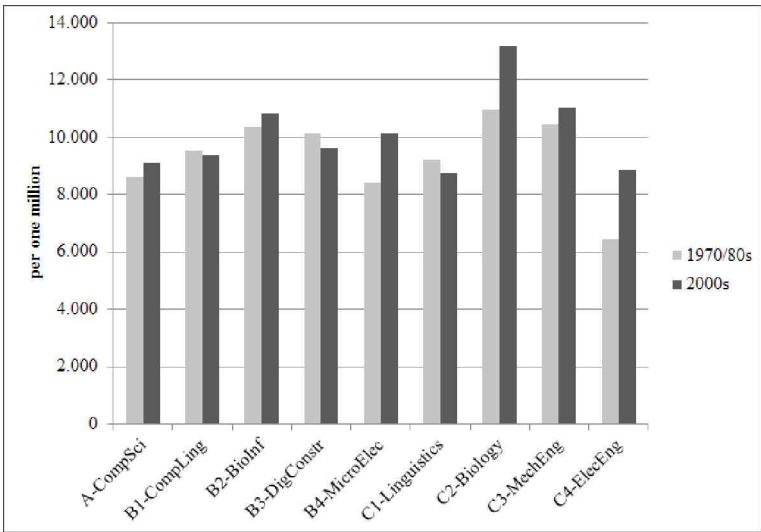


Figure 2: The noun-*of*-noun feature across registers and time

² Note that in both time slices and for all registers the noun after *number of* is most often used in the plural.

B1 – 1970/80s			B1 – 2000s		
head-N	of-phrase-N	per 1M	head-N	of-phrase-N	per 1M
<i>part</i>	<i>speech</i>	94.13	<i>part</i>	<i>speech</i>	78.52
<i>point</i>	<i>view</i>	56.80	number	<i>word</i>	62.14
number	<i>word</i>	36.52	<i>point</i>	<i>view</i>	48.02
<i>set</i>	<i>rule</i>	34.89	<i>set</i>	<i>feature</i>	29.37
<i>end</i>	<i>word</i>	23.53	number	<i>time</i>	27.68
<i>piece</i>	<i>information</i>	20.29	<i>set</i>	<i>word</i>	25.98
<i>speaker</i>	<i>English</i>	19.48	number	<i>feature</i>	23.73
<i>frequency</i>	<i>occurrence</i>	18.66	<i>type</i>	<i>information</i>	22.03
<i>set</i>	<i>filter</i>	16.23	number	<i>error</i>	22.03
<i>amount</i>	<i>training</i>	16.23	number	<i>translation</i>	20.34
<i>beginning</i>	<i>word</i>	15.42	number	<i>sentence</i>	19.21
<i>rule</i>	<i>inference</i>	15.42	<i>pair</i>	<i>word</i>	17.51
<i>set</i>	<i>word</i>	15.42	number	<i>utterance</i>	17.51
<i>model</i>	<i>language</i>	15.42	<i>amount</i>	<i>datum</i>	17.51
<i>manner</i>	<i>articulation</i>	14.61	<i>violation</i>	<i>Rule</i>	16.95
<i>subset</i>	<i>English</i>	13.80	number	<i>occurrence</i>	16.95
<i>Department</i>	<i>Computer</i>	12.98	number	<i>state</i>	16.95
<i>center</i>	<i>gravity</i>	12.98	<i>type</i>	<i>coherence</i>	16.38
<i>end</i>	<i>loop</i>	12.98	<i>set</i>	<i>rule</i>	16.38
number	<i>case</i>	12.98	<i>pair</i>	<i>sentence</i>	15.82

Table 2: Noun-of-noun across two time slices for computational linguistics

To determine whether there is a clear diachronic trend specific to the contact registers or not, we compare how many different head nouns (types) are used in comparison to all head nouns (tokens) used in the construction, i.e., we calculate the type-token ratio (ttr). Table 4 shows this comparison across all registers contained in the corpus. Except for digital construction (B3), all contact registers (B1, B2, B4) show a decrease of different head nouns used in this pattern, i.e., a reduction in their lexical range. Digital construction (B3) has the lowest ttr for the noun-*of*-noun pattern in the 1970/80s in comparison to the others; it seems to strive for the same level as B1 and B2 in the 2000s. The seed registers (C subcorpora) show the reverse trend, i.e., a greater vocabulary variation in head nouns in the 2000s compared to the 1970/80s.

In summary, the contact registers create their own trend distinct from the seed registers concerning the lexical range of head nouns in the noun-*of*-noun pattern.

B2 – 1970/80s			B2 – 2000s		
head-N	of-phrase-N	per 1M	head-N	of-phrase-N	per 1M
<i>number</i>	<i>patient</i>	46.79	<i>conflict</i>	<i>interest</i>	144.54
<i>degree</i>	<i>freedom</i>	38.53	<i>number</i>	<i>gene</i>	91.91
<i>grade</i>	<i>membership</i>	34.86	<i>number</i>	<i>cluster</i>	72.27
<i>amount</i>	<i>information</i>	31.65	<i>set</i>	<i>gene</i>	42.10
<i>function</i>	<i>time</i>	29.36	<i>number</i>	<i>protein</i>	38.59
<i>number</i>	<i>point</i>	26.60	<i>number</i>	<i>time</i>	34.38
<i>Institutes</i>	<i>Health</i>	26.15	<i>degree</i>	<i>freedom</i>	32.98
<i>number</i>	<i>time</i>	24.77	<i>group</i>	<i>gene</i>	31.57
<i>set</i>	<i>datum</i>	23.39	<i>number</i>	<i>sample</i>	31.57
<i>sum</i>	<i>square</i>	22.48	<i>number</i>	<i>parameter</i>	28.07
<i>analysis</i>	<i>variance</i>	22.48	<i>number</i>	<i>feature</i>	26.66
<i>amount</i>	<i>datum</i>	22.48	<i>set</i>	<i>protein</i>	24.56
<i>rate</i>	<i>change</i>	21.10	<i>set</i>	<i>parameter</i>	24.56
<i>point</i>	<i>view</i>	21.10	<i>number</i>	<i>sequence</i>	23.86
<i>period</i>	<i>time</i>	20.18	<i>order</i>	<i>magnitude</i>	23.15
<i>use</i>	<i>computer</i>	19.27	<i>pair</i>	<i>protein</i>	21.75
<i>number</i>	<i>sample</i>	18.81	<i>subset</i>	<i>gene</i>	0.35
<i>number</i>	<i>cell</i>	17.43	<i>set</i>	<i>sequence</i>	18.94
<i>amount</i>	<i>time</i>	16.05	<i>pair</i>	<i>sequence</i>	18.24
<i>number</i>	<i>parameter</i>	15.14	<i>prediction</i>	<i>protein</i>	17.54

Table 3: Noun-of-noun across two time slices for bioinformatics

registers	1970/80s			2000s			tendency
	tokens	types	ttr	tokens	types	ttr	
A-CompSci	123	43	34.96	226	76	33.63	=
B1-CompLing	55	35	63.64	122	44	36.07	-
B2-BioInf	150	72	48.00	82	32	39.02	-
B3-DigConstr	191	56	29.32	94	37	39.36	+
B4-MicroElec	13	11	84.62	103	28	27.18	-
C1-Ling	147	68	46.26	73	51	69.86	+
C2-Bio	169	85	50.30	86	52	60.47	+
C3-MechEng	175	75	42.86	147	76	51.70	+
C4-ElecEng	98	54	55.10	53	33	62.26	+

Table 4: Tokens vs. types of head nouns across subcorpora in SciTex

4. Discourse tenor: modals and semi-modals of obligation

Another linguistic feature potentially involved in diachronic changes is the use of modal and semi-modal verbs. According to Biber/Johannson/Leech (1999), the use of semi-modals has increased over time, which is a relatively recent development in English. Modal verbs instead have shown a decrease in use (cf. Leech 2012). In our analysis, we investigate the use of semi-modal verbs of obligation (*(had) better, have (got) to, need to, ought to, be supposed to*; see Example 3) in comparison to modal verbs of obligation (*must, should*; see Example 4). The questions we pose are whether the increase of semi-modals and decrease of modals can also be observed within scientific registers and whether the contact registers show specific tendencies in the use of semi-modals and modals of obligation in comparison to their seed registers.

- (3) *In order to deal with the functional approach we **need to** specify the behaviour of the dictionary by describing the functions that could be executed on it.*
- (4) *In other words, we advocate that dictionaries **should** actively co-operate in finding the correct translation.*

The first analysis is a diachronic comparison of semi-modals and modals of obligation. Table 5 shows their frequencies in the two time slices. The semi-modals have increased, while the modals have decreased in use. These diachronic changes are in line with the observations made by Biber et al. (1999) and Leech (2012). Thus, this recent linguistic development has also taken place within our corpus.

	1970/80s	2000s	change
semi-modals	327.82	459.55	+
modals	1285.00	799.55	-

Table 5: Semi-modals vs. modals of obligation across two time slices per 1M words

In the second analysis, we compare triples of subcorpora in order to see whether the contact registers tend towards (a) computer science, (b) the corresponding seed registers or (c) create their own variation regarding the use of semi-modals vs. modals of obligation. For this purpose, we extract the two variants according to discipline and time slice. Figures 3 and 4 show the distribution across the SciTex registers for the 70/80s and 2000s, respectively.

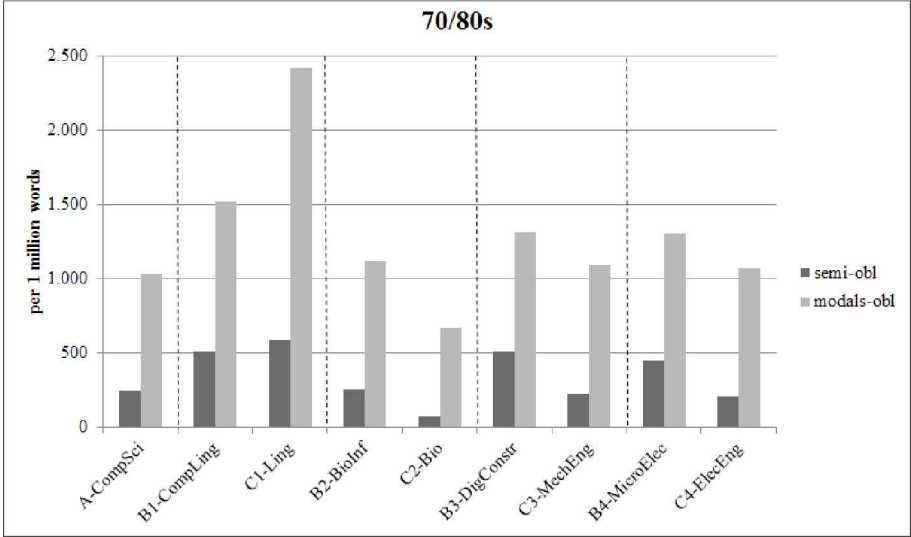


Figure 3: Semi-modals and modals of obligation across SciText in the 70/80s

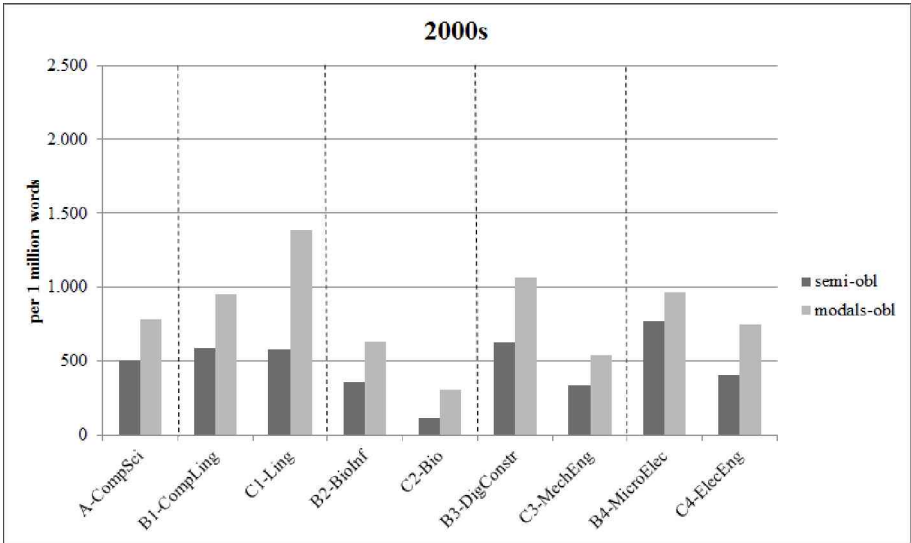


Figure 4: Semi-modals and modals of obligation across SciText in the 2000s

When comparing, for example, the A-B1-C1 triple, computational linguistics seems to be more similar to computer science than to linguistics in both time slices (see again Figures 3 and 4). To statistically test the commonalities/differences observed, we calculate the Pearson’s chi-square test for each com-

parison of B vs. A and B vs. C subcorpora. Table 6 shows the chi-square values as well as significant differences for p-values < 0.05.

	1970/80s				2000s			
	χ^2	p-value	sig.	tend.	χ^2	p-value	sig.	tend.
B1-A	24.1779	8.783e-07	s	B1→A	0.5043	0.4776	ns	B1→A
B1-C1	37.5845	8.754e-10	s		51.0215	9.136e-13	s	
B2-A	0.3194	0.5719	ns	B2→A	2.7633	0.09645	ns	B2→A
B2-C2	69.8537	<2.2e-16	s		23.2244	1.442e-06	s	
B3-A	59.0217	1.559e-14	s	B3→?	2.9315	0.08687	ns	B3→A/C3
B3-C3	113.8438	<2.2e-16	s		0.7838	0.376	ns	
B4-A	21.9584	2.786e-06	s	B4→A	15.4435	8.501e-05	s	B4→A
B4-C4	53.7949	2.226e-13	s		37.1705	1.082e-09	s	

Table 6: Chi-square for contact register comparison across time

Except for digital construction (B3), the contact registers (B subcorpora) lean towards computer science (A), a tendency that becomes more pronounced diachronically. Computational linguistics (B1), for example, shows significant differences to computer science (A) and linguistics (C1) in the 70/80s. In the 2000s, computational linguistics still differs from linguistics but does not differ from computer science. Thus, computational linguistics and computer science have converged in the linguistic behavior of this feature. This kind of convergence can be observed for all contact registers, the tendency being less pronounced for digital construction (B3).

In summary, all the registers contained in SciTex show a diachronic increase of semi-modals of obligation and a decrease in modals of obligation – a development conforming to the diachronic trend in English overall (see again Biber/Johannson/Leech 1999 and Leech 2012). However, comparing all the registers in SciTex, the contact registers show a greater distance from their corresponding seed disciplines and a convergence with computer science with regard to the use of modals and semi-modals of obligation over time. So again, some kind of standardization is taking place with regard to the contact registers (B subcorpora).

5. Summary and discussion

In this paper, we have looked at features potentially involved in recent diachronic changes in scientific registers, focusing on what we have termed “contact registers”, i.e. interdisciplinary registers having evolved at the intersection of computer science and some other discipline in the last thirty to forty years or so. Our overarching research question is how such new disciplines develop linguistically, trying to trace diachronic trends, such as, for instance, registerial diversification and linguistic standardization (in scientific writing overall as well as within and across the recently evolved disciplines). The theoretical background of our work is register theory as provided by Systemic Functional Linguistics, which offers a grid of categories for analysis at both the levels of context/situation (field, tenor, mode) and lexico-grammar (Section 2). More concretely, we have explored two lexico-grammatical features, one related to discourse field (the noun-*of*-noun pattern; Section 3.1) and one related to discourse tenor (modals and semi-modals of obligation; Section 3.2). We could show that these are two features involved in a diachronic trend of standardization in the contact registers under investigation (computational linguistics, bioinformatics, digital construction and microelectronics).

Our methodology is inherently comparative. In order to elaborate differences and commonalities across different scientific registers over time, we need to take into account several comparative perspectives – the register perspective and the temporal perspective. We are continuously exploring more features that are systematically related to field, tenor and mode of discourse, focusing on tracing standardization and diversification trends specific to the scientific domain. In terms of field, e.g., we will analyze more closely how phrases or n-grams might contribute to discipline specific trends. As we are extracting more (different kinds of) lexico-grammatical features, we are feeding their categorization back as annotations to the corpus, thus creating a resource that should be useful for other purposes as well, e.g., for teaching scientific writing or for NLP tasks such as automatic text classification. Ultimately, the SciTex corpus, enriched with multiple aspects of linguistic information, will be made available via the CLARIN-D repository (<http://clarind.net/index.php/de/>).

References³

- Banks, David (2008): *The development of scientific writing, linguistic features and historical context*. London: Equinox.
- Biber, Douglas (2006): *University language: a corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.
- Biber, Douglas/Johansson, Stig/Leech, Geoffrey (1999): *Longman grammar of spoken and written English*. London: Longman.
- Charles, Maggie (2006): *Phraseological patterns in reporting clauses used in citation: A corpus-based study of theses in two disciplines*. In: *English for Specific Purposes* 25: 310-331.
- CWB (2010): *The IMS open corpus workbench*. www.cwb.sourceforge.net.
- Degaetano, Stefania/Kermes, Hannah/Lapshinova-Koltunski, Ekaterina/Teich, Elke (2013): *Scitex – A diachronic corpus for analyzing the development of scientific registers*. In: Bennett, Paul/Durrell, Martin/Scheible, Silke/Whitt, Richard J. (eds.): *New methods in historical corpus linguistics*. (= *Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache / Corpus Linguistics and Interdisciplinary Perspectives on Language (CLIP) 2*). Tübingen: Narr, 93-104.
- Degaetano, Stefania/Lapshinova-Koltunski, Ekaterina/Teich, Elke (2012): *Feature discovery for diachronic register analysis: a semi-automatic approach*. In: *Proceedings of the LREC2012*. Istanbul, Turkey. Paris: ELRA, 2786-2790. www.lrec-conf.org/proceedings/lrec2012/pdf/268_Paper.pdf.
- Evert, Stefan (2005): *The CQP query language tutorial*. Stuttgart: Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.
- Fang, Zhihui/ Schleppegrell, Mary J./Cox, Beverly E. (2006): *Understanding the language demands of schooling: nouns in academic registers*. In: *Journal of Literacy Research* 38(3): 247-273.
- Halliday, Michael A.K. (1988): *On the language of Physical Science*. In: Ghadessy, Mohsen (ed.): *Registers of written English: situational factors and linguistic features*. London: Pinter, 162-177.
- Halliday, Michael A.K. (1993): *The analysis of scientific texts in English and Chinese*. In: Halliday, Michael A.K./Martin, James R. (eds.): *Writing science: literacy and discursive power*. London: Falmer Press, 124-132.
- Halliday, Michael A.K./Hasan, Ruqaiya. (1985): *Language, context and text: a social semiotic perspective*. (= *Language and Learning Series*). Geelong, Victoria: Deakin University Press.

³ All URLs have been checked and found valid as of late January 2015.

- Hyland, Ken (2007): Different strokes for different folks: disciplinary variation in academic writing. In: Flottem, Kjersti (ed.): *Language and discipline perspectives on academic discourse*. Newcastle: Cambridge Scholars Press, 89-108.
- Kytö, Merja/Romaine, Suzanne (2000): Adjective comparison and standardization processes in American and British English from 1620 to the present. In: Wright, Laura (ed.): *The development of standard English 1300-1800: theories, descriptions, conflicts*. Cambridge: Cambridge University Press, 171-194.
- Leech, Geoffrey (2012): Why do linguistic forms decline and disappear? – the neglected negative side of recent change in standard English. In: *Proceedings of the 4th International Conference on Corpus Linguistics – Language, corpora and applications: diversity and change (CILC2012)*, Jaén, Spain.
- Mair, Christian (2006): *Twentieth-century English: history, variation and standardization*. Cambridge: Cambridge University Press.
- O'Halloran, Kay L. (2005): *Mathematical discourse: Language, symbolism and visual images*. London: Continuum.
- Teich, Elke/Fankhauser, Peter (2010): Exploring a corpus of scientific texts using data mining. In: Gries, Stefan Th./Wulff, Stefanie/Davies, Mark (eds.): *Corpus-linguistic applications. Current studies, new directions*. Amsterdam/New York: Rodopi, 233-247.
- Ventola, Eija (1996): Packing and unpacking of information in academic texts. In: Ventola, Eija/Mauranen, Anna (eds.): *Academic writing. Intercultural and textual issues*. Amsterdam/Philadelphia: John Benjamins, 153-194.

The expression ofthetic judgments in Older Germanic and Romance

Abstract

This paper investigates the formal, syntactic and discourse-pragmatic properties of clauses in which the division into a predicational base, or topic of the utterance, and a comment on this topic, fails to apply. Since Kuroda (1972), this property has been claimed to hold for sentences representing the so-called thetic type of judgment, as opposed to categorical sentences which display a bipartite division into an aboutness topic and a comment added to this topic. This property is linked to a number of universal formal properties like the use of semantically empty copula or existential verbs, indefinite subject expressions, and verb-subject order (VS) cross-linguistically. We investigate the properties of thetic judgments in Old Germanic and Old Romance in comparison to each other and investigate the similarities between prototypical thetic sentences involving novel referents and VS-clauses with given postverbal subjects. Our claim is that the latter are instances of theticity in the broader sense of this notion.

1. Introduction

This paper demonstrates how corpus-based investigations support research on the role of information structure in syntactic variation and change. On the basis of data from Old High German (OHG) and Old French (OF), we study the formal, syntactic and discourse-pragmatic properties of clauses in which the division into a predicational base, or topic of the utterance, and a comment on this topic, fails to apply. Since Kuroda (1972), this property has been claimed to hold for sentences representing the so-called thetic type of judgment, as opposed to categorical sentences which display a bipartite division into an aboutness topic and a comment added to this topic, as in the sentence pair in (1a, b) from Portuguese (see Martins 1994: 393):

- (1a) Ptg. *O gato está lá no jardim.*
the cat is there in-the garden

‘The cat is in the garden.’

categorical judgment, topic-comment structure

- (1b) Ptg. *Está lá um gato no jardim.*
 Is there a cat in-the garden
 ‘There is a cat in the garden.’
 thetic judgment, no topic-comment structure

Prototypical instances of the thetic kind of judgment are presentational sentences and existential constructions, i.e. clauses that are used to establish novel referents into the discourse. This property is linked to a number of universal formal properties like the use of semantically empty copula or existential verbs, indefinite subject expressions, and verb-subject order (VS) cross-linguistically, as exemplified by the modern Portuguese example (1b).

Old Germanic and Old Romance have long been known to be among the languages that display VS order in declaratives used to establish novel referents into the discourse. The examples in (2) and (3), taken from OHG and OF, display the basic universal properties of presentational sentences and existential constructions as canonical representatives of the thetic type of judgments:

- (2) OHG *uuas thar ouh sum uuítua* ¹
 was there also certain widow
In thero burgi
 in this town
 ‘There was a widow too in that city.’
 Lat. *Vidua autem quaedam erat / In ciuitate illa*
 (Tatian 201, 2)

- (3) OF *A rome ot .I. empereeur qui ot*
 in Rome had one emperor that had
non dyoclecien.
 name Diocletian
 ‘There was an emperor in Rome that had the name
 Diocletian.’
 (Sept Sages 01, 001)

¹ The slash ‘/’ in data from the OHG *Tatian* translation stands for a line break.

At the same time, VS order in Older Germanic and Romance is attested in clauses that deviate from the properties of canonicalthetic sentences in several respects. First, the clauses contain predicates different from the class of existential verbs, and second, their subjects are definite descriptions referring to contextually given entities, cf. (4) and (5):

- (4) OHG *bigonda* *ther* *phariseus* [...] *quedan*
 began this Pharisee speak.INF
 ‘The Pharisee began to speak’
 Lat. *Phariseus autem coepit* [...] *dicere*
 (Tatian 126, 5-6)

- (5) OF *.I.* *jour* *apela* *li emperieres* *les .VII. sages*
 one day called the emperor the 7 wise men
 par *leur* *nons*
 by their names
 ‘One day the emperor called the 7 wise men by their names.’
 (Sept Sages 01, 005)

The question arises as to how to interpret sentences like those in (4) and (5). On the one hand, the postverbal position of the subjects can be explained on purely structural grounds, e.g., as resulting from a kind of verb-second constraint leading to VS order in cases in which the verb is fronted to C but the subject cannot be placed in SpecCP because this position is already taken by another constituent or because of other reasons. In line with this argumentation, declarative clauses with post-verbal subjects cannot be unified under any common discourse-pragmatic feature but rather behave heterogeneously with respect to discourse and information structure. On the other hand, it can be assumed that sentences with post-verbal given subjects share properties of presentational and existential constructions, if we look at the function of the clauses in the global context.

The present paper will pursue an explanation in line with the second option. It builds on the idea that albeit given, the subjects in (4) and (5) are not topics, because the respective contexts do not license utterances with a topic-comment division. Our claim, thus, will be, that VS order with both novel and given subjects in Old Germanic and Old Romance can be interpreted in a singular fashion.

But there are also differences that become obvious if we compare VS order in Germanic and Romance. First, it seems that verb-initial order, i.e., an order with no element preceding the verb at all, are more common in OHG (cf. (2) and (4) above) but rare in medieval Romance texts. In the latter, verb-initial order is only attested after conjunctions like *et* or particles like *si* etc. which are not counted as a constituent occupying the first position in the clause. Second, there is no consensus concerning the interpretation of VS order in terms of syntactic structure in Germanic and Romance. While for older Germanic, both verb-initial clauses as well as Adverb-VS orders are analysed as involving verb fronting to C, thus derivable under the well-known verb-second constraint to root clauses, this analysis is questioned for Romance by Kaiser (2002) and Rinke/Meisel (2009) arguing that in such clauses, the verb only goes to a functional projection below C, e.g. TP. Therefore, we want to address the relevance of our findings with respect to the proper interpretation of clause structure in Germanic and Romance.

2. Pros and cons of the V2-analysis

2.1 V2 in early Germanic

While in OE, strict verb-second is argued to apply in main clauses introduced by a *wh*-word, the negative particle *ne*, and temporal connectives of the type of *tha / thonne* ‘then’ (van Kemenade 1987 and later), OHG is considered a language with an already firmly established verb-second rule. Axel (2007) shows that constitutive features of the verb-second property of modern V2-languages (Fanselow 2003, 2004; Frey 2004) already apply in the earliest OHG attestation, above all i. obligatory verb fronting to C and ii. obligatory filling of SpecC. The only property of verb-second that is not well-established in OHG, is base-generation of expletive *es* in SpecC, which seems to emerge towards the Middle High German period.

- (6a) *ārstuont* *siu* *tho* *if*
 rise-3SG.PRET she then-PRT up-PRT
 ‘She arouse then’
 Lat. *et*² *surrexit*
 (Tatian 84, 14)

² In the manuscript of the Tatian translation, ‘*et*’ is used to represent the ‘*et*’-ligature.

(6b) [_{CP} C **árstuont**_i [_{VP} *siu tho úf t*_i]]

Following this, postverbal subjects in OHG may be explained as in-situ subjects in cases in which a non-subject XP, either base-generated in or moved to the leftmost position in the middle field, is moved to SpecC.³ The first option is demonstrated by (7) in which the sentence adverbial *chiiuusso*, which is base generated higher than the subject in the middle field, is moved to SpecC, while the lexical subject remains in-situ:

- (7) *endi [chiiuusso] ist christus in dheru selbun salbidhu chimeinit*
 and certainly is Christ in this same salve meant
 ‘and certainly Christ is meant in this same salve’
 Lat. *et utique christus ipsa unctioe monstratur* (Isidor 144)

But subjects remain in post-verbal position because they are non-topics. Consider (8) which predicates about the referent of the object expression, the disciples, rather than about the subject Jesus:

- (8) [...*Inti teta/ thaz uuarun zueliui*_i *mit Imo*
 = ‘and He chose **twelve men**_i to be with Him’]
 (Tatian 59, 16)

_{TOP} [*thie*]_i *namta her boton*
 DEM:acc.pl. named he apostles
 ‘These He named apostles’
 Lat. *quos & apostolos nominavit*

It is intriguing that even in cases in which no higher material is shifted to SpecC as in (8) above, the subject may remain in postverbal position. As Axel (2007) notices, these instances are not the equivalents of V2 clauses with expletive *es* in modern German, so it could not be the case that the prefield is empty because the expletive is still not available in the grammatical system of the language. There must be other reasons preventing the subject, even if it is a pronoun, to move to SpecC in these cases.

³ No intermediate functional projections such as TP/IP between VP and CP are assumed in Axel's (2007) approach (cf. also Haider 1997 and Sternefeld 2007 for modern German).

Villehardouin (13 th century)		
V1 (<i>et</i>)V(X)	(157/1180)	13.3%
SVX	(410/1180)	34.7%
V3 XP(S)/(YP)V	(63/1180)	5.3%
XVS	(263/1180)	22.2%
XV (<i>pro</i>)X(<i>pro</i>)V(<i>pro</i>)	(287/1180)	24.3%
Sept Sages (13 th century)		
V1 (<i>et</i>)V(X)	(151/1073)	14.1%
SVX	(559/1073)	52.1%
V3 XP (YP)/(S) V	(53/1073)	4.9%
XVS	(96/1073)	8.9%
XV (<i>pro</i>)X(<i>pro</i>)V(<i>pro</i>)	(214/1073)	19.9%

Table 1: Frequency of different word order patterns

Third, *unambiguous evidence against* a V2-analysis is present, e.g. *verb third clauses* (13).

- (13) OF *Maintenant touz les .VI. messages s'agenoillierent*
 Now all the six messengers REFL-kneel
a leur piez moult plorant
 to their feet much crying
 'Now all the six messengers sobbingly fell to their knees.'
 (Villehardouin, § 28, 1, in: Rinke/Meisel 2009)

In addition, postverbal DP-subjects (in contrast to pronominal subjects) always occur to the right of negation and short adverbs that mark the left edge of the *vP*. Therefore, they cannot have left the *vP*-domain. Consider example (14):

- (14) OF *Et la le fist moult bien mahi de vilaincourt*
 and there it did very well Mahi of Vilaincourt
 'And Mahi de Viliancourt did it very well there'
 (Villehardouin, §169, 03, in: Rinke/Meisel 2009: 104 [= (17d)])

Example (15) shows that subject noun phrases can occur to the right of participles in OF.

- (15) OF *Et dedenz ces. VIII. jorz furent venuz tuit*
 and within these eight days were come all
li vessel et les barons
 the ships and the barons
 ‘And within these eight days came all the ships and the
 barons’
 (Villehardouin, § 126, 03, in: Rinke/Meisel 2009: 106)

This example constitutes a typical instance of OF VS order. One could say that VS clauses are a closed group: they are mainly *triggered by temporal (and locative) adverbs* in clause-initial position. This is also observed in OE and OHG. Consider example (16):

- (16) OHG *tho uuas man In hierusalem*
 then was man in Jerusalem
 ‘There was a man in Jerusalem’
 Lat. & *ecce homo erat In hierusalem*
 (Tatian 37, 23)

3. Conditions on VS order

In the introductory section, we drew the attention to the fact that VS order in Germanic and Romance is typical for declaratives sharing all properties of presentational sentences and existential constructions, like semantically empty copula or existential verbs and novel indefinite subjects, cf. (17)=(4) and (18)=(5) from above. But at the same time, VS order is attested in clauses which do not share the properties of presentational sentences and existential constructions.

First, and most importantly, we find VS orders involving subject expressions referring to a contextually given entity. Consequently, the subject expressions are not indefinite phrases but definite lexical expressions, even pronominal ones. This is in contrast to the canonical function of the respective clauses, namely to establish a novel referent in the discourse.

- (17) OHG *bigonda ther phariseus* [...] *quedan*
 began this Pharisee speak.INF
 ‘The Pharisee began to speak’
 Lat. *Phariseus autem coepit* [...] *dicere*
 (Tatian 126, 5-6)
- (18) OF *.I. jour apela li emperieres les .VII. sages*
 one day called the emperor the 7 wise men
par leur nons
 by their names
 ‘One day the emperor called the 7 wise men by their names.’
 (Sept Sages 01, 005)

Second, the verbs that appear in clauses with VS orders in OHG and OF also differ from the semantic class of predicates established in presentational sentences and existential constructions. Here, we observe the following lexical classes of predicates: motion verbs, context verbs, perception verbs, passives, phase verbs etc. We illustrate VS order for motion verbs in OHG and OF:

- (19) OF *après se leva li seconz*;
 thereafter REFL stood up the second
 ‘Thereafter the second one stood up’
 (Sept Sages 01, 010)
- (20) OHG *quamun sie thó*
 come-3PL.PRET they then-PRT
 ‘Then they came’
 Lat. *& uenerunt*
 (Tatian 55, 27)

Another licensing condition is the beginning of episodes. As already described for older Germanic (Petrova 2008, 2011), VS tends to appear discourse-initially, e.g. at the beginning of a text or of new episodes within texts. In OF, we also observe that VS order is very common in the beginning of a text or a paragraph. Some examples are given below, the numbering “(X), 001” indicates the beginning of a new chapter.

- (21) OF *A rome ot .I. empereur qui ot
in Rome had one emperor that had
non dyoclecien.
name Diocletian*
'There was an emperor in Rome that had the name
Diocletian'
(Sept Sages 01, 001)

Additionally, as already mentioned above, we observe that clauses involving VS order prototypically contain overt temporal expressions. Remarkably, these temporal adverbials introduce new time intervals serving as the topic time of the episode to come. Consider (22) from OHG where the new episode is introduced by establishing a new indefinite topic time interval:

- (22) OHG *uuas tho giuuortan in
PASSAUX-3SG.PRET then become-PAST.PART in
anderemo sambaztag
another-DAT.SG Sabbath*
'It happened then on another Sabbath'
Lat. *Factum est in alio sabbatum autem*
(Tatian 106, 6)

In OF, we also discover this function of adverbials in VS clauses:

- (23) OF *.I. jour apela li emperieres les .VII. sages
one day called the emperor the seven wise men
par leur nons.
by their names*
'One day the emperor called the 7 wise men by their names'
(Sept Sages 01, 005)

In fact, adverbials are the most frequent sentence initial element in Old French inversion structures. This is also true for the two texts we have mentioned before:

- Villehardouin: 232 out of 263 XVS-clauses (88.2%)
- Sept Sages: 78 out of 96 XVS-clauses (81.5%)

The above mentioned adverbials also introduce sentences that report about the ongoing course of actions, either at the beginning of a paragraph or after a passage of background information or reported speech. Such sentences induce a whole focus reading because they are interpreted as an answer to a context question like “What happened afterwards?” Such sentences mainly serve to advance the storyline. They describe temporal sequences of events rather than elaborating on a topic, as also stated in Rinke/Meisel (2009). In the following paragraph we discuss the interpretation of VS sentences in more detail.

4. Interpretation

We believe that the properties described above indicate that postverbal subjects are no topics because the clause has no topic comment structure. Here are our arguments in detail.

Let us start with the lexical classes of predicates. We claim that the verbs triggering VS order involve a change in the discourse setting and establish a novel situation: motion verbs signal the arrival or withdrawal of referents, perception verbs and phase verbs convey sudden changes in the state of informedness of the protagonists or the initiation of a new state of affairs. For passives, such a pattern is difficult to establish, but if we look at the sentences, we see that they also introduce a novel situation which is relevant for the further development of the narrative (these are sudden or unexpected events or actions). To conclude, the lexical classes triggering VS order license conditions under which the clause does not predicate on a single referent but rather on the entire new situation.

Let us interpret the position of VS in discourse organisation. The fact that VS is typical for discourse-initial or episode-initial sentences suggests that there is no direct linking to the previous context, consequently, no topic is established as the predicational base of the utterance.

Finally, let us interpret the role of the adverbials. Obviously, as we have seen above, VS order correlates with clauses which signal temporal succession rather than temporal overlapping with the preceding discourse. From the point of view of rhetorical structure, discourse continuity is prototypical for relations of narration or coordination in discourse, while its opposite, elaboration, is related to simultaneity. Petrova (2011) concludes that the overall

function of VS clauses is that they are incompatible with the rhetorical relation of elaboration.

Let us look how this relates to the lack of topic-comment structure in VS clauses. According to Sasse (1995), theticity is more than just presentational sentences and existential constructions. Rather, he subsumes *all focus sentences* such as (24b) under the thetic type of judgment. Only the context in (24a) licenses an answer uttering something about a referent in the preceding question (capitalization is used to mark the constituent carrying the main accent in the clause):

- (24) a. A: *How's your neck?* B: *My neck HURTS.*
 b. A: *What happened?* B: *My NECK hurts.*

Theticity is extended to utterances which do not contain a division into topic and comment. In other words, topic-comment does not depend on the givenness of a referent but rather on the fact whether or not the context in which an utterance is set triggers an interpretation of this utterance as adding a comment on a particular referent singled out as the topic of the predication.

We want to defend the idea that this condition is not fulfilled in the contexts of our VS clauses. We argue that the clauses described so far are event reporting sentences predicating on a situation as a whole, rather than on a particular referent. If we should identify the current question under discussion (Beaver/Clark 2009) which an VS-clause provides an answer to, then this question would be of the type “What happened then, how does the story go on?” rather than “What about X?”

This also becomes obvious if we try to apply the traditional tests for identifying sentence topics. According to the literature, sentence topics are those constituents which can replace X in a semantically equivalent periphrasis of the critical sentence of the type ‘As for X, X ...’ or ‘He said about X that Y’ (cf. Reinhart 1981, Erteschik-Shir 2007: 19f.). In our opinion, both tests fail when applied to the postverbal subjects of (X)VS sentences in OHG and OF.

References⁴

Text editions

OHG:

Isidor – Der althochdeutsche Isidor. Nach der Pariser Handschrift und den Monseer Fragmenten, ed. H. Eggers. Tübingen: Niemeyer. 1964.

Tatian – Die lateinisch-althochdeutsche Tatianbilingue Stiftsbibliothek St. Gallen Cod. 56, ed. A. Masser. Göttingen: Vandenhoeck & Ruprecht. 1994.

OF:

Villehardouin – Josfroi de Villehardouin: La Conquete de Costentinoble: d'après le manuscrit no. 2137 de la B.N., C.R.A.L.; Nancy. 1978.

Sept Sages – Les sept Sages de Rome: roman en prose du XIIIe siècle; d'après le manuscrit no. 2137 de la B.N., C.R.A.L.; Nancy. 1981.

The Isidor, Tatian and Villehardouin texts are electronically available in the TITUS corpus :

Isidor: <http://titus.uni-frankfurt.de/texte/etcs/germ/ahd/isidor/isido.htm>

Tatian: <http://titus.uni-frankfurt.de/texte/etcs/germ/ahd/tatianx/tatia.htm>

Villehardouin: <http://titus.uni-frankfurt.de/texte/etcs/ital/afr/villehar/ville.htm>

Secondary literature

Axel, Katrin (2007): *Studies in Old High German syntax: left sentence periphery, verb placement and verb-second*. Amsterdam/Philadelphia: John Benjamins.

Beaver, David/Clark, Brady Z. (2009): *Sense and sensitivity: how focus determines meaning*. Malden, MA etc.: Wiley-Blackwell.

Clark, Robin/Roberts, Ian (1993): A computational model of language learnability and language change. In: *Linguistic Inquiry* 24(2): 299-345.

Costa, João (2004): *Subject positions and interfaces: the case of European Portuguese*, Berlin/New York: Mouton de Gruyter.

Erteschik-Shir, Nomi (2007): *Information structure. The syntax-discourse interface.* (= Oxford Studies in Syntax and Morphology). Oxford: Oxford University Press.

Fanselow, Gisbert (2003): *Münchhausen-style head movement and the analysis of verb second*. In: Mahajan, Anoop (ed.): *Syntax at sunset: head movement and syntactic theory*. Los Angeles: USLA Working Papers in Linguistics, 40-76.

⁴ All URLs have been checked and found valid as of late January 2015.

- Fanselow, Gisbert (2004): Cyclic phonology – syntax-interaction: movement to first position in German. In: *Interdisciplinary Studies in Information Structure (ISIS) 1*: 1-42. http://opus.kobv.de/ubp/volltexte/2006/826/pdf/isis01_1fanselow.pdf.
- Frey, Werner (2004): The grammar-pragmatics interface and the German prefield. In: *Sprache und Pragmatik 52*: 1-39.
- Haider, Hubert (1997): Projective economy. In: Abraham, Werner/van Gelderen, Elly (eds): *German: syntactic problems – problematic syntax*. Tübingen: Niemeyer, 83-103.
- Kaiser, Georg (2002): *Verbstellung und Verbstellungswandel in den romanischen Sprachen*. Tübingen: Niemeyer.
- Kuroda, Sige-Yuki (1972): The categorical and thethetic judgment. Evidence from Japanese Syntax. In: *Foundations of Language 9*: 153-185.
- Martins, Ana Maria (1994): *Clíticos na História do Português*. PhD Dissertation, Universidade de Lisboa.
- Petrova, Svetlana (2006): A discourse-based approach to verb placement in early West-Germanic. In: Ishihara, Shinichiro/Schmitz, Michaela/Schwarz, Anne (eds.): *Interdisciplinary studies on information structure (ISIS) 5*. Potsdam: Universitätsverlag, 153-185.
- Petrova, Svetlana (2011): Modeling word order variation in discourse: On the pragmatic properties of VS order in Old High German. In: *Oslo Studies in Language 3(3)*: 209-228.
- Reinhart, Tanya (1981): Pragmatics and linguistics: an analysis of sentence topics. In: *Philosophica 27*: 53-94.
- Rinke, Esther/Meisel, Jürgen M. (2009): Subject-inversion in Old French: syntax and information structure. In: Kaiser, Georg/Remberger Eva-Maria (eds.): *Proceedings of the Workshop “Null-subjects, expletives, and locatives in Romance”*. Konstanz: Universitätsverlag, 93-130.
- Roberts, Ian (1992): *Verbs and diachronic syntax*. Dordrecht: Kluwer.
- Roberts, Ian (2007): *Diachronic syntax*. Oxford: Oxford University Press.
- Sasse, Hans-Jürgen (1995): “Theticity” and VS order: a case study. In: *Sprachtypologie und Universalienforschung 48(1/2)*: 3-23.
- Sternefeld, Wolfgang (2007): *Syntax. Eine morphologisch motivierte generative Beschreibung des Deutschen*. Vol. 1. 2nd. edition. Tübingen: Stauffenburg.
- van Kemenade, Ans (1987): *Syntactic case and morphological case in the history of English*. Dordrecht: Foris.
- Yang, Charles D. (2002): *Knowledge and learning in natural language*. Oxford: Oxford University Press.

RICHARD INGHAM

Spoken and written register differentiation in pragmatic and semantic functions in two Anglo-Norman corpora

Abstract

This study shows that historical corpora can reveal the selective effect of register for favouring or disfavouring linguistic innovations. Four innovations in the history of French are considered, and show contrasting patterns as regards the adoption of new forms or meanings. Two are favoured in the spoken and two in the written register corpus. The corpora are both of Anglo-Norman, for which the existence of a corpus of courtroom dialogues allows a probably unique opportunity to study a representation of authentic spoken register use in the medieval period, in addition to a corpus of written-origin data belonging to the same discourse domain.

1. Introduction to issues and sources

The field of corpus studies has in recent years seen a remarkable proliferation of specialist corpora, for various domains such as correspondence, newspapers, spoken language, academic English, and so on, though the study of language through a very large all-purpose corpus, such as the BNC, continues to hold a central place in research: the goal of overall representativeness advocated by corpus theorists and designers, which it is intended to serve, remains a key principle. There has always been, among corpus researchers, an awareness of register and stylistic differences and more broadly of how language varies with content domains, which a large all-purpose corpus is able to address by having subdivisions distinguished by register or genre. Nevertheless, with the long-term trend towards increased specialisation in research comes a dynamic favouring specialised corpora: researchers in a given sub-domain may need access to plentiful data of a type not previously explored, or not explored in quantities they need. Now, the same considerations apply to the historical study of language: one must also ask how well a historical corpus covers past states(s) of the language (even though data limitations will often make the question difficult to answer). Likewise, if specific issues require individual treatment, specialist historical corpora providing data that control for certain external variables are as desirable as they would be in the contemporary context. In recent years, linguists have pursued issues in language

variation using the resources various types of historical corpora can provide (Curzan 2009).

Ultimately, the issue is how a corpus can handle the inescapable variability of language, either by setting out to be as comprehensive as possible, or else by restricting external parameters so that variation is reduced, allowing for the more accurate study of a putative distinct ‘variety’ through the use of that corpus, alongside others dedicated to other varieties. Where relevant data are available for the corpus study of past states of a language, variability may be profitably addressed in either of these ways: large ‘whole-language’ corpora can be envisioned insofar as this is possible, or individual specialist corpora can be created. It seems uncontroversial to say that where textual range made available by corpora needs to be as comprehensive as resources allow, without unnecessary duplication.

The aim of this paper is to show that where available data sources can be enlarged so as to capture distinct registers in historical eras of language, interesting linguistic differences emerge. The issue will be pursued with reference to differences between written and spoken registers in the earlier history of French. The terms ‘written’ versus ‘spoken’ will be used here as convenient labels for, respectively, the ensemble of formal features of language composed offline, and the informal ones of language produced in real time conditions. Variation along these dimensions is not well served by existing corpora of pre-modern French such as the *Base de Français Médiéval* (BFM), or the *Nouveau Corpus d’Amsterdam* (NCA; Stein (ed.) 2007), which either contain almost exclusively literary works or –, where they feature a better spread of texts, as is the case with the *Modéliser le changement : les voies du français* (MCVF) corpus (Martineau et al. 2009) – cover each period less extensively than the shorter timescale corpora, and do not aim to balance genres over time, so that confounds arise between time period and text type. These are not new observations on our part, but are aspects of the well-known ‘bad data’ problem in diachronic linguistics: language change typically begins in spoken language, but the textual record until recent times consists of written language, so that historical linguists and language change researchers lack access to the spoken register in which the developments they study usually took place. At first sight, the prospects for improving on the situation in this or any pre-modern language are not good. It has not of course arisen because corpus creators were inept in their sampling of older texts. In general, written material from the past has tended to be kept (and edited) only if it has some literary, legal or

religious merit. The texts corpus designers could use were thus from the start biased against ordinary language use, in particular as regards interactive spoken language, although this data deficiency can to some extent be mitigated by taking into account fictional dialogue, e.g. in drama (Kytö/Culpeper 2006), as well as dialogic interaction in trial reports (Huber 2007, Archer 2005).

As for the pre-modern period, it is usually assumed that only fictional dialogue is available for this purpose, in verse drama or direct speech in other verse and prose works, otherwise we have to wait until modern era trial reports before encountering records of words actually spoken (or claimed to have been). There is, however, an exception to this pervasive lack of authentic spoken language from the Middle Ages. It consists of debates between lawyers presented in direct speech in the English law yearbooks, beginning in the late 13th century, and continuing until the end of the Middle Ages. These debates are supposed to be the record of the pleading used by the two sides in a pre-trial hearing before a panel of judges, especially in cases concerning property. Their authors are thought to have been jurists or law students who, for the purposes of the training given to apprentice lawyers, recorded proceedings in some form of shorthand, and then reconstituted the dialogues afterwards in the form that we have; they were then copied in numerous manuscripts and eventually published by modern editors. They are in the variety of Old French used in England conventionally known as Anglo-Norman.¹

Originally created by the French-speaking settlers who came to England during and after the Norman conquest, Anglo-Norman underwent a gradual change in its status so that by the late 13th century it was almost always a second language, though one acquired in childhood in what appeared to be naturalistic circumstances (Ingham 2007, 2012b), by members of English medieval society favoured either by birth or by education. It enjoyed quite widespread use in urban communities (Short 2009), especially in the South of England. The view of some French philologists that it was no more than an artificial jargon, or indeed merely the 'bad French' of incompetent language learners, has been shown to be untenable (Rothwell 2001, Trotter 2003). Its erratic spelling, often representing insular phonology, which deviated considerably from continental French, may give the superficial impression of poorly learned

¹ Not to be confused with Law French, which was a later variety consisting of legal jargon padded out with numerous English words, and using a drastically simplified version of French grammar.

French, but close study of how its morpho-syntactic structure patterned shows that it followed continental French in most grammatical respects (Ingham 2008, 2012b). Though its morpho-phonology certainly differed from that of central French (Pope 1934), the same was true of other peripheral varieties such as Picard and Walloon. As we shall see, Anglo-Norman, although socio-linguistically somewhat unusual in that it was restricted to bilingual educated speakers, rather than being spoken as a mother tongue by a broadly monolingual population, is good for our purpose in that it showed interesting register variation, for which plentiful documentary evidence survives that bears on the matter at hand.²

The Anglo-Norman data to be used in this study are drawn from two databases, which can be referred to as corpora in that they contain texts systematically chosen to exemplify the language (variety) intended as the object of study. The design of the first corpus, the Parliament Rolls of Medieval England (Given-Wilson et al. (eds.) 2005), henceforth referred to as PROME, aimed to include all extant petitions to the English Crown in the medieval period; those in Anglo-Norman begin in around 1290 and continue into the 15th century. The second, the Anglo-Norman Yearbooks Corpus (Ingham/Larrivé 2010) was created by sampling the medieval records of pleading in English medieval King's bench courts, at periods each roughly a generation apart, i.e. 1267-1272, 1292-1305, 1340-1346 and 1385-1390.

The following extract will illustrate the interactional properties of this source, as well as the specialised terms used:

- (1) *Mallore: E ne quidez point qe pur cele aide le viscount purra destreindre quele parte il voudra?*

'And do you not think that for that aid the sheriff can distrain wherever he pleases?'

Lanfare: Noun pas en le haut estrete.

'Not in the highway.'

Mallore: Certes si purra.

'Certainly he can.'

(YB 1305)

² It is unfortunately the case that texts of a comparable nature originating in France are not extant, because the legal profession there used Latin to record all its transactions during the period in question.

Dialogic interaction such as the above does not represent unscripted conversation, of course, but just how different it was from everyday language will never be known. The question we can address here is how far it differed from written register language in the same general content domain, that of the law. The language of the yearbooks will be compared with texts which are indisputably written register in origin, petitions written in Anglo-Norman intended to be delivered to the King in Parliament during the later mediaeval period beginning at the end of the 13th century. The two data sources thus overlap closely in time, and indeed were produced by the same class of educated professionals within the legal system, cf. Dodd (2009) who considers that the petitions were drawn up by legal professionals guided by the petitioner himself or herself. The following extract illustrates features of this text type, such as the use of complex clause structure, but also the roughly comparable content material to that of the yearbooks:

- (2) *Item, prie la commune: qe quant homme est atteint a suyte de partie pur trespas faite encountre la pees, et le trespasour pris et en garde des mareschals, et les mareschals le lessent a meynprise ou aler a large; q' ils soient chargez des damages avantditz.* (PROME Edw II March 1348)

‘Also, the commons pray: that when a man is attainted at the suit of the party for a trespass committed against the peace, and the trespasser is taken into the custody of marshals, and the marshals grant him bail or to go free, they should be charged with the aforesaid damages.’

2. Linguistic variables investigated in the corpora

In this study we compare the uses in the two corpora of four linguistic expressions: two discourse markers which showed interesting sense extension in Old French, a newly lexicalized connective, and an indefinite pronoun/modifier. These elements offer a range of different linguistic items on which to contrast the two registers provided by the corpora, especially in their semantic and discursal features. As will be seen, two of them are favoured by spoken and two by written register uses.

Discourse markers and connectors are known to be features of language which often show substantial variation between written and spoken registers. Anglo-Norman was no different here. The adverb *encore*, with basically a temporal meaning, featured frequently in the debates as an adversative marker used in argumentation with the force of English *still*, *nevertheless*, e.g.:

- (3) *Sharschulle: Il est possible qe feffementz ount este faitz de tut temps de la terre, issi qe nul tenant unqes puis temps de memore murust seisi, issi qe seisine ne put il lier, et unquore la chose due.* (YB 1346)

‘It is possible that feoffments have been made of the land time after time, so that no tenant ever died seised since time of memory, and so that he could not lay seisin, and yet the thing is due.’

- (4) *Thorpe: Le Roi en sa presence demene recorda et resceit homage et fines; et si ceo soit entre en sa tresorie, et nest pas par cas mande, uncore en autre place ceo serra vouche come rrecord.* (YB 1341)

‘The King in his own presence recorded and received homage and fines; and if this is entered in his Treasury, though, perhaps, it is not sent in, yet in another place it shall be vouched as a record.’

No such uses of *encore* have been identified in the petitions at this period. Instead, it appears to have been used in this source only in a temporal sense, e.g.:

- (5) *A queu chose faire le roi ne poait adonqes, ne unqore poet entendre, par certaines causes.* (PROME Edw III March 1348)

‘Which thing the king was then unable to do, and still cannot attend to, for certain reasons.’

This register difference is informative as to the process of lexical sense extension whereby abstract senses expressing speaker attitude are added to a primary referential meaning (Traugott/Dasher 2005). They confirm that the sense extension was favoured in the spoken register rather than in that of written discourse, as we would expect if linguistic change typically occurs in spoken usage.

Another very commonly used discourse marker in the year book debates showing a similar sense extension is the expression *donques*. Its basic meaning is again temporal (‘then’), but already in Old French it gained the argumentative meaning ‘so’. In the use which concerns us, in utterance initial po-

sition, it can signal a subjective inference on the part of the speaker or writer, e.g.

- (6) *Thorpe: Donques vous ne deditez mye qe nous fumes distreint par vostre defaute.* (YB 1340)

‘So you don’t deny that we were distrained by your fault’

- (7) *Thorpe: Bien, Sire. Donques vous veiez bien coment il plede en descharge de cesti terre.* (YB 1340)

‘Very well, sir. So you can see how he pleads in discharge of this land’

It could also serve to signal a discourse move concluding the barrister’s argument:

- (8) *Stoufford: Donques demandoms nous jugement si, par la seisine nostre pere, nous qe sumes issu en la taille de cesti bref serroms ouste.* (YB 1340)

‘So we ask for a ruling if, by our father’s occupation of the property, we who are issue in tail shall be ousted from this writ.’

Such cases occurred frequently in the yearbooks. Note that *donques* had no temporal value in these instances: it was not anchored to speech time, either in the sense of ‘then’ prior to speech time, or ‘then’ subsequent to speech time, i.e. in the future. In fact, the tense of the main verb is in each case present-, not past- or future-referring. In this use, the temporal meaning ‘then’, denoting a distal point in time, has been bleached, as it has in the analogous English item.

The spoken discourse of the year books clearly favoured this speaker oriented use, with its discourse function of signalling a subjective inference, very much as is the case with utterance-initial *donc* in modern French; the subjective judgment of the speaker was engaged by the very nature of action of the pleading. Such uses, however, are almost entirely absent from PROME; the only exceptions, significantly, are cases where spoken register communication is either conveyed in direct speech (9), or approximated to in the form of a personal letter (10).

- (9) *Sur queu respons dist le chaumberleyn a les dites communes, donques vous voillez assentir au tretee du pees perpetuele si homme la puisse avoir. Et les dites communes responderent entierement et uniement, oil, oil.* (PROME Edw III April 1354)

‘At which response the Chamberlain said to the aforesaid Commons, “So you want to agree to the treaty of perpetual peace if it can be obtained?” And the aforesaid Commons replied entirely and unanimously: “Yes, yes”.’

- (10) *Car le plus grant partie de noz gentz serront hors du paiis devers vous, et devers l’Escotz, donques je sceai bien qe les orglous vileins de Londres veullent ordeiner une armee encountre vous la.*
(PROME Ric II Nov 1380 (extract from letter))

‘For most of our people will be away from the country with you, and with the Scots, so I know very well that proud wretches of London will array an army against you there.’

The third item to be considered is the connective *à cause que* (‘because’). Although in modern French it is now a substandard form, its introduction appears on the evidence of these Anglo-Norman corpora to have been an innovation that occurred in the written register, some time in the 14th century. The yearbook dialogues avoided it completely, but by the late 14th century it had become a fairly common way of providing a causal link between propositions (Ingham 2012a). Examples are:

- (11) *Vous savez bien coment au darrein parlement le roi de vostre assent respit le noun du roi de France, a cause qe son adversari avoit enfreint la pees autrefoitz afferme entre eux.* (PROME Edw III Jan 1371)
‘You know well how at the last parliament the king, with your assent, resumed the title of King of France, because his enemy had broken the peace previously established between them.’

The equivalent connectives in the yearbook debates are normally *pur ceo qe* (‘because’) or *car* (‘for’), which are also used in the petitions, so they were register neutral, unlike *à cause qe*.

The final item to be investigated is one where change also took place first in the written register. Buridant (2000) noted that the indefinite expression *aucun* (‘some’, ‘any’ in medieval French) began being used in negative contexts in continental French administrative documents, not in fictional works. Later on, it became the indefinite normally used in French negative clauses, but in the mediaeval period the most common indefinite in this context was still *nul* (‘any’). An investigation of administrative documents in Anglo-Norman by Ingham (2011) concluded that Buridant’s findings for continental French also

applied to the insular variety. We have subsequently discovered that the initial stages of a shift to *aucun* in negative clauses occurred in the written register texts in PROME, where *aucun* appears in this context around the mid-14th century:

- (12) ... *qe le ditz grevances, oppressions et damages en mesme le roialme desadonqes mes ne serroient suffertz en ascune manere.*
(PROME Edw III Feb 1351)

‘... that the said grievances, oppressions and damages in the same kingdom henceforth will not be tolerated in any way’

- (13) *Item, ordeine est qe nul autre subget du dit realme, gardant et sustenant ces ordinances, n’encourge aucune forfaiture de vie et de membre, de terres, de heritage ne des biens devers le roi.*
(PROME Edw III Jan 1365)

‘Also, it is ordained that no other subject of the said realm, keeping and observing these ordinances, shall incur any forfeiture of life and limb, lands, inheritance or goods against the king.’

Although *nul* remained by far the most common negative indefinite, *aucun* in negative clauses became fairly common in PROME after 1375, with around 50 hits in negative contexts up to 1399, e.g.:

- (14) ... *qe toute chose qe y doit estre tenuz en secret sanz découvrir ne découvriront a aucun estrange* (PROME Ric II Oct 1377)

‘... that all which ought to be kept secret without disclosure, should be disclosed to no other person.’

- (15) *Ils n’osent ne ne veullent en aucune manere grantier taillage.*
(PROME Ric II Nov. 1381)

‘They do not dare nor wish to grant tallage in any way’

- (16) *Et ceux qi vorront bailler lour peticions les baillent avant devaunt dymenge proschein venant; et apres mesme le dymenge ne soit aucune petition resceuz.* (PROME Ric II January 1380)

‘And those who wish to submit their petitions should hand them in before next Sunday; and no petition will be accepted after the same Sunday.’

- (17) *Et pur tant qe le dit Philip’ n’y dona aucun responce effectuel in celle partie, est agarde qe...* (PROME Ric II January 1380)

‘And because the said Philip gave no effective reply to this, it is decided that ...’

The use of *ne ... aucun* in negative clauses here follows what was to become the norm in French later.

However, the spoken register dialogues in the yearbooks made no use whatsoever of *aucun* in negative contexts throughout the 14th century, preferring the exclusive use of *nul* instead. In other words, the spoken register was in this instance showing conservatism as compared with written register texts, confirming the observation of Buridant (2000) but with the specific addition that the variation was a matter, not of the identity of the originators of the texts (legal professionals in both cases), nor of the content domain (legal rights in both cases), but of whether the communication was oral or written.

3. Conclusion

The value of having textual sources which clearly differentiate registers when examining the historical development of languages is illustrated by the analyses we have provided above. It has not been thought necessary to provide frequency counts of the items: the significant factor in each case is the entire absence of the form or the use in question from one or other corpus, allowing the case for register differentiation to be clearly made. Naturally a language does not always rigidly demarcate differing registers semantically or discursively. But from the investigation conducted here it has been shown that in some cases innovations began in the written medium, whereas in others they are associated initially with the spoken medium. The texts we have used do not perfectly represent the spoken/written contrast to the extent that, materially, they have all come down to us in written form. Yet the very sharp contrasts between them in terms of the variables studied have made it plain that their origin in either written or spoken form has significantly affected the kind of language used.

Although the opportunity afforded by these corpora to target register variation rather specifically, whilst holding content domain and time period constant, is certainly unusual for the pre-modern era, it is to be hoped that further investigation will reveal similar ways of studying language using historical corpora that will likewise offer fruitful perspectives for researchers to pursue.

References³

Primary sources

- BFM: Base de Français Médiéval. Lyon: ENS (laboratoire ICAR). <http://bfm.ens-lyon.fr/>.
- MCVF: Martineau, France et al. (eds.) (2010): Corpus MCVF, Modéliser le changement: les voies du français. University of Ottawa. www.voies.uottawa.ca/corpus_pg_fr.html.
- NCA : Stein, Achim (ed.) (2007): Le Nouveau Corpus d'Amsterdam. Stuttgart University. www.uni-stuttgart.de/lingrom/stein/corpus/#nca.
- PROME: Given-Wilson, Chris et al. (eds.) (2005): Parliament Rolls of Medieval England.
- YB: Ingham, Richard/Larrivée, Pierre (eds.) (2010): Narrations et dialogues: the Anglo-Norman Yearbook corpus. Birmingham City University, UK / University of Caen, France.

Secondary sources

- Archer, Dawn (2005): Historical sociopragmatics: questions and answers in the English Courtroom (1640-1760). (= Pragmatics and Beyond New Series). Amsterdam/Philadelphia: John Benjamins.
- Buridant, Claude (2000): Nouvelle Grammaire de l'ancien Français. Paris: SEDES.
- Curzan, Anne (2009): Historical corpus linguistics and evidence of language change. In: Kytö, Merja/Lüdeling, Anke (eds.): Corpus linguistics: An International Handbook. Berlin: de Gruyter, 1091-1109.
- Dodd, Gwilym (2009): Justice and Grace: private petitioning and the English Parliament in the late Middle Ages. Oxford: OUP.
- Huber, Magnus (2007): The Old Bailey Proceedings, 1674-1834. Evaluating and annotating a corpus of 18th- and 19th-century spoken English. In: Meurman-Solin, Anneli/Nurmi, Arja (eds.): Annotating variation and change. (= Studies in Variation, Contacts and Change in English 1). www.helsinki.fi/varieng/series/volumes/01/huber/.
- Ingham, Richard (2007): Bilingualism and language education in medieval England. World Universities Network, October 2007. www.wun.ac.uk/multilingualism/seminar_archive/07_08_program/ingham.html (URL no longer available).
- Ingham, Richard (2008): The grammar of later medieval French: an initial exploration of the Anglo-Norman Dictionary textbase. In: Guillot, Céline/Heiden, Serge/Lavrentiev, Akexei/Marchello-Nizia, Christiane (eds.): Corpus 7. Constitution et

³ Unless stated otherwise, all URLs have been checked and found valid as of late January 2015.

- exploitation des corpus d'ancien et de moyen français, 115-134. <http://corpus.revues.org/1506>.
- Ingham, Richard (2011): Grammar change in Anglo-Norman and continental French: the replacement of non-assertive indefinite *nul* by *aucun*. In: *Diachronica* 28(4): 441-467.
- Ingham, Richard (2012a): Anglo-Norman and the 'plural history' of French: the connectives *pourtant* and *à cause que*. In: *Revue française de linguistique appliquée* 16: 107-119.
- Ingham, Richard (2012b): *The transmission of Anglo-Norman*. Amsterdam: John Benjamins.
- Kytö, Merja/Culpeper, Jonathan (2006): *Corpus of Early Modern English dialogues (1560-1760)*. University of Uppsala. www.engelska.uu.se/corpus.html (URL no longer available).
- Pope, Mildred (1934): *From Latin to Modern French; with especial consideration of Anglo-Norman*. Manchester: Manchester University Press.
- Rothwell, William (2001): English and French in England after 1362. In: *English Studies* 82: 539-559.
- Short, Ian (2009): *L'Anglo-normand au siècle de Chaucer: un regain de statistiques*. In Thiolier-Méjean, Suzanne (ed.): *Le plurilinguisme au Moyen Age: Orient/ Occident*. Paris: L'Harmattan, 67-77.
- Traugott, Elisabeth/Dasher, Richard (2005): *Regularity in semantic change*. Cambridge: Cambridge University Press.
- Trotter, David (2003): Not as eccentric as it looks: Anglo-French and French French. In: *Forum for Modern Language Studies* 39: 427-438.

A historical linguistics corpus of Portuguese (16th-19th centuries)

Abstract

The aim of the present paper is to present and discuss a work in progress that involves:

- the creation of online editions of historical documents of a metalinguistic nature, which function both as publications and corpora, allowing for the comparison of manuscript images with the diplomatic edition and providing tools for analysis;
- the application and development of tools that can easily be manipulated by users and adapted to different kinds of historical texts.

The project is still in its first phase, which involves inventorying the metalinguistic texts held by the Évora Public Library (BPE). A survey of the texts of this nature identified in the various catalogues of the library has been carried out. Until now, 43 manuscripts and 200 printed texts with metalinguistic interest, all coming from the reserved catalogues of the BPE, have been identified. In the old reading room catalogue, further 313 works were also identified, while the modern catalogue is yet to be studied. As soon as the inventory is concluded, this will be followed by the organization and the online publication of a catalogue identifying and describing (bibliographical description) the works of a metalinguistic nature held by BPE. The texts' digital processing shall begin after these previous tasks have been completed.

1. Introduction

The Historical Linguistics Corpus (HLC) dealt with in the present paper is a markedly interdisciplinary work, promoting links between linguistics, history and literature on the one hand, and IT on the other hand. It seeks to make available, in an accessible, usable format, a significant number of historical documents of metalinguistic interest, creating tools for the user, and enabling the success of future developments, namely the extension of HLC to cover other works of a similar nature and also works of a different nature.

Providing an online corpus of (meta)linguistic texts of Portuguese between the sixteenth and the nineteenth centuries held by the Évora Public Library, HLC seeks to promote the acquisition of knowledge of some of the most important metalinguistic sources of Portuguese and foster their study, thereby contribut-

ing towards creating online resources in the field and thus advances in research into the language and its history at a national and international level.

These goals are of great importance in view of the current state of knowledge. Although there is a long written tradition of linguistic research in the field of Portuguese, one of the most widely spoken languages in the world, which enjoyed global status even before the advent of the phenomenon of globalization, not enough data or resources have been produced or made available. As a result of this work, we foresee a breakthrough in terms of the acquisition of knowledge on the metalinguistic memory of Portuguese, merging the philological tradition with technical innovation in methodological terms and making a wide range of material, which has been ignored because it is unpublished or rare, available for the first time to researchers and the general public.

The importance of and interest in the dissemination of these texts is bound up with the fact that Portuguese linguistic heritage lies in an indeterminate number of manuscripts and printed texts that often languish undiscovered in libraries and archives. There is a need for the inventorying, systematic processing and publication of such documents. It is known that much of the memory of the Portuguese language has yet to be established precisely due to the absence of corpora that make different text types representing different eras available to researchers all over the world in easily accessible formats. Despite the difficulties inherent in the construction of textual corpora, there is a great need to begin this task, since it is of crucial importance for the survey and analysis of the sources that advances in the study of the linguistic memory be made.

The choice of documents of metalinguistic nature is supported firstly by the fact that among the few corpora which exist for Portuguese there are very few that include texts of this nature. Moreover, the metalinguistic texts for Portuguese from the sixteenth century are recognized as sources with a dual interest for the history of the Portuguese language since, in addition to describing a certain historical state of the language, they are also an example of this state, in this way acting simultaneously as primary and secondary sources.

The proposal to publish documents held by the Évora Public Library is linked with the geographical proximity of the researchers, teachers at the University of Évora, and their role in providing a service to the local community. In addition, the Évora Public Library collection, in spite of its immense value, is difficult to access for researchers, so the benefits of this project go far beyond

the local community, and take on a level of importance at the national and international level.

2. The corpus

We aim to select, from the vast range of material available in the valuable collection held by the library, the works that are considered to be of the greatest value in terms of the “linguistic memory”, many of which are unique or rare, and almost all of them poorly catalogued. The project work will first involve the preliminary cataloguing of documents and works with metalinguistic interest, then texts will be selected for publication from among those that are unpublished or whose publications are incomplete and/or not easily accessible.

The criteria used for selection of the texts will be, in addition to the interest they arouse as (meta)linguistic documents, their rarity and/or the fact that they are not available in other corpora, as well as their state of preservation. Below we list only a few of the titles, manuscripts and printed editions that have already been identified and from which will be selected the texts that constitute the HLC, without prejudice to its enlargement in future developmental work:

Manuscripts:

- Apontamentos de orthographia.
- Arte da gramatica e orthographia portugueza, distinta da latina e qualquer outra língua. Dedicada ao real collegio das artes (1600?).
- Castro, P. João Baptista de: Aparato para a Rhetorica, ou Homem Rhetorico.
- Freire, Francisco José (1768): Reflexões Sobre a Lingua Portugueza, Escritas por Francisco Joze Freire da Congregaçam do Oratorio de Lisboa em 1768. (It should be noted that although the 1842 edition is already available in the online BN version, Freire’s original manuscript is in the Évora Public Library.)
- Lima, José dos Santos Baptista e (1740?): Conclusões grammaticaes, dedicadas ao Príncipe D. José por ... Professor em Macau.
- Novo methodo de grammatica portugueza, composto e offerecido ao Exm^a Sr. D. Thomás de Almeida, director Geral dos estudos, etc., por João Pi-

nehiro Freire da Cunha, professor de grammatica Latina, n'esta Corte, e natural da mesma.

- Observações do Dr. Pedro José Esteves á Orthographia Portugueza.
- Regras da orthographia portugueza.
- Vocabulario da Letra A.

Printed editions:

- Caldas Aulete, Francisco Júlio (1870): Grammatica Nacional Elementar, adoptada pelo Conselho Geral de Instrucção Publica, Additada com os elementos da língua Concani por J. M. Dias, conforme 3ª ed., Orlim, Na Typographia da India Portugueza.
- Cunha, João Pinheiro Freire da (1770): Breve tratado da orthographia para os que não os estudos ou diálogos ... Lisboa.
- Espada, João Chrysostomo Vallejo (1861): Grammatica portugueza, Lisboa.
- Fonseca, Roque da (1869): Compendio da Orthographia da Lingua Portuguesa, 2ª ed. Correcta e Augmentada com a Orthographia de princípios e varias notas, Margão, Na Typographia do Ultramar.
- Gouveia, J. F. De (1867): Noções Geraes e Elementares de Grammatica Portugueza, Adaptada na Escola Portugueza de Baretos em Cavel, Bombaim, Impressa na Typ. de Viegas & Son.
- Leal, Bento de Araújo (1734): Miscellanea gramatical. Na qual se explicam as partes da oração com todas as suas etymologias, e circumstancias (...), Lisboa, Off. Pedro Ferreira.
- Macedo, José de (com o pseud. de António de Melo da Fonseca): Antidoto da Lingua Portugueza, offerecido ao mui poderoso rei D. João V, Nosso senhor, Amsterdam, em Casa de Miguel Dias (sem indicação do ano, porém a dedicatória é de 1710).
- Pereira, Bento (1655): Florilegio dos modos de fallar e adagios da lingoa portuguesa (...), Lisboa, Por Paulo Craesbeeck.
- Pereira, Pe. José Filipe (1865): Compendio da Grammatica Elementar da Lingua Portuguesa por Systema Philosophico, para uso dos Alumnos das Escholas de Ensino Primario, Orlim, Na Typ. da Ind. Portugueza.

These texts will initially be processed in a conventional manner and will be read and transcribed (in terms of a diplomatic rendering), accurately repro-

ducing the evidence available and preserving all their relevant features (errors, omissions, spelling, word boundaries, abbreviations, etc.). In subsequent stages of the project, the texts will be digitally processed and made available online.

3. Document processing

The digital processing of documents will include the creation and use of resources and tools for natural language processing in order to obtain:

- a document text in an ASCII-like format to enable content analysis;
- electronic dictionaries that can be associated with the documents due to their specific vocabulary;
- the tagging of the ASCII text documents with part-of-speech (POS) markers. These markers enable linguistic researchers to look for word categories in their document analysis;
- the tagging of the ASCII text documents with named entities. These markers can help researchers to look for named entities across the text;
- the tagging of the ASCII documents with a view to sentence polarity. Using sentiment analysis techniques, the sentences of the documents are marked in order to enable researchers to search for sentences where the author's opinion is positive or negative with respect to a particular subject.

4. Content analysis

The process begins with the application of a text recognition system. Each book is carefully scanned. Then, for each page, the system performs a segmentation of the text areas to be analyzed, typically corresponding to paragraphs. The images corresponding to these areas are converted into text, which can subsequently be treated by natural language processing tools.

The conversion of each image to the text it contains is based on Optical Character Recognition (OCR). We have chosen the analytic OCR approach, trying to identify the individual graphemes, and then make the best interpretation of their sequence. For this interpretation, the system uses a dictionary of terms that are contemporary with the time of each book. As in previous works (Boschetti et al. 2009), our system searches for the best results by combining

the output of more than one OCR tool, such as OCRopus,¹ tesseract,² or Abby FineReader.³

The process is semi-automatic, as successfully performed by other projects—such as PaRADIIT⁴ (Ramel/Sidere 2011). The automatic recognition is complemented with human intervention to correct persistent errors. Let's consider an example is the extract below from a book (Freire 1842), with the original being shown at the top of Figure 1 and the recognition result below.

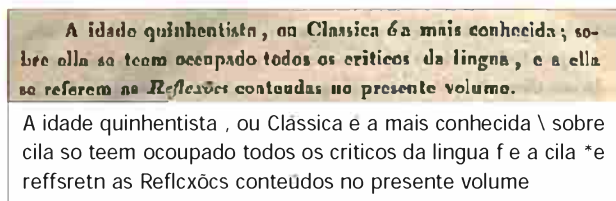


Figure 1: Text image and its automatic recognition

One of the mistakes concerns the word *ella* (feminine pronoun, third person, singular, corresponding to English *she*), appearing as *cila* in the OCRed text. Due to the poor image quality, some characters are difficult to recognize. Furthermore, the system dictionary does not contain all the words of the time in question, as is this case with *ella*. In our approach, a supporting dictionary is conceived for each time span, open to receive new terms that are encountered throughout the process of text recognition. In this semi-automatic process, a linguist marks the term *ella* as a valid word for the language of the time, and subsequent occurrences will be correctly identified.

5. Resources and tools

5.1 Electronic dictionary

The electronic dictionary will contain all the terms occurring in the HLC, as well as an ontology featuring the names of the entities mentioned in the texts and their classification.

¹ <http://code.google.com/p/ocropus/>. All URLs have been checked and found valid as of late January 2015.

² <http://code.google.com/p/tesseract-ocr/>.

³ <http://finereader.abby.com/>.

⁴ <https://sites.google.com/site/paradiitproject/>.

The dictionary will primarily be used to support the semi-automatic task of OCR as outlined above. A linguist assigned to this task will accept the dictionary suggestions or add new terms to the dictionary. The interface of the dictionary is integrated with the OCR tools.

The dictionary should contain the morpho-syntactic categories for each word in the corpus as well as other important information such as the current orthography of the word and the list of corpus documents where it occurs.

To give an example: *ella* is 'she' in Old Portuguese orthography; today it is written *ela*. The dictionary entry will look as follows:

Ella --- pronoun, feminine, 3rd person, singular, current: *ela*, used: <list of documents>

Elle --- pronoun, masculine, 3rd person, singular, current: *ele*, used: <list of documents>

Later the electronic dictionary will be made available to linguistic researchers both as an object of study and as a tool to support queries in the corpus documents.

The dictionary structure and implementation is an important issue (Hockey 2000, esp. p. 146-171). We use the WordNet structure in our electronic dictionary (Tasovac 2009).

5.2 Part-of-speech tagging

The documents tagged with part-of-speech markers can be used by linguistic researchers for (meta)linguistic analyses such as counting the use of the definite article before a possessive pronoun. This sort of analysis is important to infer the evolution of language phenomena (e.g.: *seus nomes* > *os seus nomes*).

The part-of-speech markers include a large set derived from the Brown corpus currently used for English, enlarged with Portuguese-specific categories which for some analyses can be grouped together into more basic categories such as prepositions, nouns, verbs, etc.

The part-of-speech tagging is a semi-automatic process supervised by a linguist that has to perform tasks such as deciding on the word category markers suggested by the POS-tagger and correcting the tagging of the documents. Since we do not have a training corpus, we use an unsupervised POS-tagger

that is able to infer groups of word categories (Collobert et al. 2011). These groups of word categories correspond to their role in the sentences; e.g., it will group all pronouns in a group and the linguistic supervisor must determine the groups and decide whether to include or exclude certain words in the groups.

5.3 Named entity recognition

This natural language task will give rise to a set of markers in the HLC documents where each name is tagged according to the category it pertains to, such as geographic place, person, or institution. These markers can help researchers, e.g., to count an author's citations throughout a given document in order to infer the impact of that author as a well accepted authority.

The named entity recognition process uses some mixed techniques including machine learning as well as part-of-speech markers and syntactic information.

5.4 Sentiment analysis

Sentiment analysis is a natural language task that uses the results of the part-of-speech tagger and some machine learning techniques in order to infer the topics of the sentences in the document and their polarity. This feature enables the researchers to search for an author's opinions on certain topics. The (meta)linguistic analysis of the author's recommendation for using or not using a given construction can be facilitated with this feature.

6. Conclusions

Besides making a considerable number of meta-linguistic texts dating from the 16th to the 19th century available online, the project aims at the development of tools (which can be made available to others later), beyond the usage of existing tools that allow for the manipulation of data from written texts from several historical periods. This service is really useful for linguists because it facilitates the constitution of specialized corpora and the location of certain words or syntactical structures within them, which is useful for statistical purposes, for example in the study of the diachronic evolution of a given linguistic phenomenon.

References

- Boschetti, Federico/Romanello, Matteo/Babeu, Alison/Bamman, David/Crane, Gregory (2009): Improving OCR accuracy for classical critical editions. In: Proceedings of the 13th European conference on Research and advanced technology for digital libraries (ECDL'09). Berlin/Heidelberg: Springer, 156-167.
- Collobert, Ronan/Weston, Jason/Bottou, Léon/Karlen, Michael/Kavukcuoglu, Koray/Kuksa, Pavel (2011): Natural language processing (almost) from scratch. In: Journal of Machine Learning Research 12, 2461-2505.
- Freire, Francisco José (1842): Reflexões sobre a Língua Portuguesa, escriptas por Francisco José Freire, publicada com algumas anotações pela Sociedade Propagadora dos Conhecimentos Uteis. Lisboa: Typographia da Sociedade Propagadora de Conhecimentos Uteis. <http://purl.pt/135>.
- Hockey, Susan M. (2000): Electronic texts in the humanities: principles and practice. Oxford/New York: Oxford University Press.
- Ramel, Jean-Yves/Sidere, Nicolas (2011): Interactive indexation and transcription of historical printed books. Presentation given at: Digital Humanities 2011 (DH2011): June 19-22, Stanford University (USA).
- Tasovac, Toma (2009): More or less than a dictionary: WordNet as a model for Serbian L2 lexicography. In: Infoteka – Journal of Informatics and Librarianship 10: 13a-22a.

Testing the validity of translation universals for Brazilian Portuguese by employing comparable corpora and NLP techniques

Abstract

The present study investigates the lexical features of three of the so-called translation universals, namely, *simplification*, *explicitation* and *levelling out*, for the Brazilian Portuguese language. The aim was to test their validity and to see whether the features change over time. To this end, comparable corpora were used containing nine pairs of specimens of translated and original Brazilian Portuguese narrative prose (fiction) dating from the 19th, 20th, and 21st centuries. The methodology consisted in the employment of NLP techniques and statistical measures as a complementary method of analysis. The results obtained confirmed only the validity of the *simplification* hypothesis.

1. Introduction

The quest for typical patterns of translated texts is not a new concern among scholars. The birth of the polysystem theory in the late 1970s revolutionized research on translation by shifting the emphasis of the investigations from a prescriptive approach to a more descriptive one. Gellerstam (1986) introduced the term *translationese*, which refers to the specific language of translations. Later on, influenced by the polysystem concepts, Toury (1995) put forward “laws” or “norms” of translations such as the law of standardization and interference. In 1996, Baker elaborated the four so-called translation universals, namely *explicitation*, *simplification*, *normalisation* and *levelling out*. According to Baker (1996), translation universals can be defined as hypotheses on some features that must be present in all translated texts regardless of the source and target languages, the type of translator, and the textual genre.

The present study aims to examine the lexical manifestation of three translation universals, namely *simplification*, *explicitation*, and *levelling out* as described by Baker (1996). *Simplification* postulates that translated texts are simpler, easier to understand; *explicitation* claims that translators tend to make statements more explicit rather than leave them implicit; *levelling out* hypothesizes that translated texts might be more similar to each other in comparison to original texts in the same language. Baker (1996) proposed the use of corpus linguistics techniques and tools to investigate these hypothetical features in

order to understand what really happens during the process of translation as well as to know the distinctive features of translated texts.

In order to test the validity of the translation universals, this study uses comparable corpora of Brazilian Portuguese composed of specimens from narrative prose (fiction) dating from the 19th, 20th, and 21st centuries. Some authors have discussed the changes that occur in translations over time and relate them to the literary “norms” of each period (Basnett/Lefevere 1990). Therefore, the purpose of this study is twofold: (i) to check whether the lexical features of the *simplification*, *explicitation* and *levelling out* hypotheses are in some way manifested in the translated texts, and (ii) to analyse whether these features change from one time period to another. To this end, three experiments were performed by employing NLP techniques as well as statistical measures for the analysis.

Even though the methodology adopted here had already been employed before to study the universals in other languages (e.g. Corpas Pastor 2008; Corpas Pastor et al. 2008a, 2008b; Laviosa 1997, 1998) the present study is a novelty because (i) it consists in the investigation of the lexical features of three universals at the same time; (ii) it is a first attempt to investigate the universals for the Brazilian Portuguese language and (iii) because it adopts a diachronic approach.

Knowing the distinctive features of translated texts can help in translators’ training (McEnery/Xiao 2007) and, from the NLP perspective, this knowledge can be used to enhance the output of NLP applications such as Machine Translation systems (Corpas Pastor et al. 2008b). In the following sections, we will describe work that is related to the present study, the structure of the corpora used for comparison, our experiments, and provide a discussion of our results and conclusions.

2. Related work

Initial research on translation universals focused on small-sized parallel corpora and the examination of shifts between original and translated texts (see Gellerstam 1986; Øverås 1998; Helgegren 2005; Santos 1995, 1997). Later on, investigations started to use comparable corpora in order to find patterns that differentiate translated from original texts in the same language (Laviosa 1997, 1998, 2002; Frankenberg-Garcia 2009).

The necessity to establish a more rigorous methodological status to the investigations led to the use of large amounts of textual data, believed to be more representative of a given language, together with robust NLP techniques to search for universal patterns of translated texts (see Corpas Pastor 2008; Corpas Pastor et al. 2008a, 2008b; Ilisei et al. 2010, Cheong 2006). Currently, investigations on translation universals are also adopting machine learning methods (see, e.g., Ilisei et al. 2010). The findings of some of these studies (Corpas Pastor et al. 2008b, Ilisei et al. 2010, Helgegren 2005) did validate the *simplification* hypothesis by showing that translated texts exhibit lower lexical density when compared to original texts. On the other hand, the convergence (*levelling out*) hypothesis was not confirmed by other studies (Corpas Pastor 2008, Corpas Pastor et al. 2008a). As regards the *explicitation* hypothesis, some studies (e.g. Cheong 2006) showed that both translation contraction and translation expansion co-occur throughout translated texts even though translation expansion occurs more often. These investigations also added other hypotheses to the list of translation universals such as the *transfer* and *implicitation* hypotheses which postulate, respectively, that translators tend to transfer syntactic and lexical features from the source language to the target language and that translated texts tend to leave concepts from the source texts implicit rather than explicit in order to adapt the message into a different language system (Corpas Pastor et al. 2008a, Cheong 2006).

Although translation universals have been investigated for different languages, studies on universals for the Portuguese language are still in their infancy, specifically with regard to the Brazilian Portuguese variety.

3. Resource description

The corpora we have established for the comparison are composed of two collections of texts, namely, translations from American and British English into Brazilian Portuguese and texts originally written by Brazilian authors. The two groups of texts comprise 18 prose narratives (fictional). All translations were carried out by professional translators and the original Brazilian texts were written by highly reputed authors. The texts are comparable in terms of:

- genre: all texts belong to the same genre and sub-genre. They are all literary narratives (novels), originally written in Brazilian Portuguese or translated into Brazilian Portuguese;

- size: the translations and the original texts used for comparison are roughly the same size;
- time span: the pairs of novels were published in the same period. The majority of the pairs were published in the same year for the first time; only two pairs differ by one year between the dates of their first publication, and one pair, by two years;
- type of translator: all translations were performed by professional translators;
- audience: all the texts were written for male and female intellectual adults as their target audience.

Table 1 displays the size of each text in terms of tokens.

Year	Original	Year	Translation
1875	59,648	1875	61,689
1880	45,704	1880	47,321
1882	57,960	1882	53,186
1943	49,326	1943	52,010
1966	106,845	1966	109,761
1984	69,251	1985	67,370
2003	149,534	2003	151,592
2005	40,903	2006	39,317
2008	50,378	2006	45,924
Total word-count	629,549	Total word-count	628,170

Table 1: Corpus size in tokens

4. Hypotheses to be checked

4.1 Simplification

The *simplification* universal postulates that translated texts are simpler, easier to understand. They contain simplified language compared to original texts. Low lexical density (type-token ratio) is an indicative feature of this universal (Baker 1996, Corpas Pastor 2008, Corpas Pastor et al. 2008b). Low lexical den-

sity means a less varied vocabulary (more repetition of words). Thus, it is expected that the *simplification* universal is lexically manifested in all translated texts in terms of low lexical density in comparison to the comparable original texts.

4.2 Explicitation

The *explicitation* universal can be defined as the tendency for translated texts to “spell things out rather than leave them implicit” (Baker 1996: 180). The lexical manifestation of this universal can be verified by the “use and overuse of explanatory vocabulary” (ibid., 181). To investigate this feature, Baker suggests to compare the frequency of explanatory words using corpora of original and translated texts in the same language and same domain. Thus, we expect to encounter more explanatory vocabulary (i.e. more words such as *because*, *therefore*, *in the sense that*, *consequently*, and so on) within the set of translated texts from all centuries than within the set of original texts.

4.3 Levelling out

Levelling out is defined as the tendency for translated texts to be more similar to each other in terms of lexical density (type-token ratio) and sentence length in comparison to original texts (Baker 1996). To verify whether this universal is present in the translated corpus, we propose to examine it by comparing similarities (in these cases differences) of translated texts concerning the lexical density. We expect to encounter smaller differences within a set of translated texts in comparison to greater differences within the set of original texts in terms of lexical density. All translated texts, from all three centuries studied, should consistently present a lower lexical density variance, this being indicative of their higher homogeneity in comparison to a higher heterogeneity within the set of original texts, which corresponds to a greater variance value.

5. Our experiments

In order to verify the *simplification*, *explicitation*, and *levelling out* universals, this study tested the validity of the universals by comparing the lexical features of translated and original in terms of lexical density, frequency of explanatory vocabulary, and lexical density variance.

5.1 Lexical density

The lexical density was extracted from the corpora using the *Word List* tool from *Wordsmith Tools*¹ software, which provides statistical data of texts selected for analysis. The texts were split into sections and the type-token ratio was calculated for each section of each text. In order to compare the lexical density between the two collections of texts, we calculated the overall average of the individual results obtained for the type-token ratio from the sections of each group of texts, and the partial averages from each time period (Table 3). The unpaired two tailed t-test was also used to calculate the statistical differences between the translated vs. the original texts in terms of lexical density (Table 4). The calculation was performed for each time period in order to compare the p-values obtained.

5.2 Frequency of explanatory vocabulary

To compare the frequency of explanatory words, we built our own list of explanatory vocabulary of Brazilian Portuguese language (Table 2). A program was written using the *Python* programming language in order to extract the frequency of those words from both collections of texts. Then we calculated the ratio of total occurrences of explanatory words for each text. The frequency of the explanatory vocabulary was compared by calculating the overall average of the ratios for the translated and original collections and the average of the ratios for each time span.

5.3 Lexical density variance

The homogeneity of the translated texts versus the heterogeneity of original texts in terms of lexical features was examined by measuring the lexical density variance of the two collections of texts. The results obtained were compared, whereby time span differences were taken into consideration (Table 7).

¹ www.oup.com/elt/catalogue/guidance_articles/ws_form?cc=global (URL no longer available).

Explanatory vocabulary in Portuguese	Translation
<i>a razão pela qual</i>	<i>the reason why</i>
<i>assim</i>	<i>hence, thus</i>
<i>dado que</i>	<i>given that</i>
<i>de modo que,</i>	<i>given that</i>
<i>devido a</i>	<i>due to</i>
<i>já que</i>	<i>since, once</i>
<i>na medida em que</i>	<i>as</i>
<i>no sentido em que</i>	<i>in the sense that</i>
<i>pela simples razão de</i>	<i>for the simple reason that</i>
<i>pelo fato</i>	<i>by the fact</i>
<i>pois</i>	<i>because</i>
<i>por causa de</i>	<i>because of</i>
<i>por isso</i>	<i>therefore</i>
<i>por motivo de</i>	<i>for the reason</i>
<i>porque</i>	<i>because</i>
<i>portanto</i>	<i>therefore</i>
<i>posto que</i>	<i>given that</i>
<i>sendo assim</i>	<i>hence, thus</i>
<i>sendo que</i>	<i>since, because</i>
<i>uma vez que</i>	<i>since, because</i>
<i>visto que</i>	<i>since, because</i>

Table 2: List of explanatory vocabulary

6. Results

6.1 Simplification

The investigation of the *simplification* hypothesis compared two groups of results obtained from the sections of the translated and original texts of the three centuries under concern. The assumption was that the lexical density of the

translated texts would present smaller values in comparison to the values obtained from the comparable original texts, regardless of the time period.

Table 3 displays the lexical density average calculated for sections of the texts from each century. The results reveal that all translated texts present a smaller average for the type-token ratio in comparison to the average obtained from the comparable original texts, i.e., all translated texts from all centuries present a lower rating result in relation to the corresponding original texts.

Table 3 also reveals that the lexical density average of the texts from the 19th century is greater than the average of the texts from the 20th and 21st century. A greater value obtained for the type-token ratio of a given text implies that this text contains more varied vocabulary than a text that presents a smaller type-token ratio. More varied vocabulary means more complexity to the reader. Thus, the translations from the 19th century contain more varied vocabulary in comparison to the later time spans but less varied vocabulary in comparison to the comparable original texts.

To compare the statistical differences between the translated and original texts we applied the student’s t-test. The confidence level chosen for this study was 0.05. The p-values in Table 4 reveal that there is a statistically significant difference between the translated texts versus the original texts from all centuries. However, this difference is smaller in the 19th century texts than in the 20th and 21st century texts. Thus, the p-value reveals that the translations from the 19th century are statistically more similar to the comparable original texts in terms of lexical density than the translations from the 20th and 21st centuries.

Century	Original Texts	Translated Texts
19 th	41.43	39.28
20 th	44.96	33.73
21 st	41.869	31.55

Table 3: Lexical density average of sections

Century	p-values	Result
19 th	0.02	SD ²
20 th	0.00000000000000129	SD
21 st	0.000000000232	SD

Table 4: Lexical density: Translated texts versus original texts

² SD= Statistically different.

6.2 Explication

Table 5 shows the normalized individual results, which confirm that all translated texts from the 19th century contain, proportionally to their length, more explanatory vocabulary than the comparable original texts. However, for the 20th century this feature was not borne out by all texts. The translated text dating from 1943 presents a smaller amount of explanatory vocabulary in relation to its comparable original text, but the other two translated texts from the same century present a larger amount of explanatory words.

Among the translated texts from the 21st century, only one of three contains proportionally more explanatory vocabulary when compared to the comparable original text. The overall average of the translated texts from the 21st century reveals a lower value than the overall average obtained for the set of original texts from the same time span (Table 5). Therefore, the hypothesis formulated as the *explication* universal could only be confirmed for all texts from the 19th and 20th centuries if we consider the time spans individually (Table 6). However, the overall average of the results of all texts from all centuries is greater for the collection of translated texts, i.e., the overall rate reveals a greater proportion of explanatory vocabulary within the set of translated texts (Table 5).

Century	Original Text	Explanatory vocabulary proportion	Translated Text	Explanatory vocabulary proportion
19 th	1875	0.0045	1875	0.0092
	1881	0.0065	1881	0.0068
	1882	0.0054	1882	0.0097
20 th	1943	0.0066	1943	0.0053
	1966	0.0047	1966	0.0055
	1984	0.0035	1985	0.0051
21 st	2003	0.0074	2003	0.0052
	2005	0.0036	2006	0.0061
	2007	0.0066	2006	0.0053
Overall average		0.0054		0.0065

Table 5: Proportion of explanatory vocabulary per text

Century	Explanatory vocabulary proportion: Original texts	Explanatory vocabulary proportion: Translated texts
19 th	0.0054	0.0085
20 th	0.0049	0.0053
21 st	0.0058	0.0055

Table 6: Average of results per time span

6.3 Levelling out

The starting assumption for the validation of this universal was that the collection of translated texts will present consistently lower values of variance when compared to the variance values obtained for the original texts. Thus, the verification of this hypothesis required the application of a variance test.

Table 7 shows the results. The variance is lower within the set of translated texts dating from the 20th and 21st centuries if we compare it to the results obtained for the original texts from the same periods. However, if we compare the results of variance between translations and original texts from the 19th century, the translated texts present a greater variance. These results reveal that the translated texts dating from the 19th century are less similar to each other in terms of lexical density (vocabulary variety) than the comparable set of non-translated texts. Thus, the *levelling out* hypothesis cannot be confirmed since we did not find consistently lower variance values for the collection of translated texts regardless of the time span.

Century	Lexical density variance: Original texts	Lexical density variance: Translated texts
19 th	21.87	38.5
20 th	87.21	41.22
21 st	68.33	46.11

Table 7: Results for *levelling out*. Values obtained for lexical density variance

7. Discussion of results

The results obtained with the aim to validate the starting assumptions could not be confirmed for all universals here examined. However, the results reveal that the features of translated texts change from one time period to another. While the 20th and 21st centuries present similar results, the 19th century differs

from these periods in terms of lexical density average, proportion of explanatory vocabulary and lexical density variance. The results show that the 19th century possessed more varied vocabulary, more explanatory words and they are less similar to each other than the comparable original texts.

The *simplification* universal was the only universal investigated whose results revealed a predictive probability of its validity, given the consistency of the results for all the translated texts. The average of values obtained for all sections from all translated texts as well as the values obtained with the application of the t-test show, respectively, that translated texts present less varied vocabulary than original texts in the same language, and there is a statistically significant difference between the set of translated and original texts in terms of lexical density. However, the 19th century presents a greater lexical density average in comparison to the 20th and 21st centuries.

The results obtained for the investigation of the *explicitation* universal did not reveal consistency. The overall average of the results obtained from all translated texts indicates that these texts contain more explanatory vocabulary in comparison to original texts; however, the individual results do not present a greater amount of explanatory vocabulary for all translated texts in comparison to the comparable original texts. A consistently larger amount of explanatory vocabulary was found only within the set of texts from the 19th century. Hence, we argue that the manifestation of this universal can be neither validated nor discarded. Actually, the investigation of this universal deserves more attention in future work, especially on the basis of a larger amount of data.

The similarity within a set of translated texts *versus* the dissimilarity within the set of comparable original texts could not be confirmed at all. The results obtained for the heterogeneity test revealed inconsistency. This situation implies that not all sets of translated texts present lower variance for the lexical density parameter, i.e. not all translated texts under analysis are similar to each other when compared to original texts. Lower variance values in comparison with greater values obtained for the original texts could be found within the set of translated texts dating from the 20th and 21st centuries, but a higher value was found within the set of translated texts dating from the 19th century. Therefore, we can say that the universal *levelling out* could not be verified for the lexical density parameter in the corpora under analysis even though the results obtained for the translated texts dating from the 20th and 21st centuries are lower. This universal would be confirmed if the heterogeneity test had presented re-

sults consistently lower for all texts from all centuries for the translated collection in comparison to the original collection.

Regarding the variance value obtained for the texts from the 19th century, a possible explanation for this higher level of heterogeneity within the texts from this period might be related to different strategies adopted by the translators.

The t-test revealed differences between the two groups of texts, too. The p-values obtained for the 20th and 21st centuries shows that the translated texts are statistically very different from original texts in terms of lexical features. However, the p-value obtained for the 19th century reveals a smaller statistical difference when compared to the p-values from the other two centuries. Hence, the variance test and the t-test proved that the texts from the 19th century exhibit different and peculiar lexical features when compared with the texts from the 20th and 21st centuries, due to a higher value obtained for variance and a smaller p-value obtained using the t-test. These results mean that they are statistically different from the original texts in terms of lexical features, but in a smaller proportion, and that they are less similar to each other than the translated texts from the 20th and 21st centuries. The explanation of these phenomena will probably be found in future research in which a corpus alignment would help to understand and differentiate the strategies adopted by translators from one period to another.

8. Conclusion

The validity of the universals examined in the present study was confirmed only for the *simplification* hypothesis. As regards the *explicitation* and *levelling out*, in general, the results obtained do not provide enough support to confirm the validity of these universals. These findings are consistent with previous studies on translation universals. However, the corpora used in the experiments presented here are limited by the restricted number of texts and by the genre. Thus, the validity was tested but the results obtained could not provide conclusive answers to the hypotheses elaborated with respect to the manifestation of the universals in translated Brazilian Portuguese. Therefore, there are many other possibilities of investigation remaining for this language. The present study aimed to be a starting point for investigations into the translation universals for Brazilian Portuguese, but improvements as to the methodology used as well as studies using more textual genres, other source languages and other universal features remain important desiderata.

Acknowledgements

Research for this paper was funded by a European union grant, under the *Erasmus Mundus Programme*, within the International Master in Natural Language Processing and Human Language Technology. I offer my gratitude to Ph.D. student Iustina Ilisei for invaluable assistance in bringing this study to its final form. Special thanks go to professor Jorge Baptista from the University of Algarve and professor Pieter Seuren from the Max Planck Institute for Psycholinguistics in Nijmegen for having contributed to the realization of the present study.

References

- Baker, Mona (1996): Corpus-based translation studies: the challenges that lie ahead. In: Somers, Harald (ed.) (1996): Terminology, LSP and translation: studies in language engineering. In honour of Juan C. Sage. Amsterdam/Philadelphia: John Benjamins, 175-186.
- Bassnett, Susan/Lefevere, André (1990): Translation, history and culture. London/New York: Pinter.
- Cheong, Ho-Jeong (2006): Target text contraction in English-into-Korean translations: a contraction of presumed translation universals? In: *Meta. Translator's Journal* 51, 2: 343-367.
- Corpas Pastor, Gloria (2008): Investigar con Corpus en Traducción: Los Retos de un Nuevo Paradigma. Frankfurt a. M. etc.: Peter Lang.
- Corpas Pastor, Gloria/Mitkov, Ruslan/Afzal, Naveed/Garcia, Moya Lisette (2008a): Translation universals: do they exist? A corpus-based and NLP approach to convergence. In: Proceedings of the LREC'2008 Workshop on Building and Using "Comparable Corpora", LREC-08, Marrakesh, Marroco.
- Corpas Pastor, Gloria/Mitkov, Ruslan/Afzal, Naveed/Pekar, Viktor (2008b): Translation Universals: Do they exist? A corpus-based NLP study of convergence and simplification. In: Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas (ATMA), Waikiki, Hawaii, 21-25.
- Frankenberg-Garcia, Ana (2009): Are translations longer than source texts? A corpus-based study of explicitation. In: Beeby, Alison/Rodríguez, Patricia/Sánchez-Gijón, Pilar (eds.): Corpus use and translating. Amsterdam/Philadelphia: John Benjamins, 47-58.
- Gellerstam, Martin (1986): Translationese in Swedish novels translated from English. In: Wollin, Lars/Lindquist, Hans (eds.): Translation studies in Scandinavia. Lund: CWK Gleerup, 88-95.

- Helgegren, Sofia (2005): Tracing translation universals and translator development by word aligning a Harry Potter corpus. Master's thesis, University of Linköping.
- Ilisei, Iustina/Inkpen, Diana/Corpas Pastor, Gloria/Mitkov, Ruslan (2010): "Identification of Translationese: a machine learning approach". In: Gelbukh, Alexander (ed.): CICLing 2010, LNCS 6008. Heidelberg: Springer, 503-511.
- Laviosa, Sara (1997): Investigating Simplification in an English Comparable Corpus of Newspaper Articles. In: Klaudy, Kinga/Kohn, Janos (eds.): *Transfere Necesse Est. Proceedings of the 2nd International Conference on current trends in Studies of Translation and Interpreting 5-7 September 1996*, Budapest, Hungary. Budapest: Scholastica, 531-540.
- Laviosa, Sara (1998): Core patterns of lexical use in a comparable corpus of English narrative Prose. In: *The corpus-based approach. Special issue of Meta*. Montreal: Les Presses de L'Universite de Montreal, 557-570.
- Laviosa, Sara (2002) *Corpus-based translation studies. Theory, findings, applications*. Amsterdam/New York: Rodopi.
- McEney, Anthony/Xiao, Zhonghua (2007): Parallel and comparable corpora: What is happening? In: Anderman, Gunilla/Rogers, Margaret (eds.): *Incorporating corpora: the linguist and the translator*. Clevedon, UK: Multilingual Matters, 18-31.
- Øverås, Linn (1998): In search of the third code: an investigation of norms in literary translation. Cand. phil. degree thesis, University of Oslo.
- Santos, Diana (1995): On grammatical Translationese. In: *Short papers presented at the Tenth Scandinavian Conference on Computational Linguistics*, Helsinki, 29-30th May 1995. Compiled by Kimmo Koskenniemi. Helsinki: University of Helsinki, 59-66.
- Santos, Diana (1997): O tradutês na literatura infantil publicada em Portugal. In: *Actas do XIII Encontro da Associação Portuguesa de Linguística, Vol. II* (Lisboa, 1-3 October 1997). Lisboa: Associação Portuguesa de Linguística, 259-274.
- Toury, Gideon (1995): *Descriptive translation studies and beyond*. Amsterdam/Philadelphia: John Benjamins.

Structuring a diachronic corpus

The Georgian National Corpus project

Abstract

The paper deals with the structural premises of diachronic corpora that are meant to represent specimens of a given language throughout its historical stages and to provide a diachronic cross-century retrieval. On the basis of the Georgian National Corpus project, it discusses ways to cope with variation caused by the use of different scripts and by language change, as well as requirements of annotating the different layers (chronological, dialectal, sociolectal, genre-based, etc.) the text materials pertain to, including a critique of the concepts of the ISO 639-6 standard.

1. Introduction

Corpora that are designed to embrace a given language throughout its historical stages and to provide diachronic access to its features present special challenges as to their structuring. Among these challenges, we may mention the problem of linguistic variation with all its facets, including phonetic change and its (ortho-)graphical representation, morphological, syntactical, and semantic change, but also the necessity of balancing between well and less well attested text genres. Until the present day, only a few projects have successfully attempted to establish corpora that cover a time-span of more than a few centuries. In the present paper, we discuss some of the peculiar requirements of a large-scale diachronic corpus on the basis of the Georgian National Corpus project,¹ which has to cope with most of the problems addressed above. After outlining the project and its background, the paper focusses first on the problem of the various scripts used for Georgian throughout its history and their handling, and second, on the question of how to annotate the linguistic varieties to be subsumed in the corpus with a view to differentiated retrieval.

¹ The initial spark of the project was the foundation of a coordination council in Tbilisi on July 19, 2011 (see http://geocorpus.blogspot.de/p/blog-page_21.html – all URLs quoted here were last checked on January 28, 2015). The project started, with kind support by the Volkswagen Foundation, in autumn, 2012.

2. The Georgian National Corpus project and its background

The plan to establish a Georgian National Corpus (hereafter: GNC) that covers the complete time range from the earliest attestations of written Georgian in the 5th century C.E. up to the present day has evolved from several corpus building initiatives that have been realized since the late 1980s, mostly in joint endeavours of German and Georgian partners. This is true, first of all, for the text database of the TITUS and ARMAZI projects in Frankfurt,² which covers nearly all published text materials from the periods of Old and Middle Georgian (roughly 5th-13th and 13th-18th cc.) as well as a minor collection of Modern Georgian texts (19th c.; mostly grammatical treatises and poetic works). These materials (ca. 6 Mio. tokens), most of which were electronically prepared since 1987 via OCR, with manual correction and formatting, have been thoroughly preindexed and are searchable via a word-form based retrieval system, which reflects the chronological order of the attestations in its output. For the time being, a lemmatization function has not yet been implemented; however, the retrieval engine provides a lexicon-based word analysis for nominal forms (cf. Figure 1).³

The second main pillar of the GNC is the GEKKO corpus run by Paul Meurer in Bergen / Norway,⁴ which has been compiled, mostly via data harvesting, from free online resources in Georgian, among them many newspapers and journals, but also literary texts (both autochthonous and translated) as well as pages from several official and semi-official websites in Georgia. The corpus thus established comprises ca. 125 Mio. tokens; about one fifth of it (20 Mio. tokens) has already been equipped with a full morphological annotation and a lemmatization function which includes the extremely complex verbal system of the Georgian language (cf. the sample output in Figure 2).⁵

A third pillar of the GNC is the extensive corpus of dialectal varieties of spoken modern Georgian ('Georgian Dialect Corpus', GDC) compiled under the direction of Marina Beridze at the Arnold Chikobava Institute of Linguistics

² Cf. <http://titus.uni-frankfurt.de/texte/texte2.htm#georgant> and <http://armazi.uni-frankfurt.de>.

³ For the example see <http://titus.fkidg1.uni-frankfurt.de/database/titusinx/titusinx.asp?LXLANG=517&LXWORD=uplisatws&LCPL=1&TCPL=1&C=A&T=0&LMT=100&K=0&MM=0&QF=1>; the search engine (see <http://titus.fkidg1.uni-frankfurt.de/search/query.htm>) accepts both Georgian and Romanized input.

⁴ Cf. <http://clarino.uib.no/gekko>.

⁵ For the example see <http://clarino.uib.no/gekko/simple-query> (subcorpus: Georgian – disambiguated (ქართული-დის.); query input: [lemma="ყოფნა"]).

in Tbilisi.⁶ This corpus, which also includes varieties of Georgian spoken outside of Georgia (in Iran, Turkey and Azerbaijan), has recently been made accessible for online retrieval via a word-based search engine (cf. the sample output in Figure 3).⁷

The GNC project further builds upon an extensive amount of recordings of spoken varieties of Georgian that were prepared within the project ‘The sociolinguistic situation of present-day Georgia’ (2005-2009).⁸ Many of these materials have been fully transcribed (ca. 1.5 Mio. tokens) and are, for the time being, accessible via the TITUS server,⁹ the Language Archive at the MPI Nijmegen,¹⁰ and the GDC project (cf. Figure 4).¹¹

The integration of all these data and functionalities, which is the main object of the two-year start phase of the GNC project begun in 2012, will bring together an unparalleled diachronic corpus extending over a time-span of about 1600 years and including chronological as well as dialectal and sociolectal variation.¹²

⁶ Cf. www.mygeorgia.ge/gdc/About.aspx and Beridze/Lortkipanidze/Nadaraia, this volume.

⁷ For the example see www.mygeorgia.ge/gdc/Default.aspx (query input: აზობს).

⁸ The project was kindly supported by the Volkswagen Foundation from 2005 to 2009.

⁹ Cf. <http://titus.fkidg1.uni-frankfurt.de/ssgg/ssgg.htm>.

¹⁰ Cf. http://corpus1.mpi.nl/ds/imdi_browser?openpath=MPI663243%23. To access the data users will have to register with the Language Archive; see http://dobes.mpi.nl/access_registration/ for instructions and http://corpus1.mpi.nl/ds/RRS_V1/RrsRegistration for the required form.

¹¹ For the example (recording in the Atcharian dialect by N. Surmava, 2006) see http://corpus1.mpi.nl/ds/imdi_browser?openpath=MPI696092%23.

¹² The project web site will be www.gnc.ge; for the time being cf. <http://armazi.uni-frankfurt.de/gnc/gnc.htm> and <http://clarino.uib.no/gnc>.

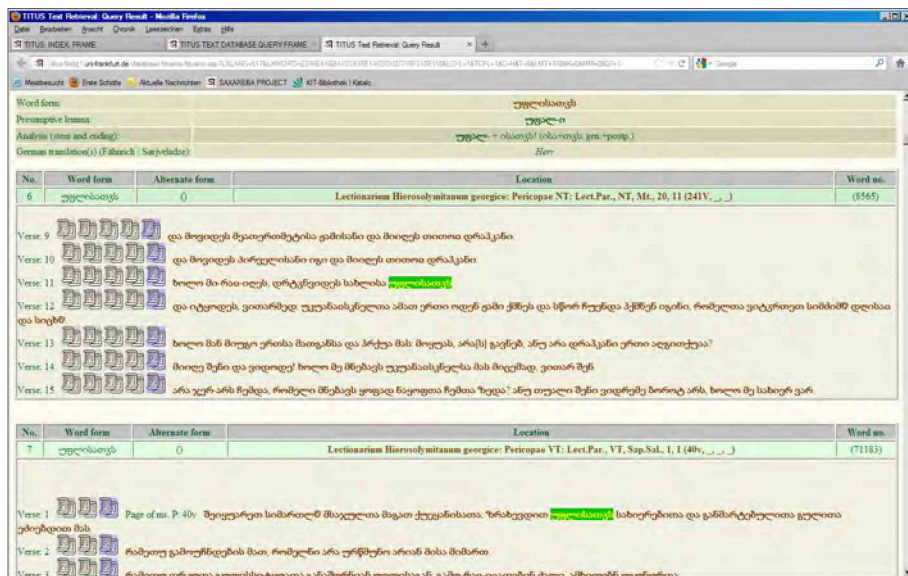


Figure 1: Query output of the TITUS search engine (uplisatw ‘for the Lord’; cf. footnote 3)

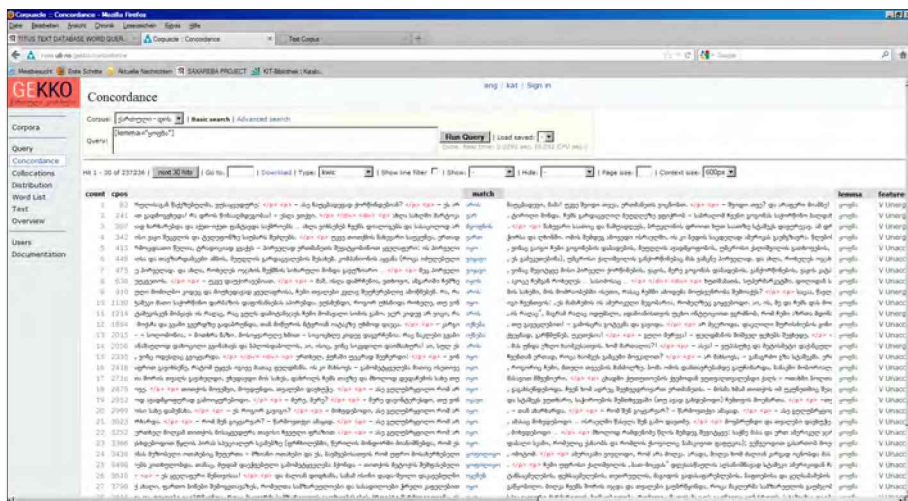


Figure 2: Search output of the GEKKO retrieval engine (lemma qopna ‘to be’; cf. footnote 5)

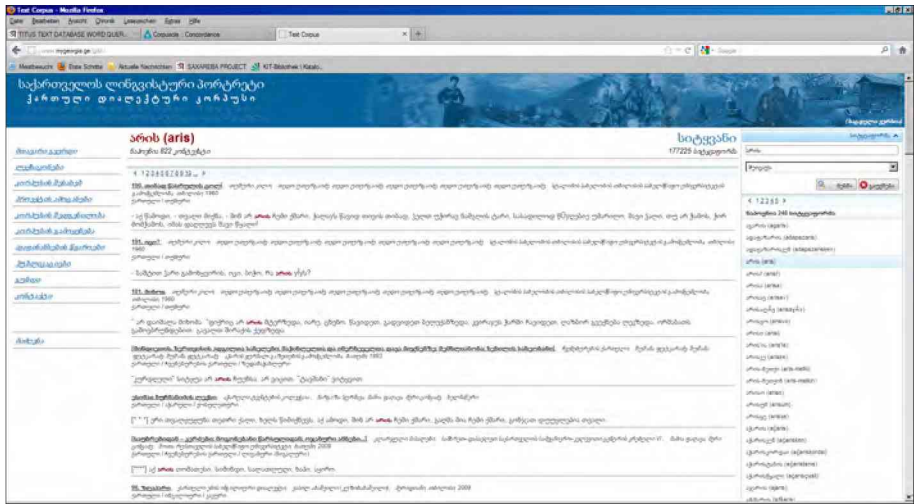


Figure 3: Search output of the GDC retrieval engine (word-form *aris* 'he/she/it is'; cf. footnote 7)

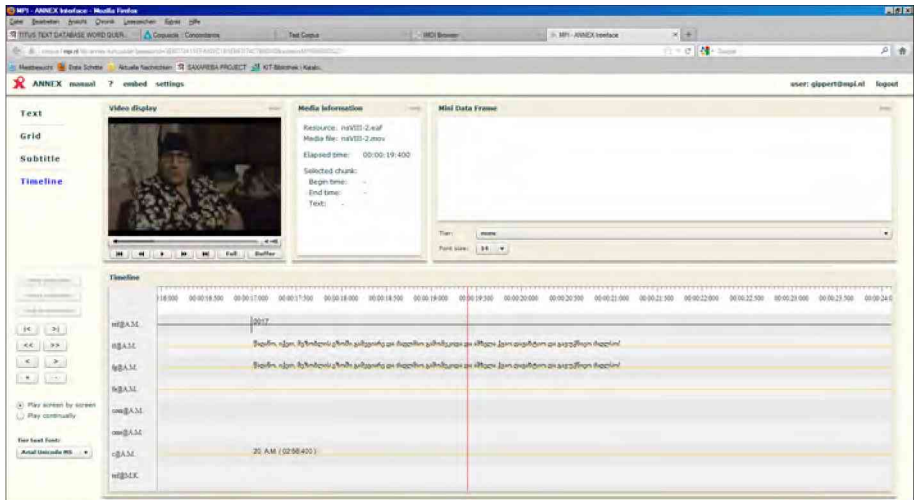


Figure 4: Dialect text from the SSG project (via the ANNEX interface of the Language Archive at the MPI Nijmegen; cf. footnote 11)

3. Scripts and encoding

As a matter of fact, Georgian is a near-to ideal showcase to develop and test a “true” diachronic corpus, even though it has changed much less than other languages since it was first written; consider, e.g., a common word-form of today like *gmadlob* ‘I thank you’ which has not changed at all since its first attestation in a palimpsest manuscript of about the 6th c. C.E.,¹³ in spite of the peculiar consonant clusters it contains. However, the literary history of Georgian was anything but homogeneous, the language having been written with three different scripts in the course of time: *Asomtavruli*, the Old Georgian majuscule script (ca. 5th-10th cc.), *Nusxa-Xucuri*, the minuscule script used in manuscripts of ecclesiastical content (ca. 9th-19th cc.), and *Mxedruli*, the mi-

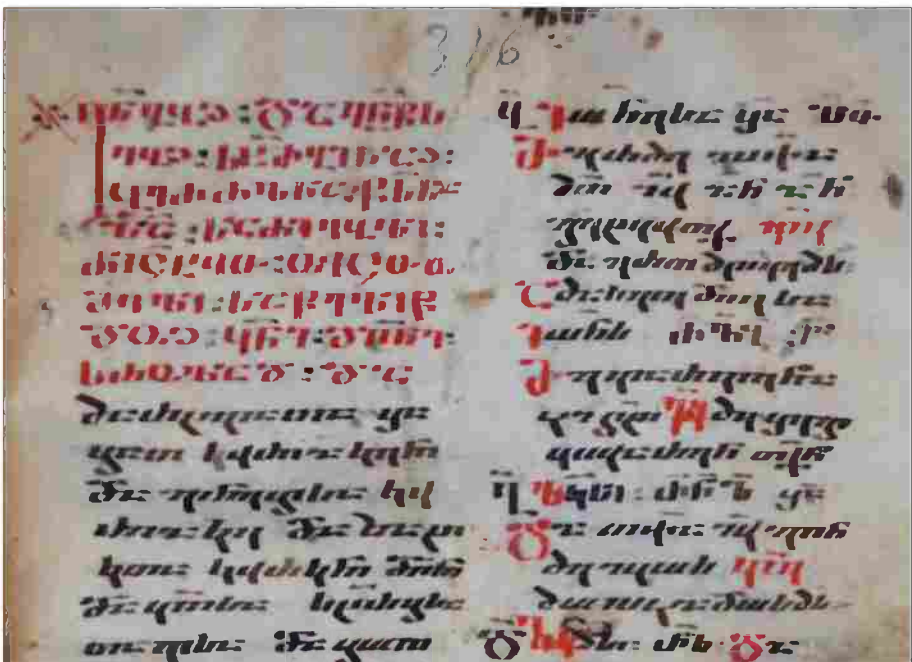


Figure 5: Old Georgian manuscript page (excerpt)

¹³ In the so-called *Khanmeti* version (cf. 3.1 below) of the legend of St. Christina, preserved in the Codex georgicus no. 2 of the Austrian National Library, Vienna; see [http://titus.fkidg1.uni-frankfurt.de/texte/etcc/cauc/ageo/xanmeti/vienna/vienn010.htm#Coll.Hag._Mart._Christin._11_8__022r-019v_22rb_1_\(1\)_186,19_f360,27](http://titus.fkidg1.uni-frankfurt.de/texte/etcc/cauc/ageo/xanmeti/vienna/vienn010.htm#Coll.Hag._Mart._Christin._11_8__022r-019v_22rb_1_(1)_186,19_f360,27) (users who are no members of the TITUS project will have to register via the form provided on <http://titus.uni-frankfurt.de/titusstd.htm>).



Figure 6: Modern Georgian press page (excerpt)

nuscule cursive used since the Middle Ages until the present day (cf. Figure 5 showing an Old Georgian manuscript written partly in *Asomtavruli* – in red ink – and partly in *Nusxa-Xucuri* script,¹⁴ and Figure 6 showing a modern print in *Mxedruli* script).¹⁵

All three Georgian scripts have been assigned separate code points in the Unicode standard¹⁶ so that it is possible today to encode the text materials of all periods as they were written originally. Under these conditions, an integrative approach towards diachronic retrieval across scripts presupposes the establishment of equivalences, which is straightforward for most letters. In a few cases, however, there are systematical discrepancies.

¹⁴ Manuscript no. 16 of the (old) collection of Georgian manuscripts of St. Catherine's monastery on Mt. Sinai, fol. 316r (photograph J.G., 2009).

¹⁵ From an article by H. Kurdadze on the Georgian alphabet in the *Inflight Magazine of Georgian Airways* 5, 2006/7: 4.

¹⁶ *Asomtavruli*: U+10A0 – U+10C5; *Nuska-Khutsuri*: U+2D00 – U+2D25; *Mkhedruli*: U+10D0 – U+10F5; see www.unicode.org/charts/PDF/U10A0.pdf and www.unicode.org/charts/PDF/U2D00.pdf, resp.

3.1 The notation of *u* and *v*

One discrepancy is determined by the fact that the *Asomtavruli* script inherited a peculiarity from its model, the Greek alphabet of Hellenistic times, in that it had no character for the vowel [u], which was written with a digraph <ΟΥ> equivalent to Greek <OY> instead. This digraph developed to a single letter, <ჟ>, in the minuscule script and is still a single letter, <უ>, in the *Mxedruli* script of today. In the rendering of Old Georgian manuscripts, it has been usual practice for long to transcribe *Asomtavruli* into *Mxedruli*, and most scholarly editions of Old Georgian texts are printed in the modern script. The <ow> digraph is usually replaced by the single <უ> letter in these editions, albeit it could as well be represented by the corresponding digraph, <ოჟ>, in *Mxedruli* script. Thus, the sequence სოჟნელები და სხოვანი <sowlnelebi, da sxovani> ‘aromatic spices, and others’¹⁷ is usually transcribed as სულნელები და სხოვანი <sulnelebi, da sxvani>, not transliterated as სოჟნელები და სხოვანი <sowlnelebi, da sxovani>. This, now, is problematic in a diachronic perspective as both nouns have slightly changed meanwhile, *sulnelebi* having been replaced by *surnelebi* while *sxvani* ‘others’ is written with the <v> character, <ვ>, instead of <უ> today (სხვანი). It is true that the replacement of <ow> by <v> is near-to regular in the given environment (between a consonant and a vowel); however, there are cases where <u> has been maintained in the same constellation (e.g., in Modern Georgian ჭკუა <čkua> ‘intellect’) or, conversely, <v> was used in the same position already in Old Georgian (e.g., in ღვანი <kvani> ‘stone’) so that the application of an automatic substitution rule may fail. For the change leading from *sulnel-* to *surnel-*, there is no “automatic” rule at all as the dissimilation involved here is sporadic, not regular.¹⁸

¹⁷ From Lc. 24.1 in the so-called “Pre-Athonian” redaction of the Old Georgian NT (9th-10th cc.), first attested in the so-called “Graz Lectionary” (manuscript Gr. 2058/2 of the Graz University Library, fol. 8r), a Sinai codex of ca. the 8th c. mixing Khanmeti and Haemeti features; see http://titus.uni-frankfurt.de/texte/etcs/cauc/ageo/xanmeti/grlekt/grlek.htm?grlek017.htm#Gr._2058/2_8r_3_Lk_24_1.

¹⁸ In contrast to the regular dissimilation rule of Modern Georgian which changes a sequence of *r – r* into *r – l* as in the adjective formation suffix *-ur-* (see, e.g., *čex-ur-i* “Czech”) appearing as *-ul-* in *rusuli* “Russian” or *german-ul-i* “German”. Modern Georgian does admit of sequences of *l – l* as in *alubali* “cherry”. – The Old Georgian stem *sulnel-* can still be found used (as an obsolete form) in religious contexts today.

3.2 The notation of *wi*

The second element of the *Asomtavruli* digraph <QԿ>, the letter *vie*, <Կ>, is problematic in other contexts, too. As the descendant of Greek <Υ>, it usually stands for a diphthong-like [wi] sequence (resulting from or replacing Greek [ū]); the same is true for its *Nusxa-Xucuri* equivalent, <վ>. In such cases, modern transcriptions replace <Կ> either by *Mxedruli* <ვო>, i.e. <vi>, in accordance with the modern pronunciation, or by transliterative <ვ> = <w>, as in *ჟღღჟღღღ* <gwrwgwni> ‘crown’ rendered by either *გვირგვინი* <gvirgvini> or *გვრგვინი* <gwrwgwni>. Again, these replacements are not straightforward as they are not applicable when <Կ> or <վ> follows or precedes a vowel. What is more, the orthographic rules of Old Georgian manuscripts differ to a considerable extent in the use of the character. For instance, we often meet with the sequence [wi] being represented by the digraph <QԿ>, i.e. <ow>, instead of plain <Կ> = <w>, or [v] in post-vocal position being rendered by <Կ> = <w> or <QԿ> = <ow> instead of <Վ> = <v>; cf., e.g., *Asomtavruli* spellings like <xowdodit> instead of “normal” <xwdodit>, <itqows> instead of <itqws>, <simšowidita> instead of <simšwdita>, or <moaowlina> instead of <moavlina>, all appearing in the lower layer of the palimpsest pages of the Kurashi Gospel manuscript.¹⁹

3.3 Diplomatic rendering vs. diachronic retrieval

All these graphical discrepancies must be taken into account if the corpus is meant to reflect the manuscript heritage of Old (and Middle) Georgian as neatly as possible (in the sense of a “diplomatic” rendering of hand-written sources) and yet to provide diachronic access to its linguistic contents. To cope with these demands, it is desirable to envisage a multilevel annotation format that is able to store authentic spellings, period-conformant normalizations, and diachronic surrogates side by side. A similar approach has been worked out for the project “Referenzkorpus Altdeutsch” (cf. Figure 7), which is to be diachronically aligned with corpora of later stages of German to yield a diachronic corpus of all periods of German.²⁰ In such an annotation system, an Old Georgian spelling variant like *სქვანი* (= <sxwani>) should be stored as-is (i.e., in *Asomtavruli*) alongside its “normalized” Old Georgian equiva-

¹⁹ Cf. Gippert (2013: 113).

²⁰ Cf. www.deutschdiachrondigital.de for the project of a “sprachstufenübergreifendes tiefenannotiertes Korpus historischer Texte des Deutschen”.

lent, 𐌸𐌹𐌺𐌿𐌸𐌹𐌺𐌹 (= <sxowani>), as well as its “modern” adaptation, 𐌸𐌹𐌺𐌿𐌸𐌹𐌺𐌹 (= <sxvani>), and its lemmatic basis, the stem 𐌸𐌹𐌺𐌿- (= <sxva->), the latter representing the entry point for diachronic queries.²¹

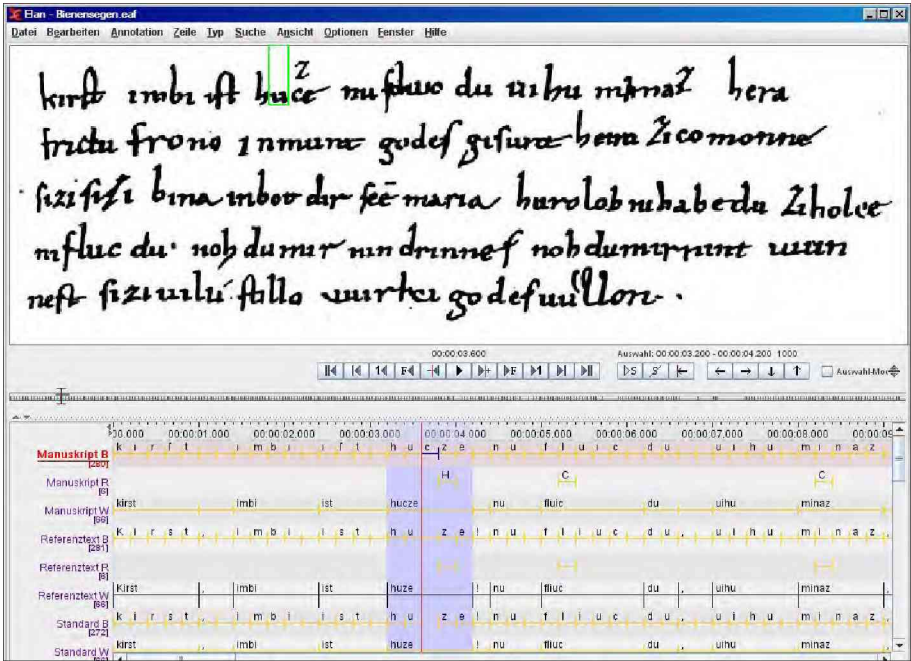


Figure 7: Old High German manuscript with multilevel annotation distinguishing diplomatically rendered and normalized spellings (from the “Referenzkorpus Altdeutsch” project)²²

4. The annotation of linguistic varieties

While unified access to linguistic elements across the history of a language is one fundamental task of a diachronic corpus, the differentiation of the individual varieties comprised in it is another one. For more detailed investigations into the historical diversification of a given language, it is necessary to distinguish the different layers the textual materials pertain to, especially

²¹ It is true that the plural form *sxvani* is rare today, the regular plural form being *sxvebi*. The so-called “old” plural forms require special treatment in the corpus.

²² The so-called “Lorscher Bienensegen” contained in the Vatican manuscript Cod.lat.Pal. 220, p. 58r; see http://titus.uni-frankfurt.de/texte/etcs/germ/ahd/klahddkm/klahd.htm?klahd091.htm#Kl.ahd.Dkm._Bienenseg._1_S396_._Vat.lat._220_58r_47.

with a view to lexicographic analyses. This concerns not only the successive chronological layers – in the given case Old, Middle and Modern Georgian – but also other layers that are distinguishable in the data, among them dialects, sociolects, and registers determined by text genres (i.e., “styles”) and communication modes (e.g., “spoken” vs. “handwritten” vs. “printed” vs. “electronic”, etc.). In the case of Georgian, this is crucial indeed, as the diversification of discernible layers begins as early as the Old Georgian period.

4.1 Layers of Old Georgian

As a matter of fact, a large set of layers must be distinguished for Old Georgian with respect to chronological, regional, and other properties. Chronologically, the set begins with the so-called *Khanmeti* and *Haemeti* varieties, which represent the earliest strata of Georgian literacy (ca. 5th-7th and 7th-8th cc.), with a “mixed” variety attested in the famous Graz lectionary,²³ and which are clearly distinguishable by peculiar morphological features. Within the subsequent period of “standard” or “classical” Old Georgian (ca. 9th-12th cc.), we may distinguish several locally-based varieties mostly established by Georgian writers in the monastic diaspora, on Mt. Sinai, Mt. Athos, or in Palestine, but also within Georgia as in the case of the “Gelati” school of the 11th-12th c. C.E.²⁴ Albeit most of the textual material of Old Georgian is religious, there are still some genre-specific peculiarities that force us to distinguish authentic from translated sources, and among them Biblical, hagiographical, homiletic, historiographical, philosophical, documentary, and other styles. A peculiar layer of Old Georgian is met with in documents that emerged later than the 12th c., in an attempt to maintain the religiously determined Old Georgian standard of literacy alongside the developing “Middle” Georgian vernacular which mostly manifested itself in secular literature; this layer, which had its impact up to the 18th c., may be called “Late Old Georgian”.

4.2 Layers of Middle and Modern Georgian

Different from Old Georgian, the Middle Georgian period was much less characterized by chronological or local differentiation. Instead, it was marked by greater genre-specific differences between, e.g., poetic, epic, historiographic, or documentary texts, manifesting themselves mostly in lexicographic fea-

²³ Cf. footnote 17 above.

²⁴ Cf. <http://armazi.uni-frankfurt.de/armaz1m.htm>.

tures (e.g., in an increasing impact of Persian) but also in the degree of grammatical conservativeness. Thus, the Old Georgian phenomenon of verbal tmesis (e.g. *mo-vinme-vida* ‘someone came’, lit. ‘hither-someone-went’, with the preverb *mo-* being split from the verbal root *-vid-* by the inserted indefinite pronoun *vinme* ‘somebody’) is still attested frequently in the 13th c. epics (prose and verse) but only exceptionally later. A similar distinction of genre-based registers is applicable to written Modern Georgian, too; here we would have to distinguish, right from the beginning, poetic and prose genres, the latter including belletristic literature as well as journalistic, juridical, scientific, or other styles. And of course, there were changes in the orthographical standards and the grammar within the period of Modern Georgian, too, which can roughly be divided into three subperiods in this respect, viz. Tzarist, Soviet, and contemporary.

4.3 Dialectal and sociolectal variation

As in other languages with a strong literary standard, dialectal and sociolectal variation comes into play mostly in spoken manifestations of Georgian. Roughly speaking, the Georgian dialects form two subgroups, a western and an eastern one, with Kartlian, the dialect of the central eastern part of the country and its capital, Tbilisi, being closest to the written language of today. Among noteworthy sociolects, we may mention that of the Georgian Jewry, which is characterized by a peculiar terminology (not necessarily of Hebrew origin) and a special intonation, or the argots of thieves or drug dealers, which have characteristic lexical features, too.

4.4 Annotation of layers

How, then, to account for all these divergent layers in a diachronic corpus of Georgian? Traditionally, the pertinence of a text to a given layer has been regarded as meta-information that is best stored in a (TEI) header. This, however, has a big disadvantage as it cannot account adequately for mixed texts such as, e.g., prose texts containing verse passages, journalistic texts containing quotations from argot speech, or even hagiographical texts containing quotations from the Bible or passages in foreign languages.²⁵ The annotation of information concerning chronological, dialectal, sociolectal and other layers would in

²⁵ A striking example of the latter is an Early New Persian sentence quoted, in Georgian script, in the “Life of St. Nino” (see Gippert 1992: 10).

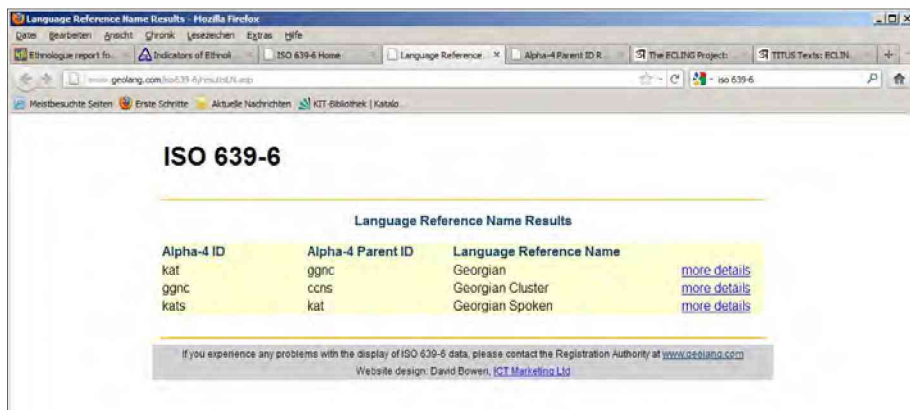
such cases better be stored word-by-word in order to facilitate layer-based queries and indexation. This can again be achieved via a multilevel annotation scheme; cf. Figure 8 illustrating this with another example from the “Referenzkorpus Altdeutsch” project where Old High German and Latin words are annotated accordingly using the respective three-letter codes of the ISO 639-3 standard²⁶ (“goh” = “German-Old-High” and “lat” = “Latin”).

Layer	00:00:28.000	00:00:29.000	00:00:30.000	00:00:31.000	00:00:32.000	00:00:33.000	00:00:34.000	00:00:36.000
Manuskript B (190)	d	e	m	o	v	e	l	l
Manuskript W (R1)	d	e	m	o	v	e	l	l
Referenztext B (190)	d	e	m	o	v	e	l	l
Referenztext R (R1)	d	e	m	o	v	e	l	l
Standard B (190)	d	e	m	o	v	e	l	l
Standard W (R1)	d	e	m	o	v	e	l	l
Lemma (R3)	der	fel	in	dese	tulli	ter	pater	noſter
Übersetzung (R1)						dreimal	Vater, Ahnherr, Schw	unser (angsh.-pte ve
Sprache (R3)	goh	goh	goh	goh	goh	lat	lat	lat

Figure 8: Old High German manuscript with multilevel annotation distinguishing Old High German (“goh”) and Latin (“lat”) words (from the “Referenzkorpus Altdeutsch” project)²⁷

²⁶ For the standard see the official site of the ISO 639-3 Registration Authority at www.sil.org/iso639-3/.

²⁷ The so-called “Wurmsegen” (no. 1) contained in the Tegernsee manuscript Clm 18524b, p. 203v of the Bayerische Staatsbibliothek, Munich; see http://titus.uni-frankfurt.de/texte/etcs/germ/ahd/klahddkm/klahd.htm?klahd077.htm#Kl.ahd.Dkm._Wurms.1_3_S374__Clm_18524b_203v_47.

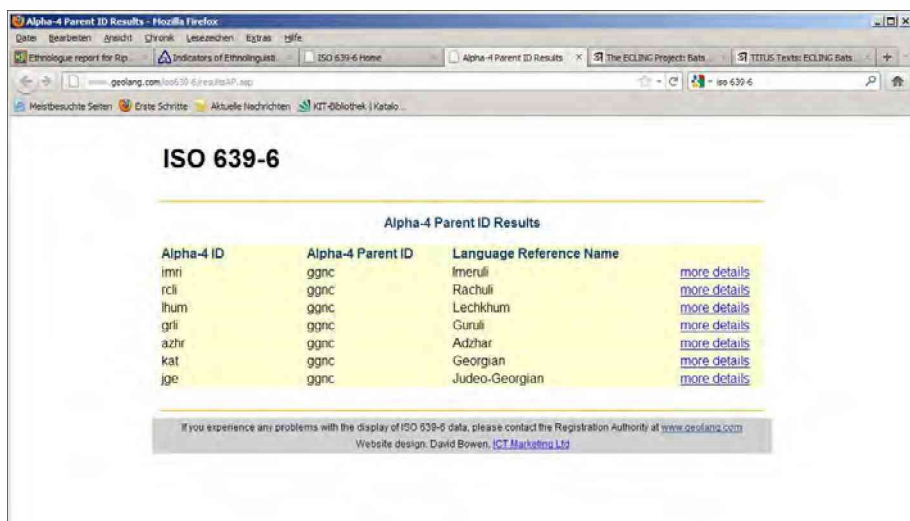


The screenshot shows a web browser window displaying the ISO 639-6 database results for the query 'Georgian'. The page title is 'ISO 639-6' and the sub-header is 'Language Reference Name Results'. The results are presented in a table with three columns: Alpha-4 ID, Alpha-4 Parent ID, and Language Reference Name. There are three rows of data, each with a 'more details' link.

Alpha-4 ID	Alpha-4 Parent ID	Language Reference Name	
kat	ggnc	Georgian	more details
ggnc	ccns	Georgian Cluster	more details
kats	kat	Georgian Spoken	more details

Below the table, there is a footer note: 'If you experience any problems with the display of ISO 639-6 data, please contact the Registration Authority at www.iso.org. Website design: David Bowen, ICT Marketing Ltd.'

Figure 9a: The ISO 639-6 database (query result for “Georgian”)



The screenshot shows a web browser window displaying the ISO 639-6 database results for the query 'GGNC'. The page title is 'ISO 639-6' and the sub-header is 'Alpha-4 Parent ID Results'. The results are presented in a table with three columns: Alpha-4 ID, Alpha-4 Parent ID, and Language Reference Name. There are eight rows of data, each with a 'more details' link.

Alpha-4 ID	Alpha-4 Parent ID	Language Reference Name	
imri	ggnc	Imeruli	more details
rcii	ggnc	Rachuli	more details
lhum	ggnc	Lechikhum	more details
grli	ggnc	Guruli	more details
azhr	ggnc	Adzhar	more details
kat	ggnc	Georgian	more details
jge	ggnc	Judeo-Georgian	more details

Below the table, there is a footer note: 'If you experience any problems with the display of ISO 639-6 data, please contact the Registration Authority at www.iso.org. Website design: David Bowen, ICT Marketing Ltd.'

Figure 9b: same (query result for subnode GGNC)

It goes without saying, however, that a three-letter-code of this type is in no way sufficient to cover the diversity of chronological, dialectal, and other layers we have to deal with in the GNC project, all the more since ISO 639-3 distinguishes nothing but “Georgian” (= “kat”, ← *kartuli*, the self-designation of the language) and “Old Georgian” (= “oge”). The reduced amount of possible codes in this standard ($2^3 = 17,576$ possible combinations of three basic letters) has recently led to the foundation of a successor standard, ISO 639-6,

Alpha-4 ID	Alpha-4 Parent ID	Language Reference Name	
katf	kats	Kharthuli-Formal	more details
kali	kats	Kharthuli	more details
khur	kats	Kakhuri	more details
igib	kats	Ingilo	more details
bash	kats	Tush	more details
khvr	kats	Khevsur	more details
mikhv	kats	Mokhev	more details
psav	kats	Pshav	more details
mtul	kats	Mtiul	more details
fjdn	kats	Ferejdan	more details

If you experience any problems with the display of ISO 639-6 data, please contact the Registration Authority at www.geolang.com
Website design: David Bowen, [ICT Marketing Ltd](http://www.ictmarketing.com)

Figure 9c: same (query result for subnode KATS)

which operates with four letters, yielding a total of $(26^4 =) 456,976$ new codes.²⁸ In its first stage of development, this standard comprised²⁹ a set of 20 codes related to Georgian and its varieties, arranged as parent-child relations in a tree-like structure. The picture thus achieved was anything but convincing, however, let alone sufficient for our purposes. First, there were no codes available concerning older stages of Georgian, not even the “oge” code of ISO 639-3, albeit the codes of this standard were declared to be maintained in the new proposal and “kat” for “Georgian” was still present (cf. Figure 9a showing the output of a query for “Georgian” in the database of the site that has been responsible for the registration since 2009).³⁰ Second, there was no differentiation in the codes as to dialectal and sociolectal layers; thus, “jge” for “Judeo-Georgian” (again taken over from ISO 639-3) was registered on the same level as, e.g., the Rachian dialect of western Georgia (“rcli” = “Rachuli”; cf. Figure 9b). Third, the tree structure remained enigmatic, given that nine Georgian dialects (plus “katf” = “Georgian formal”) were subsumed as children under

²⁸ Cf. www.iso.org/iso/catalogue_detail?csnumber=43380 for a rough outline and Gippert (2012: 21-23) for a preliminary account of the standard.

²⁹ The official Registration Authority for the standard was until 2014 the British company *GeoLang* (now *Ascema*; <http://www.geolang.com/>).

³⁰ The URL in question (www.geolang.com/iso639-6/) was still available on January 28, 2015, but not working properly. It seems that the process of further developing ISO 639-6 has been interrupted.

“kats” = “Georgian spoken” (cf. Figure 9c), whereas six other dialects (plus “jge” = “Judeo-Georgian”) were children of “ggnc” = “Georgian cluster”, in its turn the parent of “kat” and the grand-parent of “kats” (cf. the schematic illustration in Figure 10). The very fact that the nine first-mentioned dialects pertain to the eastern group and the six other ones, to the western group, is in no way a satisfactory explanation why only the former depended on “kats” = “Georgian spoken” (and, further up, “kat” = “Georgian”).

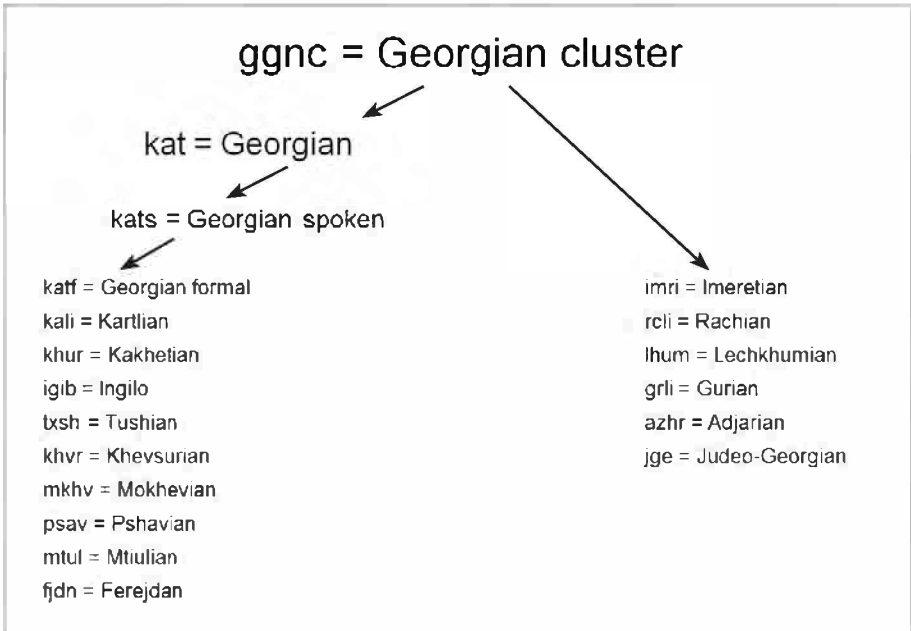


Figure 10: Dependencies of Georgian varieties in ISO 639-6

As a matter of fact, it is more than doubtful that the complex interrelationship between the chronological, dialectal, and other “lectal” layers of Georgian can at all be depicted adequately in a flat tree structure of the given sort. Instead, we should rather conceive this as a set of matrixes, among them one of spoken dialects interacting with sociolects, one of chronological layers interacting with genre-based variants (as illustrated in Table 1), and others.

	Old Georgian						Middle	Modern Georgian		
	Khanneti	Haemeti	Standard	Sinai	Athontic	Late	Middle	Early	Soviet	Contemporary
Biblical	×	×	×	×	×	×		×		×
Hagiographical	×		×	×	×	×		×		
Homiletic	×	×	×	×	×	×		×		×
Theological			×	×	×	×		×		×
Historical			×			×	×	×	×	×
Philosophical			×			×	×	×	×	×
Documentary			×				×	×	×	×
Epigraphic	×	×	×	×		×	×	×	×	×
Scientific			×	×		×	×	×	×	×
Journalistic								×	×	×
Belletristic							×	×	×	×
Poetical			×	×		×	×	×	×	×

Table 1: Matrix of chronological layers of written Georgian and text genres attested in them

In such a system of matrixes, the information that a given word belongs to a biographical text in a record of spoken Judeo-Georgian from Kutaisi in West Georgia would not be covered by a mere three- or four-letter-code such as “jge” but by a set of specifications (structured, e.g., as a sequence of codes for “language – chronological layer – mode – dialectal area – sociolect – genre”, in the given example quasi “Georgian – Modern – Spoken – Imeretian – Jewish – Biographical”).³¹ This concept would not only allow us to keep different types of “lects” apart but also to assign common “layers” (modes, genres, and, to a certain extent, sociolects) cross-linguistically. The development of a repertoire of annotation codes for these purposes, ideally to be standardized, is a

³¹ A similar approach has been outlined by the WordWideWeb consortium (www.w3.org/International/articles/language-tags/Overview.en.php), which proposes *language-extlang-script-region-variant-extension-privateuse* as a sequence of “types of subtag” (cf. also the discussion in www.rfc-editor.org/rfc/rfc5646.txt). This sequence has the shortcoming that there is no clear distinction between chronological, dialectal, sociolectal, and genre-specific layers, all to be covered by the “extended language” (*extlang*), *region*, and *variant* subtags.

task of high priority indeed. The GNC project will contribute to this in elaborating and schematizing the distinctions surfacing in the text materials it covers. This process will also be the basis for determining the necessary extensions of the corpus with a view to an optimal balancing between the text genres and “lects” reflected in them.

References

- Gippert, Jost (1992): Zum Status des Mittelpersischen im südlichen Kaukasus. Paper read on the conference “Bilingualism in Iranian Cultures”, Bamberg, July 1992. <http://titus.uni-frankfurt.de/personal/jg/pdf/jg1992b.pdf>.
- Gippert, Jost (2012): Language-specific encoding in endangered language corpora. In: Seifart, Frank/Haig, Geoffrey/Himmelmann, Nikolaus P./Jung, Dagmar/Margetts, Anna/Trilsbeek, Paul (eds.): Potentials of language documentation: methods, analyses, and utilization. Honolulu: University of Hawai’i Press, 17-24. <http://hdl.handle.net/10125/4512>.
- Gippert, Jost (2013): The Gospel Manuscript of Kurashi. A preliminary account. In: *Le Muséon* 126: 83-160.

The Georgian Dialect Corpus: problems and prospects

Abstract

The *Georgian Dialect Corpus* (GDC) covers a significant segment of the spoken language of Georgia. It is conceived as a sub-corpus of the *Georgian National Corpus*¹ and is designed for wide interdisciplinary research. Since 2006, the project has been funded by the *Shota Rustaveli National Science Foundation*.

With its structure, the GDC represents a wide spectrum of regional, temporal and stylistic variations of the Georgian linguistic reality. It contains texts from all Georgian dialects (including the dialects spread in *Iran, Turkey, and Azerbaijan*); intensive work on a corpus of *Laz* texts is underway.

Currently, we are working on the elaboration of a morphological annotation concept. In this process, the first step is lemmatization. While automatic lemmatization is an easily solvable and trivial problem in corpora of standard languages with exhaustive morphological descriptions, it is a rather difficult task in a dialect corpus containing a comprehensive collection of texts from up to twenty dialects. Therefore it is undertaken manually in most dialect corpora. In our concept, we effectively apply a lexicographical datapool and a standard language parser within a semi-automatic annotation process. The lemmatization process is then based on the standard form, dialect lemmata and standard lemmata being “deemed equal”. The implementation of this presupposes the manual lemmatization of a certain amount of dialect texts.

1. Introduction

The Georgian Dialect Corpus (GDC) was initiated as part of the comprehensive project *Linguistic Portrait of Georgia*. The team started working on the project in the late 20th century and, initially, it was mainly aimed at a large-scale computational documentation of Georgia’s linguistic diversity. To this end, we chose a corpus strategy as the most effective means for language documentation (see Beridze/Nadaraia 2011). The GDC was conceived to represent a wide range of regional, temporal and stylistic variations of the Georgian linguistic reality (including more than twenty dialects of three Kartvelian languages, historical varieties of Georgian, etc.).

¹ See Gippert/Tandashvili, this volume.

The project work has been conducted with the support of various foundations. Since 2006, the project has been funded by the *Shota Rustaveli National Science Foundation*.

2. The concept of GDC

Presently, two approaches are identifiable in the corpus representation of dialect data. One of them aims to build a fragmented corpus with general characteristics, thus being mostly illustrational and designed to give an impression about the diversity of the language under concern rather than to provide comprehensive linguistic knowledge. This approach has been sustained in the *Russian National Corpus* (see Letuchiy 2005 and Rakhilina 2009). According to the second approach, the dialect data are meant to become a new type of scholarly resource to represent and study not only the language but also the communicative patterns of the linguistic community in question (see Kryuchkova/Goldyn 2008). Since its inception, the idea of our corpus has been oriented towards the second approach.

In the GDC, each dialect (or language variety) is represented as an individual sub-corpus. This provides the opportunity to investigate linguistic phenomena and/or cultural objects both within the individual communicative space or a regional cultural area.

Presently, the GDC contains texts from all Georgian dialects (including those spread in Iran, Turkey, and Azerbaijan); intensive work on the processing of Laz texts are underway.

3. Representativeness and interdisciplinary accessibility of the corpus

In order to accomplish the main task of the corpus – to create a comprehensive model of the Georgian linguistic and cultural domain – it is necessary to solve the problem of representativeness in the corpus. Representativeness can be achieved if a maximum of dialect data is included in the corpus. Only predefined narrative data are not sufficient to represent the dialect vocabulary. We have attempted to solve this problem by integrating “non-textual” vocabulary into the corpus.

Thus, in order to achieve completeness of dialect vocabulary data in the GDC, the following tasks were addressed:

- the consideration of a maximum of text diversity (thematic and genre diversity; social diversity concerning age, sex, occupation, migration background, etc.);
- the integration of non-textual components into the corpus (lexicographic and encyclopedic components, data from scholarly studies, etc.).

In order to provide interdisciplinary applicability of the corpus, we had sometimes to “divert” from standards established for corpora of spoken languages. For instance, in the GDC, punctuation almost always follows the written norm, and hesitation, pauses, false starts etc. are not marked. We do not use a complete phonetic record either and avoid the marking of specific intonational and prosodic characteristics. It is true that information of this type is significant for a thorough linguistic analysis of a text, but we believe that a keen researcher can find it in the multimedia part of the corpus, the integration of which is part of our future plans. Our approach thus protects us from narrow disciplinary marginalization.

The structure of the GDC has been entirely determined by the fact that its technological background comprises a whole chain of text processing, beginning from the recording of the text data up to their integration in the database of the corpus.

4. The textual and lexical bases

The text database of the GDC consists of dialect texts that were recorded at various periods of time and preserved at various stages of philological processing. They comprise:

- published dialect texts;
- archival dialect records (assembled since the 1920s);
- audio recordings (assembled since the 1960s);
- new digital video recordings, collected by us during various expeditions (since 2005).

As already mentioned, in order to fully represent the vocabulary data, the corpus integrates lexicographic and encyclopedic information.

4.1 Lexicographic data in the corpus

We first converted the printed dictionaries of various Georgian dialects into a text database. The dictionaries are integrated into the corpus by means of a special editor. The editor is designed to extend the features of the dictionaries: it can add meanings, dictionary examples, comments, and references; besides, it provides the opportunity to mark the foreign origin of a given word. The dictionary included in the corpus is both accessible separately for the corpus user and, simultaneously, its content is represented as a word list in the corpus concordance.

Every entry is “distributed” in the following way in the corpus: the headwords of the printed dictionaries appear among the word forms in the word list. Besides, they are represented as text fragments in the corpus, so that each word form can be searched according to the concordance (see Beridze/Nadaraia 2009).

Besides the printed dialect dictionaries, the corpus integrates dictionary data that were collected by means of dialect questionnaires at various periods of time, as well as data attested in the scholarly literature.

4.2 Encyclopedic data in the corpus

For the GDC, we have obtained the data from a unique interdisciplinary expedition that was carried out in 1935. During a period of two months, the expedition recorded ethnographic data at large within Georgia and among ethnic Georgians in Azerbaijan (the *Saingilo* speaking province). Parts of the data were published later; however, linguistic features did not appear in the publication. The full set of data, preserved as manuscript copies at the Arn. Chikobava Institute of Linguistics, has been digitized (more than 25 volumes), and following a special processing (differentiation of scholarly comments and speakers’ texts, etc.), it will be gradually integrated into the corpus. The data in question are unique both in terms of thesaurus modeling concerning ethnographic realities, and of an appropriate reflection of encyclopedic discourse in the construction of a common communicative and cultural model of a language.

The integration of lexicographic and encyclopedic components into the corpus along with the textual data allows not only for a full representation of the linguocultural portrait of Georgia but also generates prospects to create new corpus-based cross-dialect dictionaries.

5. Meta-text annotation

In compiling meta-annotation features, three principal “sets” were identified: personal, geographic, and linguistic. In the corpus, the information sets appear as individual databases. The text editor allows keeping predefined dialect text as a base; the editor selects various kinds of information from the feature sets and attaches them to a given text.

The Georgian ethnosomal reality is represented in the GDC by the following databases (see Fig. 1):

- a list of administrative units (country, region, district, village/city);
- a list of names of nationalities;
- a list of names of ethnic origins (Georgian, Russian, Chinese ...);
- a list of provincial backgrounds (Svan, Gurian, Megrelian, Kakhertian, Imeretian ...);
- a list of languages and dialects.

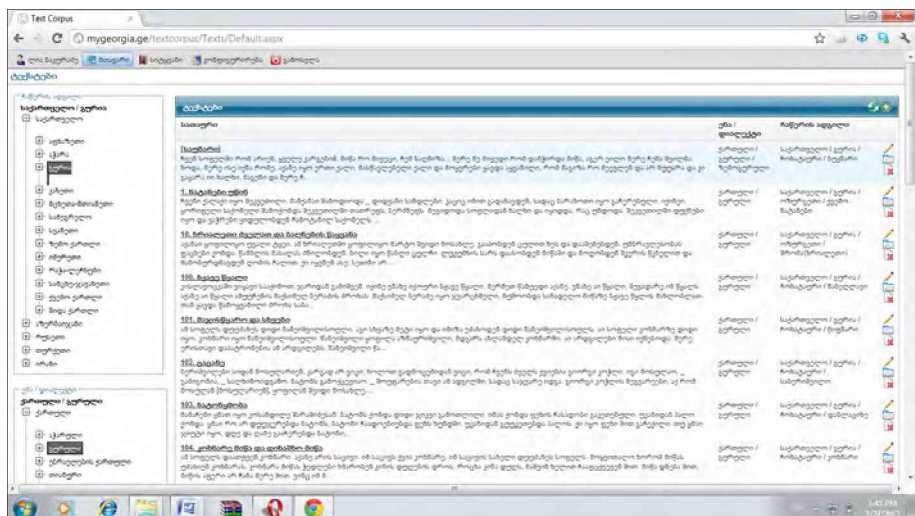


Figure 1: Text description according to language/dialect

The base of administrative units is activated at various stages while adding the texts to the corpus:

- 1) concerning the place where the text was recorded;
- 2) concerning the place(s) of birth and residence of a speaker (or his/her parents and spouse);
- 3) concerning the initial and terminal points of a speaker's (or his/her family's) migration.

In order to demonstrate the Georgian ethnosocial reality, the corpus also comprises the following features: *ethnic origin* and *provincial background*; for instance, *ethnic origin: Georgian / provincial background: Svan*; *ethnic origin: Georgian / provincial background: Imeretian*, etc. Linguistic features (language > dialect) transforms this picture into a “linguistic reality”, showing which of the three *Kartvelian* languages a text pertains to and how it is localized within the linguistic and cultural space of Georgia.

Furthermore, the corpus can be applied for interdisciplinary purposes due to the fact that the texts it contains can be “filtered” at various levels and according to various properties, among them thematic, chronological, typological, bibliographic data; information about the recorder of the text and the author of a scholarly publication, about the medium of the text (manuscript, audio, video ...), and other features.

6. Principles of morphological analysis

Currently, we are working on the elaboration of a morphological annotation concept. In this process, the first step is lemmatization. While automatic lemmatization is an easily solvable and trivial problem in corpora of standard languages with exhaustive morphological descriptions, it is a rather difficult task in a dialect corpus containing a comprehensive collection of texts from up to twenty dialects. Therefore it is undertaken manually in most dialect corpora.

In our concept, we effectively apply a lexicographical datapool and a standard language parser within a semi-automatic annotation process

6.1 Dialect dictionaries in the GDC as a means of partial morphological annotation

By means of integrating dictionaries into the corpus, we attempt not only to demonstrate, with a novel approach, the close link between dictionaries and text corpora, but also to use dictionaries that are compiled in accordance with

traditional lexicographic principles in the corpus building itself. For this purpose, we decided to use the “left side” of a dialect dictionary for partial lemmatization and part-of-speech annotation. This small experiment was rather successful and demonstrated that lexical items from dialect texts which coincide with a headword given in a dialect dictionary and which have an exact equivalent in the written standard may be described quite accurately by primary morphological markers.

6.2 Lemmatization and part-of-speech tagging in the GDC

In the GDC, the lemmatization process is based upon the written standard, dialect lemmata and standard lemmata being “deemed equal.” The implementation of this presupposes the manual lemmatization of a certain amount of dialect texts. Simultaneously, we work on the part-of-speech tagging.

Lemmatization and equalization with standard forms are undertaken in the following way (see Fig. 2):



Figure 2: Dialect word form equalization with a standard word form

We decided to use the lemmatic forms of the dialect dictionaries for partial lemmatization and part-of-speech markup in the following way:

- a) each dictionary entry (headword) is transferred to a database as a lexical entry and, simultaneously, as an identifier of lexemes in the form of dictionary entries attested in the text. All headwords attested in the texts that coincide with a lemma in the dictionary are automatically tagged as lemmas;
- b) in the lemmatization of a verb form, we indicate both the *masdar* and the future third person singular form as the headword. The same principle is

followed in the Explanatory Dictionary of the Georgian Language (Chikobava 1951) and most of the Georgian dialect dictionaries. Therefore, the activation of this function allows us to automatically identify verb forms in the future third person singular;

- c) along with information sets that reflect the lexicographic peculiarities of a given dictionary, the editor for adding dictionary data contains primary morphological annotation tags for nouns (substantive, adjective, numeral, pronoun), verbs, uninflected words (adverb, postposition, interjection, particle, conjunction). Thus, the tagged entry not only allows to identify forms that are attested in the text data but also provides a means for primary morphological tagging;
- d) in case a dialect form in the dictionary is explained by means of a simple equivalent from the written standard and the word is attested in the form of the dictionary entry in a text, we can automatically ascribe both its lemma and its standard equivalent.

The process thus allows for the lemmatization and partial morphological annotation of data that are attested in their lemmatic form in a given text; however, it does not facilitate the complete automatic lemmatization of the corpus yet.

6.3 A morphological processor for the Georgian standard language and the morphological annotation of the GDC

Our concept of a morphological annotation of the dialect corpus focuses on the application of a morphological processor of standard Georgian, with a view to enhancing it with additional “morphological knowledge” and, thus, to providing a means of semi-automatic identification of all tokens in the corpus (via lemmatization, surface and deep annotation).

The widening of the “knowledge base” of the morphological processor implies the addition of morpheme variants to the system of productive grammar rules.

For instance, in the analysis of preverbal forms of verbs with the purpose of automatic lemmatization, the processor must be provided with “knowledge” about all the phonetic variants of the preverbs evidenced in Georgian dialects. Besides, simple rules must be created which reflect the variations occurring in a dialect (i.e., rules of phonetic change).

Morphological annotation can then be carried out in several stages.

A. Tentative analysis. The following word forms are identified on the basis of the word list:

1. words shared with the standard language;
2. dialect words different from their equivalents in the standard language.

Words in A.1 can undergo full morphological annotation immediately, while those in A.2 can be annotated by applying affixation rules.

B. The list of forms (already POS-annotated) that are shared with the standard language are subdivided into two parts (see Figure 3):

1. words without homonymy;
2. words with homonymy.

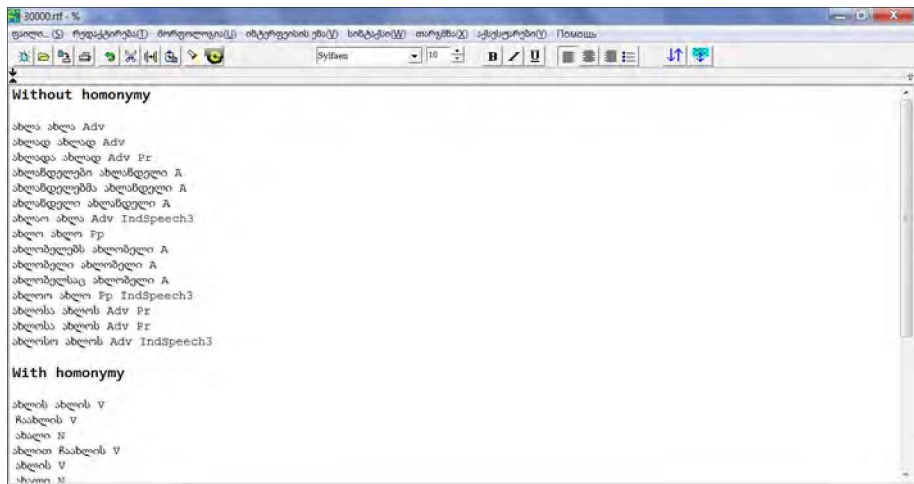


Figure 3: Two types of word forms shared with the standard language

C. Disambiguation.

Words with homonymy are listed separately (in B.2) because the system has no automated disambiguation component yet in this step. Hence, for each form, the complete set of part-of-speech markers is listed with the corresponding lemmas.

However, the system includes partially automatic word sense disambiguation. During the process, the rules for word sense disambiguation are added to the analyzer. For example, in some Georgian dialects, nouns with consonant-final stems frequently lose the case markers of the nominative and dative cases, thus appearing as pure stems. This ambiguity is excluded within other dialects. The corresponding rule is applied in order to correctly attach case markers to nouns appearing as stems in various dialects.

The advantage of our concept is that in the process of the widening and developing of the knowledge base of the morphological processor, professional linguists can be involved without acquiring any additional skills. Basing themselves upon linguistic issues, they are able to pose a specific problem to the processor and to solve it.

7. Outlook

The morphological annotation of the GDC is the most recent task of the project. It is not the only work that is undertaken within the framework of the project, but the lexical base of the GDC is being built in such a way that it covers all printed dictionaries and represents a platform for a new, integrated and universal dialect dictionary.

The GDC is a monitor corpus; new data are permanently added to the corpus. It is planned to create a new sub-corpus of the GDC in the nearest future. The sub-corpus will be aiming at the documentation of the speech of migration people from various regions of Georgia.

References²

Beridze, Marina/Nadaraia, David (2009): The Corpus of Georgian Dialects. In: Levičká, Jana/Garabík, Radovan (eds.): NLP, Corpus linguistics, Corpus Based Grammar Research. Fifth International Conference, Smolenice, Slovakia, 25-27 November 2009. Proceedings. Bratislava: Tribun 25-34. http://korpus.juls.savba.sk/~slovko/2009/Proceedings_Slovko_2009.pdf.

Beridze, Marina/Nadaraia, David (2011): Словарь как текстовый компонент корпуса (Корпус грузинских диалектов) [Dictionary as a textual component of a corpus (Georgian Dialect Corpus)]. In: Труды международной научной конференции «Корпусная лингвистика – 2011» [Proceedings of the International Scientific Conference “Corpus Linguistics – 2011”], St. Petersburg, 92-97 [in Russian]. http://corpora.phil.spbu.ru/Works2011/Беридзе_92.pdf.

² All URLs have been checked and found valid as of late January 2015.

- Chikobava, Arnold (1951): ქართული ენის განმარტებითი ლექსიკონი [Explanatory dictionary of the Georgian language]. 8 vols. Tbilisi: Georgian Academy of Sciences. [in Georgian].
- Kryuchkova, Olga/Goldyn, Valentin (2008): Текстовый диалектологический корпус как модель традиционной сельской коммуникации [Textual dialect corpuses as a model of traditional rural communication]. In: Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной международной конференции «Диалог». Периодическое издание / Computational linguistics and intellectual technologies. Papers from the Annual International Conference "Dialogue" 7(14), Moscow: Институт проблем информатики РАН, 268-273 [in Russian]. www.dialog-21.ru/digests/dialog2008/materials/html/41.htm.
- Letuchiy, Alexander (2005): Корпус диалектных текстов: задачи и проблемы [A corpus of dialectal texts: tasks and problems]. In: Национальный корпус русского языка: 2003-2005. Результаты и перспективы [National Corpus of the Russian Language: 2003-2005. Results and perspectives]. Moscow: Indrik, 215-232 [in Russian]. <http://ruscorpора.ru/sbornik2005/13letuchy.pdf>.
- Rakhilina, Ekaterina (2009): Корпус как творческий проект [Corpus as a creative project]. In: Национальный корпус русского языка: 2006-2008. Новые результаты и перспективы [National Corpus of the Russian Language: 2006-2008. New Results and Perspectives]. St. Petersburg: Nestor-Istoria, 7-26 [in Russian]. <http://ruscorpора.ru/sbornik2008/01.pdf>.

Integrating annotated ancient texts into databases

Technical remarks on a corpus of Indo-European languages tagged for information structure

Abstract

This paper shows, on the basis of corpora of Indo-European languages tagged for information structure, how technical restrictions of two freely available programs (EXMARALDA and ANNIS) influence the work with these corpora. It explains both the requirements of the used languages (Latin, Ancient Greek, Old Indic, Avestan, Hittite, and Luwian) and the technical restrictions posed by the XML editor concerning these requirements. Furthermore, it outlines workarounds to deal with the mentioned restrictions. Finally, it demonstrates how this affects the work with the corpora in the database. Additionally, a few remarks on handling certain syntactical aspects and example queries for the database are presented, which can be tested on samples available online.

1. Introduction

Within the scope of the projects ‘Information Structure in Older Indo-European Languages’ and ‘Information Structure in Complex Sentences – Synchronic and Diachronic’,¹ coherent texts from Latin, Ancient Greek, Old Indic, Avestan, Hittite, and Luwian are annotated. In order to give a broad perspective on the behavior of these languages concerning information structure and to explore diachronic developments, we analyze a variety of genres and authors from different periods.

This paper introduces the languages in question along with a few selected tiers of their annotation. At first, I will explain the choice of programs used. Subsequently, in chapter 3, the main part of this paper, I will outline the specific requirements of each annotated language on the basis of an example. The first point of each sub-item presents peculiarities of the language. Then restrictions of the programs and their influence on the graphic appearance of each language will be explained, which sometimes leads to deviation from the standard

¹ The projects were started at the Friedrich Schiller University Jena in 2009 and 2011 respectively and are conducted under the supervision of Prof. Dr. Rosemarie Lühr (Humboldt University Berlin).

look found in modern text editions. The following sub-items demonstrate, via the given examples, the special handling of certain elements, which also affects retrieval from the database. Those specifics are usually to be applied to all languages; exceptions will be stated. Hence, this paper is not concerned with the information structure of those texts but with basics concerning the layout and ways of data retrieval.

2. Choice of programs

Even though stand-off architecture is often used nowadays, it is inconvenient if the original text has to be altered, be it for, e.g., eradication of editing errors, separation of words, or insertion of elliptical forms.² It is also helpful to have a certain level of computational skills, which does not necessarily apply to the standard linguist. These are two of the reasons why inline annotation with a program containing a user-friendly surface was chosen. Further reasons are, since the main focus of the two projects is on information structure which also includes grammatical structure, stylistic devices, saliency etc., that it is essential to have an overview of the whole sentence and all levels of annotation. Additionally, the projects desire for their data to be compatible for broad usage and, thus, want to avoid reinventing the wheel.³ Therefore, we decided to use already existing programs that meet both these requirements and our demand for the tools to be versatile for numerous languages. The EXMARaLDA⁴ Partitur Editor is used to compile the data that are then integrated into the ANNIS⁵-database, which is developed by the SFB 632⁶ in Potsdam and Berlin.

On the one hand, EXMARaLDA is a valuable annotation tool for these purposes. Partitions can be easily adjusted with respect to the amount of columns and tiers to accommodate the specific requirements of each sentence, for example, by adding tiers for more subordinate structures or by inserting columns

² Cf. Dipper (2005: 2).

³ On the importance of interchangeability and accessibility of digitized corpora, see Babeu (2011: 41). She also gives an overview of ongoing projects digitizing texts from numerous ancient languages, including work on image databases, POS-taggers, treebanks, and more.

⁴ EXMARaLDA = Extensible Markup Language for Discourse Annotation (current version of the Partitur Editor: 1.5.2), for further information see www.exmaralda.org.

⁵ ANNIS = ANNotation of Information Structure (current version: 3.0.1), more information under www.sfb632.uni-potsdam.de/d1/annis/ and in Zeldes et al. (2009).

⁶ For further information see www.sfb632.uni-potsdam.de/.

for elliptical forms.⁷ Although EXMARaLDA was designed primarily for the transcription of oral texts, it can be used for the analysis of written texts as well. The XML format also ensures exchangeability and makes the data archivable on a long-term basis.

The database ANNIS, on the other hand, is a great tool to retrieve both actual tagged information as well as secondary information by combining various search variables. Additionally, ANNIS is constantly being developed and improved.⁸ It also supports further modification of the data if wished for.

The gathered data will be made publicly accessible once the projects are completed to allow other researchers to use them. Until then, the examples given in this paper are available online as a sample corpus from the projects and are found on the ANNIS homepage.⁹ All queries given in the following can be used in the database.

3. Language requirements vs. implementation vs. retrieval

With the languages covered by the projects comes a variety of different writing systems: the Greek and Avestan scripts, Devanāgarī script in Sanskrit and Vedic, cuneiform writing in Hittite, and hieroglyphs in Luwian. For our purpose, the texts are used in their standard transliteration to simplify database usage and to make them accessible to a broad audience. Thus, a variety of additional characters reproducing the original versions have to be dealt with. To provide uniformity within the corpora, which is essential for data retrieval, special characters are entered with the virtual keyboard integrated into EXMARaLDA. Explanations on divergences in the transliteration will be given in the sections dealing with the respective languages.

To begin with, the tier labels of the examples shown in this paper are explained in Table 1. It also comprises references to tables which contain further information:

⁷ This is an advantage over other programs, e.g., ELAN, where adding new columns is not so easily done.

⁸ Our gratitude goes to Amir Zeldes and the ANNIS team, who try to implement all suggestions and needs.

⁹ See www.sfb632.uni-potsdam.de/annis/corpora.html.

Tier label	Content
[text]	The text is presented with every word separated as a token (cf. Tables 2, 4, 5, 7). It is stripped of any punctuation or other purely graphical information (cf. Table 8).
[lem]	Shows the lemma forms of each token (cf. Table 5).
[glos]	The translation of each word (in German in the actual corpus) and the grammatical analysis is given. Solely for display purposes, the line is split up into two. Hence, in the corpus, [glos2] is an integral part of [glos] (cf. Tables 6, 7, 8 for true visualization).
[glos2]	
[pos]	Denotes the part of speech of each token (with indication of certain properties such as correlation or compounds, cf. Table 5).
[orig]	Shows the sentence as it is found in the editions (Greek script for Greek, cf. Table 3, transliteration for the other languages). Missing parts of the texts and line breaks in poetry are indicated here.
[transl]	Contains the translation of the sentence as faithful to the original as possible.
[MCclause-st]	States the type of main clause, e.g., decl(arative), imper(ative), inter(rogative), and parataxis (cf. Table 6).
[MCgrfunct]	Presents the syntactical analysis of the main clause with the grammatical function of its elements.
[MCword_order]	Indicates the position of the verb, type of clitic (cf. Table 2), and word-order peculiarities such as dislocations.

Table 1: Explanation of tiers used in this paper¹⁰

3.1 Latin

3.1.1 Special characters

Latin, providing the basis for our own writing system, does not pose much of a challenge. It has no additional characters or any special way of text rendering. Line breaks of poetic texts are indicated with a slash in the [orig] tier.

¹⁰ Please note that this is only a selection of the information tagged in the corpora. The projects work on at least 27 categories for each sentence. The annotations include, but are not limited to, verbal semantics, saliency, particle usage, topic and focus, type of discourse, style, and the number of syllables. Furthermore, the syntax of subordinate clauses, if they occur, is sorted out in detail; cf. also footnotes 14 and 21.

3.1.2 [text], [MCword_order]: treatment of clitics

The only technical obstacle that Latin features is the occurrence of clitics (e.g., *-que*, *-ne*, *-ve*), which are to be analyzed separately; but in order for the data to be transformed into ANNIS, the texts have to be properly tokenized.¹¹ This means that the tier used for tokenization (in our corpora: [text]) has to contain a single cell per column. Only the cells in the subsequent tiers can be merged. This is illustrated in Table 2:

	wrong				right			
[text]	de	senatusque		consulto	de	senatus	-que	consulto
[lem]	de	senatus	que	consultum	de	senatus	que	consultum
[glos]	of	senate(M):	and	decision(N):	of	senate(M):	and	decision(N):
[glos2]		GEN.SG		ABL.SG		GEN.SG		ABL.SG
[pos]	prep	noun	conj	noun	prep	noun	conj	noun
[orig]	de senatusque consulto				de senatusque consulto			
[transl]	'and of the senate's decision'				'and of the senate's decision'			
[MCword_order]	#	#	en-clitic	#	#	#	en-clitic	#

Table 2: Tokenization (Caesar *De Bello Gallico* 7.1.1)

As can be seen, the original *senatusque* cannot be maintained if its single constituents are to be analyzed. Hence, the parts are split and a hyphen preceding the enclitic form marks it as such (cf. the right side of Table 2). This can also be used in ANNIS to search for enclitics: a query 'tok=/ - .*'¹² will return all enclitic forms within the corpora.¹³

¹¹ See Chiarcos/Ritz/Stede (2009) on the problem of different ways of tokenization.

¹² The queries must be used without the quotation marks in ANNIS. The [text] tier from EXMARaLDA is labeled 'tok' in ANNIS, so it refers to the same tier. The hyphen is the element distinguishing enclitics from other words; the dot is the wildcard for any single kind of character; the asterisk marks an unlimited number of characters. See Zeldes (2013) on how to conduct queries in ANNIS.

¹³ See also 3.3.2 and 3.5.3 for further information concerning the tokenization of the texts.

Another way to find clitics is to use a tag within the [MCword_order] tier: ‘proclitic’ or ‘enclitic’; to find both use ‘MCword_order=/. *c l i t i c /’. But because this tier discriminates between matrix clause and subordinate clause levels,¹⁴ it will only return clitics within one level.

3.1.3 All tiers: treatment of empty cells

The last tier of Table 2 displays ‘#’ in cells without tagged information. This applies to almost all tiers in all corpora. The advantage of this procedure is to be able to also find things that are not there, e.g., it is possible to search for subjects that are not topics. The hash is also inserted to the [text] and [lem] tiers of elliptical forms restored in the texts (tagged as *pro* or *PRO* in [pos]).

3.2 Greek

3.2.1 Special characters

For Greek it is sufficient to use the established transliteration patterns, e.g., $\psi > ps$ or $\eta > \bar{e}$ etc. While inserting accents and length markers via the virtual keyboard from EXMARaLDA is very convenient, this does not work for transliterated *iota subscriptum*, as the program provides neither subscription nor *iota subscriptum* as a default character; the Unicode code point for subscript *i* (U+1D62) does not work properly at the moment either. Therefore, it is put as a common *i* behind the character with which it usually occurs as a subscript letter. This makes it harder to distinguish, e.g., δi going back to ωi (as in *Τρωικός* ‘Trojan’) or ω (as in *κατωκίζω* ‘to found a town’), but can be solved by looking at the occurrence in the [orig] tier. Fortunately, this is not a distinctive feature concerning grammatical forms. Thus, a Greek sentence looks as displayed in Table 3. Here, δi appears twice in the first tier, going back to different letters (in bold), as can be seen in the [orig] tier:

¹⁴ Subordinate clauses, subdivided into levels according to dependency relations, have corresponding tiers named [SC1word_order], [SC2word_order] ..., currently up to [SC7...]. The sub-levels include not only true subordinate clauses (such as conditional or relative clauses), but also elements similar to subordinate clauses (such as *participia conjuncta*, appositions which do not contain verbs but rather adjectives with a strong verbal character, and so on).

[text]	tōn	dē	póleōn	hósai	mēn	neōtata
[lem]	ho	dé	pōlis	hōsoi	mēn	nēos
[glos]	the:	but	city(F):	all those:	indeed	the newest:
[glos2]	GEN.F.PL		GEN.PL	NOM.F.PL		ACC.N.PL
[pos]	art.def	part	noun	prrel	part	adj.superlat
[orig]	Τῶν δὲ πόλεων ὅσα μὲν νεώτατα					
[transl]	‘All of the cities, which have been recently’					

[text]	oikisthēsan	kai	ēdē	plōimōtērōn	ōntōn [...]
[lem]	oikēō	kaí	ēdē	plōimos	eimí
[glos]	being inhabited:	and	already	fitter for sailing:	being:
[glos2]	AOR.IND.PASS3PL			GEN.F.PL	GEN.F.PL
[pos]	vfin	conj	adv	adj.compar	ppt.prns.act
[orig]	ὠκίσθησαν καὶ ἤδη πλωιμωτέρων ὄντων, [...]				
[transl]	‘inhabited and, being already fitter for sailing, [...]’				

Table 3: Greek sentence (Thucydides *The Peloponnesian War* 1.7.1)

3.2.2 [text], [lem], [orig]: the virtual keyboard

With the latest release of ANNIS3 came a virtual keyboard that solved many problems of retrieving special characters. Right now we are working on creating layouts for a virtual keyboard fitting our needs of special characters. It will be available by the time the corpora go online. If all fails, wildcards can be used for those letters (‘.’, ‘..’, or ‘.*’) or the Unicode code points.¹⁵

3.3 Sanskrit and Vedic

3.3.1 Special characters

As with Greek, the transliterated Old Indic special characters (cf. Tables 4 and 5) are created with the virtual keyboard in EXMARaLDA. When the virtual keyboard containing all those characters is implemented into ANNIS, they will be easily retrievable.

¹⁵ This method is rather laborious. A simple search for Greek *καὶ* ‘and’ would then look like ‘or ig=/ .*03ba03b11 f76 .*/’, with the reversed backslash as escape character introducing each Unicode code point.

3.3.2 [text], [lem], [pos]: sandhis and compound words

The frequent sandhis¹⁶ in Sanskrit and Vedic are handled as follows: in the [orig] tier the text is shown in transliteration including all sandhis, while in the [text] tier those sandhis are dissolved into their respective pausa forms; cf. Table 4:

[text]	kiṃgotraḥ	nu	somya	asi	#	iti
[lem]	kiṃ+gotra-	nu	somya-	as	#	iti
[glos]	from which family:	now	dear(M):	be:	you	#
[glos2]	NOM.M.SG		VOC.SG	PRS.IND.ACT2SG		
[pos]	comp/adj	part	noun	vfin	pro	QUOT
[orig]	kiṃgotro nu somyāsiti					
[transl]	'From which family are you now, dear?'					

Table 4: Sanskrit sentence (*Chāndogya-Upaniṣad* 4.4.4)

Compound words, on the other hand, are treated together as one token, especially since they function as single words and have only one inflected ending. The word stems of which the compounds consist are annotated in the [lem] tier, separating the simplex forms by '+'¹⁷; cf. Table 5:

[text]	tamālatarukṛtanilayau
[lem]	tamāla-+taru-+kṛ+nilaya-
[glos]	having made a nest on a tamāla-tree: NOM.M.DU
[pos]	comp/adj

Table 5: Sanskrit compound word (*Pañcatranta* 1.15.1)

¹⁶ Sandhis are a phonological phenomenon where adjoining morphemes or lexemes show assimilation. Although they occur – mainly orally – in other languages as well, the Old Indic languages reproduce them in writing as a standard feature. Sandhis must not be confused with compound words or clitics.

¹⁷ This applies to the other languages as well, although none of them has a system as complex as the Old Indic languages. In Latin, for example, due to the strong lexicalization, only those words are given with two lemmas where the single parts were still somewhat transparent to the speaker. Hence, a noun derived from a compound verb is not marked as 'comp/'; but the verb itself is.

There are two ways to find compounds within the database. The first way is to search for the annotation as shown in the [pos] line in Table 5. The ‘comp’ before the slash denotes a compound. Behind the slash the word class is specified, which enables us to search, e.g., for compound adjectives only. Because the slash also functions as a regular expression in the ANNIS query language (AQL), it cannot be used in the query.¹⁸ The dot as single character wildcard is an apt replacement (e.g., ‘pos=/comp .adj/’).

The second way is to search for the plus sign within the [lem] tier. The first idea for a retrieval request might look like ‘lem=/.*+.*/'. However, the output of this query is zero, because ‘+’ functions as a regular expression for at least one occurrence of the previous character. The backslash ‘\’ has to be used, which works as an escape expression indicating that the following character is to be used as such and not as a regular expression, resulting in ‘lem=/.*\+.*/'. The same solution applies to the search for a period ‘.’, a question mark ‘?’, brackets ‘()’, ‘[]’, ‘{}’,¹⁹ and almost all other meta-characters used because they have special functions in the AQL.

3.3.3 [pos]: prefixed verbs

Old Indic verbal prefixes, if they are not truly composed and written together with their verbs, are annotated separately; they are retrievable via a query ‘pos=/pfx/’. This allows, for instance, searching for prefixes that are adjacent to their respective finite verbs, as opposed to those that stand apart from them. To find words that are placed next to each other, the elements of a query in ANNIS have to be linked to one another: ‘pos=/pfx/ & pos=/vfin/ & #1 . #2’.²⁰ If a prefix is separated from its verb by at least one element, ‘tmesis’ is tagged in the [style] tier for both forms. Hence, the search for prefixes in tmesis is also possible via ‘pos=/pfx/ & style=/.*tmesis.*/ & #1 __ #2’. To conduct a search within [style], it is strongly recommended to always use wildcards as there are often multiple stylistic devices annotated.

¹⁸ Cf. 3.5.1 and footnote 23 for more information on handling of the slash.

¹⁹ In the corpora, round brackets are used for better text comprehension in the translations; square brackets indicate restored text elements that are missing in the language source; curly brackets are used in Hittite texts instead of ‘<>’ because elements within angle brackets are not displayed in the database due to their use in the programming language.

²⁰ Here the dot stands for: part one (the prefix) directly precedes part two (the finite verb).

3.4 Avestan

3.4.1 Special characters

The rendering of transliterated Avestan special characters deviates merely in a few details from that found in text editions. As with subscript letters, EXMA-RaLDA neither provides superscript characters nor is the work with the corresponding Unicode code points (U+1D5B, U+1D58, U+1D4A for superscript ν , u and ϑ) straightforward. Therefore, x^ν is written as plain $x\nu$ and η^ν as $\eta\nu$, cf. Table 6:

[text]	vīsaitiuuā	asti	miθrō	
[lem]	vīsaitiuuant-	ah	miθra-	
[glos]	twentyfold:	be:	contract(M):	
[glos2]	NOM.M.SG	PRS.IND.ACT3SG	NOM.SG	
[pos]	adj.num	vfin	noun	
[orig]	vīsaitiuuā asti miθrō /			
[transl]	‘Twentyfold is a contract’			
[MCclause-st]	main:decl			
[MCgrfunct]	pred	v.copul	subj	
[text]	aṅtarə	haša	suptidarənga [...]	
[lem]	aṅtarə	haxāii-	supti- +darənga-	
[glos]	between	companion(M):	shouldering an obligation:	
[glos2]		ACC.DU	ACC.M.DU	
[pos]	prep	noun	comp/adj	
[orig]	aṅtarə haša suptidarənga / [...] /			
[transl]	‘between two companions shouldering obligations, [...],’			
[MCclause-st]	~			
[MCgrfunct]	attr/subj			
[text]	aštaiθiuuā	aṅtarə	zāmātara	xvasura [...]
[lem]	aštaiθiuuant-	aṅtarə	zāmātar-	xvasura-
[glos]	eightyfold:	between	son-in-law(M):	father-in-law(M):
[glos2]	NOM.M.SG		ACC.DU	ACC.DU

[pos]	adj.num	prep	noun	noun
[orig]	aštaiθiuuā antarə zāmātara xvasura / [...] /			
[transl]	‘eightyfold between son-in-law and father-in-law, [...]’			
[MCclause-st]	main:decl/compound sentence			
[MCgrfunct]	pred	attr/subj		

Table 6: Avestan sentence (Yt X 116)

3.4.2 [MCclause-st]: syntactical analysis

In the course of the two projects, we found that the simple denomination of paratactic clauses as ‘parataxis’ was not sufficient. In all languages under concern, there are self-sufficient paratactic main and subordinate clauses containing all information that is structurally necessary, as well as clauses where crucial elements are only expressed once but with their scope extending over several clause units. To take those elliptic elements into account without always assuming a drop of some kind, those clauses were labeled as ‘compound sentence’²¹ in [MCclause-st]. Thus, the use of this term deviates from the standard usage as a synonym for complex sentences. The sentence in Table 6 above is an example of such an occurrence: the copula and the subject *contract* are only expressed once at the very beginning of the verse, followed by a climactic succession of the worth of contracts between certain people (twentyfold, thirtyfold, fortyfold, and so forth). Only the changing elements are expressed again; verb and subject are still valid from the first mentioning.

3.5 Hittite

3.5.1 Rendering of a Hittite text

In the corpora, the Anatolian languages have undergone by far the biggest graphical transformation. A standard piece of transliterated Hittite text appears in editions as in example (1):

- (1) Standard transliteration of Hittite (CTH 321 – *The Myth of Illuyanka* 3)
- ma-a-an*^{DIM}-*aš*^{MUS}*il-lu-ya-an-ka-aš-ša I-NA*
^{URU}*ki-iš-ki-lu-uš-ša ar-ga-ti-i-e-er ...*

²¹ This also applies to the already mentioned subordinate structures (tier names being [SC1clause-st], [SC1grfunct] ... up to [SC7...]); for further details see footnote 14.

It contains logograms from both Sumerian (e.g., IM) and Akkadian (e.g., *I-NA*), determinatives (e.g., ^Dx), syllabograms (e.g., *ma-a-an*), and also grammatical markers from both Sumerian (e.g., x^{MES}) and Akkadian (e.g., x^{LIM}). As already mentioned, EXMARaLDA does not provide superscript rendering. It does not support italics either. Hence, simply copying the text would result in example (2):

- (2) Unformatted Hittite sentence (CTH 321 – *The Myth of Illuyanka 3*)
 ma-a-an DIM-aš MUŠil-lu-ya-an-ka-aš-ša I-NA
 URUki-iš-ki-lu-uš-ša ar-ga-ti-i-e-er ...

As can be seen, it is impossible to differentiate between the various representations of language sources. So we developed a workaround involving the slash for italics (e.g., */I-NA/*), the underscore for superscripts (e.g., MUŠ), and their combination for superscript italics (e.g., /LIM/). Therefore, the example sentence appears within the corpora as in Table 7²² (changes in bold):

[text]	ma-a-an	<u>D</u> IM-aš	<u>MUŠ</u> il-lu-ja-an-ka-aš-	-ša
[lem]	mān	IM	Illuyankaš	-ša
[glos]	when	weather god(C): NOM.SG	Illuyanka(C): NOM.SG	and
[pos]	conj	noun	noun	part
[orig]	ma-a-an <u>D</u> IM-aš <u>MUŠ</u> il-lu-ja-an-ka-aš-ša			
[transl]	‘When the weather god and (the serpent) Illuyanka’			
[text]	<i>/I-NA/</i>	<u>URU</u> Ki-iš-ki-lu-uš-ša	ar-ga-ti-i-e-er	...
[lem]	<i>/ina/</i>	Kiškilušša	argatiya-	
[glos]	in	Kiskilussa	fight: IPF.IND.ACT3PL	
[pos]	prep	noun	vfin	
[orig]	<i>/I-NA/</i> <u>URU</u> Ki-iš-ki-lu-uš-ša ar-ga-ti-i-e-er ...			
[transl]	‘were fighting in the city of Kiskilussa, ...’			

Table 7: Adjusted Hittite sentence (CTH 321 – *The Myth of Illuyanka 3*)

²² There might be a way to improve the layout in ANNIS and restore the standard display. It remains to be seen if and how this works sufficiently. Concerning the usage of brackets within Hittite texts, cf. footnote 19.

Let us assume that we want to look for Akkadian grammatical markers, in transcriptions rendered as superscript italics. As explained above, the workaround in our Hittite corpus is ‘_/x/_’ (expressing ‘^x’). Searching for any of those lexemes using the given methods, a retrieval request may look like ‘tok=/ .*_ .*_ .*_ */’ (to express ‘tok=/ .*_ / .*_ / .*_ /’). Unfortunately, the results would include all superscriptions, not only those in italics. Hence, this method is not sufficient. The approach using the escape character backslash ‘\’ (resulting in ‘tok=/ .*_ \ / .*_ \ / .*_ /’) returns a syntax error message.²³ The solution is to use the Unicode code point for the slash instead, which is U+002F. Thus the query is ‘tok=/ .*_ \u002f .*_ \u002f .*_ /’.²⁴

3.5.2 [pos]: preverbs, adverbs, and postpositions

In Hittite and Luwian there are words which maneuver between being preverbs, postpositions, or simple adverbs. Very often it is unclear in the texts into which category a specific occurrence falls. Hence, in order to minimize interpretations, no distinction is made regarding the annotation of their part of speech: they are tagged neutrally as adverbs (‘adv’) in the [pos] tier.

3.5.3 [text]: clitics

In our Anatolian corpora, words preceding enclitics are treated a little differently than in the other languages under concern. Due to the writing of syllables and the numerous clitic chains, the word followed by an enclitic form is indicated by a hyphen at the end in addition to the hyphen preceding the actual enclitic; cf. Table 2 vs. Table 7 above.

3.6 Luwian

3.6.1 Rendering of a Luwian text

Luwian has even more peculiarities than Hittite. See example (3) for the graphics of a transliterated Hieroglyphic Luwian text in standard editions:

²³ This workaround was chosen long before the problems arising with it were known. By the time of the discovery, it was too late to change all the annotated texts; so we maintained the practice. The problem of retrieving the slash lies within the underlying parser of the AQL and cannot be solved at the moment. This might change over the course of time.

²⁴ The reverted slash is the escape character and the lower-case *u* tells the database that a Unicode code point follows and no actual numbers are to be retrieved.

(3) Standard transliteration of Hieroglyphic Luwian (Maraş 1 1a-1d)

- 1a EGO-*wa/i-mi-i* 'TONITRUS.HALPA-*pa-ru-ti-i-ia-sa* [...]
- 1c 'TONITRUS.HALPA-*pa-ru-ti-ia-si-sà* || HEROS-*li-sa*
|(INFANS.NEPOS)*ha-ma-si-sá*'
- 1d *mu-wa/i-ta-li-si-sà* |("SCALPRUM+RA/I.LA/I/U")*wa/*
i+ra/i-pa-li-sa |(INFANS.NEPOS)*ha-ma-su-ka-la-sá* [...]

It contains logograms given in Latin (e.g., EGO) or, if the word is known, in Luwian (e.g., HALPA), determinatives (e.g., (INFANS.NEPOS) or the personal marker '), syllabograms (e.g., -mi-i), vowel alternatives (e.g., -wa/i, LA/I/U), line dividers (||), and occasional markers for the beginning of a word (|x), word end signals (x-'), and logogram markers for hieroglyphs ("x"). The last three markers are maintained as such in the [orig] tier while they are not displayed within the [text] tier.

Whereas the Latin logograms remain as they are, we adjusted the Luwian logograms, determinatives, and line dividers to the rendering already used in Hittite, resulting in '/HALPA/', '_INFANS.NEPOS_' and '/'. Although the modern convention for Luwian determinatives is to use round brackets, they are set in underscores in order to 1) facilitate the comparative work on Hittite (and thus follow the older convention; cf., e.g., Werner 1991), and 2) avoid confusion because round brackets may be used for insertions (cf. fn. 19).

Words with different vowel realization such as -wa/i, -ra/i, la/i/u are found in the database with the vowel options but without the dividing slash – in other words, the different vowels appear as diphthongs or as sequences like *aiu*. This solution was necessary because the slash already fulfils two functions, namely as an indicator of italics and line breaks. It was a feasible solution because there are no syllables with diphthongs, which ruled out potential ambiguities. These adjustments result in the display shown in Table 8 (examples of graphical changes in bold):

²⁵ Clauses of this type require the reflexive pronoun; it remains untranslated.

²⁶ *HEROS-li-sa*, *warpalis*, and *tarwanis* could be either *i*-stems in the GEN.SG or possessive adjectives ending in -*iy(a)* in the NOM.C.SG; cf. Payne (2010: 81).

[text]	EGO-	-wai-	-mi-i ²⁵	<u>I</u> _TONITRUS./ HALPA /-pa-ru-ti-i-ia-sa [...]
[lem]	amu	-wa	-mi	halparuntiya-
[glos]	I	#	myself	Halparuntiyas(C): NOM.SG
[pos]	prpers	QUOT	prrefl	noun
[orig]	/ EGO- wai -mi-i <u>I</u> _TONITRUS./ HALPA /-pa-ru-ti-i-ia-sa [...]			
[transl]	I am Halparuntiyas, [...]			
[MCgrfunct]	subj		pred	

[text]	<u>I</u> _TONITRUS./ HALPA /-pa-ru-ti-ia-si-sà	HEROS-li-sa ²⁶
[lem]	halparuntiyasa	*hastaliya-
[glos]	of Halparuntiyas: NOM.C.SG	of the hero: NOM.C.SG
[pos]	adj	adj
[orig]	<u>I</u> _TONITRUS./ HALPA /-pa-ru-ti-ia-si-sà / HEROS-li-sa	
[transl]	the heroic Halparuntiyas'	
[MCgrfunct]	attr/app/pred-i	

[text]	<u>INFANS.NEPOS</u> _ha-ma-si-sà	mu-wai-ta-li-si-sà
[lem]	hamsai-	muwataliyasa
[glos]	grandchild(C): NOM.SG	of Muwatalis: NOM.C.SG
[pos]	noun	adj
[orig]	<u>INFANS.NEPOS</u> _ha-ma-si-sà-' mu-wai-ta-li-si-sà	
[transl]	grandson, Muwatalis'	
[MCgrfunct]	app/pred-i	attr/app/pred-j

[text]	<u>SCALPRUM</u> +/ <u>RAI.LAIU</u> /_wai+rai-pa-li-sa
[lem]	warpaliya-
[glos]	of the brave: NOM.C.SG
[pos]	adj
[orig]	<u>"SCALPRUM</u> +/ <u>RAI.LAIU</u> /"_wai+rai-pa-li-sa
[transl]	the brave
[MCgrfunct]	~

[text]	<code>_INFANS.NEPOS_ha-ma-su-ka-la-sá</code>	[...]
[lem]	hamsukala-	
[glos]	great-grandchild(C): NOM.SG	
[pos]	noun	
[orig]	<code>_INFANS.NEPOS_ha-ma-su-ka-la-sá</code> [...]	
[transl]	great-grandson, [...]	
[MCgrfunct]	app/pred-j	

Table 8: Adjusted hieroglyphic Luvian sentence (Maraş 1 1a-1d)

3.6.2 [MCgrfunct]: index characters

The sentence in Table 8 is the beginning of a stele inscription. It shows the typical scheme of an introductory nominal clause ('I [am] X') followed by a multitude of asyndetic appositions, which in turn can have attributes (here: 'son of Y, grandson of Z', and so on). In order to mark which attribute belongs to which apposition, index characters are used starting with *-i* and then going along the alphabet (with the exception of *-o*, which only designates objects). This procedure applies to all elements in all languages that belong closer together (e.g., also verb forms in tmesis). Index characters can occur in many tiers. This is the reason why we always recommend searching with the wildcard '*' at the end of a query.

4. Summary

Digitizing ancient languages is not a simple task, especially, if different languages with various underlying scripts are to be rendered in a transliterated, uniform, commensurable format in order to annotate and compare them. This is done in the projects 'Information Structure in Older Indo-European Languages' and 'Information Structure in Complex Sentences – Synchronic and Diachronic'. The present paper laid out the special needs of the annotated languages Latin, Ancient Greek, Old Indic, Avestan, Hittite, and Luvian. Graphical transformations concerning special characters were explained and their influences on the work with the corpora outlined. Introductory remarks towards the syntactical annotation of those languages were made, including information on part of speech, clause types, grammatical function, and word

order. Additionally, example queries to retrieve the data from the ANNIS database were given, which can be tested on the sample corpus deposited on the ANNIS homepage. It was shown that, with a few adjustments, all languages can be adapted to the capabilities of freely available programs and, therefore, the creation of new annotation devices and database structures, which hinders broader accessibility, can be avoided.

References

- Babeu, Alison (2011): "Rome wasn't digitized in a day": building a cyberinfrastructure for digital classics. (= CLIR 150). Washington DC: Council on Library and Information Resources. www.clir.org/pubs/reports/pub150/pub150.pdf (last accessed: October 27, 2013).
- Chiarcos, Christian/Ritz, Julia/Stede, Manfred (2009): By all these lovely tokens ... merging conflicting tokenizations. In: Stede, Manfred/Huang, Chu-Ren/Ide, Nancy/Meyers, Adam (eds.): Proceedings of the Third Linguistic Annotation Workshop (LAW III) at ACL-IJCNLP 2009. Singapore: Association for Computational Linguistics, 35-43. www.aclweb.org/anthology/W/W09/W09-3005.pdf (last accessed: October 27, 2013).
- Dipper, Stefanie (2005): XML-based stand-off representation and exploitation of multi-level linguistic annotation. In: Eckstein, Rainer/Tolksdorf, Robert (eds.): Proceedings of Berliner XML Tage (BXML 2005). Berlin, 39-50. http://pub.sfb632.uni-potsdam.de/publications/D1/D1_Dipper_2005a.pdf (last accessed: October 21, 2013).
- Payne, Annick (2010): Hieroglyphic Luwian – An introduction with original texts. 2nd. rev. ed. (= *Subsidia et Instrumenta Linguarum Orientis* 2). Wiesbaden: Harrassowitz.
- Werner, Rudolf (1991): Kleine Einführung ins Hieroglyphen-Luwische. (= *Orbis Biblicus et Orientalis* 106). Göttingen: Vandenhoeck & Ruprecht.
- Zeldes, Amir (2013): ANNIS User Guide – Version 3.0.0. www.sfb632.uni-potsdam.de/annis/download/ANNIS_User_Guide_3.0.0.pdf (last accessed: October 22, 2013).
- Zeldes, Amir/Ritz, Julia/Lüdeling, Anke/Chiarcos, Christian (2009): ANNIS: a search tool for multi-layer annotated corpora. In: Mahlberg, Michaela/González-Díaz, Victorina/Smith, Catherine (eds.): Proceedings of the Corpus Linguistics Conference, CL2009, University of Liverpool, UK, 20-23 July 2009. http://ucrel.lancs.ac.uk/publications/cl2009/358_FullPaper.doc (last accessed: October 30, 2013).

Managing and annotating historical multimodal corpora with the eHumanities desktop

An outline of the current state of the LOEWE project “Illustrations of Goethe’s Faust”

Abstract

Text corpora are structured sets of text segments that can be annotated and interrelated. Expanding on this, we can define a database of images as an iconographic multimodal corpus with annotated images and the relations between images as well as between images and texts. The Goethe-Museum in Frankfurt holds a significant collection of art work and texts relating to Goethe’s Faust from the early 19th century until the present. In this project we create a database containing digitized items from this collection, and extend a tool, the ImageDB in the eHumanities Desktop, to annotate and provide relations between resources. This article gives an overview of the project and provides some technical details. Furthermore we show newly implemented features, explain the challenge of creating an ontology on multimodal corpora and give a forecast for future work.

1. Introduction

Art museum collections are widely varied. Factors such as finance, politics and aesthetics play a significant role in their formation. In any case, the act of assembling a collection is a historical process resulting in a group of art objects that are somehow related to each other, if in no other way than by being held in the same place. Often, a museum is devoted to a particular purpose, and this is the common ground. The collection of the Goethe-Museum in Frankfurt am Main¹ emphasizes the Age of Goethe and houses a significant collection of illustrations of Goethe’s works. Among them are about two thousand drawings and prints of ‘Faust’ from the early 19th century until the present. Funded by the LOEWE² research cluster “Digital Humanities”³ the project “Illustrations of Goethe’s Faust”, an interdisciplinary cooperation between the *Frankfurter*

¹ www.goethehaus-frankfurt.de/. All URLs have been checked and found valid as of late January 2015.

² “Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz” of the state of Hesse; <https://hmkw.hessen.de/loewe>.

³ www.digital-humanities-hessen.de/.

Goethe-Haus and the *Text Technology Lab*⁴ at the University of Frankfurt, aims to digitize the collection of Faust illustrations and to create a research environment for viewing and annotating the digital representations. The objective of the project is to build a framework which permits users to describe inter-pictorial relations as indicated above and moreover relations between images and text.

This paper explains how an annotated image collection can be seen as a historical corpus and how it can be technically represented and automatically processed within the ImageDB, a module of the eHumanities Desktop,⁵ a research environment developed by the *Text Technology Lab*. The term corpus or text corpus is generally taken to mean a structured set of texts subjected to annotations either defining text segments or the relations between them. Annotating the internal structure of “any sign aggregate (e.g., a sentence, a paragraph or a whole text) [...], we deal with intra-aggregate relations. [...] Alternatively, we may interrelate the aggregate as a whole with other aggregates and their segments. In this case we deal with inter-aggregate relations [...]” (Gleim/Warner/Mehler 2010: 2). From the point of view of semiotics an annotated image collection that links images and their constituents to a text and vice versa relates information of two different modalities.

In this paper, we propose the following definition of a historical multimodal corpus: a set of intra- and interrelated signs of the same or a different modality where the relations can be annotated in a way that allows users to represent historical changes in the corpus. If we take the corpus of Faust illustrations as an example, the history of an art collection can be plotted – what it contained and when, what has been acquired or lost. Furthermore, it can be observed how the iconography of the artworks and the artists’ interest for particular themes and motifs in the text change over time.

There are several image database solutions offering the annotation of image interrelations (e.g. the web-based solution ConedaKOR)⁶ and the segmentation of images (e.g. HyperImage).⁷ The combination of relation-based management and image segmentation has been worked out by Meta-Image (Dieckmann/Kliemann/Warnke 2012), which is integrated in the image archive

⁴ www.hucompute.org.

⁵ <http://hudesktop.hucompute.org/>.

⁶ www.coneda.net.

⁷ www.uni-lueneburg.de/hyperimage/hyperimage/.

Prometheus.⁸ Other research projects like CLAROS⁹ or The WissKI Project¹⁰ developed effective techniques for information integration and retrieval. They rely on the CIDOC¹¹ Conceptual Reference Model (CRM)¹², an ontology for cultural heritage also used in the course of our project. Although much work has been done in the field of image annotation, none of the projects other than the eHumanities Desktop provides as a whole the functionalities needed for the challenge of multimodal historical corpus management: a collaborative, web-based framework, management, segmentation and annotation of images, image-text-linking, multiple annotation layers, and ontology integration based on RDF.

The eHumanities Desktop is a platform independent web-based system which “allows users to upload, organize and share resources” (Gleim/Mehler/Ernst 2012). These resources “can be [...] annotated and analyzed in various ways” (ibid.). These ways include, for example, powerful analytical tools for grouping, searching, sorting, and filtering images or documents. Another powerful tool is the flexible and easy-to-use access permission system. This permits users to open or restrict access to their projects, and can be used to work on copyrighted material without infringement of the copyright. The ImageDB has been in existence for several years, and has gone through several major iterations. It has significantly enhanced its functionality compared to earlier versions, with changes incorporated based on extensive user feedback.

In section 2 we describe the technical structure of the eHumanities Desktop, in section 3 we explain the transforming of the existing metadata, in section 4 we give an overview on image segmentation and finally in section 5 we take a look at the ontological representation of images in the database.

2. The technical structure of corpus annotation

In order to build up a multimodal corpus, the integration of images, texts and their annotations in a complex network (Mehler 2008) is a central requirement. Typically, a large corpus is annotated by a research group rather than a

⁸ www.prometheus-bildarchiv.de/.

⁹ www.clarosnet.org/.

¹⁰ <http://wiss-ki.eu/node/23/>.

¹¹ International Committee for Documentation of the International Council of Museums (ICOM).

¹² www.cidoc-crm.org/.

single expert. Within the ImageDB, images are managed in repositories that are shared with users or groups. The data is organized in a graph-structure (Neo4J).¹³ We delimit generally two kinds of objects in our data-structure: *authorities* and *resources* (Figure 1).

Authorities represent a group or a user. Groups contain users and users belong to one or more groups. It is possible to have users without a group, but because most projects are collaborative, it makes sense to use groups primarily for sharing and restricting access to resources. Users can share their resources with other authorities.

Resources are represented primarily as documents,¹⁴ in our case as images (FileDocuments). Resources can also be *repositories*, *annotations* or *annotation schemas*. Figure 1 (from Gleim/Mehler/Ernst 2012) shows the structure of the master data. We can, when sharing, set permission levels on a resource (read, write, delete, grant). With repositories we can display a hierachical structure similar to folders in a computer file system (ibid.).

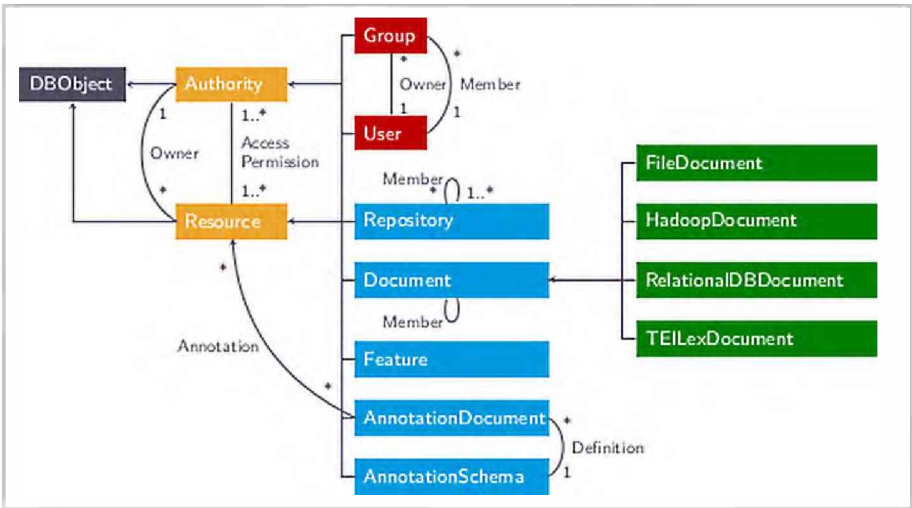


Figure 1: Class diagram of the master data

Annotations based on *annotation schemas* are resources and are another important element in the data structure. The main challenge of the project is, after the digitisation of the illustrations, to annotate the images to build a complete content reference.

¹³ <http://neo4j.org/>.

¹⁴ Many kinds of documents are possible.

We annotate resources based on an annotation schema. The schema is based on RDF¹⁵ and is very flexible. It is possible for the users who have the requisite permissions to create schemas or edit them to append or modify annotation fields. Each resource can then be annotated with any schema. It is also possible to annotate a resource with the same schema more than once.

3. Information integration in annotation schemas

The extensible annotation architecture of the ImageDB enables knowledge representation on multiple levels. In our case, the Faust illustrations are being annotated with metadata and conceptualized in a formal ontology. On the first level the metadata lists the core information about an artwork, e.g. its title, who created it and when, the physical description, and so on. The second level is the integration of the annotated information in an ontology based on the CIDOC Conceptual Reference Model (CRM), a high-level ontology devised to describe “concepts and relations relevant to the documentation of cultural heritage” (Signore 2008: 11). Although there have been many attempts to establish a common metadata set for libraries, museums and scholars (e.g. Dublin Core), “the number of metadata vocabularies will continue to grow as individual communities seek to structure their own information for their own purposes” (Doerr/Hunter/Lagoze 2003). In the early stages of the project, when a metadata standard had to be chosen, this situation cut both ways. First, we had a wide range of metadata standards from which to choose, but second, we knew that a decision could limit future interoperability. A metadata set satisfying the following constraints had to be found: an XML-based syntax to ensure compatibility with the cataloguing system of the Goethe-Museum, the ability to describe objects of different kinds (e.g. drawings, books, manuscripts) and relations between them, the increasing acceptance and use in the museums sector and finally a structural affinity to the CIDOC CRM. The decision process led to the XML harvesting schema LIDO¹⁶ released in 2010 in version 1.0 by the *ICOM-CIDOC Working Group Data Harvesting and Interchange*. Being the summation of three widespread standards (CDWA Lite, museumdat and SPECTRUM) and specifically designed as an exchange format, it is intended to overcome the difficulties of sharing information with other portals or institutions. LIDO is multilingual, supports the “full range of descriptive information about museum objects” (ICOM-CIDOC Working Group Data Harvesting and

¹⁵ www.w3.org/TR/swbp-vocab-pub/.

¹⁶ www.lido-schema.org.

Interchange 2010) and reserves a set of structural elements to express relations to other objects or thematic subjects. Most importantly, it implements the CI-DOC CRM by using the concept of events: “the creation, collection, and use of an object are defined as events that have associated entities such as dates, places and actors” (ibid.). For this reason, the data mapping to the ontology is much facilitated.

The second level of annotation, the knowledge representation with an ontological approach, is the base of processing the whole of the images and annotations as a multimodal corpus. As Signore pointed out, the “metadata level cannot exploit the full richness of possible associations among different information items” (Signore 2008: 28). Keeping in mind all possible sorts of interrelations, a collection should be viewed as a totality rather than as a list of individual items. Furthermore, there are cultural (e.g. art historical, textual, social) contexts to be respected and annotated. By creating semantically defined relations, the ontological approach permits users to represent this richness of interconnections and knowledge accumulated by scholars. The integration of images and information in a network of semantic relations, i.e., a multimodal annotated corpus, is the next task to be performed within the project. Once accomplished, research in cultural studies can be supported by automatically processing corpora. Section 5 demonstrates how an historical development in the iconography of Faust illustrations could be interpreted by an intelligent agent (i.e., a computer program). Above all, we need to identify meaningful parts within the images in order to annotate the semantic content of an illustration. How this is technically realized will be explained in the next section.

4. Image segmentation

To annotate more specifically and to connect parts of images with text (image-text-relation) we support the creation of subimages from other images. In order to allow precise annotations, it was necessary to support cropping images not only as rectangles. The first implementation of the crop-method in the ImageDB contained only the rectangle shape. For this project, we extended the functionality, so users are now able to crop images as rectangles, circles, ellipses and complex polygons.

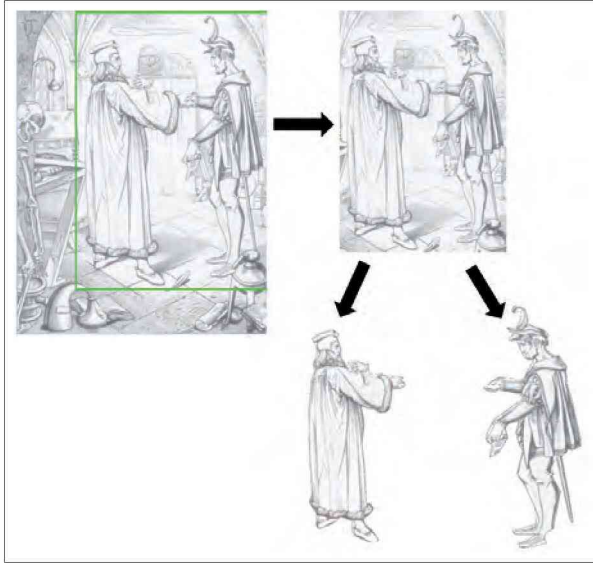


Figure 2: The crop sequence

After cropping, each new image becomes a new document owned by the current user. As the subimage is an image itself it is possible to crop this subimage again to get finer details for annotations. Each subimage, independent of the depth of cropping, knows its ‘parentage’, so to speak, and contains a reference to the original prototype image. Thus we make it possible to create a network of subimages within the prototype image and find inherent connections. By inverting the arrows in Figure 2 one can see the way a subimage is linked to its ‘parents’ and finally the prototype image.

The selection of the shape will be done manually by the user, using the mouse. The first click marks the starting point of the shape. After the first click the user can see the shape according to the position of his cursor and can scale it while moving. After the last click, depending on the shape, the shape is finished and the image can be cropped. After finishing the shape-creation it is possible to move the whole shape by clicking on it. The shapes are displayed with a green border (see Figure 4 for the example of an ellipse). This is realized with the JavaScript library *jsGraphics*.¹⁷

¹⁷ www.walterzorn.de/jsgraphics/jsgraphics.htm.

FORM: $\langle \text{Rectangle, Circle, Ellipse, Polygon} \rangle$

XYCoordinate: $[x, y]$

$x, y \in \mathbb{N}$

SHAPEFORM: *XYCoordinate**

MaxPoints(*FORM*): *Rectangle* $\rightarrow 2$; *Circle* $\rightarrow 2$; *Ellipse* $\rightarrow 3$; *Polygon* $\rightarrow n$

For the polygon-shape there is no maximal number of points for the function (n). The user clicks around the selected shape to create the polygon. In this case the user must, on complex forms, make many points to create the shape. The image is cropped with the minimal and maximal bounds of the selected shape, first as a rectangle (Figure 4). The cropping is done by the HTML5 element *Canvas*.¹⁸ By loading the image it is converted to a PNG. To create the transparency the shape is drawn on the image and any pixels that are not part of the shape are deleted (Figure 3).

cropImage(*FORM*_{*shape*}):

$$\min_{i=0}^n(\text{XYCoordinate}(x)), \min_{i=0}^n(\text{XYCoordinate}(y)), \\ \max_{i=0}^n(\text{XYCoordinate}(x)), \max_{i=0}^n(\text{XYCoordinate}(y))$$

The method is illustrated in Figure 4, using the ellipse as an example. The function requires four values to crop the image, the minimum and maximum x / y values of the shape: (1) the smallest x -value, (2) the smallest y -value, (3) the largest x -value and (4) the largest y -value.

After the shape is analysed, the picture will be cropped as a rectangle, then created with the values of the function. The values of the shape (*XYCoordinate**) will be saved as an annotation of the image. After that the image will be cropped on the fly when it is loaded in the selected shape-form (Figure 3).

5. The ontological representation of historical changes

With the aid of the segmentation functionality of the ImageDB, the image is divided into parts in order to single out particular motifs, persons and objects. These segments are annotated with an iconographical topic, linked to the text or to another image.

¹⁸ www.w3.org/TR/html5/scripting-1.html#the-canvas-element.



Figure 3: The cropped image, displayed in the ImageDB Repository-View

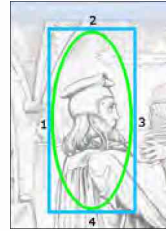


Figure 4: Schema to get the minimal and maximal values of the selected x/y values

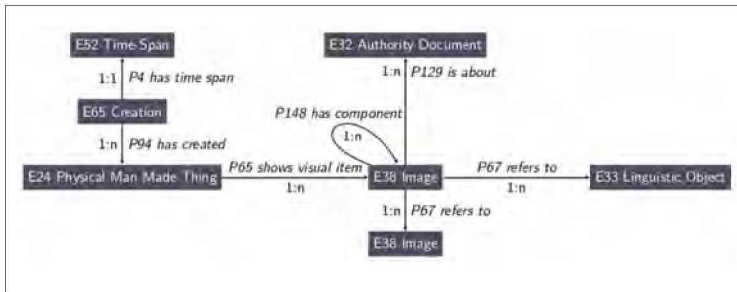


Figure 5: Ontological representation of interrelations

The ontological representation of these relations based on CIDOC CRM classes is illustrated in Figure 5. The entities (e.g. E38 Image, E33 Linguistic Objects) are related with each other via properties (e.g. P67 refers to, P129 is about). For instance, an intrapictorial relation between image and subimage is expressed by the triple “E38 Image *P148 has component* E38 Image”. A link to the text would be “E38 Image *P67 refers to* E33 Linguistic Object”. To refer to an iconographic theme (e.g. the poodle) of a Faust illustration, an extensible topical thesaurus has been set up. It is represented as *E32 Authority Document*, the CIDOC class that “comprises [...] documents that define terminology or conceptual systems for consistent use” ((SIG) 2011). Unlike text and image relations, it is linked by the property *P129 is about* since we are referring to a thematic subject and not to another artifact. The temporal dimension of the knowledge representation is gained by the event-based concept underlying the CIDOC CRM. Such an event is the creation or production of the object that brings into existence the material side of the artwork (E65 *Creation P94 has created* E24 *Physical Man Made Thing*). The limited extent of time in which this event happens is represented in the Entity E52. As a result of the annotation by means of an ontology, the image collection can be automatically pro-

cessed as a historical corpus. Questions of visual traditions or iconographic changes over time can be analysed with an RDF reasoner or a similar solution. If we take iconography as an example, we notice that early illustrations tend to be simple pictorial renditions of the text, in contrast to the expression on an abstract level sought by modern illustrators. The shift from a narrative to a psychological illustration could be proven by counting the text-image-links in combination with the knowledge of what has been illustrated – an object or a person. The more text items that have been illustrated, the closer an illustration is to the text. Seen from another side, it can show how the interest for a defined topic develops in the history of Faust illustrations. Such an automatic processing could lead to the discovery of relations that have as of yet gone unnoticed. Thus, when examined critically, the approach of this example could turn out to be too limited. In that case, reasoning over the ontology has to be refined in a sort of hermeneutic process.

6. Conclusion and prospects

With the ImageDB we can upload, manage and share images and text documents to create a multimodal corpus and annotate it. Scalable access permissions allow users to share their projects without copyright infringement. To allow precise annotations, we created the possibility to crop images in varied forms. The cropped images have a recursive backlink to parent/ancestor images, including the prototype image. This functionality will be expanded in future development iterations. For instance, we plan eventually to include other multimodal content like audio and video files. The value of the applied ontology will become clearer as the size of the corpus grows. A platform combining all kinds of resources (e.g. letters, images, music, movies) with relevance to a literary work or to a specific period is an exciting vision, and a potential goal for the ImageDB. After having integrated the images and imported the corresponding metadata in the Image DB, the main goal is currently to prepare the annotation schema to connect images with the corresponding text from the Faust edition.¹⁹

¹⁹ <http://faustedition.uni-wuerzburg.de>.

References

- Dieckmann, Lisa/Anita Kliemann/Warnke, Martin (2012): Meta-Image: Forschungs-umgebung für den Bilddiskurs in der Kunstgeschichte. In: cms-journal. Computer- und Medienservice 35, 2012: 11-17.
- Doerr, Martin/Hunter, Jane/Lagoze, Carl (2003): Towards a core ontology for information integration. In: Journal of Digital Information 4, 1. <https://journals.tdl.org/jodi/index.php/jodi/article/download/92/91>.
- Gleim, Rüdiger/Mehler, Alexander/Ernst, Alexandra (2012): SOA implementation of the eHumanities desktop. In: Proceedings of the Workshop on Service-oriented Architectures (SOAs) for the Humanities: Solutions and Impacts, Digital Humanities 2012, Hamburg, Germany. www.hucompute.org/data/pdf/dhc2012.pdf.
- Gleim, Rüdiger/Warner, Paul/Mehler, Alexander (2010): eHumanities Desktop – An architecture for flexible annotation in iconographic research. In: WEBIST 2010, Proceedings of the 6th International Conference on Web Information Systems and Technologies, Volume 2, Valencia, Spain, April 7-10, 2010. Valencia: INSTICC, 214-221.
- ICOM-CIDOC Working Group Data Harvesting and Interchange (2010): LIDO – Lightweight Information Describing Objects. Version 1.0. www.lido-schema.org/schema/v1.0/lido-v1.0-specification.pdf.
- Mehler, Alexander (2008): Large text networks as an object of corpus linguistic studies. In: Lüdeling, Anke/Kytö, Merja (eds.): Corpus linguistics. An international handbook of the science of language and society. Berlin/New York: de Gruyter, 328-382.
- (SIG), CIDOC CRM Special Interest Group (2011): Definition of the CIDOC conceptual reference model. Version 5.0.4. www.cidoc-crm.org/docs/cidoc_crm_version_5.0.4.pdf.
- Signore, Oreste (2008): The semantic web and cultural heritage: ontologies and technologies help in accessing museum information. In: Robering, Klaus (ed.): Information technology for the virtual museum: museology and the semantic web. (= Semiotik der Kultur / Semiotics of Culture). Berlin/Münster: LIT, 1-32.

A web-based application for editing manuscripts

Abstract

The present paper's ambition is to briefly present the current state of a web-based linguistic application for annotating and adding meta-data (e.g. commentary) to manuscripts. The tool provides a wide set of possibilities to link parts of images to any kind of textual information, as well as comprehensive search options on that information according to a variety of research needs.

In order to condense the paper to the relevance and advantages for research in the Humanities, the application's features are described herein without detailing their technical background.

1. Introduction

The project "Caucasian Palimpsests" as part of the LOEWE research cluster "Digital Humanities" at the Goethe-University Frankfurt is aimed at analysing and editing palimpsests with the aid of a special web-based application. According to the project description, the application should be versatile and therefore applicable irrespective of the particular investigator's demands. The development of the utility is the project's first phase; its state of April 2014 is described in chapters two and three. In order to point out the benefits of the web-based tool, only basic information about the technical background (e.g. programming languages, database system, etc.) is provided in chapter four.

Prior to my work on this project, the Caucasian palimpsests and their different layers of text were digitized using a special multispectral camera that takes pictures using different wavelengths of visible and non-visible light. Afterwards, the pictures were digitally edited using special software to provide the best readable version of the page.

2. The application

To achieve the project's aims of creating an ideal tool for analysing and editing manuscripts, i.e., one that can be used by different investigators regardless of their objectives, and which hence offers wide usability, the web-based applica-

tion allows researchers to annotate simple and complex manuscripts (e.g. letters, palimpsests), engravings, paintings and pictures of objects of any kind (e.g. amphorae, sheet music, whole archaeological sites, etc.) with textual information in a comfortable and user-friendly way. This is done by assigning text comments to respective parts of the image.

In order to do this, different shapes (rectangles, circles, polygons) can be drawn onto the image by hand (via mouse-clicks) or by copying existing shapes. Then, each shape can be related to one annotation field to which multiple lines with additional information can be related. Each of these sublines can be named individually to provide wide usability for different needs: e.g., original transcription, transliteration, glossing, translation, alternative translation, comment, etc. This is partly shown in Figure 1, which illustrates the main part of the application in its April 2014 status.

Selecting either the shape or the text field highlights both simultaneously, thus visually directing one's attention to their relationship (see Fig. 1). Additionally, every part of the image that is not highlighted by a shape will appear slightly transparent. Besides the mentioned graphical features, a tool for zooming is also included, allowing for a closer look at parts of the image that are otherwise hard to read.

The application also supports right-to-left alphabets, wild-cards for searching for word-forms, morphemes, etc., and grouping of shapes.

Providing a diversity of usability also means that resizing, moving, adding and deleting shapes is possible at any time. Similarly, changing textual information within any field is easily done, as well as adding and deleting fields and sublines. The whole application works with the Unicode character set in *UTF-8* representation and therefore supports over 110,000 characters and signs,¹ further satisfying anyone's needs regarding editing textual information independent of country- or language-specific character encodings. Thus, there is almost no restriction to the type of project that can be realized with this application, because Unicode supports even symbols for musical notes, etc.

All annotations and shapes, as well as their relationships to one another, are stored in a central database. This allows researchers to search for whole words, word forms, allomorphs and/or phrases throughout the whole corpus within

¹ www.unicode.org/.

seconds.² Depending on the chosen search option, one is able to perform either a full-text search or a search only within the pre-selected subline(s) (e.g., looking for matches in a subline labeled “alternative translation” only, etc.). This feature saves a significant amount of time and allows one to find otherwise unnoticeable relationships in and between manuscripts.

To use the application, one needs only to create a new project with the included project management screen and upload all the pictures one intends to work with into the project folder. Afterwards, the software stores the files onto the hard disk. Once multiple images have been uploaded to the system, a specific image can be located and chosen by aid of automatically created thumbnails which simplify selecting the desired picture for annotation in the following concrete steps of editing. This also simplifies administrating the edited work later by allowing for browsing through the project, since a quick preview of the image’s smaller version can be viewed alongside other images of the same project.

To prevent viewers from editing or deleting data by accident, a user rights management system secures that only project members are allowed to edit the data, while others only have the right to read/view the data. This allows researchers either to restrict the access to their work to selected groups or to publish the data to general public.

Additionally, users are able to publish a read-only version of the annotation that could be used for citation in publications. This citable version is completely isolated from the annotation’s working version, and its content can therefore neither be changed nor deleted by anybody. It can only be accessed by its unique link that is unlikely to be guessed casually.

3. User-friendly usage

Since the application is web-based, users do not need to install any software or plug-ins. The cross-browser layout ensures that major browsers display and handle the application and its parts in the same fashion, so that users of, e.g., different Linux-distributions are not limited or restricted in using it compared to users of other operating systems.

² The database consisting of individual projects and their contents constitutes a corpus as per the definition of such as “a collection of texts or parts of texts upon which some general linguistic analysis can be conducted [...]” (Lüdeling/Kytö (eds.) 2008: 1).

Annotations can be typed in without the knowledge of XML or other markup languages. As described in section 2, shapes can easily be drawn by hand. These two aspects are advantages over other annotation programs and allow users of any field of research to use the application in a very user-friendly way. The application handles processing and saving the data automatically, as well as importing/exporting from/to XML.

Furthermore, the relation between one part of the image and the related text part (i.e. annotation) can be set automatically while drawing a new shape, since the application recognizes the first text column that is not yet linked to any shape. This feature saves time during annotation whenever text is already present (i.e., after import or adding it by hand).

For an easier input of different keyboard-layouts (e.g. the official one for Egyptian Arabic), one can be selected and linked to the annotation's text fields. So, installing the keyboard layout on the computer is not necessary. Thus, i.e., using the German keyboard and selecting Egyptian Arabic from the list will print out Arabic <ﺀ> whenever German <ö> is typed. As of April 2014 there are four layouts available for input (German, English (US), English (UK), and Portuguese (Portugal)) and three for output (Modern Georgian, Old Georgian, Egyptian Arabic). More can be added upon request.

4. Technique

The front-end (i.e., the user-interface) is realized in *HTML5*,³ which itself is primarily generated and modified by *JavaScript* in this application.⁴ The web

³ HTML is the basic language that organizes and displays the contents of websites. Its current standard is version 4.01 (www.w3.org/standards/webdesign/htmlcss). The upcoming version, HTML5, is currently under development and only partly supported by modern major browsers (www.w3.org/TR/html5). It is a very powerful new version of this language, supporting several multimedia file types and actions that currently are not usable in pure websites (i.e. without plug-ins). Nevertheless, the most important feature for the application is already supported by all browsers, viz., displaying the element *canvas* for working with images and drawing shapes on them. HTML5 will become a standard and all browsers will support the new elements and features some short time after the release of the standard.

⁴ JavaScript is a (usually browser-inside) script language that was created to extend HTML (www.w3.org/standards/webdesign/script). By aid of JavaScript it is possible to modify existing HTML elements on a website dynamically, as well as to add them.

pages are all written in *PHP*⁵ and contain the output of the basic HTML-code and some JavaScript statements, as well as *SQL*-statements (in *MySQL*)⁶ and file system operations for uploading, moving, and generating images.

Since all major modern browsers support HTML and JavaScript, and since the application does not need any other plug-in to operate, there is no need for these or other software on the client. The client does not need to be a high-end computer, since HTML and JavaScript are processed quite fast (on average) by computers made since 2008. This is one main advantage over other image annotations software that require plug-ins and/or higher computational power. Another advantage is that also the programming languages used on the server are open source, well documented, and supported by a huge community.

As web server, *Apache*⁷ is used for providing and processing the data. For development and testing, it is installed locally on a standard office-computer that serves as client and server. The system has an Intel i5-660 CPU with 3.33 GHZ and 4 GB of RAM. As the operating system, Windows 7 Professional 64bit is used. There was no need for buying a special computer as a development system and/or dedicated (test-)server, since the application does not need much computing power in its current state and usage. The client (i.e., the computer used for accessing the application) need not be a particularly modern high-end machine either. For further developments and especially for hosting more projects online, a dedicated server would be necessary to ensure high computing power and provide an accurate service.

In a nutshell, the software and hardware components of the application are neither high-end nor expensive. Thus, there are no or only very low overhead costs.

⁵ PHP is a very popular script language that is available on almost every web hosting service (<http://php.net>). It can be used for generating and submitting HTML and JavaScript-code, accessing databases and the file system, generating PDFs, images, etc.

⁶ MySQL is a dialect of SQL, which is a language for creating databases and their entries as well as for accessing them. It is very common since it is free, easy to learn and administrate, and very powerful (<http://dev.mysql.com>).

⁷ Apache is a very popular and powerful free web server that enables one to involve database languages, programming languages, security techniques and self-created programs for websites and web-applications (<http://apache.org>).

If the need for adopting the application to other applications exists, this could be easily realized since (popular) open source software has been used to develop it. The programming languages are able to successfully interact with others, too. For importing/exporting data, XML would be the format of choice.

5. Public demo

The application is temporarily named *ImAnTo* (*Image Annotation Tool*). A public demo is available at <http://imanto.manuelraaf.de>. After a short registration one is able to test most of the application's functions. A short user manual is also provided for downloading. The current layout is meant for developing and testing purposes only and will be replaced by a user-friendly one at the end of the project. It is planned to offer the application for free on a dedicated server after the end of the project's current phase. Additionally, interested users are also allowed to set up the application on their local computer or intranet.

References

Lüdeling, Anke/Kytö, Merja (eds.) (2008): *Corpus linguistics: an international handbook*, Vol. 1. (= *Handbücher zur Sprach- und Kommunikationswissenschaft* 29.1). Berlin etc.: de Gruyter.

All web-sources were accessed and checked last on February 16, 2015.

Text mining in the Humanities – A plea for research infrastructures

Abstract

Research infrastructures for the Humanities can help to share digital resources and content services. In particular, they can help researchers in the Digital Humanities to save time and efforts when developing software to deal with specific research issues. Web services and web applications can be used to build a research infrastructure for sharing data and algorithms. However, the development of such infrastructures and their key software components is a software engineering task that increasingly also poses interesting and challenging research problems for Computer Science.

1. Text, knowledge, and the Humanities

As manifold as the usages of language are the purposes of text. But when looking at text in the Humanities, it looks to me as a Computer Scientist that we are, broadly speaking, always assuming that the texts we are interested in are encodings of knowledge (of a culture at a time). And this is what makes texts the subject of analysis: By looking at texts (and sometimes also at their context of origin) we intend to decipher the knowledge that they are encoding.

Looking at texts from a bird's eye view or taking a close reading perspective has always been the core business of text oriented Humanities. With the advent of Digital Humanities, however, we can scale up this task by using new analysis tools derived from the area of information retrieval and text mining. Thereby all kinds of historically oriented text sciences as well as all sciences that work with historical or present day texts and documents are enabled to ask completely new questions and deal with text in a new manner. In detail, these methods concern, amongst others,

- the qualitative improvement of the digital sources (standardization of spelling and spelling correction, unambiguous identification of authors and sources, marking of quotes and references, temporal classification of texts, etc.);
- the quantity and structure of sources that can be processed at scale (processing of very large amounts of text, structuring by time, place, authors, contents and topics, comments from colleagues and other editions, etc.);

- the kind and quality of the analysis (broad data driven studies, strict bottom-up approach by using text mining tools, integration of community networking approaches, contextualization of data, etc.).

While Computer Science and Humanities so far have acted in their working methodologies more as antipodes rather than focusing on the potential synergies, with the advent of Digital Humanities we enter a new area of interaction between the two disciplines. For the Humanities the use of computer based methods may lead to more efficient research (where possible) and the raising of new questions that without such methods could not have been dealt with. For Computer Science, turning towards the Humanities as an area of application may pose new problems that also lead to rethinking present approaches hitherto favored by Computer Science and developing new solutions that help to advance Computer Science also in other areas of media oriented applications. But most of these solutions at present are restricted to individual projects and do not allow the scientific community in the Digital Humanities to benefit from advances in other areas of Computer Science like Visual Analytics.

In consequence, I think it is important that we distinguish between two important aspects:

- (1) the creation, dissemination, and use of digital repositories, and
- (2) the computer based analysis of digital repositories using advanced computational and algorithmic methods.

While the first has originally been triggered by the Humanities and is commonly known as Digital Humanities, the second implies a dominance of computational aspects and might thus be called Computational Humanities.

To distinguish between both aspects has substantial implications on the actual work carried out. Considering the know-how of researchers and their organizational attachment to either Humanities or Computer Science departments, their research can either be more focused on just the creation and use of digital repositories, or on real program development in the Humanities as an area of applied Computer Science, as is illustrated by Figure 1.

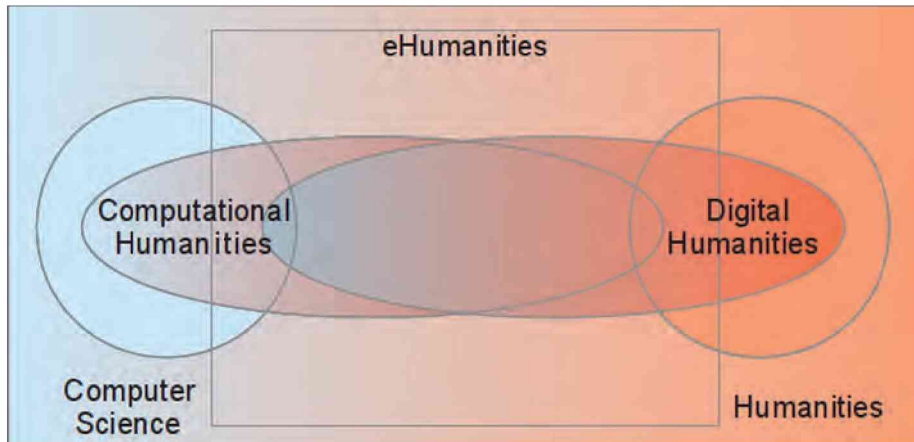


Figure 1: Positioning of Computational and Digital Humanities in the context of Computer Science and Humanities

A practical consequence also in organizational terms of this way of looking at things would be to set up research groups in both scientific communities, Computer Science and Humanities. The degree of mutual understanding of research issues, technical feasibility and scientific relevance of research results will be much higher in the area of overlap between the Computational and Digital Humanities than with any intersection between Computer Science and the Humanities.

2. Research infrastructures for the Humanities

Now, in order to use such computational methods, an individual researcher can proceed by employing two strategies depending on his, or her, own degree of computer literacy. On the one hand, there is the individual software approach. Given a selection of digital text data, the research question is being transferred into a set of issues and methods that can be dealt with by a number of individual programs. This approach allows for a highly dynamic and individual development of research issues. It requires, however, a high degree of software engineering know-how. On the other hand, there is the approach to use standard software. For well-defined and frequently encountered tasks, a Research Infrastructure will offer solutions that provide the users with data and analysis tools that are well understood, have already delivered convincing results, and can be learnt without too much effort.

Both approaches are interdependent. Probably good solutions in one domain of text oriented Humanities can be transferred to other domains by just applying these methods to different kinds of text. A good infrastructure must be capable of making such solutions accessible as best practices.

Research infrastructures are concerned with the systematic and structured acquisition, generation, processing, administration, presentation, reuse, and publication of contents. Content services make available the resources and programs needed for that. Public digital text and data resources are linked together and made accessible by common standards. New software architectures integrate digital resources and processing tools to develop new and better access to digital contents. A good example at hand is the ESFRI initiative CLARIN that aims at “Providing linguistic data, tools and services in an integrated, interoperable and scalable infrastructure for the social sciences and humanities”¹. CLARIN-D, the German sub-project, is primarily designed as a distributed, centers based project. This means that centers are at the heart of an infrastructure that aims at providing persistent data services (see also Boehlke/Heyer/Wittenburg 2013). Different types of resource centers form the backbone of the infrastructure and provide access to data and metadata and/or run infrastructure services. Access to data, metadata and infrastructure services is usually (but not solely) based on web services and web applications. The protocols and formats of infrastructure services like persistent identifiers or metadata systems and standards have already been agreed upon in an early stage of the project. Additional infrastructure or discipline specific services are built upon those basic infrastructure services. The usage of general services like registering and resolving persistent identifiers, however, is not limited to CLARIN itself. The usage of such common services by other infrastructure initiatives is intended and already in place.

Research infrastructures for the Humanities can also help to reconcile the current debate of where the Digital Humanities should be institutionalized – individual Humanities departments or, more central, at Computer Science. On the one hand, there will clearly be no gain for the scientific community in Digital Humanities as a whole when know-how will be duplicated at different Humanities departments. On the other hand, it will be difficult to foster Digital Humanities applications in the Humanities, and to develop new ones, unless the digital research is driven by the Humanities themselves. Deciding for one

¹ <http://de.clarin.eu/en/home-en.html> (last accessed: January 29, 2015).

or the other alternative is no solution, as either way will be hampered by massive drawbacks. However, the building, furnishing, and maintaining of a research infrastructure for the Humanities clearly is a task for Computer Science, while the use of the research infrastructure clearly must be left to the individual researchers and communities in the Humanities themselves. This way we obtain a division of labour that has proven to be most useful in other areas of applied Computer Science and that can lead to substantial and rich contributions of the Humanities to the field of Digital Humanities without getting involved in programming and Computer Science beyond necessity.

3. Infrastructure text mining services

Making language resources and language processing tools available as services enables researchers and developers to use exactly the amount and kind of data that is needed for a specific application. As an example from the area of linguistics, let us finally consider the Webservice access to digital text and lexical data as well as NLP algorithms and tools that was established at the Natural Language Processing Department of Leipzig University already in 2004 (Biemann et. al. 2007, Böhler/Heyer 2009). These services, Leipzig Linguistic Services (LLS) for short, comprise, amongst others,

- a very large, frequency sorted dictionary of German word forms including POS information, sample sentences and co-occurrences,
- monolingual corpora of standard size for currently 48 different languages,
- a tool for sentence boundary detection,
- graph based clustering,
- co-occurrence statistics,
- synonyms and similar words computed on co-occurrence profiles of words,
- automatic terminology extraction, and
- named entity recognition.

Let us assume that a scientist wants to use one of these tools, viz. the named entity recognition, on specific parts of a collection of texts that is encoded in TEI-P5. The task at hand is to extract the needed information from the TEI-P5 document collection, encode it in a way the Named Entity Recognizer web service is able to work on, fetch and interpret the results and probably also to perform a manual correction. In the end, the result is intended to be published

in a way that allows other scientists to validate or reuse the process through reiteration.

Research infrastructures give support on multiple levels depending on the know how of the researcher. Instead of installing, configuring and using a set of offline tools, a scientist who is part of the Digital Humanities community is able to work with programmes that are provided in the form of existing web applications facilitating the usage of low level functionality like converting TEI-P5 documents into simple text or other formats, invoking a Named Entity Recognizer webservice, sending the results to an online annotation platform and archiving the results in a repository. Researches with this level of know-how may be limited in the usage of research infrastructures since they mostly have to rely on the existence of all those tools that make up the single steps of the workflow described above. But they do not need to maintain their own local software stack and they are also able to work in an environment that allows their research to be reproducible.

A scientist who works in Computational Humanities may tap the full technical potential of research infrastructures by creating himself new web services and bundling existing and new application specific algorithms in web applications. This process results in new or alternative workflows becoming available for the whole scientific community. When doing so, the planning, implementation, and deployment process is getting more efficient due to the fact that it is possible to build upon the basic functionality that the research infrastructure provides. Just to name a few:

- Instead of implementing, deploying and hosting one's own version of a simple storage facility that allows to store intermediate results, a common workspace concept of the research infrastructure can be used that is compatible with other tools and services deployed in the infrastructure.
- Exhaustive documentation on how to generate metadata and how to plug services into the infrastructure is available, reducing the time needed to interact with other components.
- Basic concepts and services that allow to make workflows reproducible are in place (e.g. PID systems).
- The question on where to host services that cannot be run by the researcher himself (due to lack of hardware, legal reasons, lack of time to provide long-term support, ...) is answered.

References

- Biemann, Chris/ Heyer, Gerhard/Quasthoff, Uwe/Richter, Matthias (2007): The Leipzig Corpora Collection: monolingual corpora of standard size. In: Proceedings of Corpus Linguistics 2007, Birmingham, UK. www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2007/190Paper.pdf (last accessed: January 29, 2015).
- Boehlke, Volker/Heyer, Gerhard/Wittenburg, Peter (2013): IT-based research infrastructures for the Humanities and Social Sciences – Developments, examples, standards, and technology. In: *it – Information Technology* 55(1): 26-33.
- Büchler, Marco/Heyer, Gerhard (2009): Leipzig Linguistic Services – A 4 years summary of providing linguistic web services. In: Heyer, Gerhard (ed.): *Text Mining Services – Building and applying text mining based service infrastructures in research and industry.* (= Leipziger Beiträge zur Informatik XIV). Leipzig: Universität Leipzig, 55-65.