Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache

Corpus Linguistics and Interdisciplinary Perspectives on Language

Bd./Vol. 3

Paul Bennett / Martin Durrell
Silke Scheible / Richard J. Whitt (eds.)

New Methods in Historical Corpora

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.

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

Layout: Andreas Scholz, Essen; www.a-shots.de
Redaktion: Dr. Jacqueline Mullender, Stourbridge UK
Printed in Germany

ISSN 2191-9577
ISBN 978-3-8233-6760-4

Contents

III. Historical linguistic corpora: Architecture, annotation, and tools

Preface

The papers in this volume constitute a selection from the presentations given at a conference on *New Methods in Historical Corpora* held at the University of Manchester, UK on 29th and 30th April 2011, which was attended by nearly sixty colleagues from ten countries, with four plenary speakers, twenty-six session papers and five posters.

The occasion for the conference was given by the completion of the *GerManC* corpus project at the University of Manchester. This was a three-year project starting in September 2008 which was funded jointly by the UK Economic and Social Research Council (ESRC) and the Arts and Humanities Research Council (AHRC) (grant no. RES-062-23-1118) and led by Professor Martin Durrell and Dr Paul Bennett of the School of Languages, Linguistics and Cultures, with Dr Silke Scheible and Dr Richard J. Whitt as post-doctoral Research Associates. It followed on from a one-year pilot project funded by the ESRC from March 2006 to March 2007 (grant no. RES-000-22-1609), with Professor Martin Durrell as Principal Investigator, Dr Paul Bennett as Co-Investigator, and Dr Astrid Ensslin as a post-doctoral Research Associate.

The ultimate goal of this project was to compile a representative historical corpus of written German for the years 1650-1800 and develop tools for its analysis.[1] This is a crucial period in the development of the language, as the modern standard was formed during it, and competing regional norms were finally eliminated. A central aim was to provide a basis for comparative studies of the development of the grammar and vocabulary of English and German and the way in which they were standardized. The lack of such a corpus for this period of German to facilitate such comparative studies had become apparent in a

[1] General accounts of the project are to be found in: Durrell, Martin/Ensslin, Astrid/Bennett, Paul (2007): The GerManC project. In: Sprache und Datenverarbeitung 31: 71-80 and in: Scheible, Silke/Whitt, Richard J./Durrell, Martin/Bennett, Paul (2011): Investigating diachronic grammatical variation in Early Modern German. Evidence from the GerManC corpus. In: Konopka, Marek/Kubczak, Jacqueline/Mair, Christian/Sticha, Frantisek/Waßner, Ulrich H. (eds.): Grammatik und Korpora 2009. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache 1). Tübingen: Narr, 539-549.

number of doctoral projects in Manchester over the previous few years,[2] and the structure and design of the *GerManC* corpus was specifically intended to parallel that of similar historical linguistic corpora of English, notably the ARCHER and Helsinki corpora.[3] Following these models, the completed *GerManC* corpus contains nearly a million words in total consisting of 2000 word samples from eight genres: drama, newspapers, sermons, personal letters, narrative prose (fiction and biographies), academic, scientific and legal texts, covering the five principal regions in the German speech-area (North, West Central, East Central, South-West ands South-East). It is freely available in a number of versions through the Oxford Text Archive.[4]

In the course of preparing and completing the project the team became increasingly aware of the immense strides which have been made in compiling and developing historical corpora in recent years – and also of the importance of maintaining contact with other allied projects in order to avoid duplication of effort, ensure the optimal use of scarce resources and keep up with the rapid advances in technological development. It is in the nature of historical corpora that they involve methodological problems which can differ substantially from those presented by the compilation of corpora of living languages, and the tools used for analyzing a modern language may be quite unsuitable for the historical stages of the same language. Indeed, it is frequently the case that cross-linguistic perspectives and comparisons with diachronic projects in other languages can be the most beneficial.

The conference in Manchester was thus organized with this in mind, in order to provide a setting where colleagues working on historical corpus projects involving a number of languages could come together, exchange ideas and experiences and establish potentially productive contacts in a relatively small forum in a way not always possible at larger corpus-oriented gatherings, where historical projects can easily be submerged. In this respect it appears to have

[2] In particular those subsequently published as: Storjohann, Petra (2003): A Diachronic Constrastive Lexical Field Analysis of Verbs of Human Locomotion in German and English. Frankfurt et al.: Peter Lang, and Auer, Anita (2009): The Subjunctive in the Age of Prescriptivism. English and German Developments during the Eighteenth Century. Basingstoke: Palgrave Macmillan.

[3] Cf. http://www.llc.manchester.ac.uk/research/projects/archer/ (last accessed 30 June 2012) for information on the latest development of the ARCHER corpus initiated by Douglas Biber and Edward Finegan, and: Kytö, Merja (1996): Manual to the Diachronic Part of the Helsinki Corpus of English Texts: Coding Conventions and Lists of Source Texts. Helsinki: University of Helsinki, available at http://icame.uib.no/hc/ (last accessed 30 June 2012).

[4] URL: http://www.ota.ox.ac.uk/desc/2544 (last accessed 30 June 2012).

been a successful initiative; the work presented in the present volume testifies to the range, vitality and quality of current research in the field and the way in which quite new methodological advances have been achieved in a relatively short period of time.

Like all historical study, the investigation of the history of a language depends on the inevitably fragmentary sources which have come down to us. An electronic historical linguistic corpus thus promises the possibility of at least alleviating this notorious problem of 'bad data'.[5] However, it cannot overcome it, and the crucial questions thus arise – among others – of the optimal architecture for such a corpus, the problem of how representative even a large corpus can be of actual language use at a particular time, and how a historical corpus can best be annotated and provided with tools to maximize its usefulness as a resource for future researchers. In practice, the papers in this volume all attempt to address these central issues, either directly or by illustrating how corpora can be exploited to investigate specific questions in the development of a particular language, and this is reflected in the three major sections into which the papers are divided.

The first section consists of three of the longer papers which were given as plenaries and deal with fundamental issues of corpus structure, analysis or annotation. **David Denison** ("Grammatical mark-up: Some more demarcation disputes") discusses how problems can arise because of the inflexibility of some standard tagsets, which cannot cope with ambiguous or underdetermined forms – in effect, as he puts it, "what you see is what your theory can handle" – and this can be particularly difficult in a historical corpus, where the function of a particular form may be in flux. **Terttu Nevalainen** ("English historical corpora in transition: from new tools to legacy corpora?") looks back over the twenty years since the first Helsinki corpus of English and, with examples, discusses the respective merits and uses of small-scale corpora and more recent "mega-corpora" and the trade-off between corpus annotation and corpus size. **Klaus-Peter Wegera** ("Language data exploitation: design and analysis of historical language corpora") discusses, with specific reference to the Bonn corpus of Early New High German, the fundamental distinction between a "corpus" and a "text archive", and outlines the fundamental importance of a systematically structured representative corpus for the investigation of the historical development of a language.

[5] Cf. Labov, William (1994): Principles of Linguistic Change. Internal Factors. Oxford: Blackwell, p. 11.

Section two, *Analysis, projects, and results*, contains ten papers which principally describe ongoing or proposed corpus projects and discuss empirical findings from them in a theoretical context. Although many also raise the technical issues which are central to the papers in section three, this is not their major focus. They can be introduced here thematically, in terms of the issues they raise. It can be seen, first of all, that they provide a clear illustration of the distinction between historical corpora which have been compiled to investigate specific research questions and those which are more general. The paper by **Anita Auer** and **Tony Fairman** ("Letters of artisans and the labouring poor (England, *c.* 1750-1835)") is a prime example of the first of these, as an account of a corpus of writing by a group which has hitherto been neglected in accounts of the historical development of English, as well as the problems involved in developing adequate tools for searching a corpus which exhibits considerable linguistic variation. The paper by **Britta Juska-Bacher** and **Cerstin Mahlow** ("Phraseological change – a book with seven seals? Tracing the diachronic development of German proverbs and idioms by a combination of corpus and dictionary analyses") outlines a project which aims to investigate changes in the structure of set phrases and idioms in German combining a study of available corpora with data gleaned from dictionaries. A characteristic of the earlier stages of western European languages is the competition between Latin and the vernaculars, and **Olga Timofeeva** ("Anglo-Latin and Old English. A case for integrated bilingual corpus studies of Anglo-Saxon registers") outlines the desirability of compiling a corpus of Latin from Old English sources in order to gain a more rounded picture of the extent of linguistic contact phenomena in pre-conquest English. Finally, **Stefania Degaetano-Ortlieb et al.**, ("*SciTex:* a diachronic corpus for analyzing the development of scientific registers") present a genre-specific English corpus covering writing in a range of scientific fields from computer science to micro-electronics from the 1970's to 2000 which is analyzed using a Hallidayan theoretical framework.

A further group in this section consists of studies based on larger corpora. Bible translations have long been exploited for diachronic linguistic studies, and **Andrés Enrique Arias** ("On the usefulness of using parallel texts in diachronic investigations. Insights from a parallel corpus of Spanish medieval Bible translations") shows how valuable insights can be obtained by incorporating a set of Bible translations from different periods of Castillian Spanish in an electronic corpus. A possibility for corpus-based diachronic study of English is now giv-

en by the fact that synchronic corpora are now available for discrete periods in the twentieth century, and **Melanie Röthlisberger** and **Gerold Schneider** ("*Of*-genitive versus *s*-genitive. A corpus-based analysis of possessive constructions in 20th century American English") use the various stages of the Brown corpus to trace the variation over time of the alternative means of expressing possessives in written American English.

A significant and welcome development in recent years has been the compilation of corpora of less widely spoken languages or regional varieties. **Tomaž Erjavec** and **Alenka Jelovšek** ("A corpus-based diachronic analysis of Slovene clitics") describe the annotated historical corpus of Slovene and how it can be searched to analyze the development of Slovene clitics. The linguistic situation in the Grand Duchy of Luxembourg is notoriously complex, and **Peter Gilles** and **Evelyn Ziegler** ("The *Historical Luxembourgish Bilingual Database of Public Notices*") show how a parallel corpus of public notices can throw light on the historical development of this situation, provide a resource for studies in contrastive and contact linguistics, and demonstrate the value of parallel corpora. Using the methodology of "lexical bundles" developed by Douglas Biber, **Joanna Kopaczyck** ("Formulaicity in Scots historical corpora and the lexical bundles method") investigates formulaic patterns in legal and administrative texts in Scots on the basis of a collection of legal documents from medieval and early modern Scotland, compiled from three available electronic corpora. Finally, **Javier Ruano-García et al.** ("Past tense BE forms in Late Modern Lancashire English. A preliminary corpus-based approach") introduce the dia-chronic corpus of dialectal English currently being compiled in Salamanca and examine the variation between *was* and *were* in Lancashire English in the eighteenth and nineteenth centuries.

The third section, *Historical linguistic corpora: Architecture, annotation and tools*, consists of six papers which are principally concerned with technical aspects of corpus compilation and analysis. Although some of them also discuss empirical linguistic findings, this is not their major focus in the way it is for the papers in section two. Three of the papers deal with projects involving earlier stages of German. **Mathilde Hennig** ("The *Kassel Corpus of Clause Linking*") presents a German diachronic corpus project which includes texts exemplifying "immediacy" and "distance" from the seventeenth and nineteenth centuries and shows how these could be annotated to identify correlations between single grammatical features and types of clause linking. **Bryan Jurish et al.** ("Constructing a canonicalized corpus of historical German by text align-

ment") address one of the major problems encountered in annotating historical corpora, i.e. that of adequately lemmatizing a corpus of historical texts with hugely variable spellings, which they are able to achieve automatically by aligning the historical texts with current editions of the same texts. **Sonja Linde** and **Roland Mittmann**, ("Old German reference corpus: Digitizing the knowledge of the 19th century. Automated pre-annotation using digitized historical glossaries") outline how digitizing nineteenth century glossaries of older German texts can be exploited in order to expedite annotation in terms of morpho-syntactic features.

Two further papers in this section deal with Latin, which, interestingly, was the first historical language for which an electronic corpus was compiled. **Barbara McGillivray** and **Adam Kilgarriff** ("Tools for historical corpus research, and a corpus of Latin") give an account of the *LatinISE* corpus with some thirteen million tokens. They explain how it has been automatically lemmatized and tagged, and how the Sketch Engine search tool, into which it has been uploaded, has been adapted to to meet the needs of historical corpus research. **Alexander Mehler et al.** ("Inducing linguistic networks from historical corpora. Towards a new method in historical semantics") take the nineteenth century collection *Patrologia Latina*, a corpus of Late Latin texts and show how a systematic study of the networks of association with a particular word (in this case *virtus* 'virtue') using sophisticated mathematical models can throw light on its diachronic semantic development. Finally, **Achim Stein** and **Sophie Prévost**, ("Syntactic annotation of medieval texts. The *Syntactic Reference Corpus of Medieval French* (*SRCMF*)"), demonstrate how a syntactic corpus, annotated according to the principles of dependency grammar, has been compiled from two earlier text corpora of Old French.

These papers clearly illustrate the rapidity of the progress which has been achieved in respect of the compilation and annotation of historical corpora since the earliest days of simple digitization of complete texts, and it is indicative that many of them explicitly state that they involve work in progress. In this way, the present volume, like the conference on which it was based, not only constitutes a snapshot of current development, but also points the way forward to future advances.

It remains for the editors to acknowledge with gratitude the help and assistance provided by all those involved in the organization of the conference and the production of the volume. First and foremost, naturally, there are the Eco-

nomic and Social Research Council (ESRC) and the Arts and Humanities Research Council (AHRC) who provided the core funding for the conference within the framework of the *GerManC* project, as well as the School of Languages, Linguistics and Cultures at the University of Manchester for providing a beneficial and productive research environment. We must also thank all the individual authors for meeting unrealistically tight submission deadlines, as well as those colleagues in Europe, North America and Asia who refereed the contributions and made so many helpful suggestions (but who must naturally remain anonymous). We must also give particular thanks to Professor Ludwig M. Eichinger, of the *Institut für Deutsche Sprache* (IDS) in Mannheim, the general editors of the CLIP series, in particular Dr Marc Kupietz, and the team in the publication office at the IDS for their support in the production of this volume.

Paul Bennett
Martin Durrell
Silke Scheible
Richard J. Whitt

Manchester, June 2013

I. Historical linguistic corpora: Some fundamental issues

DAVID DENISON

Grammatical mark-up: Some more demarcation disputes[1]

Abstract

A selective tour of annotation in historical corpora begins with extra-linguistic mark-up: how far can it alert the corpus user to usage which is atypical of the variety being sampled? Several syntactic fossils are discussed, and a playful use of foreign and pseudo-foreign words. Are they a kind of code-switching? In the former case the answer No is given, in the latter a partial Yes.

As for grammatical mark-up, with few exceptions a given scheme must privilege one particular analysis for each word, sentence or other unit of analysis. Special tags are available in the *CLAWS* and *Penn Treebank* tagsets for cases which remain ambiguous but which are in principle decidable. 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. Historical corpora like the *Penn* family aim also for consistency of analysis. I argue that both principles can be problematic.

Consider first the push towards a unique POS tag for every word. I propose that certain kinds of word are vague as to their word class not because of a failure of analysis but because they are genuinely underdetermined. Vagueness is not ambiguity, so ambiguity tags would be inappropriate – at least with their currently intended values. Secondly, the desideratum of consistency does not allow for patterns which arguably have dual analyses synchronically, nor for items which are in transition or which have changed over the time-span of a historical corpus. Among the data discussed are the POS-tagging and parsing of adjectives derived from passive participles (*interested, amused*), multi-word prepositions (*on behalf of*), phrasal and prepositional verbs (*run over*), proper-to-common-noun conversions and noun-to-adjective transitions (*BandAid*), countable-to-mass conversions (*He looked at me across a vast expanse of table*) and the converse (*two coffees*). A brief conclusion argues that while some of the problems considered are statistically unimportant, others demand greater flexibility of mark-up.

[1] I am grateful to Hans Martin Lehmann, Gerold Schneider and Nick Smith for help in the preparation of the oral version of this paper, and to Marianne Hundt for commenting on a written draft as well. The usual disclaimers apply.

1. Introduction

In a previous paper (Denison 2007) I offered five detailed case studies of morphosyntactic tagging in English corpora. The focus there was on areas of English lexis and grammar which posed problems for tagging because items fell near category boundaries. Here I will briefly take up similar issues with different data and corpora, and extend the discussion to metadata or extra-linguistic annotation. I look at kinds of variation which are not generally well served by corpus mark-up, and ask how – or whether – the annotation could be made more helpful. Most of the problems identified are consequences of language change, but even corpora specifically designed for diachronic research are not immune.

2. One variety at a time?

Texts in corpora are generally labelled by date, genre, and so on, and information may be given on dialect, speaker, etc. Nevertheless, this hides much variation within a text. For example, speakers may tell jokes in an accent other than their own, novelists may attempt to recreate a period, sometimes earlier (or later!) than the present, and so on. The extra-linguistic mark-up cannot follow all such twists and turns. Most will admittedly be of minor importance for studies looking for statistical effects across a large corpus, but non-native analysts may be misled in the discussion of individual examples.

2.1 Date

Language changes over time, but not homogeneously. Corpus texts, just like everyday speech, can be littered with novel usages which go beyond the norms of their time, and equally they may harbour usages which are – strictly speaking – no longer current. Two related concerns, then, are the effect on our understanding of linguistic history, and how far linguistic mark-up can or even should reflect such chronological layering. Consider this simple sentence from the *British National Corpus* (*BNC*):

(1) How goes it, Bruce? (AB9 7)

This apparent example of V2 syntax appears in a text dated "1985-93" in the *BNC*. It is clearly a fossil – a self-conscious archaism or perhaps foreignism, now established as a kind of salutation. The usage may well be supported by another *How V ...* inversion type:

(2) How come you're homeless anyway? (A0F 1551)

Nevertheless the V2 pattern for interrogatives is one which has generally disappeared for most lexical verbs since the seventeenth century, and both (1) and (2) are idioms which are too idiosyncratic to tell us much about the productive syntax of Present Day English (PDE).

Fossil syntax is surprisingly common. An apparent example of the so-called sentence brace is seen in:

(3) A chemical does not a product make (PV 564)

Example (3) is a creative variation on a fossil, a familiar proverb, (4), in turn a fairly common variant of the more normal (5) (ignoring spelling variations), which was translated from Aristotle into English by the 16th century; see here Speake (ed.) (2003):

(4) One swallow does not a summer make.

(5) One swallow doth not make/does not make/maketh not (a) summer.

Starting, then, from some form of the proverb like (5), the variant (4) is probably a misquoted poetic archaism[2] of long standing, and example (3) is what is now styled a "snowclone" (Pullum 2004) – that is, an adaptation of a voguish phrase (whether archaic or not) by the substitution of different lexical items in a fixed template. It is far from obvious how to mark snowclones linguistically in a corpus, as it is the template that is in effect a prefab rather than any one idiomatic string.

The point here is that (3) is somewhat inconvenient. The sentence brace was current in prose until the early Middle English period, still fairly common in later Middle English but in steep decline in prose by the 16th century (van der Wurff/Foster 1997). Corpus users surely expect to find a clear marking of date for the examples in a corpus, but the existence of such diachronic layers within a synchronic grammar adds an undesirable complication which is not easily conveyed in metadata.

[2] *EEBO* records "Yet the old prouerbe long agoe thus spake, | One swallow yet did neuer summer make" from William Painter, *Chaucer newly painted* (1623), while *LION* has "One swallow (they say) no Sommer doth make. | Some swallow (I say) till great heat they take" from John Davies, *The scourge of folly* (1611).

2.2 Code-switching

Switching from the base language into a foreign language is routinely marked in many corpora, for example the *Helsinki Corpus of English Texts*:

> [a word or phrase] in languages other than English was annotated by surrounding it with the code (\...\) in the original version. In the TEI XML Edition, this code is replaced by the foreign element. (Marttila 2011: section 3.2.4)

This is obviously helpful. If at some point the language stops being English, users need to know – whether in order to discount the foreign word(s) or to study the process of code-switching. However, it is not always straightforward to add such annotation. This example comes from a small corpus I directed:

(6) I think if I can work that incident up a little it will form a very fitting dénoument to my unhappy "Mme de V." wh: <foreign>(en passant)</foreign> I may mention is likely to be fair copied about the A.D. 1900. This must stand, <foreign>mon cher</foreign>, for the Sunday edition & entreats an answer. (1890 Ernest Dowson, from *Corpus of lModE Prose* [1994], mark-up altered to XML type)

Dowson playfully Frenchifies his English, and as corpus compilers we had to decide which of his lexical choices, and indeed which of his sometimes fanciful spellings, to code as "foreign". How much mark-up is appropriate?

Arguably some fossils and the kinds of creative usage to be discussed in Section 5 below could be marked as code-switching too. Could switching out of 1980s English into what is apparently a different English be seen as the same in principle as switching into a foreign language? Probably not: unlike normal code-switching, comprehensibility for the wider speech community is maintained, not just for the immediate interlocutor. Anyway, given that language is **always** a mixture of rule-governed productivity, prefabs and creative extensions of rules, it is a reasonable abstraction to say that overall a corpus text "**is**" (an example of) the language of a certain date, genre, dialect – that is, that it can be taken to represent the range of possibilities of what is essentially one variety. (We should note too that the advent of World Englishes makes it even more impractical to treat creativity as code-switching.)

3. Underspecification

3.1 Vagueness vs. ambiguity

Grammatical tagging aims to assign word classes precisely; in fact tagsets routinely label forms even **more** specifically than the usual parts of speech. *CLAWS* C5 has 57 basic tags, for example, and C8 rather more. Ambiguity is the situation when the hearer/reader cannot be sure which of two or more readings was intended by the speaker/writer but does know that it must have been one or other, and the distinction affects the interpretation of the sentence. Now taggers are like reader/hearers in that they too have to figure out the correct interpretation and analysis of a sentence, and sometimes they cannot be sure. Some tagsets allow for this eventuality. The *BNC* has 30 **ambiguity tags** (28 listed), including AJ0-NN1 and NN1-AJ0 (adjective or noun), AJ0-AV0 and AV0-AJ0 (adjective or adverb), but these are intended as stopgaps, for use "when the probabilities assigned by the *CLAWS* automatic tagger to its first and second choice tags were considered too low for reliable disambiguation" (Leech/Smith 2000). The detailed discussion of **disambiguation** suggests that in principle, manual post-editing could replace an ambiguity tag with the correct single tag. Apparently similar in concept are the **multiple tags** in the *Penn Treebank* tagset (Marcus/Santorini/Marcinkiewicz 1993: 316).

In Denison (in prep.) I am proposing that some words and longer grammatical strings do not have a unique word class, not because of a failure of analysis but because they are genuinely underdetermined: they are syntactically **vague**. Examples include certain occurrences of

(7) diverse, various, certain, several (A ~ D)

(8) (look) sad, (look) sadly, ... (Adv ~ A)

(9) near, worth, like, ... (A ~ P)

(10) fun, key, draft, genius, ... (N ~ A)

In the appropriate contexts the word class of the above items is underdetermined between the two classes indicated in the brackets, so the analysis of the containing sentence is also vague. Whereas the producer of an ambiguous sentence must have intended one or other of the possible readings, a vague sentence is syntactically underdetermined for both producer and recipient. Vagueness and ambiguity are quite distinct.

It looks at first as if the *Penn Treebank* does recognize vagueness:

> We do not distinguish between verbal and adjectival uses of present and past participles, tagging both uses as VAG and VAN, respectively. (Santorini 2010)

But the fuller quotation implies that this is more likely to be avoidance of ambiguity resolution than a claim that two analyses are indistinguishable in principle:

> We have tried to plan our system so that at each stage of the annotation, information is added in a monotonic way. In particular, we want any future revisions of the bracketed structures always to add information, never to change it. This goal requires us to avoid subjective judgments since they are extremely error-prone. So, for example, we do not distinguish adjectival from verbal passive participles, nor do we attempt to implement the argument-adjunct distinction.

Here are two analyses from the *Penn Parsed Corpus of Modern British English* (*PPCMBE*) with, respectively, a verbal and an adjectival use of *pleasing*, both marked with the POS tag "VAG":

(11) and devoted herself to **pleasing** and entertaining him (YONGE-1865, 180.535)
[$_{PP}$ [$_{P}$ to] [$_{IP\text{-}PPL}$ [$_{VAG}$ [$_{VAG}$ pleasing] [$_{CONJ}$ and] [$_{VAG}$ entertaining]] [$_{NP\text{-}OB2}$ [$_{PRO}$ him]]]]

(12) with the most **pleasing** astonishment (GIBBON-1776,1,357.31)
[$_{PP}$ [$_{P}$ with] [$_{NP}$ [$_{D}$ the] [$_{ADJP}$ [$_{QS}$ most] [$_{VAG}$ pleasing]] [$_{N}$ astonishment]]]

The distinction is made in parsing at the phrasal level – IP-PPL vs. ADJP – rather than by tagging at the word level.

3.2 Vagueness of word class

I now turn to an example of word class vagueness. In the *BNC*, *dinosaur(s)* is always tagged as a common noun, either NN1 (sg) or NN2 (pl) (except for the post-punk band *Dinosaur Jr*, which is correctly marked as NP0, a proper noun, when it appears!). The nominal tag NN1 seems perfectly reasonable even for an example like (13):

(13) Are they secretly debunking today's short-sighted rave fashions by reviving the **dinosaur** antics of Tangerine Dream and Focus? (*BNC* CK5 1043)

The syntactic slot occupied by *dinosaur* in (13), premodifier of a noun, is one which can be filled by a noun.

What then would the *CLAWS* tagger have made of the following example, had it occurred in the *BNC*?

(14) Richard represents views that myself and those who work in the business of football find **totally dinosaur**. (2011 Karren Brady, *London Evening Standard*)

Here we see a recent, perhaps nonce development of clear Adjective syntax for *dinosaur*. I argue that N > A changes of this kind come about through stepwise changes, not abrupt, involving "bridge examples" which are systematically vague in category and cannot be definitively assigned either to N or to A (Denison 2001, 2008, in prep.). The word *dinosaur* in the incipient new sense 'embarrassingly outdated' is a suitable candidate. Example (14) is not a bridge example: the N > A trajectory has reached a clear endpoint. If the wholly adjectival use of (14) spreads to more speakers, they would no longer have clear grounds for deciding whether *dinosaur* as premodifier in the *BNC* example, (13), was Adjective or Noun. When using attributive *dinosaur* 'embarrassingly outdated', such speaker/writers and their hearer/readers would not need to decide between the N and A classifications, as nothing at all hinges on the distinction. In short, the existing pattern (13) would become morphosyntactically **vague**, at least for speakers who have both N and A entries for *dinosaur* in their lexicon.

There are two important points being made here. One is that corpus mark-up does not recognize word class vagueness even in principle – and maybe it should. The other is that there may be unique analyses, previously uncontroversial, which ought to be revisited and retrospectively reclassified as vague when a new possibility enters the grammar.

4. Alternative analyses

Corpora with grammatical mark-up do not generally offer alternative analyses of the same sentence within a given annotation scheme.[3] The aim in principle is to find "the" correct analysis. Unique analyses may not always capture the whole truth about the syntax of a sentence, however. I discuss two such pat-

[3] There is also a quite different (and irrelevant) situation, namely where a whole corpus has been processed more than once by different tagging programs. The *ANC* is supplied with three different stand-off tag schemes, while members of the English Department in Zurich can view certain corpora with a choice of tagsets and parses.

terns and only briefly raise the question of whether alternative structural analyses can involve vagueness rather than ambiguity.

4.1 Prepositional verbs

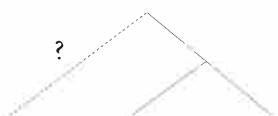
Here once more are two examples from the *PPCMBE*:

(15) I **sent for** one of these doctors (Reade 1863)

(16) But no Doctor was **sent for** till Monday (Nightingale 189x)

This time the tagging is not controversial, but the parsing is open to question. Most syntactic tests suggest that P is a constituent of PP. The *PPCMB*E stays with the PP analysis:

(17)



[$_{NP\text{-}SBJ}$ I] [$_{VBD}$ sent] [$_{PP}$ for [$_{NP}$ one of these doctors]]

(18)

[$_{NP\text{-}SBJ\text{-}1}$ no Doctor] [$_{BED}$ was] [$_{VAN}$ sent] [$_{PP}$ [$_{P}$ for] [$_{NP}$ *-1]]

The tags and parses shown are those of the *PPCMBE*, with partial trees added to draw attention to the constituency of the preposition *for*. The 2nd edition of the *International Corpus of English, Great Britain* (*ICE-GB2*) analyses prepositional verbs in a similar way.

That is not the only possible analysis. The lexical unity of the V + P pair and the existence of a passive lead some scholars to suggest reanalysis (for example Mitchell 1958; Vestergaard 1977; Denison 1985; Quirk et al. 1985):

(19)

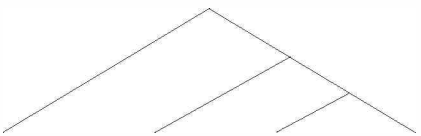
[$_{NP\text{-}SBJ}$ I] [$_{PREP\ VB}$ sent for] [$_{NP}$ one of these doctors]]

As far as I know, such an analysis (or reanalysis) of prepositional verbs has not been used in corpus parsing schemes.

4.2 Complex prepositions

In the *PPCMBE* the phrase *on behalf of his country* (1888 Trollope) is parsed as follows:

(20)



[$_{PP}$ [$_{P}$ on] [$_{NP}$ [$_{N}$ behalf] [$_{PP}$ [$_{P}$ of] [$_{NP}$ his country]]]]

That is, the preposition *on* is head of a PP, with an NP headed by *behalf* as complement. In contrast *ICE-GB2* treats *on behalf of* as a complex preposition with the three words *on, behalf* and *of* "ditto-tagged" (because they function grammatically as a single unit):

(21)

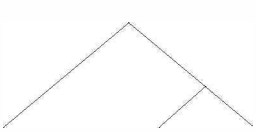
[$_{PP}$ [$_{PREP}$ on behalf of] [$_{NP}$ the people of Britain]]

This is a familiar dilemma, discussed by many scholars (especially Hoffmann 2005). Quirk et al. claim that eight out of nine indicators support a complex preposition analysis (1985: 670-673). Hardliners, on the other hand, find no **syntactic** grounds for recognizing such strings as complex prepositions (e.g. Huddleston/Pullum et al. 2002; Aarts 2007). Perhaps this is another context where a certain principle can usefully be appealed to (Denison 2010: 122):

(22) WYSIWYTCH
What You See Is What Your Theory Can Handle

Now the *on behalf* construction has a variant with a possessive before *behalf*. Even *ICE-GB2* treats *behalf* as a noun in that case, with no ditto-tagging:

(23) engaged **on our behalf** in military action (S2B-030 099)



[$_{PP}$ [$_{P}$ on] [$_{NP}$ [$_{PRON}$ our] [$_{N}$ behalf]]]

Here there is no choice of analysis and no complex preposition.

Returning to the *PPCMBE*, we find that its creators insist on consistency as a guiding principle:

> Although our treatment of fused forms generally reflects their phrasal origin, certain such items must be treated as unitary because of their syntactic distribution. For instance, UNDERHAND must be treated as an adjective because it can appear as a prenominal modifier. [...] Once an item is treated as unitary in one context, it is treated that way consistently. (Santorini 2010)

In their corpus *behalf* is always treated as a noun.

Table 1 lists the relevant occurrences of *behalf*:

Pattern	N
in behalf of X	6
in the behalf of X	2
in X's behalf	6
in that behalf	16
on behalf of X	11
on the behalf of X	1
on X's (own) behalf	6
Total	**48**

Table 1: *behalf* in the *PPCMBE*

Now as it happens, the 11 occurrences of the string *on behalf of* constitute less than a quarter of the 48 occurrences in the corpus. Whatever the motivation, Table 1 suggests that it may have been a good decision not to give *on behalf of* a multi-word analysis in this corpus but always to analyse *behalf* as a separate lexical item: not only is there a choice between *of-X* and *X's*, there is no single fixed form for the *of* pattern.

In the *BNC*, there is even a rare plural (*behalfs* ×2, *behalves* ×3) as against 4014 singular *behalf*. However, the string *on behalf of* occurs 2708 times in the *BNC* and vastly outnumbers *on the behalf of, in behalf of,* etc. The pattern *on X's behalf* (including *on my/our/his behalf*) occurs over 1100 times. Does this too argue against the complex preposition analysis? After all, we could simply be observing the usual choice between poss-s and poss-of constructions (as in *the book's cover* vs. *the cover of the book*), which would be the null hypothesis here. However, as I have argued elsewhere (Denison 2010: 118-22), the variation between poss-s and poss-of in the case of the *on behalf* string is not free variation, because common nouns prefer *of X*, while the examples with *X's* nearly all

involve possessive determiners and proper nouns. The incipient complementary distribution is confirmed in the spoken part of the *BNC* and in the *Diachronic Corpus of Present-Day Spoken English* (*DCPSE*). The two alternative patterns (*on behalf of X* and *on X's behalf*) are increasingly dissociated from each other, and there is indeed increasing lexicalization of the fixed string *on behalf of*.

What kind of mark-up should be used? The *BNC* is in my opinion particularly good here. Every occurrence of the string *on behalf of* is tagged in two different ways at different levels of XML mark-up:[4]

(24) a. [$_{PRP}$ on] [$_{NN1}$ behalf] [$_{PRF}$ of]

b. [$_{PRP}$ on behalf of]

That is, in (24)a we find three words tagged individually as PRP (preposition) + NN1 (singular common noun) + PRF (preposition *of*), whereas in (24)b the whole string is treated as a "multiword" (Leech/Smith 2000) and tagged as a preposition.

The *Corpus of Historical American English* (*COHA*) runs from 1800 or so to the present. It uses the same *CLAWS* tagger as the *BNC* but without the same post-processing, and the tagging of *on behalf of* that is displayed online is the multiword type of (24)b. The *PPCMBE* does not cover much of the twentieth century, stopping at 1914. As we have seen, it effectively tags *on behalf of* analogously to (24)a. In my view, a diachronic corpus covering lModE to the present day or the near future should not be required to apply the same tagging/parse to *on behalf of* throughout the period, *contra* Santorini's principle of consistency quoted above (2010), since the evidence in favour of a multiword analysis has been increasing over time.

Some underdetermined (vague) syntactic patterns – typically the locus of change – merely involve underdetermination of word class (and therefore also of phrasal projection). In other cases, however, I argue for dual analyses (cf. dual inheritance in a Construction Grammar framework). This cannot easily be accommodated in mark-up. The two synchronic situations correspond to diachronic changes that do not and do involve structural change, respectively.

[4] I am grateful to Sebastian Hoffmann for clarification of this point (p.c. 1 May 2012).

5. Language change

One crude dichotomy in diachrony is between abrupt and gradual change. On the whole, grammatical mark-up copes better with abrupt change.

5.1 Abrupt change: count nouns and mass nouns

Here we have a different problem: a kind of rapid linguistic change involving an important morphosyntactic distinction which is rarely traceable via linguistic mark-up. A count noun can be singular or plural and when singular cannot normally form a grammatical NP without a determiner. A mass (non-count) noun has no plural and can form an NP without an overt determiner. The syntax and semantics are significantly different. However, as is well known, there is productive conversion of certain mass nouns to count:

(25) Bring me two **coffees**. (*BNC* A73 2535)

The converse is also found. Here are some *BNC* examples of count nouns with mass noun syntax, following the hints in Matthews (1979: 29-31):

(26) It was real **mood-swing**. (C86 479)

(27) who did not give the impression of a mind of exceptional ability – there was not enough **knife** in the mind (A68 1139)

(28) He knew his son was all **mouth and trousers** (FBG 265)

(29) 'It's slit up each side,' she said showing an expanse of **thigh**. (ACK 604)

Given the possibility of nouns switching allegiance between count and mass subcategories, and given that many NP contexts do not serve to distinguish them at all, it is not surprising that corpus annotation schemes generally do not attempt to mark countability on nouns. Here is what is said about the *BNC* tagset:

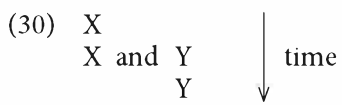
> We make no special distinction between common nouns that can be mass (or "non-count") nouns (eg *water, cheese*), and other common nouns. All are tagged NN1 when singular and NN2 when plural. (Leech/Smith 2000: §2)

(Nouns are marked NN0 if they are morphologically invariant for number, as with *sheep*.) Other tagsets are similar. The *CLAWS* tagger used for the *BNC* subcategorizes nouns by verb concord, or the potential for it, which is easier to operationalize than a distinction based on NP syntax.

The change proper noun > common noun is an abrupt change, a kind of conversion. The *BNC* tags *Xerox* (*Corporation*) as NP0 = proper noun, *xerox* as NN1 = singular common noun. With *Band-aid*, *band-aid* 'wound dressing' it generally uses NN1, even though it is a proprietary name dating back to 1924. This is the familiar process whereby certain brand-names get turned into common nouns. For *Band-Aid* or *BandAid* referring to charity fundraising concerts, it sometimes uses NP0, which is a curious chronological reversal.

5.2 Gradual change

Gradual change is often represented as

(30)


The two co-existent states *X* and *Y* are generally thought of as different forms, but they may equally be underlying analyses, identical in surface form. Mark-up is often less sensitive to gradual change of this type (cf. complex prepositions, N > A, etc.), until the earlier pattern *X* has almost disappeared.

We have already looked at the development of a common noun usage for *band-aid*. A further development gives the word an adjective use. Examples (31) are internet data from *WebCorp* dated 2005-2009:

(31) a. Keeping the heater core for "cooling" is a very **bandaid** approach to [...]

b. it's a very **bandaid** solution to a big problem.

c. OMG..that is so **bandaid**!

Unlike the nonce example of adjectival *dinosaur* in (14), *bandaid* is more firmly established in full adjectival use, as illustrated in (31)a and b, at least. Mostly it is a noun, but in some examples it can only be an adjective. As suggested earlier, the simultaneous existence of two different word class analyses for such a word has consequences for the "bridge examples" where the two word classes are neutralized. They become systematically vague in category and should not be definitively assigned either to N or A.

How do corpora of present-day English and reference works deal with such matters? *OED* recognizes *Band-Aid* as an adjective (though in fact all its examples are premodifiers that are vague between an N and an A analysis). The *BNC* calls it a noun:

(32) the sort of **band-aid** solution (HHX 3069)
[$_{NN1}$ band-aid] [$_{NN1}$ solution]

So does *COCA* and the other Brigham Young University corpora, which use a crude form of the tagging applied to the *BNC* – though *Bandaid* is only tagged as an Adjective in instances where in fact it has been converted to a **verbal** participle!

(33) a. All we're doing is **Band-Aiding** ourselves
(1986 *Time Magazine*)

b. I have **band-aided** it up (*COCA* 2006 *Iowa Review*)

We encounter similar problems with phrasal verbs. Most tagsets have a special tag for the particle of a phrasal verb, e.g. "RP" in the *PPCMBE*, "AVP" in the *BNC*:

(34) and the Gib was **run up** (*PPCMBE* holmes-trial-1749) [*gib* = *jib* (sail)]
[$_{NP\text{-}SBJ}$ [$_{D}$ the] [$_{N}$ Gib]] [$_{BED}$ was] [$_{VAN}$ run] [$_{RP}$ up]

As Santorini explains, the VP is flat: "The trees in the corpora are simply underspecified" (2010). Now tagging always distinguishes phrasal verbs from prepositional verbs, and parsing at least distinguishes absence/presence of PP. However, diachronically they are not always distinct. Compare the treatment in the *PPCMBE* of passive *run through* and *run over*:

(35) and all things being **run through** which I think necessary to be premised (*PPCEME* boethpr-e3-p2)
[$_{NP\text{-}SBJ\text{-}2}$ [$_{Q}$ all] things] [$_{BAG}$ being][$_{VAN}$ run] [$_{PP}$ [$_{p}$ through] [$_{NP}$ *-2]]

(36) he'd been nearly **run over** by a hackney coach (*PPCMBE* dickens-1837)
[$_{NP\text{-}SBJ}$ [$_{PRO}$ he]] [$_{HVD}$ 'd] [$_{BEN}$ been][$_{ADVP}$ [$_{ADV}$ nearly]] [$_{VAN}$ run] [$_{RP}$ over]

This is perhaps modern intuition: ***rún*** *through* (prepositional) vs. *run* ***óver*** (phrasal). *ARCHER* (tagged by Nick Smith with *CLAWS* and the *Template Tagger*) and *DCPSE* treat passive *run over* in essentially the same way, although their tagsets are different:

(37) Mr. Kenyon Parker, Q.C. [...] was **run over** by a Hansom cab yesterday afternoon in Chancery-lane, and seriously injured. (*ARCHER* 1866pal2.n6b)
[$_{VABDZ}$ was] [$_{VVN}$ run] [$_{RP}$ over] by a Hansom cab

(38) each of them looks as if they've been **run over** by a steam roller (*DCPSE* DI-B78 0048)
[$_{PRON}$ they] [$_{VP}$ [$_{AUX}$ 've [$_{AUX}$ been] [$_{V}$ run]] [$_{AVP}$ [$_{ADV}$ over]] [$_{PP}$ by a steam roller]]

In writing, both *run through* and *run over* are ambiguous syntactically. Historically, *run over* started off as a prepositional verb, as in the following example:

(39) I wish you had been poked into cells, and black holes, and **run over** by rats and spiders and beetles. (1865 Dickens, *Our Mutual Friend*, II.ii.268)

In the context of road accidents, it was reanalysed as a phrasal verb. The syntactic reanalysis corresponds to a semantic change. Earlier, as a preposition, *over* referred to the trajectory of a vehicle or horse passing over a victim; later, as a particle, *over* came to be resultative, referring to the position of the victim.[5] Once again, therefore, it is not obvious that consistent tagging and parsing of the *run over* combination is desirable right across a diachronic corpus.

Here is another case, the participle. Past participles like *interested, amused, concerned* used to be verbal, as shown by the typical co-occurrence with intensifier *much*. Examples (40)-(42) from the *PPCMBE* illustrate this:

(40) Once I sat between him and Miss Ellen Tree after dinner, and was much **amused** at their conversation and his stories (FAYRER-1900)
[$_{BED}$ was] [$_{NP\text{-}MSR}$ [$_{Q}$ much]] [$_{VAN}$ amused][$_{PP}$ at their conversation and his stories][6]

(41) He will be very much **interested** to hear of you. (YONGE-1865)
[$_{ADJP}$ [$_{QP}$ [$_{ADV}$ very] [$_{Q}$ much]] [$_{VAN}$ interested] [$_{IP\text{-}INF\text{-}SPE}$ to hear of you]]

(42) Woke early , much **vexed** at having to go away again. (BENSON-190X)
[$_{IP\text{-}PPL}$ [$_{NP\text{-}MSR}$ [$_{Q}$ much]] [$_{VAN}$ vexed] [$_{PP}$ [$_{P}$ at] ...]][7]

More recently they have come to be adjectival, modified by *very*.

[5] Note that with example (38) there is a mismatch between the older semantics and the PDE syntax, since the point of the comment, about figures in certain artists' paintings, is not that they look prone and injured but that they look **flattened**, as if a steamroller has passed over them!

[6] In (40) NP-MSR = measure noun phrase, VAN = passive participle (verbal or adjectival).

[7] In (42) and (44) IP-PPL = participial clause, but ?not complement of V.

As for *Ving*, it can be a clear adjective – and be so tagged in corpora. Consider example (43), from *ARCHER*, which some users have tagged with several different programs. The first two taggings mark it with the code for adjective, but the third does not.

(43) It pays, though it may seem **boring**. (1961evan.j8b)
it [$_{VM}$ may] [$_{VVI}$ seem] [$_{JJ}$ boring] – *CLAWS* (Nick Smith)
it [$_{MD}$ may] [$_{VB}$ seem] [$_{JJ}$ boring] – ZH TREETAG (also *willing, unwilling, uninteresting, surprising*)
it [$_{Vmod}$ may] [$_{inf}$ seem] [$_{ING}$ boring] – ZH ENGCG2 (also *surprising, willing*, whereas *unwilling, uninteresting* are tagged as adjectives)

If we bring historical knowledge to the question, we find that certain verbal *Ving* forms were once able to occur where now only adjectives can (allegedly) appear:

(44) we began to Clamber up those Hills, which **seem hanging** over the Road of Gombroon (*PPCMBE* FRYER-E3-H,II)
which [$_{VBP}$ seem] [$_{IP\_PPL}$ [$_{VAG}$ hanging] [$_{PP}$ over the Road of Gombroon]]

(45) The long crisis in Laos **appeared nearing** a showdown today. (*Brown* A21)
The long crisis in Laos [$_{VBD}$ appeared] [$_{VBG}$ nearing] a showdown today. (TREETAG annotation)
The long crisis in Laos [$_{Vpast}$ appeared] [$_{ING}$ nearing] a showdown today. (ENGCG2 annotation)

(46) Large and agonizing drops **seemed forcing** their way to his [eyes] (*ARCHER* 1799lee-.f4a)

(47) the shrill shrieks of owls, the loud cries of the wolf, and mournful screams of panthers, which were redoubled by distant echoes as the terrible sounds **seemed dying** away (1797blee.f4a)

(48) I have tried to remember its teachings, but of late they **seemed slipping** from my mind. (1876roe-.f6a)

What does all this tell us? Participles – both present and past – show many changes over the last 300-400 years, both in word class and distribution. Attempts to be consistent in tagging mask such changes, and uncorrected tagging can produce bizarre results.

6. Does it matter?

Two answers can be given:

Arguably, No. Some of the problems discussed are fairly peripheral. Mark-up is an aid, not an end in itself, and mark-up that is "good enough" – allowing the user to find patterns most of the time with adequate precision and recall – is a reasonable aim.

Arguably, Yes. What's convenient for the POS tagger is not necessarily convenient for the user. I take the position that it **does** matter. The God's Truth fallacy, whereby a corpus "may easily create the erroneous impression that it gives an accurate reflection of the entire reality of the language it is intended to represent" (Rissanen 1989: 17), applies to grammatical mark-up too: misclassified examples will mislead students. Experienced researchers can find misclassified examples if they already have suspicions, but if not, relevant examples may be missed.

For a word of vague (that is, underdetermined) class, I would prefer tagsets to include tags that explicitly signal indeterminacy between two categories; they could be something like an ambiguity tag in form. In other cases, I wish tagging could make distinctions that are deliberately avoided in corpora with which I am familiar. Stand-off tagging allows different mark-up schemes for the same material, as with *Zurich Corpus Navigator 2.0* (Hans Martin Lehmann) or *American National Corpus 2*, but these are essentially different tagsets and taggers and not simultaneously available. Software which offers "layers" of user mark-up (cf. Julia Richling's and Anke Lüdeling's papers at the New Methods conference) might allow alternative mark-up to be exploited more easily. The way that the *BNC* can offer alternative taggings of multiword lexical items is pleasing (section 4.2 above), but it is not clear how that would translate to parsing, and in any case it would break down when faced with multiply overlapping prefabs like *those sort of, those sort, what sort, some sort of, sort of thing, that sort of thing*, etc. (Denison 2007: Section 2.4).

The balance between too much and too little in corpus annotation is always a delicate one. My brief survey of metalinguistic and grammatical mark-up suggests to me that it is the latter where it would be particularly worth aiming for something more – and indeed something different.

References

Corpora and databases

ANC = American National Corpus
ARCHER = A Representative Corpus of Historical English Registers
BNC = British National Corpus
COCA = Corpus of Contemporary American English
COHA = Corpus of Historical American English
DCPSE = Diachronic Corpus of Present-Day Spoken English
EEBO = Early English Books Online (26 Mar. 2012)
ICE-GB2 = International Corpus of English, Great Britain
LION = Literature Online (ProQuest) [http://lion.chadwyck.co.uk] (26 Mar. 2012)
lModE Prose = A Corpus of late Modern English Prose
OED = Oxford English Dictionary
PPCEME = Penn Parsed Corpus of Early Modern English
PPCMBE = Penn Parsed Corpus of Modern British English
WebCorp = Linguist's Search Engine (web concordancer)

Secondary references

Aarts, Bas (2007): Syntactic gradience. The nature of grammatical indeterminacy. Oxford: Oxford University Press.

Denison, David (1985): Why Old English had no prepositional passive. In: English Studies 66: 189-204.

Denison, David (2001): Gradience and linguistic change. In: Brinton, Laurel J. (ed.): Historical linguistics 1999. Selected papers from the 14th International Conference on Historical Linguistics, Vancouver, 9-13 August 1999. (= Current Issues in Linguistic Theory 215). Amsterdam/Philadelphia: John Benjamins, 119-44.

Denison, David (2007): Playing tag with category boundaries. In: Meurman-Solin, Anneli/Nurmi, Arja (eds.): VARIENG e-Series 1, Annotating variation and change (= Proceedings of ICAME 27 Annotation Workshop) Helsinki: Research Unit for Variation, Contacts and Change in English (VARIENG). http://www.helsinki.fi/varieng/journal/volumes/01/denison/

Denison, David (2008): Category change in English with and without structural change. Paper presented at NRG4, Universiteit Leuven.

Denison, David (2010): Category change in English with and without structural change. In: Traugott, Elizabeth Closs/Trousdale, Graeme (eds.): Gradience, gradualness and grammaticalization. (= Typological Studies in Language 90). Amsterdam/Philadelphia: John Benjamins, 105-28.

Denison, David (in prep.): English word classes: categories and their limits. (= Cambridge Studies in Linguistics). Cambridge: Cambridge University Press.

Hoffmann, Sebastian (2005): Grammaticalization and English complex prepositions: a corpus-based study. (= Routledge Advances in Corpus Linguistics 7). London/New York: Routledge.

Huddleston, Rodney/Pullum Geoffrey K., et al. (2002): The Cambridge grammar of the English language. Cambridge: Cambridge University Press.

Leech, Geoffrey/Smith, Nicholas (2000): The British National Corpus (Version 2) with improved word-class tagging (BNC2 POS-tagging Manual). http://ucrel.lancs.ac.uk/bnc2/bnc2postag_manual.htm

Marcus, Mitchell P./Santorini, Beatrice/Marcinkiewicz, Mary Ann (1993): Building a large annotated corpus of English. The Penn Treebank. In: Computational Linguistics 19: 313-30.

Marttila, Ville (2011): Manual to the Helsinki Corpus TEI XML Edition.

Matthews, Peter H. (1979): Generative grammar and linguistic competence. London: George Allen & Unwin.

Mitchell, Terence F. (1958): Syntagmatic relations in linguistic analysis. In: Transactions of the Philological Society 1958: 103-106.

Pullum, Geoffrey K. (2004): Snowclones. Lexicographical dating to the second. http://itre.cis.upenn.edu/~myl/languagelog/archives/000350.html

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.

Santorini, Beatrice (2010): Annotation manual for the Penn Historical Corpora and the PCEEC. http://www.ling.upenn.edu/hist-corpora/annotation/index.html

Speake, Jennifer (ed.) (2003): The Oxford dictionary of proverbs. Oxford: Oxford University Press.

Vestergaard, Torben (1977): Prepositional phrases and prepositional verbs: a study in grammatical function. (= Janua Linguarum series minor 161). The Hague: Mouton.

van der Wurff, Wim/Foster, Tony (1997): Object-verb order in 16th century English: a study of its frequency and status. In: Hickey, Raymond/Puppel, Stanislaw (eds.): Language history and linguistic modelling: A Festschrift for Jacek Fisiak on his 60th Birthday. Vol. 1. (= Trends in Linguistics. Studies and Monographs 101). Berlin/New York: Mouton de Gruyter, 439-453.

TERTTU NEVALAINEN

English historical corpora in transition: from new tools to legacy corpora?

Abstract

The first multigenre historical corpora of the English language were published in the early 1990s, almost thirty years after the first Present-Day English corpus was released in 1964. *The Helsinki Corpus of English Texts* (*HC*) came out in 1991, and the *Helsinki Corpus of Older Scots* (*HCOS*) in 1995. The introduction to the latter justifiably called it a 'new tool' (Meurman-Solin 1995).

These tools were new in several respects. They provided systematically selected data on historical varieties of English, comprising closely matching genres from consecutive periods of time. They also made it possible to search texts using an extensive set of metadata, including period-, variety-, and writer-specific information.

However, twenty years is a long time in the life of electronic data sources – long enough in fact to make the first Present-Day English corpora in the *Brown Corpus* family 'historical'. Like these first synchronic corpora, the diachronic corpora of the 1990s were carefully designed but small. *The Helsinki Corpus*, for example, amounts to c. 1.5 million running words. Twenty years on, corpora of this kind are sometimes called 'bijou' corpora, in contrast to the hundreds of millions of words contained, for example, in *COHA*, a monitor corpus of historical American English (for more details on English historical corpora, see *CoRD*).

This paper considers the various material and methodological issues in English historical corpus linguistics that have changed since the pioneering days twenty years ago. I will suggest a division of labour between 'legacy' corpora and their mega-sized successors, and discuss the trade-off between corpus annotation and corpus size.

1. Introduction[1]

Over the last ten to fifteen years corpora have become mainstream in linguistic research. In the wake of the digital turn in the humanities, corpora and other digital databases have grown in size and become accessible over the internet, and their use has increased accordingly (see Nevalainen/Fitzmaurice (eds.) 2011). In this paper I want to look at how these developments have affected

[1] My research for this paper was supported by the Academy of Finland Professorship scheme 2010-2014.

historical corpora, which were a novelty twenty years ago when resources like the *Helsinki Corpus of English Texts* were first released, and a million-word corpus was considered the norm. Back in 1989, Matti Rissanen, a pioneer in the field, articulated three major problematic issues related to the use of historical corpora. Despite the advances made in recent years, two of these problems are still with us today.

Rissanen (1989) called the first of them "the philologist's dilemma", by which he meant the "risk that corpus work and computer-supported quantitative research methods will discourage the student from getting acquainted with original texts". This lack of familiarity with the texts they are studying would undermine students' understanding of the language variety represented by the corpus, and hence their ability to access and interpret their findings. This concern is related to the second problem that Rissanen referred to as the "God's truth fallacy", meaning that "an authoritative corpus may easily create the erroneous impression that it gives an accurate reflection of the entire reality of the language it is intended to represent". The third problem he raised was a particularly acute one in the early days of corpus linguistics, namely "the mystery of vanishing reliability", by which he acknowledged the small size of historical corpora, pointing out that "it may be difficult to maintain the reliability of the quantitative analysis of less frequent syntactic and lexical variants". Of these three problems the third is perhaps less of an issue today with the increasing variety and growing size of historical corpora. Paradoxically, the downside of this development is that "the philologist's dilemma" and "God's truth fallacy" have not disappeared to the same extent, "the philologist's dilemma" having perhaps become even more acute with the advent of megacorpora.

In this paper I discuss the transition of historical corpora from what were exciting new resources twenty years ago to something that may appear inadequately small and antiquated by current standards. I will argue for the middle ground, for both breadth and depth, and suggest that, to escape the "God's truth fallacy", there is a need for a variety of corpora that complement each other in various ways. I realize that the use of multiple corpora may aggravate "the philologist's dilemma", as the number and variety of historical databases will grow automatically with time, not only because more and more data are becoming available in digital form, but also because corpora that were contemporary thirty or forty years ago have become historical from the perspective of their present-day users, and so add to the original texts to be acquainted with.

My discussion focuses on three aspects of these developments, which make the corpus user's work an exercise in *data-source triangulation* (cf. Hammersley/ Atkinson 1983: 192). I begin by discussing some recent trends in English historical corpora in Section 2, and then move on to two issues in historical corpus linguistics I would suggest merit more attention: genre continuity in diachronic databases, and making complementary use of diverse corpora. Section 3 focuses on the continuity and compatibility of data sources in historical corpora, and Section 4 discusses the complementarity of historical corpora, and illustrates data-source triangulation in practice.

2. Increase in historical corpora over time

2.1 General- vs. special-purpose corpora

When talking about historical corpora, the Labovian bad-data problem looms large (Labov 1994: 11): historical research is always constrained by the kind and amount of data available. Diachronic "proto-corpora" such as the *Helsinki Corpus of English Texts* (*HC*) aim to provide for maximal chronological continuity and, as far as possible, to match genres over time. In doing so they cover a wide range of linguistic variation, including both official and private, even speech-based, genres. Just like many general-purpose synchronic corpora of Present-Day English, these historical corpora are typically quite small. This is also the case with the original version of the *Helsinki Corpus*, which comprises 1.5 million words and covers a thousand years of the history of English from the 8th to the 18th century.

Special-purpose historical corpora may consist of a single genre such as drama, newspapers or correspondence, or of multiple genres which share a *literary form* such as dialogue (e.g. dramatic dialogue, witness depositions, dialogue in didactic works) or a *subject domain*, such as medical writing (surgical texts, recipes, regimens, health guides, etc.).[2] Depending on the genres included, these corpora tend to cover a shorter time period than a general-purpose corpus such as the *Helsinki Corpus*, and provide data for the study of specific usage-based or user-based variation over time. For this reason, special corpora come with detailed and systematic metadata inventories. For example, the

[2] For examples, see the entries for the *Corpus of Early English Correspondence* (*CEEC*), a *Corpus of English Dialogues* (*CED*) and the *Corpus of Early English Medical Writing* (*CEEM*), in the Corpus Resource Database (CoRD) at http://www.helsinki.fi/varieng/CoRD/index.html.

Corpus of Early English Correspondence, which covers the period from c. 1400 to 1800, was designed for historical sociolinguistic research and includes a separate database with a wide range of information about the letter writers and the recipients of their letters, including their social status, domicile, education, family background, and migration history.

Although *grammatical annotation* has its problems with historical data, it is now increasingly added not only to general-purpose historical corpora but also to special corpora. A *modern spelling version* can facilitate grammatical annotation and the application of corpus tools to historical data (Baron/Rayson/Archer 2009). As the conversion of an original spelling corpus to a modern spelling format can be carried out by customizing the program to suit the corpus, this option is available at least for the more recent periods, in the case of English, from Early Modern English (1500-1700) on. Having multiple versions of a corpus is one way of making these carefully selected data serve diverse research purposes, without sacrificing the original.

2.2 Growth of corpora in number and size

Corpora of contemporary or present-day language are elusive: it is always only a matter of time until they come to be classified under historical databases. This is what has happened, for example, to the two pioneering English language corpora, the *Brown University Corpus of American English* and the *Lancaster-Oslo/Bergen Corpus of British English*, which both record the language of 1961. The passage of time also makes it necessary to keep updating "contemporary" data sources. The *Brown Corpus* and its British English counterpart have grown into a family of corpora, consisting of the 1990s updates of the original *Brown* and *LOB*, the *Freiburg Brown* and *Freiburg LOB* corpora. The number of areal historical corpora would increase further if text collections such as the *Kolhapur Corpus of Indian English* (1986; Shastri 1988) were replicated using more recent materials.[3]

A further development has taken place as the *Brown Corpus* family has started to move back in time with the compilation of the 1930s *Brown* and *LOB*, and there is a project to stretch the timeline further back by extending the *Brown* family to the beginning of the 20th century (Leech/Smith 2005). A 2006 version of the *LOB* corpus also exists (Baker 2009).

[3] The *Brown* and *LOB* corpora, the *Kolhapur Corpus*, and their manuals are available from the ICAME Corpus Collection (http://icame.uib.no/newcd.htm). For *Brown*, see Francis/Kučera (1979).

The *Helsinki Corpus* has also been significantly extended as its subcorpora have been enlarged and provided with grammatical annotation. The latest addition to the *Helsinki Corpus* family is the TEI compliant XML version of the corpus, which came out in the autumn of 2011. A Late Modern British English corpus (*PPCMBE*; Kroch/Santorini/Diertani 2010) was compiled following the genre division and 70-year subperiods of the Early Modern English section of the *Helsinki Corpus*.

A major development in historical corpus linguistics is the remarkable growth in the size of corpora and databases in recent years. *The Corpus of Historical American English* (*COHA*), which represents American English from 1810 to 2010, consists of 400 million running words and covers four genres (fiction, magazine, newspaper, and other non-fiction). Its compiler, Mark Davies, compares it with the *Google Books* database, which includes 155 billion[4] words, a resource which he is also harnessing as a corpus to be accessed from his website. He finds that the results obtained using the two data sources are quite similar, suggesting that for a variety of research purposes the new megacorpus might not provide a significant added value.

Returning to Rissanen's concerns (1989) discussed in my introduction, it is obvious that the great advantage of megacorpora is that "the mystery of vanishing reliability" by and large disappears. The one caveat that remains is, of course, the uneven amount of printed text available from different periods. As far as English is concerned, the 15th and even the 16th century are poorly represented both in terms of the number and range of available genres, in comparison with the 17th century and subsequent periods.

However, "the philologist's dilemma" surfaces with megacorpora when accessed through a user interface that allows searching but no reading or downloading of the texts themselves. No full-text access can be provided if technical, copyright or other restrictions are imposed on the amount of text available online. The "God's truth fallacy" takes on a different form with megacorpora such as *Google Books*, which contain a huge amount of text and are therefore representative of the books in print stored in libraries at a given time, but do not include all print genres. Newspapers, for example, need to be sampled from different sources.

Moreover, a historical linguist may find it hard to date all the material included in databases such as *Google Books*: a book may remain in print over an ex-

[4] US 'billion', so 1,000,000,000.

tended period of time, or be reprinted with a new publication date but the original contents. By blurring the distinction between the present and the past, such publication histories create what could be called "a real-time problem", which can widen the margin of dating error in the analysis of the texts in the database. With small corpus families, the problem does not emerge, except with early data and undated ego documents.

3. Genre continuity as an empirical issue

With both kinds of corpora, large and small, the degree of comparability of data over time remains an issue. Theoretically-oriented literature treats the continuity of historical records with caution, acknowledging the gaps in the material that has come down to us and, reminiscent of Rissanen's "God's truth fallacy", warning against equating direct observations based on any extant records with language change. Janda/Joseph (2003: 12-14), for example, make a three-way distinction between *diachronic correspondences* (juxtaposing two potentially non-adjacent times and, one might add, incompatible source materials), *linguistic innovation* (initiated by an individual at one particular time), and *language change* proper (adoption of the innovation, over time, by a group of individuals). While the moment of innovation is usually beyond empirical research, its spread, i.e. language change, is not, provided we have diachronically balanced material to study. Apart from dating their texts, corpus linguists are naturally concerned about the comparability and relative stability of their materials over time – something that the compilers of the *Brown* family of corpora have taken great pains to maintain.

While genre continuity was relatively easy to accomplish over the short stretch of thirty years that separates the original 1960s *Brown* and *LOB* corpora and their 1990s Freiburg updates, *Freiburg Brown* and *Freiburg LOB*, it becomes harder to maintain a matching genre structure with the 1930s versions, and increasingly problematic with their turn of the 20th-century counterparts. The question also arises whether the 1960s genre selection does justice to later periods, such as the 2006 version of the *LOB* corpus (Baker 2009).

Looking at the *Helsinki Corpus* family, all these issues are multiplied: it is obvious that, despite the long continuity of several religious genres, for example, there is no way of matching genres over a period of a thousand years, and even if that were the case, there could be no guarantee that their text-type characteristics would remain unaltered over time. In fact, research on the *Brown* family

of corpora suggests that this is not necessarily the case even with matching corpora (Hundt/Mair 1999). Fully mindful of this fundamental problem, the compilers of historical corpora look for broad continuities within the major period divisions of the corpus, or take recourse to some higher-order classificatory criteria (e.g. Kytö/Rissanen 1993: 13, Nevalainen/Raumolin-Brunberg 1993: 61-64).

In practice, the question of genre stability and continuity becomes an empirical issue. Research has shown long-term shifts in the linguistic properties of written genres. Processes of *colloquialization* have been detected in drama, diaries, and fiction, which have been shown to have become linguistically more involved in more recent times, while the opposite trend of *economization* has been detected, for example, in newspapers (see further e.g. Biber/Finegan 1997, Hundt/Mair 1999, Biber/Clark 2002, Nevalainen 2008, Szmrecsanyi/Hinrichs 2008).

What makes this issue relevant in practice is that, unlike corpus compilers, corpus users often tend to rely on the "null hypothesis" that part-of-speech frequencies, for example, remain constant in their corpus throughout the period of observation. As we have seen, this need not be the case. One would therefore like to see more empirical work on various aspects of corpus stability over time. To that end, Säily/Nevalainen/Siirtola (2011) carried out a study on the variation in noun and pronoun distribution in the *Parsed Corpus of Early English Correspondence* (*PCEEC*, 1400-1680). The results indicate a slow but statistically significant trend towards the use of fewer nouns in the corpus over the centuries. Female letter writers used more pronouns and fewer nouns than male writers did in each subperiod, which suggests stable sociolinguistic variation with respect to this variable. The findings support early multivariate studies such as Biber/Finegan (1997), which detect genre drifts over time, based on general-purpose corpora, accommodating a number of genres but with a limited number of texts representing each of them. More research would obviously be needed to answer the intriguing question that emerged in Säily/Nevalainen/Siirtola (2011), based on a larger special-purpose corpus, of whether English personal letters become less focused on information as early as the 17th century, while at the same time projecting stable sociolinguistic variation over time. The issue of genre continuity thus becomes a matter of data granularity, or, in terms of the systemic functional grammar, of the degree of delicacy at which registers are identified (Matthiessen/Teruya/Lam 2010: 177).

4. Complementarity of corpora: a case study on politeness

As small corpora are usually carefully structured and come with rich metadata, for reasons of research economy, large corpora usually contain less metadata and may be less structured in terms of content. Data sources like the *Google Books* database are of course a case in point in both respects, but their sheer size allows the study of low-frequency linguistic features, typically of lexical phenomena. Research can benefit from the complementary strengths of corpora by making parallel use of both large and small corpora.

My case study in data-source triangulation presents the diachronic evolution of three sets of 'polite' words in English, which constituted "buzz words" in different periods. The earliest of the three is *courteous*, *courtesy* and their derivatives, which were originally associated with courtly behaviour, and appeared in the late Middle Ages. The second set, *civil* and *civility*, originally pertained to citizens, and was promoted in the Renaissance, and the third, consisting of *polite* and *politeness*, was first used with reference to smoothed or polished objects, and came into vogue in the Enlightenment.

The *Google Books* record for the three adjectives is shown in Figure 1. One of the shortcomings of this vast database becomes immediately obvious: the amount of data available from the early periods is considerably smaller than data from the later ones. Figure 1 ranges from 1600 to 2000, and so misses out the late medieval period, which is particularly relevant for the study of *courteous*, a word which is attested only at low frequencies throughout the later periods. In the graph for *civil* (the top one), the 17th-century part is jagged, suggesting that the books dated to that century are still relatively few, and unevenly distributed in terms of content. The fact that this was also the century during which the English Civil War broke out suggests a problem that arises from blanket searches for lexical items that may be, as in this case, polysemous. Judging by the titles listed by the *Google Books* n-gram viewer, another sense prevalent in the sources pertained to civil law (cf. Michel et al. 2010). For all its deficiencies, Figure 1 nevertheless indicates that it was in the 18th century that *polite* (the middle graph) had its heyday in the written language, only to decline in the subsequent centuries. The polysemous *civil* also declined but managed to retain a larger relative share of the ever-increasing English lexicon than the two monosemous politeness words.

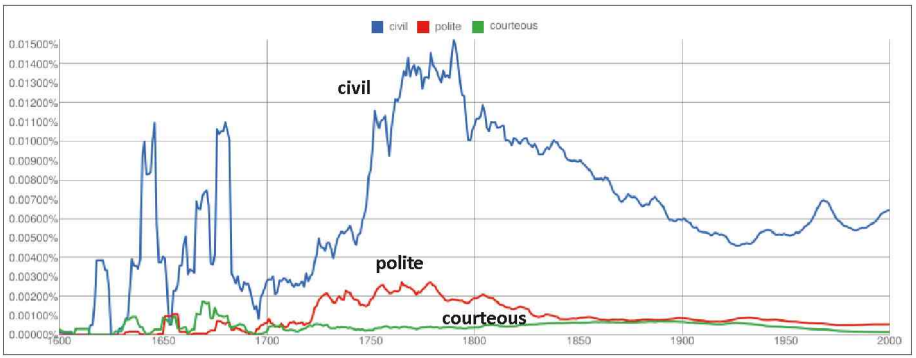


Figure 1: *Civil*, *polite*, and *courteous*, 1600-2000.

By way of comparison, Figure 2 shows the *Google Books* record for three related German adjectives, *artig*, *gefällig*, and *höflich* between 1800 and 2000. It suggests that *höflich*, which is derived from *Hof* ('court') and etymologically related to courtly behaviour, gains currency in the 20th century.

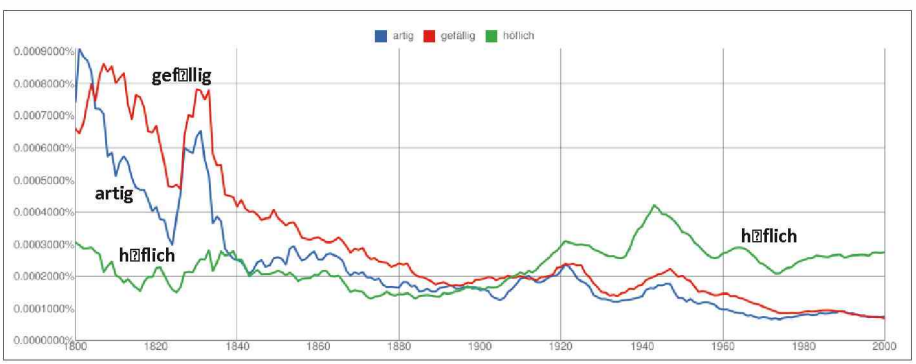


Figure 2: *Artig*, *gefällig*, and *höflich*, 1800-2000.

A small special-purpose corpus can give a fuller picture of the distribution and contexts of use of the three sets of polite words in the 18th century. *The Corpus of Early English Correspondence Extension* (*CEECE*), 1681-1800, consists of 2.2 million words, 75 letter collections and over 300 letter writers. The personal information recorded for each writer includes, for example, gender, profession, social status, education, domicile, and migration history, which makes it possible to group writers according to these variables on the aggregate level. The corpus returns altogether 412 'polite' words between 1700 and 1800: 279 instances of *civil/civility*, 118 of *polite/politeness*, and 15 of *courteous/courtesy* (Nevalainen/Tissari 2010). The example in (1) illustrates the use of two of them by Mary Wollstonecraft in 1786.

(1) … agreeable companion – a young Clergyman, who was going to settle in Ireland, in the same capacity as myself. He was intelligent and had that kind of **politeness**, which arises from sensibility. My conductor, was beyond measure **civil** and attentive to me, he is a good sort of a man, I was, at first, at a loss to guess what department he filled in the family; but I find now he is the Butler, and his wife the housekeeper. (A 1786? FN MWOLLSTON 123)

Figure 3 shows the distribution of these three sets of words over three subperiods in the 18th century. As in the *Google Books* data, the *civil* set dominates throughout the period, although only the 'polite' senses were included in the analysis. The use of the *polite* set increases until the latter half of the century but declines in the last twenty years, 1780-1800. The use of the third set is negligible in the 18th-century correspondence corpus.

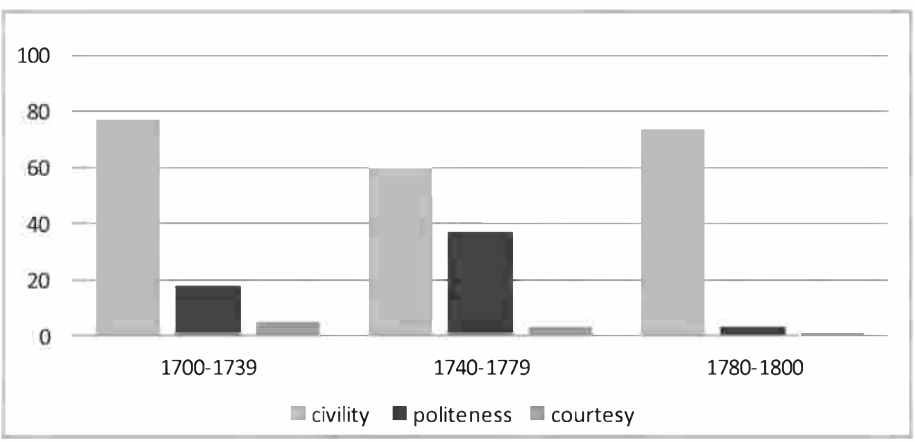


Figure 3: Relative frequencies (%) of the three sets of politeness words in *CEECE*.

A closer variationist analysis reveals that the *civil* set is preferred by upper ranking women. Male writers also use it more than the *polite* set throughout the social spectrum available in the corpus. The *polite* set in fact appears to be promoted by the lower gentry and professional writers, women more than men (Nevalainen/Tissari 2010: 144-145). Lower ranking women did not use it at all but, overall, they were sparsely represented in the corpus.

These findings can be supplemented by another database, the 18th-century *Old Bailey Proceedings*, which consists of some ten million words. Shoemaker (2004: 291-296) reports that *civil* and *polite* are both marginal in this vast body of trial accounts, where the non-elite section of the London population, men

and women alike, is also well represented. Where *polite* and *politeness* occur in 39 trials out of a total of 45,000 trial accounts, *civil* and *civility* are found in 91. Shoemaker adds that these cases typically involved the middle and upper classes, and concludes that the Londoners who were dealt with in these court records rarely used the language of politeness themselves, nor were they described in those terms by their peers, or by the magistrates they came into contact with.[5]

Including the relevant sections of the *Old Bailey Proceedings*, the recent *London Lives* database provides an even more comprehensive record of primary sources on 18th-century London, as the website specifies, "with a particular focus on plebeian Londoners". It gives access to 240,000 manuscript and printed pages from a variety of London archives and other related databases. This fully searchable resource yields a mere 62 instances of *polite/politeness* from the period between 1730 and 1819. Not all come from records of spoken interaction, but some occur in advertisements or magazine titles, as in *THE CONVIVIAL MAGAZINE, AND POLITE INTELLIGENCER*. Examples (2) and (3) illustrate contexts in which these politeness words appeared in cross-examinations given in the *Old Bailey Proceedings*.

(2) How long was it before you scraped acquaintance? – I believe about three quarters of an hour; she was drinking with another man, and then came and sat by me.

I presume you was **polite** enough to ask her to drink with you? – She might ask herself, but I do not recollect. (*Old Bailey Proceedings*: Accounts of Criminal Trials, 11th September 1776)

(3) I presume there was no great **politeness** or **civility** passed between you and the prisoner at the bar, when you went to his shop on the Saturday, and apprised him that he was suspected of being the person to whom stolen goods had been sold? – I did not tell him any such thing. (*Old Bailey Proceedings*: Accounts of Criminal Trials, 22nd February 1781)

It is possible to proceed further into the Late Modern English period and explore areal corpora such as the *Corpus of Historical American English* (*COHA*, 1810-2000) to find out more about the distribution of politeness terms in the 19th and 20th centuries. Figures 4 and 5 show the three adjectives, *civil*, *polite*, and *courteous* in the American collection of the *Google Books* database and

[5] See Huber (2007) for an introduction to the *Old Bailey Corpus* (*OBC*), based on *The Proceedings of the Old Bailey*, a corpus project currently in progress.

COHA, respectively. The *Google Books* graphs in Figure 4 are based on a moving annual average and are thus smoother than the ten-year *COHA* averages in Figure 5. The comparison is not completely straightforward because of the different sizes of the corpora, and the way in which the frequencies are calculated in each of them (as a percentage, with a moving average, of the unigram total in the *Google Books* collection, and normalized to a million words in *COHA*). Despite these differences, both Figures show a general downward trend for *civil* until a peak in the 1960s, which is likely to be a reflection of the African-American Civil Rights movement, and is thus unrelated to the politeness sense of the word.

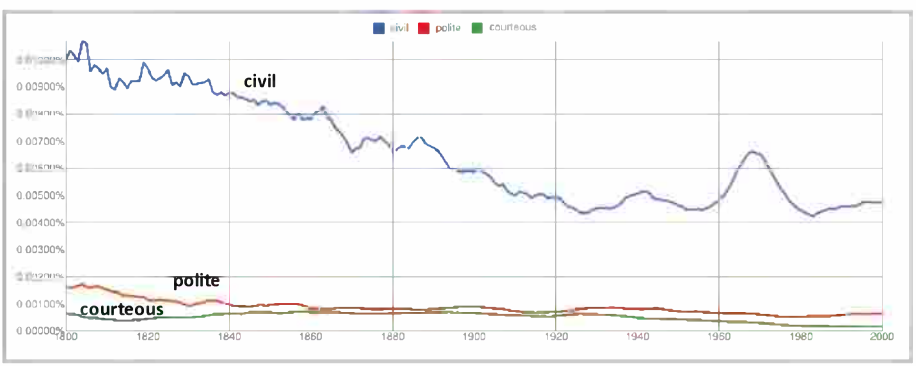


Figure 4: *Civil*, *polite*, and *courteous* in American English books, 1800-2000.

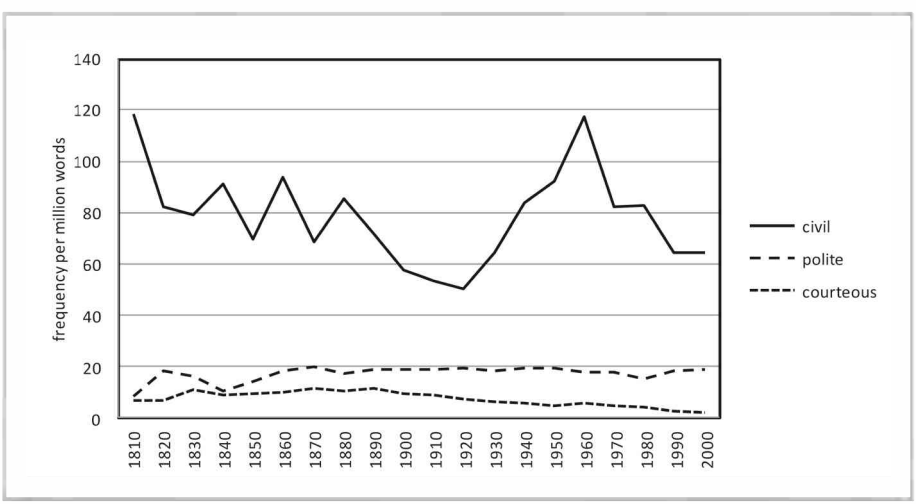
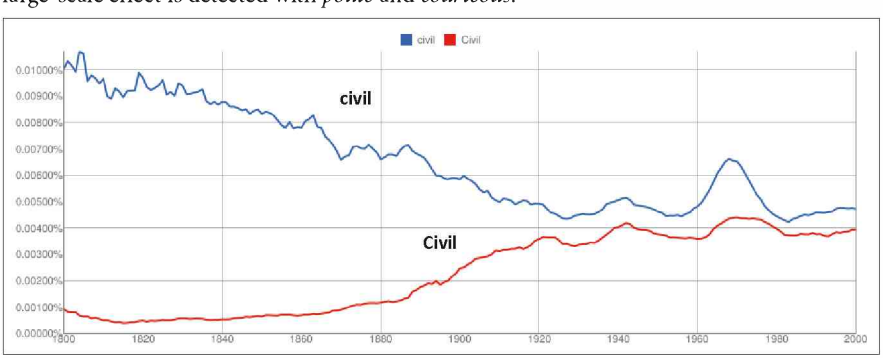


Figure 5: *Civil*, *polite*, and *courteous* in *COHA*, 1810-2000 (normalized to a million words per decade).

Both data sources suggest a low, relatively even distribution for *polite* throughout the two centuries, and a clear drop for *courteous* in the last few decades. Unless the senses of *civil* are disambiguated, little can be said about the development of its politeness sense over time.[6] Polysemy of course remains a problem with nearly all data sources, but is manageable with small corpora, which can serve as useful windows on diachronic trends.

This lexical case study of 'polite' words illustrates the benefits of data-source triangulation. A rough overall picture can be obtained using a database like *Google Books*. However, *Google Books* is an unannotated and unlemmatized data source consisting of printed books, and its n-gram viewer does not allow in-depth searches of extralinguistic information apart from what language or language variety these books were printed in and when. Provided that there are suitable special-purpose corpora available, zooming in on a particularly interesting period can take the study closer to the social reality of the change in progress. I used three complementary data sources to explore the extent to which the three sets of 'polite' words had permeated, or failed to permeate, 18th-century English society: a corpus of personal correspondence, which has plenty of information on the letter writers, and two fully searchable databases consisting of trials and other localized archival sources.

[6] The fact that the *Google Books* n-gram viewer is case-sensitive weeds out a large number of cases unrelated to the politeness sense of *civil*. The American data shown in Figure 4, above, are supplemented below with cases in which *civil* is spelled with a capital letter (lower curve) and typically collocates with Civil War and Civil Rights. The same is true of Figure 1, where what looks like a gap in the use of *civil* between 1650 and 1750 in fact contains a large number of capitalized instances of *civil* (*Civil War*, *Civil Government*, *Civil Society*, etc.). No similar large-scale effect is detected with *polite* and *courteous*.



5. Concluding remarks

With a dramatic increase in the digitization of texts, triangulation of data sources is a useful method for making the best use of both small, well structured corpora and large but potentially messier databases. Small corpora have often been compiled with particular research questions in mind and are based on detailed sampling frames. Many of them come with rich metadata and are grammatically annotated, whereas grammatical annotation is less commonly attached to large databases, which rarely contain metadata at a level of detail that would be useful, for example, for historical sociolinguistic studies. A fruitful way forward, in terms of both research economy and reliability, will include more work on both kinds of resource, and the creation of platforms where they could be easily accessed and flexibly compared.[7]

References

Baker, Paul (2009): The BE06 Corpus of British English and recent language change. In: International Journal of Corpus Linguistics 14: 312-337.

Baron, Alistair/Rayson, Paul/Archer, Dawn (2009): Word frequency and key word statistics in historical corpus linguistics. In: Anglistik: International Journal of English Studies 20 (1): 41-67.

Biber, Douglas/Clark, Victoria (2002): Historical shifts in modification patterns with complex noun phrase structures: How long can you go without a verb? In: Fanego, Teresa/Pérez-Guerra, Javier/López-Couso, María-José (eds.): English historical syntax and morphology. Selected papers from 11 ICEHL. Amsterdam: John Benjamins, 43-66.

Biber, Douglas/Finegan, Edward (1997): Diachronic relations among speech-based and written registers in English. In: Nevalainen, Terttu/Kahlas-Tarkka, Leena (eds.): To explain the present: studies in the changing English language in honour of Matti Rissanen. Helsinki: Société Néophilologique, 253-275.

Common Language Resources and Technology Infrastructure (CLARIN). http://www.clarin.eu/external/index.php?page=about-clarin&sub=0

Corpus of Early English Correspondence Extension (*CEECE*). Compiled by Samuli Kaislaniemi, Mikko Laitinen, Minna Nevala, Terttu Nevalainen, Arja Nurmi, Minna Palander-Collin, Helena Raumolin-Brunberg and Anni Sairio. Department of English, University of Helsinki.

Corpus of Historical American English (*COHA*): http://corpus.byu.edu/coha

[7] In the European context, the *Common Language Resources and Technology Infrastructure* project (*CLARIN*), aims at such shared infrastructure resources.

Corpus Resource Database (*CoRD*): http://www.helsinki.fi/varieng/CoRD/index.html

Francis, W. Nelson/Kučera, Henry (comp.) (1979 [1964]): Brown Corpus Manual. Providence, Rhode Island: Department of Linguistics, Brown University. http://khnt.aksis.uib.no/icame/manuals/brown/INDEX.HTM

Freiburg-Brown Corpus ('*Frown*'), original version. Compiled by Christian Mair, Albert-Ludwigs-Universität Freiburg.

Freiburg-LOB Corpus ('*F-LOB*'), original version. Compiled by Christian Mair, Albert-Ludwigs-Universität Freiburg.

Google Books Corpus: http://googlebooks.byu.edu/compare-googleBooks.asp

Google Books Ngram Viewer: http://books.google.com/ngrams/graph

Hammersley, Martyn/Atkinson, Paul (1983): Ethnography: Principles in practice. London: Tavistock Publications.

Helsinki Corpus of English Texts (*HC*) (1991): Department of English, University of Helsinki. Compiled by Matti Rissanen (Project leader), Merja Kytö (Project secretary); Leena Kahlas-Tarkka, Matti Kilpiö (Old English); Saara Nevanlinna, Irma Taavitsainen (Middle English); Terttu Nevalainen, Helena Raumolin-Brunberg (Early Modern English). The Helsinki Corpus TEI XML Edition (2011). Helsinki: VARIENG.

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, Volume 1), Helsinki: VARIENG. http://www.helsinki.fi/varieng/journal/volumes/01/huber

Hundt, Marianne/Mair, Christian (1999): 'Agile' and 'uptight' genres. The corpus-based approach to language change in progress. In: International Journal of Corpus Linguistics 4(2): 221-242.

ICAME Corpus Collection. http://icame.uib.no/newcd.htm

Janda, Richard D./Joseph, Brian D. (2003): On language, change, and language change – or, of history, linguistics, and historical linguistics. In: Joseph, Brian D./Janda, Richard D. (eds.): The handbook of historical linguistics. Malden, MA/Oxford: Blackwell, 3-180.

Kroch, Anthony/Santorini, Beatrice/Diertani, Ariel (2010): Penn Parsed Corpus of Modern British English. http://www.ling.upenn.edu/hist-corpora/PPCMBE-RELEASE-1/index.html

Kytö, Merja/Rissanen, Matti (1993): General introduction. In: Rissanen/Kytö/Palander-Collin (eds.), 1-17.

Labov, William (1994): Principles of linguistic change. Vol. 1. Internal factors. Oxford: Blackwell.

Lancaster-Bergen/Oslo Corpus (*LOB*), original version (1970–1978): Compiled by Geoffrey Leech, Lancaster University, Stig Johansson, University of Oslo (project leaders), and Knut Hofland, University of Bergen (head of computing).

Leech, Geoffrey/Smith, Nicholas (2005): Extending the possibilities of corpus-based research on English in the twentieth century: a prequel to LOB and FLOB. In: ICAME Journal 19: 82-98.

London Lives, 1690 to 1800. http://www.londonlives.org/index.jsp

Matthiessen, Christian M.I.M./Teruya, Kazuhiro/Lam, Marvin (2010): Key terms in systemic functional linguistics. London/New York: Continuum.

Meurman-Solin, Anneli (1995): A new tool: the Helsinki Corpus of Older Scots (1450-1700). In: ICAME Journal 19: 49-62.

Michel, Jean-Baptiste/Shen, Yuan Kui/Presser Aiden, Aviva/Veres, Adrian/Grey, Matthew K./The Google Books Team/Pickett, Joseph P./Hoiberg, Dale/Clancy, Dan/Norvig, Peter/Orwant, Jon/Pinker, Steven/Nowak, Martin A./Lieberman Aiden, Erez (2010): Quantitative analysis of culture using millions of digitized books. Science 331(6014): 176-182. doi:10.1126/science.1199644.

Nevalainen, Terttu (2008): Variation in written English: Grammar change or a shift in style? In: Kermas, Susan/Gotti, Maurizio (eds.): Socially-conditioned language change: diachronic and synchronic insights. Lecce: Edizioni del Grifo, 31-51.

Nevalainen, Terttu/Fitzmaurice, Susan M. (eds.) (2011): How to deal with data: Problems and approaches to the investigation of the English language over time and space (= Studies in Variation, Contacts and Change in English 7). Helsinki: VARIENG. http://www.helsinki.fi/varieng/journal/volumes/07

Nevalainen, Terttu/Raumolin-Brunberg, Helena (1993): Early Modern English. In: Rissanen/Kytö/Palander-Collin (eds.), 53-73.

Nevalainen, Terttu/Tissari, Heli (2010): Contextualizing 18th-century politeness: social distinction and morphological levelling. In: Hickey, Raymond (ed.): Eighteenth century English: Ideology and change. Cambridge: Cambridge University Press, 133-158.

Parsed Corpus of Early English Correspondence (*PCEEC*), tagged version (2006), annotated by Arja Nurmi, Ann Taylor, Anthony Warner, Susan Pintzuk, and Terttu Nevalainen. Compiled by the CEEC Project Team. York: University of York and Helsinki: University of Helsinki. Distributed through the Oxford Text Archive.

Rissanen, Matti (1989): Three problems connected with the use of diachronic corpora. In: ICAME Journal 13: 16-19.

Rissanen, Matti/Kytö, Merja/Palander-Collin, Minna (eds.) (1993): Early English in the computer age: explorations through the Helsinki Corpus (= Topics in English Linguistics 10). Berlin/New York: Mouton de Gruyter.

Säily, Tanja/Nevalainen, Terttu/Siirtola, Harri (2011): Variation in noun and pronoun frequencies in a sociohistorical corpus of English. In: Literary and Linguistic Computing 26(2): 167-188.

Shastri, S. V. (1988): The Kolhapur Corpus of Indian English and the work done on its basis so far. In: ICAME Journal 12: 15-26.

Shoemaker, Robert (2004): The London mob. Violence and disorder in eighteenth-century England. London: Hambledon and London.

Szmrecsanyi, Benedikt/Hinrichs, Lars (2008): Probabilistic determinants of genitive variation in spoken and written English: a multivariate comparison across time, space, and genres. In: Nevalainen, Terttu/Taavitsainen, Irma/Pahta. Päivi/Korhonen, Minna (eds.): The dynamics of linguistic variation: Corpus evidence on English past and present. Amsterdam/Philadelphia: John Benjamins, 291-309.

The Proceedings of the Old Bailey, 1694-1913. http://www.oldbaileyonline.org

KLAUS-PETER WEGERA

Language data exploitation: design and analysis of historical language corpora

Abstract

Crucial features in the theory of how to elicit data from linguistic corpora are, first of all, establishing the distinction between a 'corpus' as properly understood and a 'text archive', and secondly, the classification of different types of corpora (i.e. finely structured as opposed to large corpora). 'Corpus' can be defined very narrowly and contrasted with 'text archive'. The advantages of a clearly structured corpus of historical stages of a language with relatively limited extant records are demonstrated using the example of the corpus of the new *Middle High German grammar*. Central questions raised in acquiring data from a corpus concern the status of the data and the strategies to be employed in their analysis. The notion of "representativeness" will be re-evaluated and methods outlined to illustrate how a comprehensive analysis of a corpus may be undertaken.

1. What is a corpus?

Any metalinguistic statement about a linguistic phenomenon is made on the basis of linguistic data, which may be obtained in two different ways, either by introspection – all competent speakers of a language can introspectively access linguistic data[1] of their language, state and verify them, etc. – or by external language material. Historical linguistics is dependent on external data, which means that many theoretical problems do not arise in the first place.

This paper thus deals only with the external exploitation of linguistic data. For this purpose, the theoretical status of external data collections has first to be clarified. In the past, examples from grammars, dictionaries or material collections in the form of card indexes often served as data collections, which were commonly called corpora. In Germany, a serious academic discourse about corpora has developed only in the last 25-30 years, and has now matured there

[1] Unlike Lehmann (2007: 4f.), I do not include a distinction between examples and data, because even invented examples have been recorded as data, being a part and the result of linguistic knowledge in the process of language acquisition. However, methodologically, one has to distinguish between examples that are invented and those that are based on external data.

into the independent methodology known as "corpus linguistics".[2] However, aspects concerning specific characteristics of corpora featuring historical language data (text corpora, material corpora), which require a theoretical and methodological foundation, have been largely neglected in the process.[3]

A novel practical dimension arose when the possibilities of electronic data processing gradually appeared in the 1970s, and it rapidly became clear that digitized texts could be dealt with much faster, more precisely and in a shorter amount of time. First, text collections were compiled, manually digitized, lemmatized and at least partially annotated, and these were then available as digitized corpora for various research objectives. These corpora were compiled from single texts at first, and only then digitized and analyzed electronically for a specific research aim.[4]

Since then, corpora have become closely associated with electronic data processing. However, digitization is not part of the definition of a corpus.[5] Admittedly, digitized corpora do represent today's methodological standard. Additionally, annotated and, if possible, lemmatized corpora supplied with headers will soon become standard. Similarly, texts are not a defining property of corpora. Texts are only another form of corpora, in addition to those of single phenomena like words (with or without context), grammatical structures, single sentences. However, text corpora have become the common form in many domains because they allow the extraction of data and focus on smaller units such as sentences, smaller grammatical structures or fixed phrases. This results in text corpora being assumed to be self-evident in definitions.[6]

The beginnings were quickly supplemented by a further aim, i.e. to digitize texts without any kind of defined research objective, merely for their own sake, often with the intention of later analysis, regardless of its nature or user. Since

---

2 Cf. for instance Bergenholtz/Schaeder (1979), Lenz (2000), Köhler (2005), Schwitalla/Wegstein (2005), Scherer (2006), Lüdeling/Kytö (2008/09), Mukherjee (2009), Lemnitzer/Zinsmeister (2010), Perkuhn/Keibel/Kupietz (2012), some contributions in: Kallmeyer/Zifonun (2007), Kratochvílová/Wolf (2010), as well as articles in http://www.linguistik-online.de/38_09/ and http://www.linguistik-online.de/39_09/.

3 This is also true for international research as far as I have reviewed it. First approaches can be found in Klein (1991), Hoffmann (1998), Wegera (1990; 2000), Schwitalla/Wegstein (2005), Curzan/Palmer (2006), Curzan (2009), Claridge (2008), Rissanen (2008).

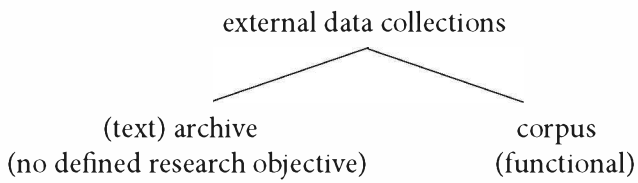
4 Cf. Johansson (2008) on the history.

5 Cf. for instance Lemnitzer/Zinsmeister (2010: 40).

6 For instance Mukherjee (2009: 20f.).

then, the notion has become quite widespread that the main thing is that the texts are at least in the can. Even though such text collections are often called corpora, they are at worst unsystematic and opportunistic, and at best text archives.

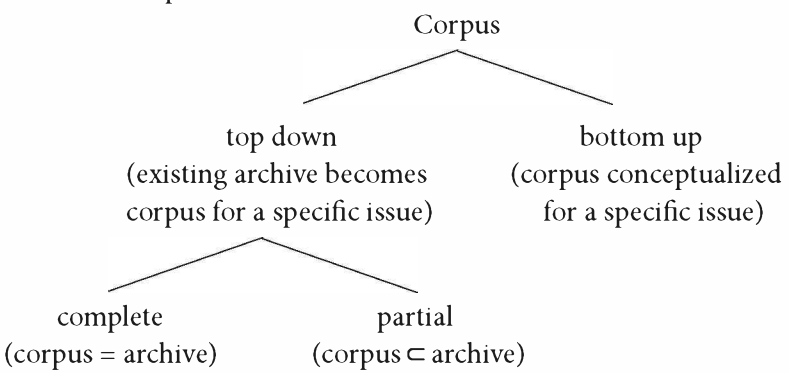
In order to prevent a confusion of terms, and to illustrate my position: a corpus is always functional, in so far as an object of research and a research objective are always the motive and point of origin for the construction of a corpus, for instance the verification of a hypothesis, the analysis and description of a grammatical topic, etc.[7]



This conceptual limitation seems necessary because corpus criticism (even positive) often misses the point. In this sense, the *Bonn Corpus of Early New High German* (*Bonner Frühneuhochdeutsch-Korpus*) and the *Bochum-Bonn Corpus of Middle High German* (*Bochum-Bonner Mittelhochdeutsch-Korpus*; now *MiGraKo*) are, strictly speaking, not corpora but simply well-designed text archives. However, they were regarded as corpora at the time of their origin and already functionally related to the grammatical description which was based on them. The *Bonn Corpus of Early New High German* was compiled in the mid-1970s, manually digitized, lemmatized and annotated as the basis for research on the inflectional morphology of Early New High German nouns, verbs and adjectives. Exclusively for this purpose, the corpus was of sufficient size, even taking into account the state of technology then available. Criticism of this corpus which tries to identify its weaknesses, or even its strengths and its previous function as a corpus, must take this fact into account. In most other respects, this archive is only partially suitable as a corpus, or not at all. The same applies to a certain extent to the *Bochum-Bonn Corpus of Middle High German*, too, which has served as a corpus for most new research on Middle High German grammar since the mid-1990s.

[7] This is similar to Scherer (2006: 5): "Korpora sind prinzipiell zweckgebunden."

One can compile a whole new corpus which is specifically conceptualized to address a certain issue (bottom up). However, the dangers of too many theoretical pre-suppositions influencing the design of the corpus are high "[ ], so-dass man letztlich die Ostereier zur allgemeinen Überraschung dort findet, wo man sie vorher versteckt hat" ("so that to everyone's surprise you end up finding the Easter eggs where they had been hidden").[8] Or one can resort to an already existing archive and make it the basis for one's corpus (top down). Here, both theoretical factors, archive and corpus, correspond with each other: if the entire archive is used as the basis, then the archive obtains the theoretical status of a corpus. Meanwhile, if only parts of the archive are used as the basis, then the text archive is preserved as such and only the selected part obtains the status of a corpus.



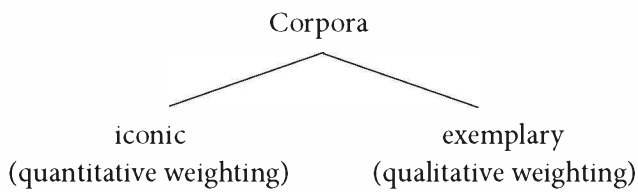
The bottom up alternative is an option for historical linguistics as long as not all surviving historical linguistic sources are generally available in digitized form. However, sources will be digitally available in the foreseeable future for the period up to about 1200. Additionally, this may be the case in the long term for later centuries, so that eventually there will only be top down corpora.

For a long time, the sheer size of a corpus had been a fetish, and was therefore not negotiable. Large corpora certainly have great advantages, for instance in the search for rare events or infrequent features, especially words, although it seems that even the largest corpora always leave something to be desired. When one thinks about the principle of mass corpora further, thetoutcome is foreseeable,twhere all historical texts will be recorded (theoretically) in the collection, which then becomes a comprehensive text archive. For now, small and structured corpora are valued for defined research objectives.[9]

[8] Wolf (2010: 18).

[9] Heikkinen/Mikko (2008); see also de Haan (1992).

Corpora can attempt to represent the tradition of a certain period in a quantitatively weighted manner (like, for example, the *LIMAS* corpus,[10] which was orientated towards the *Brown* corpus). These corpora can be described as being representative in an iconic sense. However, corpora can also be qualitatively weighted by trying to find the most suitable texts respectively for specified, preassigned parameters. Such corpora can be described as exemplary.



The second alternative appears more practicable for historically orientated corpora, because there is not – at least until 1500 – enough existing material, and thus, a rather simple definition arises for the term "corpus", i.e. "A corpus (in the linguistic sense) is a specified external amount of data, which serves as the basis of linguistic analysis". The digital processing of data, their exploitation by means of further processes, their embeddedness in continuous text, and their analysis, are all subsequent and subsidiary processes.

2. How can a corpus be structured? Example of an exemplary (qualitatively weighted) corpus

Corpora can (and have to) be structured very differently depending on the research aims,[11] and I can only refer to an example which we worked with in Bochum, Bonn and Halle. The (structured) corpora which we had favoured from the beginning, in the 1970s in Bonn[12] and then later in the 1990s in Bochum, Bonn, and Halle,[13] are of manageable dimensions. They contain selected and qualified material (here: texts) for certain parameters. I would like to

[10] http://www.korpora.org/Limas/ and http://icame.uib.no/brown/bcm.html

[11] The majority of literature in corpus linguistics to this day is concerned with lexicographic or lexicological problems and related corpus questions.

[12] http://ww.korpora.org/fnhd/

[13] http://www.degruyter.com/view/supplement/9783111844206_Quellenkorpus_Uebersicht.pdf

exemplify this procedure, its limits and problems, based on the corpus of the new *Middle High German Grammar*.[14]

The design of a historical corpus of German, which covers a time span of about 300 years (1050-1350) and is intended to show diachronic development, must first establish appropriate temporal parameters. Traditionally, Middle High German is divided into three phases: before 1170, 1170-1250 and after 1250. These mechanical divisions have little connection to the divisions used by a theory of linguistic development. A division based on linguistic development is often possible only after the work has been finished (in this case, on the grammar). Therefore, the application of a procedure in advance, which makes only a few assumptions and is pragmatic, was sensible in the project. It had already been successfully practised in the course of the *Grammar of Early New High German* (*Grammatik des Frühneuhochdeutschen*) and is still being practised in the *GerManC* project.[15] For Middle High German, the time span between 1050 and 1350 is mechanically divided into fifty-year periods. This procedure is also appropriate because German manuscripts were seldom dated within the manuscript itself until the fourteenth century, and therefore have to be classified with the help of external criteria, especially by palaeographic means. Such age determinations are not exact to the year, and normally only allow the assignment of those textual witnesses for certain decades or quarter centuries, sometimes even only half-centuries. On the one hand, these fifty-year periods are long enough to assign manuscripts which can only be dated vaguely. On the other hand, they are short enough to still perceive diachronic developments. In addition, more subtle distinctions can be taken into account at any time by means of manuscripts that are dated more precisely.

There is now a first – diachronic – dimension for the structure of the corpus, i.e. the time periods:

I 1050-1150[16]
II 1150-1200
III 1200-1250
IV 1250-1300
V 1300-1350

[14] Klein/Solms/Wegera (2009). See also http://www.degruyter.com/view/supplement/9783111844206_Quellenkorpus_Uebersicht.pdf.

[15] http://www.llc.manchester.ac.uk/research/projects/germanc/

[16] The period between 1050 (more precisely 1070) and 1150, which only contains limited material, will be referred to as a single period.

The formation of regional varieties is also of fundamental importance for the history of German. Accordingly, every historically oriented corpus structure needs to include a spatial dimension. The following horizontal (regional) dimension results in the corpus structure of "Middle High German" (with a lot of subtleties that may be omitted here):

Middle Franconian Hessian-Thuringian (later: East Central German)
Rhine Franconian-Hessian East Franconian
Alemannic Alemannic-Bavarian crossover area Bavarian

Thus, we already have two dimensions that have to be taken into account during the selection of texts.

Middle Franconian
1070-1150
1150-1200
1200-1250
1250-1300
1300-1350

Hessian-Thuringian
1070-1150
1150-1200
1200-1250
1250-1300
1300-1350

Rhine Franconian	East Franconian
1070-1150	1070-1150
1150-1200	1150-1200
1200-1250	1200-1250
1250-1300	1250-1300
1300-1350	1300-1350

Alemannic	Alemannic-Bavarian	Bavarian
1070-1150	1070-1150	1070-1150
1150-1200	1150-1200	1150-1200
1200-1250	1200-1250	1200-1250
1250-1300	1250-1300	1250-1300
1300-1350	1300-1350	1300-1350

Figure 1: Temporal and spatial structure of the *Bochum-Bonn Corpus of Middle High German*[17]

[17] Cf. footnote 14.

Hence, in terms of figures, the corpus has more than 35 grid fields that have to be filled with texts. Here, the first problems arise because there are not enough surviving texts for each grid field which can confidently be assigned to the correct region and period. This brings us to the question of which principle should be preferred: completeness or absolute quality. In the first case compromises will be necessary, while in the second case evidential gaps might occur.

This problem becomes more aggravated with each subsequent dimension. However, the additional dimension of text type seems to be indispensable. Many linguistic phenomena cut across several text types; some show certain phenomena never, more rarely, or more frequently than others. They can also react more innovatively or more conservatively to certain linguistic phenomena (alterations) than others. Nonetheless, even a rough zoning into a few types of texts not only brings a sharp increase in the number of grid fields, but results in most grid fields remaining blank.

A theory of text types for Middle High German exists only in its initial stages, although the tradition is of a manageable size. The sources used in the corpus of the *Middle High German grammar* are only distinguished according to three forms of presentation: "verse", "official documents", and "(other) prose". Even this provisional and not very satisfactory distinction means that not all grid fields can always be filled, let alone evenly.

Texts in verse are at first sporadic and there is a significant number only from the second half of the twelfth century. German official documents have been preserved in substantial numbers only from the second half of the thirteenth century. Only prose texts are represented relatively constantly throughout the entire period. A further distinction of text types such as Bible translation/Bibleystories, factual texts, legal texts etc. may be possible, but only as an additional, subordinated feature for individual texts, no longer as a structural feature.

This raises the question of a hierarchical structure for these parameters. If, for historical research, a reasonable scope per grid is already not possible in the third dimension until the early modern period, a decision about what the first two criteria should be can only be made depending on the topic (target-oriented). The order of time, region, (text type) proved to be advisable for the analysis and representation of the grammar. However, in structuring the corpus for the *Middle High German grammar* according to the categories of "verse", "official documents" and "(other) prose", further subdivision in terms of text type

was not fully feasible. Nevertheless, a sensible distribution of the text types was always aspired to if the amount of material allowed a selection.

Another feature of structured corpora is their organization in terms of the size of the individual texts as well as in terms of the size of the whole corpus. Initially, it seems evident that one should always take complete texts. On closer inspection however, this approach is not necessarily ideal, because texts vary in length. In order to ensure the comparability of individual texts, it is helpful to create equivalent lengths and accordingly limit the length of longer texts. Comparability can also be ensured with texts of unequal length by using an arithmetic operation, i.e. each figure is related to the total number of all tokens or all types (for lemmatized material) and the quotients are then compared. There is another factor, though. If time and funding are limited (as is the case with many projects), it is often better to manually digitize only as much text as appears optimal for a specific research question. At this point, one enters a minefield. So far, there is no reliable knowledge about how extensive the material for a particular study has to be.[18] We can say from approximately 35 years of experience that a text length of 12,000 word forms is the minimum level for inflectional morphology, below which too much targeted additional data collection is necessary. The maximum is 28,000 word forms, above which something new is rarely to be found. The optimal level of acceptable expense and result turns out to be exactly 20,000 word forms. Something similar is true for word formation, though only inthe case of highly productive word formation patterns. Rare meanings are only occasionally to be found in further text sequences, though. For graphemic and phonological analysis, a much smaller amount of text is required. We have relatively little experienceswith syntax, where the required scope is strongly dependent on individual syntactic phenomena. However, this also means that many interesting, datable and localizable texts are too short. From time to time, here too, compromises are necessary.

The large-scale so-called *Historical Reference Corpus of German* (*Historisches Referenzkorpus des Deutschen*), better: *Reference Archive* (formerly *DDD* = "Deutsch-Diachron-Digital"), found a pragmatic solution, having encountered tension between the demands of comprehensive text archives which aim if possible to be comprehensive, and manageable, planned and structured archives which have already served as corpora. The *Reference Archive* has a double orientation due to the nature of the written records. A corpus of all German

[18] Cf. also de Haan (1992) and Lauer (1995).

writing up to around the year 1200 has already been compiled (i.e. manually digitized, lemmatized and annotated). For the period after 1200, given the currently available technical aids and a limited time frame for which funding is available, only a selection can be made. The question of text length and selection only arises at this point. Since both demands – the one for the broadest possible coverage and the one for the most significant, structured selection – should be met, we decided on a hybrid solution: as many texts as possible are considered, it will be basically possible to add further texts later (in so far as the same formal criteria are maintained). However, structured subcorpora are marked by a header and can be retrieved as a subset, such as, for instance, the corpus of the new *Middle High German grammar* or the *Early New High German grammar*.

3. What do corpora stand for?

For a start, corpora stand only for themselves. Here, the data are hard and the results have a firm foundation. Corpora do, however, always stand for more, in that they imply something more than themselves. In the 1970s and 1980s, there was a short but lively debate in Germany on the question of the representativeness of corpora.[19] It quickly became clear that linguistic corpora can never be representative in a strictly statistical sense for a language or a historical stage of a language, because there is a lack of precise knowledge of the so-called overall population.[20] There has been an agreement that statements about language based on historical linguistic corpora are only exemplary, referring beyond themselves, but not in a representative sense. Regarding historical corpora, this view can be modified.

[19] Cf. Schank (1973), Nikitopoulos (1974), Bungarten (1979), Rieger (1979), König (1982), Bergenholtz/Mugdan (1989).

[20] Clarified once again by Köhler (2005: 5): "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."

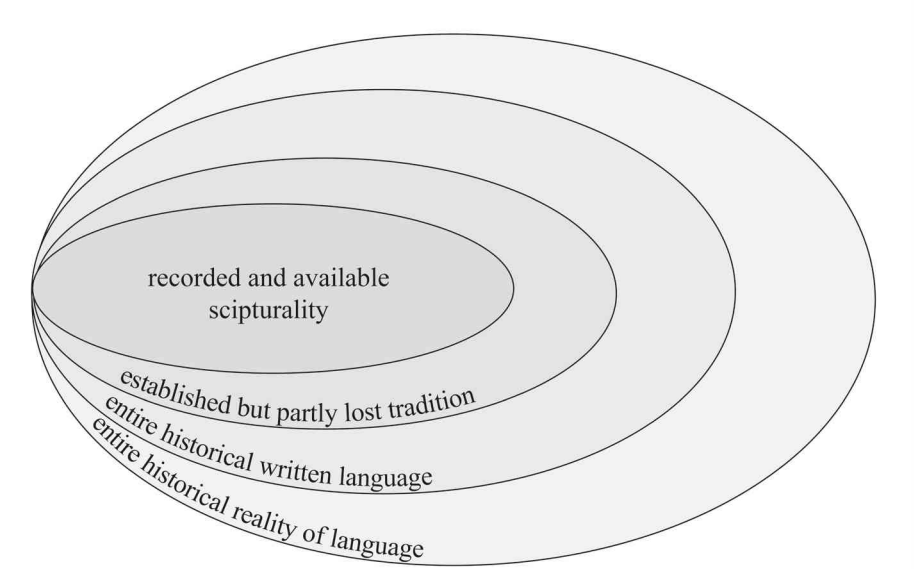


Figure 2: Entire historical reality of language and tradition

Historical corpora like the one for the *Middle High German grammar* may not be representative of **the** Middle High German language and most probably not of **the** written Middle High German language either. However, they can certainly be representative (even in a strict statistical sense) of the known and accessible written records of Middle High German (which have by now been largely recorded completely). Furthermore, structured corpora increase the demand for representativeness (though without accomplishing it) by including, for example, particular variables in the corpus design.[21]

I would like to illustrate this with an example showing that even results of analyses made on the basis of this tradition do not allow a straightforward interpretation.

[21] For such strategies in corpus generation see also Biber (1993).

Period	Region						
	Upper German				Central German		
					West CG	Hess.-Thur.	
	Alem.	Alem.-Bav.	E.Franc.	Bav.	C.Franc.	Rh.Franc.-Hess.	East CG
$^{2}11/^{1}12$	Single attestation				–		
$^{2}12$	<5%	<10%	–	5%	0	0	
$^{1}13$	<5%	31% 14/55%	–	<10%	0	Single attestation	
$^{2}13$	13% <5 - 22%	17% <5 - 59%	–	34% 13 - 69%	<5%	<5%	<5%
$^{1}14$	15% <5 - 31%	63% <5 - 100%	59% 12 - 84%	89% 31 - 100%	<10%	<5%	<10%

Table 1: Percentages of apocope with dative *-e* (former masculine *a*- and *i*-stem nouns). NB that $^{2}12$ indicates the 2nd half of the 12th century etc. The figures should be understood as the average of all texts. In the case of greater variation, the texts with the highest and the lowest percentage are given as a range, e.g. in Bavarian in $^{1}14$ the average is 89% with a range of 31%-100%.[22]

Table 1 shows the development of the apocope of *-e* in the dative singular of so-called strong masculine nouns. There are relatively hard data where there is no evidence for any such cases of apocope. It is highly probable that little would change even if further texts from the same language area and the same time period were to be available and taken into account. This also applies in general for results with percentages below 5% (mostly these are only 2-3%). In contrast, the quantitative results for Upper German in the thirteenth and fourteenth century are problematic, and Bavarian in the first half of the fourteenth century can serve as an example. Here, we have four texts, including one text in which the apocope of the dative *-e* has already been completely implemented. Apart from this we have one text which shows the apocope of *-e* in fewer than one-third of examples. Here, we have a hard datum, too: there are texts in the first half of the fourteenth century in Bavarian with complete apocope of *-e*. But what do these texts generally stand for? If the weighted arithmetic mean was generated by regionally homogenizing the texts, a figure of 89% apocope of *-e* follows. This percentage masks the variability of apocope in Bavarian in the first half of the fourteenth century. Only the honest option is left, namely to offer a variable grammar, pointing out the (documented) frame of the variables (in this case between 31% and 100%) besides the mean. This has an effect

[22] Following Klein/Solms/Wegera (forthcoming).

on the method of the description. The form of the grammar has to refrain from overly general rules if possible, and has to be oriented methodologically towards grammars that make diachronic developments and variability their principle of representation. The only valid statement is thus one that refers directly to the results of the corpus, i.e., for example:

> In the corpus, an average percentage of 89% for the apocope of the masculine dative *-e* is documented for Bavarian in the first half of the fourteenth century. At the same time, the percentage varies in all four texts between 31% and 100%.

As is generally known, the paradox occurs with statements that go beyond the corpus, so that the statement becomes more correct the less precisely it is formulated.

The situation could, of course be formulated as follows:

> In Bavarian in ^{1}14, there is an average percentage for the apocope of the masculine dative *-e* of more than 50%

or even:

> In Bavarian, apocope already occurs frequently in ^{1}14 with masculine nouns in the dative case.

However correct such formulations might be, they are not very illuminating, in that they hardly go beyond known statements of older grammars. On the other hand, a statement like the following would be wrong because every additional text would alter the figures:

> In Bavarian, the average percentage for apocope of the masculine dative *-e* in ^{1}14 is 89%.

A statement that is probably at least 90% correct can, after all, be made using a calculation of the confidence interval (here according to Wilson):[23]

> In Bavarian, the average percentage for apocope of the masculine dative *-e* in ^{1}14 is between 85% and 92% (85,31%-92,04% to be precise).

4. How can corpora be used?

Corpus material can be utilized in quite different ways. In the German philological tradition (and to a certain extent in other branches of scholarship), until a few decades ago data from material collections (mostly in the form of

[23] Wilson (1927).

card indexes) were often used after the fact. Thus, statements were supported, if not proven, which had previously been extracted on the basis of wider knowledge which could not be retraced. The search strategy in the analysis was such that only positive evidence was chosen. Evidence which deviated from this, or which was not fully compliant with the statement, was not explicitly looked for. At worst it might even be withheld, or at best listed in a selection of exceptions. Typical statements made on the basis of such extracted material can be found in great numbers in older historical grammars. This procedure has justifiably been compared to the excavation of a quarry. Today, it is noticeably different in that corpora are mostly evaluated exhaustively. Yet, there is a high risk of only selecting evidence which confirms a particular exception.[24] The distinction between a corpus-based and a corpus-driven approach[25] is a purely theoretical construct, in so far as one does not simply associate the terms with older or more modern procedures respectively,[26] but considers them basically equal for the exploitation of language material. Working with a corpus can be more corpus-based (better: 'theory-driven') or more corpus-driven, but insisting on only one methodological procedure is impractical. No study approaches a corpus completely naively, without any previous knowledge, and no seriousstudy prefers theoretical presuppositions to actual results.

[24] It would seem strange, however, if this distinction were associated with different approaches, and distinctions like "Philology vs. Linguistics" (Lehmann 2007) or even *Sprachwissenschaft* vs. *Linguistik* (Wolf 2010: esp. 24) associated with different preferences or common practices of corpus exploitation.

[25] Cf. Tognini-Bonelli (2001), Storjohann (2005), Mindt (2010). The term 'corpus-based' could better be replaced by the term 'theory-driven' for the purpose of this distinction. For the distinction 'theory-driven' vs. 'corpus-driven' in computational linguistics, see Dipper (2008).

[26] Most of the time the term 'corpus-driven' is identified with older procedures and thus unjustifiably associated with negative connotations, or only seen as progress compared to work with 'invented examples' such as: "Das Korpus wird als Fundus für authentische Sprachbeispiele angesehen, auf die anhand ausgewählter Beispiele zurückgegriffen wird [...]. Die Sprachdaten aus dem Korpus werden weder in ihrer Gesamtheit noch nach einheitlichen Kriterien in systematischer Art und Weise untersucht [...]" (Mindt 2010: 53f. with reference to Tognini-Bonelli 2001). Instead, the 'corpus-driven' procedure is hailed as progress: "Demgegenüber kann konstatiert werden, dass der korpusgeleitete Ansatz der deutlich innovativere und auch der arbeitsintensivere ist. Er kann sich nicht auf eine Auswahl der Daten stützen, sondern erfordert die Berücksichtigung aller Fälle."(Mindt 2010: 61).

One special procedure is the so-called *gesamthaft* (roughly: 'fully comprehensive') analysis,[27] which has been used within the framework of the *Grammatik des Frühneuhochdeutschen* (*Early New High German Grammar*). *Gesamthaft* here does not just mean that a corpus is evaluated completely (which goes without saying); neither does it mean 'exhaustive' (which also goes without saying). *Gesamthaft* means a slightly different search direction, which not only takes account of the positive evidence for a phenomenon, but also all the competing evidence. This method may seem obvious, but it is still applied comparatively rarely.

This approach may be clarified by an example. In order to analyze and describe the diachronic development of the noun plural suffix *-er*, all documented *-er* plurals can be looked up and listed exhaustively for the entire corpus. By taking the evidence into account, first assumptions can then be made about the diachronic development, too. Such an analysis also records all "negative" cases beyond this evidence, in other words all deviating or differently documented forms (competitors). The overall population in question is not formally determined anymore, but is a mixture of formal and functional aspects. This is comparatively easy, if one points out all plural allomorphs in terms of their relative percentages and diachronic changes. However, this proves to be far more complicated if *-er* is compared with its direct competitors, namely the *-e* plural or the unmarked plural:

MHG. pl. *diu kint-* NHG. *die Kind-er*
MHG. pl. *diu wort-* NHG. *die Wört-er* **or** *die Wort-e*

In this case, the positive evidence can be easily identified, because *-er* is straightforwardly segmentable. Determination of the competing forms can be achieved from different perspectives. The total number of former *a*-stem neuter nouns and Old High German *-er* plurals can be established as the overall population from which the respective percentage and the development of the plural in *-er* can be measured. This approach is appropriate for Middle High German because no masculine nouns yet show plurals in *-er*. The development of the plural allomorph *-er* and its relation to the other formatives at whose expense it expands is the focus of the investigation.

However, the plural marker *-er* can also be related to alternative exponents of plural. In this case, all documented segmentable plural allomorphs constitute the total overall population without exception. Here, the chief interest is di-

[27] This term is often equated with 'exhaustive', but I mean something different.

rected at the relative proportion of the plural marker *-er*, compared to that of the other exponents of plural number on the noun. In this case the first procedure is to be preferred, as the increase in this relatively small group of nouns is marginalized by the profound changes in the highly frequent plural markers when taking into account all plural forms.

I conclude from the reasons set out above that corpus analysis should always specify with respect to the subject matter:

- the degree of **completeness** (entire corpus vs. subcorpora)
- the degree of **exhaustivity** (all evidence, evidence of a sequence, or every x^{th} piece of evidence)
- the degree of ***Gesamthaftigkeit*** (two, several, or all competitors of a phenomenon).

References

Bergenholtz, Henning/Mugdan, Joachim (1989): Korpusproblematik in der Computerlinguistik: Konstruktionsprinzipien und Repräsentativität. In: Bátori, István S./Lenders, Winfried/Putschke, Wolfgang (eds.): Computational linguistics / Computerlinguistik. An international handbook on computer-oriented language research and applications / Ein internationales Handbuch zur computergestützten Sprachforschung und ihrer Anwendungen. (= Handbücher zur Sprach- und Kommunikationswissenschaft 4). Berlin/New York: de Gruyter, 141-149.

Bergenholtz, Henning/Schaeder, Burkhard (eds.) (1979): Empirische Textwissenschaft. Aufbau und Auswertung von Text-Corpora. Königstein: Scriptor.

Biber, Douglas (1993): Representativeness in corpus design. In: Literary and Linguistic Computing 8: 243-257.

Bungarten, Theo (1979): Das Korpus als empirische Grundlage in der Linguistik und Literaturwissenschaft. In: Bergenholtz/Schaeder (eds.), 28-51.

Claridge, Claudia (2008): Historical corpora. In: Lüdeling/Kytö (eds.), 242-259.

Curzan, Anne/Palmer, Chris C. (2006): The importance of historical corpora, reliability and reading. In: Facchinetti, Roberta/Rissanen, Matti (eds.): Corpus-based studies in diachronic English. (= Linguistic Insights 31). Bern, etc.: Peter Lang, 17-34.

Curzan, Anne (2009): Historical corpus linguistics and evidence of language change. In: Lüdeling/Kytö (eds.), 1091-1109.

Dipper, Stefanie (2008): Theory-driven and corpus-driven computational linguistics, and the use of corpora. In: Lüdeling/Kytö (eds.), 68-96.

Häcki Buhofer, Annelies (ed.) (2009): Fortschritte in Sprach- und Textkorpusdesign und linguistischer Korpusanalyse I u. II / Proceedings in language and text corpus design and linguistic corpus analysis I and II. http://www.linguistik-online.de/38_09/ [part I]; http://www.linguistik-online.de/39_09/ [part II] (both last visited: 24 June 2013).

Heikkinen, Vesa/Mikko, Lounela (2008): Small corpus, great institution – and an attempt to understand them. http://scidok.sulb.uni-saarland.de/volltexte/2008/1690/pdf/Heikkinen_Lounela_form.pdf (last visited: 24 June 2013).

Hoffmann, Walter (1998): Probleme der Korpusbildung in der Sprachgeschichtsschreibung und Dokumentation vorhandener Korpora. In: Besch, Werner et al. (eds.): Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung. Vol. 1. 2. Ed.(= Handbücher zur Sprach- und Kommunikationsforschung 2.1). Berlin/New York: de Gruyter, 875-889.

de Haan, Pieter (1992): The optimum corpus sample size? In: Leitner, Gerhard (ed.): New directions in English language corpora. Methodology, results, software developments. (= Topics in English Linguistics 9). Berlin: Mouton de Gruyter, 3-19.

Johansson, Stig (2008): Some aspects of the development of corpus linguistics in the 1970s and 1980s. In: Lüdeling/Kytö (eds.), 33-53.

Kallmeyer, Werner/Zifonun, Gisela (eds.) (2007): Sprachkorpora – Datenmengen und Erkenntnisfortschritt. Jahrbuch 2006 des Instituts für Deutsche Sprache. Berlin: de Gruyter.

Klein, Thomas (1991): Zur Frage der Korpusbildung und zur computerunterstützten grammatischen Auswertung mittelhochdeutscher Quellen. In: Wegera, Klaus-Peter (ed.): Mittelhochdeutsch als Aufgabe. (= Zeitschrift für deutsche Philologie (Sonderheft) 110): Berlin: Erich Schmidt, 3-23.

Klein, Thomas/Solms, Hans-Joachim/Wegera, Klaus-Peter (2009): Mittelhochdeutsche Grammatik. Vol. III: Wortbildung. Tübingen: Niemeyer.

Klein, Thomas/Solms, Hans-Joachim/Wegera, Klaus-Peter (forthcoming): Mittelhochdeutsche Grammatik. Vol. II: Flexionsmorphologie. Berlin/New York: de Gruyter.

Köhler, Reinhard (2005): Korpuslinguistik. Zu wissenschaftstheoretischen Grundlagen und methodologischen Perspektiven. In: LDV-Forum 20/2, 1-16.

König, Werner (1982): Probleme der Repräsentativität in der Dialektologie. In: Besch, Werner u.a. (eds.): Dialektologie. Ein Handbuch zur deutschen und allgemeinen Dialektforschung. Vol. 1. (= Handbücher zur Sprach- und Kommunikationsforschung 1.1). Berlin/New York: de Gruyter, 463-485.

Kratochvílová, Iva/Wolf, Norbert R. (eds.) (2010): Kompendium Korpuslinguistik. Eine Bestandsaufnahme aus deutsch-tschechischer Perspektive. (= Germanistische Bibliothek 38). Heidelberg: Winter.

Lauer, Marc (1995): How much is enough? Data requirements for statistical NLP. In: Proceedings of the second conference of the Pacific Association for Computational Linguistics. Brisbane/Australia, 1-9. http://xxx.lanl.gov/PS_cache/cmp-lg/pdf/9509/9509001v1.pdf (last visited: 24 June 2013).

Lehmann, Christian (2007): Daten – Korpora – Dokumentation. In: Kallmeyer/Zifonun (eds.), 9-27.

Lemnitzer, Lothar/Zinsmeister, Heike (2010): Korpuslinguistik. Eine Einführung. 2. ed. Tübingen: Narr.

Lenz, Susanne (2000): Korpuslinguistik. (= Studienbibliographien Sprachwissenschaft 32). Tübingen: Stauffenburg.

Lüdeling, Anke/Kytö, Merja (eds.) (2008): Corpus linguistics: an international handbook. Vol. 1. (= Handbücher zur Sprach- und Kommunikationswissenschaft 29.1). Berlin/New York: de Gruyter.

Lüdeling, Anke/Kytö, Merja (eds.) (2009): Corpus linguistics: an international handbook. Vol. 2. (= Handbücher zur Sprach- und Kommunikationswissenschaft 29.2). Berlin/New York: de Gruyter.

Mindt, Ilka (2010): Methoden der Korpuslinguistik: Der korpus-basierte und der korpus-geleitete Ansatz. In: Kratochvílová/Wolf (eds.), 53-65.

Mukherjee, Joybrato (2009): Anglistische Korpuslinguistik. Eine Einführung. (= Grundlagen der Anglistik und Amerikanistik 33). Berlin: Erich Schmidt.

Nikitopoulos, Pantelis (1974): Vorgriffe auf eine Thematisierung der Repräsentativität eines Corpus. In: Deutsche Sprache 1/74: 32-42.

Perkuhn, Rainer/Keibel, Holger/Kupietz, Marc (2012): Korpuslinguistik. (= UTB 3433). Paderborn: Fink.

Rieger, Burghard (1979): Repräsentativität. Von der Unangemessenheit eines Begriffs zur Kennzeichnung eines Problems linguistischer Korpusbildung. In: Bergenholtz/Schaeder (eds.), 52-70.

Rissanen, Matti (2008): Corpus linguistics and historical linguistics. In: Lüdeling/Kytö (eds.), 53-68.

Schank, Gerd (1973): Zur Korpusfrage in der Linguistik. In: Deutsche Sprache 4/73: 16-26.

Scherer, Carmen (2006): Korpuslinguistik.(= Kurze Einführungen in die germanistische Linguistik 2). Heidelberg: Winter.

Schwitalla, Johannes/Wegstein, Werner (eds.) (2005): Korpuslinguistik deutsch: synchron – diachron – kontrastiv. Würzburger Kolloquium 2003. Tübingen: Narr.

Storjohann, Petra (2005): Corpus-driven vs. corpus-based approach to the study of relational patters. In: Proceedings of the Corpus Linguistics Conference 2005 in Birmingham. http://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference

-archives/2005-journal/PhraseologyandPatterns/paper045PetraStorjohann.doc (last visited: 24 June 2013).

Tognini-Bonelli, Elena (2001): Corpus linguistics at work. (= Studies in Corpus Linguistics 6). Amsterdam: Benjamins.

Wegera, Klaus-Peter (1990): Mittelhochdeutsche Grammatik und Sprachgeschichte. In: Besch, Werner (ed.): Deutsche Sprachgeschichte. Grundlagen, Methoden, Perspektiven. Festschrift für Johannes Erben. Frankfurt a.M. etc.: Peter Lang, 103-113.

Wegera, Klaus-Peter (2000): Grundlagenprobleme einer mittelhochdeutschen Grammatik. In: Besch, Werner et al. (eds.): Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung. 2nd edition. Vol. 2. (= Handbücher zur Sprach- und Kommunikationsforschung 2.2). Berlin/New York: de Gruyter, 1304-1320.

Wilson, Edwin B. (1927): Probable inference, the law of succession, and statistical inference. In: Journal of the American Statistical Association 22: 209-212.

Wolf, Norbert Richard (2010): Korpora in der Korpuslinguistik. In: Kratochvílová/ Wolf (eds.), 17-25.

Corpora:

http://www.korpora.org/fnhd/ (last visited: 24 June 2013).

http://www.degruyter.com/view/supplement/9783111844206_Quellenkorpus_Uebersicht.pdf (last visited: 24 June 2013).

http://www.llc.manchester.ac.uk/research/projects/germanc/ (last visited: 24 June 2013).

II. Historical linguistic corpora: Analysis, projects, and results

ANITA AUER / TONY FAIRMAN

Letters of artisans and the labouring poor (England, *c.* 1750-1835)

Abstract

The majority of language corpora available to date that cover the Late Modern English period (1700-1900) contain samples of writing by the classically educated layers of society. It is this kind of data that the 'standard' history of the English language has been based on. The labouring poor formed the greater part of the population (60-70%) during the Late Modern English period and, though many of them could not write (compulsory elementary schooling was only introduced in 1870), by about 1800 so many could write something that they formed the majority of those normally called 'literate'. A unique insight into the language use of the labouring poor has been provided through the laws for poor relief, which gave paupers the opportunity to apply for relief from parish funds during the period 1795-1834. For the last 18 years Tony Fairman has collected poor relief application letters from archives of English County Record Offices. This paper describes the data and the compilation principles of the letter collection, as well as the challenges involved in the conversion of the letter collection into a searchable corpus.

1. Introduction

In the field of English historical linguistics, and in other languages, we have seen an enormous increase in the compilation of letter corpora over the last decade.[1] In periods for which only written data are available to the linguist, letters may be seen as a text type that is closer to the spoken medium and therefore more likely to foster linguistic innovation, cf. Elspaß (2003) and Nevalainen/Raumolin-Brunberg (2003). Moreover, letters can allow us to catch a glimpse of the language used by people who are rarely represented in any of the other text types.[2] For instance, the 'standard' history of the Late Modern English period (1700-1900) in England has been largely based on ma-

[1] For an overview of existing diachronic letter corpora in English, as well as historical corpora of other genres, see the *Corpus Resource Database* (*CoRD*) hosted by the Research Unit for Variation, Contacts and Change in English at the University of Helsinki, Finland. http://www.helsinki.fi/varieng/CoRD/index.html, accessed on 26 October 2011.

[2] Other text types that provide an insight into the language use of the lower classes are plays, diaries, and court depositions, as for instance *The Old Bailey Corpus*.

terial that was written by the classically educated[3] layers of society, e.g. *A Representative Corpus of Historical English Registers*, (cf. Biber et al. 1994), and individually compiled corpora of selected educated letter writers, such as letters contained in the *Network of Eighteenth Century English Texts* (*NEET*) by Fitzmaurice (2007), *Language and letters of the Bluestocking network* by Sairio (2009) and *The Bishop's grammar: Robert Lowth and the rise of prescriptivism* by Tieken-Boon van Ostade (2011).[4] While the available studies of educated letter writers certainly reveal interesting patterns of language variation within their correspondence, the group of educated writers cannot be considered representative of the population at the time. In fact, the educated only formed a small part of the population, as opposed to the labouring poor who constituted between 60 and 70% of the population. While compulsory elementary schooling was only introduced in 1870 (the Elementary Education Act), by about 1800 so many, including also the labouring poor, could write something that they formed the majority of those normally called "literate" (See Cressy 1980 and Fairman 2012). A unique opportunity to gain insight into the language use of the labouring poor has been provided through the laws for poor relief, which legalized the payment and receipt of out-relief from parish funds during the period 1795-1834. More precisely, the poor law passed in 1795 entitled people "in distress" to apply for "out-relief", if they lived outside the parish in which they had formal settlement. If the officials accepted the applicants' claims, assistance was sent, which could either mean removal from the current domicile to the parish of legal settlement, or money being sent (cf. Whyte, 2004: 280). Many application letters for poor relief survived in archives of County Record Offices. For the last 18 years the independent scholar Tony Fairman has collected a great number of these application letters from all over England. These unique data supplement, and may even lead to a revision of,

[3] Whenever we use the term 'educated', it means classically educated henceforth.

[4] Over the last decade there has been an increasing interest in historical linguistics in the language of letters of the lower social classes in different varieties of English. See for instance *The Corpus of Older African American Letters* (*COAAL*) compiled by Edgar Schneider, Lucia Siebers, and Michael Montgomery (now publicly accessible); *The Hamburg Corpus of Irish English* (*HCIE*) compiled by Peter Siemund, Lukas Pietsch and associates at Hamburg University (not publicly accessible), as well as in different languages, e.g. the special issue of Multilingua on lower class language in the 19th century, Vandenbussche/Elspaß (eds.) (2007). While we are aware of ongoing lower-class letter projects, none of these corpora have yet been made publicly available; in this paper it is therefore not possible to draw comparisons with respect to any other similar corpus compilation projects and the challenges involved.

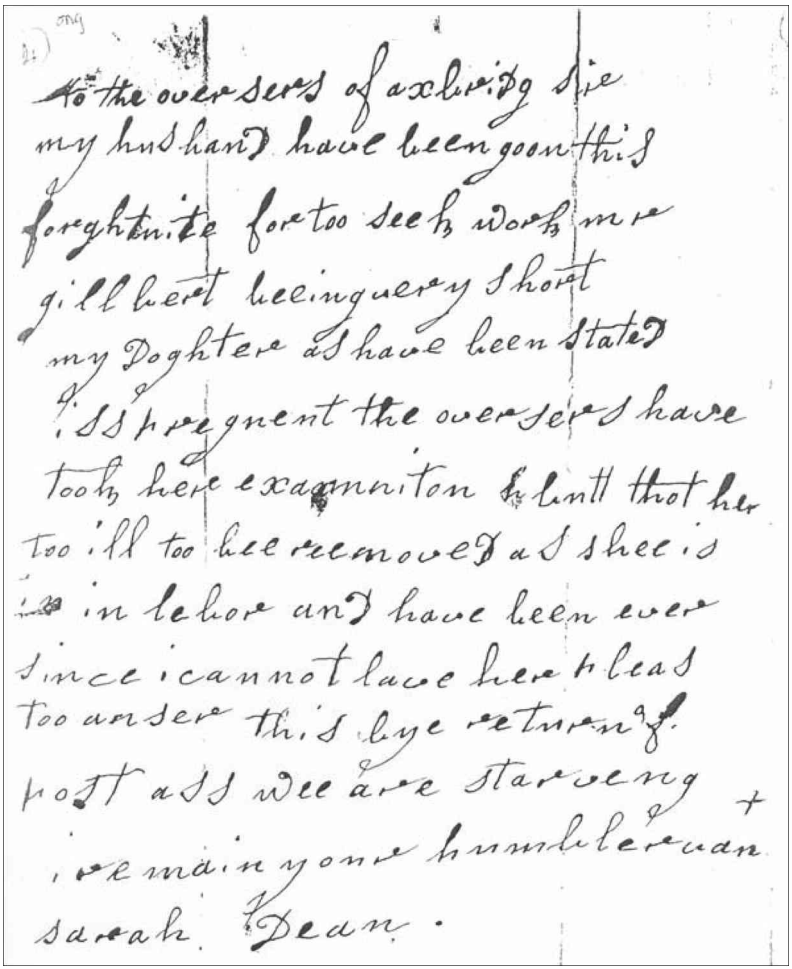
what has hitherto been assumed to be a history of the English language, by adding data "from below", in this case data from the lower social stratum, and will therefore be of great interest to linguists as well as social historians.

The main aim of this paper is to introduce the *Letters of Artisans and the Labouring Poor* (*England, c. 1750-1835*) and to describe the conversion of this letter collection into a corpus. The paper is divided into two parts. The first part (2) describes the data and the compilation principles of the letter collection. The second part (3) is concerned with the challenges involved in the conversion of the letter collection into a searchable corpus. Finally, the current state of the corpus project is briefly outlined.

2. The data and compilation principles

In corpus linguistics, the view is commonly held that a corpus is "not simply a collection of texts" (Biber et al. 1998: 246). This implies that the creation of a corpus is based on a detailed plan that carefully considers the criteria of sampling, balancing and representativeness. While the creation of many corpora is indeed based on a detailed plan that tries to consider the above-mentioned criteria, the starting point of some corpora is an existing collection of texts representing a particular language/variety/genre, that does not and cannot fulfil these criteria. This is precisely the case with the *Letters of Artisans and the Labouring Poor* (*England, c. 1750-1835*), for which we use the acronym *LALP*.

The collection of poor relief application letters from every English county consists of 2046 letters containing around 303473 orthographic units [January 2012], groups of graphs that were deliberately separated by the letter writers. When Fairman started collecting poor relief application letters, his aim was to find out more about lower-class writing during the period 1750-1835. This was also the time when the English language had been codified in grammars and dictionaries, but schooling and therefore literacy had largely been a prerogative of the upper layers of society. Nevertheless, as pointed out earlier, it may be assumed that many people (in all layers of society) were able to write something by about 1800. Given that the poor law passed in 1795 entitled people "in distress" to apply for "out-relief", the poor, even though they may not have received much training in writing, needed to write these letters of application in order to survive. The opportunities for schooling, and therefore the levels of literacy, differed greatly, which is clearly reflected in the letters contained in the collection, as illustrated by examples 1 and 2 below.

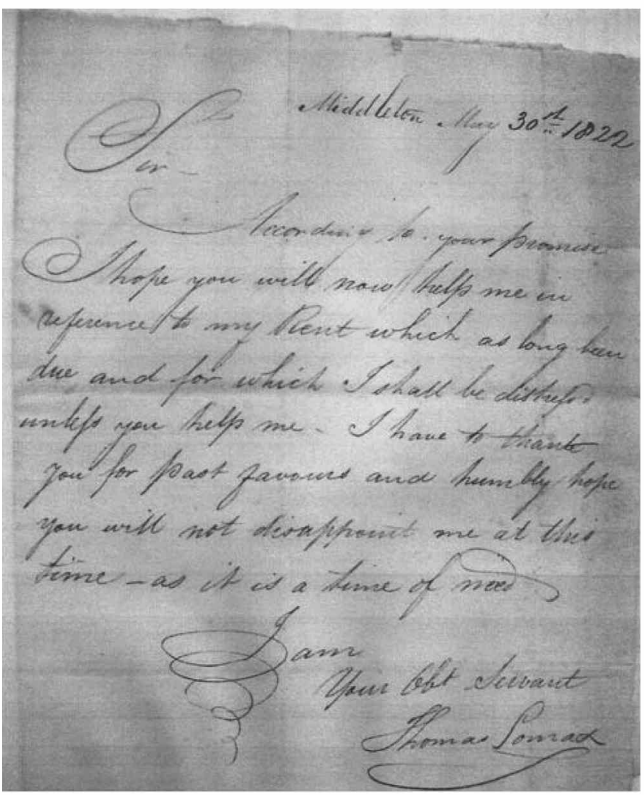


Letter 1: Letter by Sarah Dean, 5 July 1830, parish Axbridge
(Somerset Record Office (Taunton): D/P/ax/13/3/6)

to the oversers of axbridg sir
my hushanD have been goon thi
forghtnite for too seek work mr
gillbert beeing very short
my Doghter as have been stateD
iss pregnent the oversers have
took her exaamniton h butt that her
too ill too bee reemoveD as shee iS

```
in in lebor anD have been ever
since i cannot lave her pleas
too anser this bye return of
post ass wee are starveng
i remain your humblervant
sarah Dean
```

Apart from variant spelling and grammar and lack of punctuation, the handwriting in the letter shows that the letter writer wrote down or drew one graph after the other. The spaces between all the graphs, and thus the lack of joined-up writing, indicate that the writer was not particularly experienced in writing, as opposed to the writer of letter 2 below.



Letter 2: Letter by Thomas Lomax, 30 May 1822, parish Holcombe, Bury (Greater Manchester County Record Office: GB127.L21/3/13/5)

```
    Middleton, May 30st 1822
Sir
      According to your promise
I hope you will now help me in
Reference to my Rent which as long been
due and for which I shall be distreſs'd
unless you help me. I have to thank
you for past favours and humbly hope
you will not disappoint me at this
time - as it is a time of need
      I am
           Your obt Servan
                      Thomas Lomax
```

Letter 2 greatly differs from letter 1, apart from the lack of punctuation. The spelling of letter 2 is very close to modern English standard spelling and grammar. The only deviation from the standard form is the dropped initial *h* in *has*, e.g. *as long been due*, which appears to be a reflection of speech. Apart from the near-standard language usage, indicating a high degree of training in writing and education in general, the fluent and joined-up handwriting similarly reveals that the letter was written by an experienced writer.

In fact, at the time, a distinction was made between two different types of training in writing, namely mechanical schooling and grammatical schooling. The contemporary use of these terms is illustrated by school advertisements in local newspapers, (see also Fairman 2007):

> "**WRITING** in all the most useful **Hands**" (i.e. mechanical schooling, at Allfree's boys' and girls' Boarding School, Herstmonceux, Sussex, 1771; Caffyn 1998: 132)
>
> "particular attention will be paid to teach the English language **grammatically**" (Button's English Classical Academy, Lewes, Sussex, 1792; Caffyn 1998: 176)

The upper layers of society, who found themselves in an advantageous position with respect to education, were not only taught the mechanical aspect of writing, but also had the opportunity to receive a classical education, which laid the foundation for learning grammatical English. Artisans and the labouring poor, on the other hand, whose access to education was restricted, were often merely taught how to write mechanically, i.e. how to draw graphs. The poorer layers of society received education in Sunday schools, charity schools, dame schools and/or they taught themselves (Lawson/Silver 1973: 189-195; 238-250). The difference between grammatical and mechanical schooling was not only reflected in school advertisements, but can even be found in contempo-

rary literature, as illustrated in an extract from the poem *The Parish Register* (1807) by the parson poet George Crabbe:

> how strange that men,
> Who guide the plough, should fail to guide the pen.
> For half a mile the furrows even lie;
> For half an inch the letters stand awry.

In line with the school advertisements, the poem extract suggests that the labouring poor had difficulties writing, which in this particular case refers to mechanical writing problems. In practice, the upper-lower-class-division was not so severe in terms of education, and the boundaries were often blurred. As there was no national curriculum, nor any cross-national quality comparisons at the time, the individual's education depended on the opportunities for education and the actual teaching received, as well as on the person's ambition. As every letter writer received different training, s/he put this knowledge into practice differently, so that the differing usage may be charted on a continuum (of spelling, morphology, syntax, and lexicon), extending up to the modern standard. On his quest to collect letters written by paupers and artisans, Fairman was on the lookout for letters that deviated from modern Standard English as much as possible, which is well illustrated in letter 1 (original and transcription above). One may thus want to argue that "non-probability" sampling, also referred to as "convenience" or "opportunistic" sampling (cf. Nelson 2010: 57), has been applied for the creation of this letter collection. The current content of the collection is listed in Table 1, in county order.

Even though *LALP* is based on opportunistic sampling, the question still arises as to how representative the letters contained in the collection are of application letters actually sent. After all, it is striking that the county record offices of Kent hold at least 769 letters, while Bedfordshire has only 5.

Information on the total number of out-relief application letters sent to parishes is difficult to gather, as not all of the records survived and/or can be traced. If the records survived, light can only be shed on the total number of application letters by tracing relief-books (account books) of the overseers from all the different parishes. These relief-books list not only the rates paid to the applicants, but also give details on the recipients and the reasons for their relief (see also the *First Annual Report of the Poor Law Commissioners for England and Wales* 1835). This information is available for selected parishes, where the records survive, but it is impossible to discover how many application letters for out-relief had actually been sent to parishes during the period 1795-1834 in the whole of England. Similarly, it is difficult to trace how many of these letters

County*	Letters (parishes)	Orthographic Units
1 Northumberland	11 letters from 3 parishes	3504
2 Cumberland	20 letters from 4 parishes	2541
3 Lancashire	56 letters from 7 parishes	8132
4 Westmorland	88 letters from 2 parishes	18469
5 Durham	38 letters from 3 parishes	3256
6 Yorkshire	138 letters from 14 parishes	17809
7 Cheshire	19 letters from 4 parishes	2509
8 Derbyshire	21 letters from 7 parishes	3693
9 Nottinghamshire	16 letters from 4 parishes	2296
10 Lincolnshire	17 letters from 7 parishes	2083
11 Shropshire	53 letters from 4 parishes	10881
12 Staffordshire	30 letters from 6 parishes	7988
13 Leicestershire	10 letters from 3 parishes	862
14 Rutland	2 letters from 1 parish	211
15 Norfolk	15 letters from 8 parishes	1926
16 Herefordshire	29 letters from 10 parishes	4498
17 Worcestershire	11 letters from 4 parishes	2217
18 Warwickshire	23 letters from 5 parishes	4429
19 Northamptonshire	14 letters from 2 parishes	1807
20 Huntingdonshire	12 letters from 1 parish	1188
21 Cambridgeshire	35 letters from 5 parishes	4049
22 Suffolk	42 letters from 6 parishes	10000
23 Bedfordshire	5 letters from 3 parishes	760
24 Gloucestershire	23 letters from 5 parishes	3550
25 Oxfordshire	15 letters from 5 parishes	2074
26 Buckinghamshire	29 letters from 4 parishes	3466
27 Hertfordshire	23 letters from 4 parishes	2764
28 Essex	166 letters from 24 parishes, cf. Sokoll (2001)	32963
29 Somerset	25 letters from 9 parishes	3393
30 Wiltshire	43 letters from 6 parishes	6920
31 Berkshire	30 letters from 4 parishes	6716
32 Middlesex [London]	20 letters from 6 parishes	4588
33 Surrey	14 letters from 3 parishes	2314
34 Kent	769 letters from 26 parishes	88754
35 Cornwall	12 letters from 3 parishes	2197
36 Devon	17 letters from 6 parishes	2523
37 Dorset	81 letters from 12 parishes	13371
38 Hampshire	48 letters from 10 parishes	8525
39 Sussex	26 letters from 9 parishes	4247
TOTAL	**2046 letters from 249 parishes**	**303473 words**

* The numbers preceding the individual counties correspond with the numbers in the map titled "The counties of England and Wales in the nineteenth century" in Williams (ed.) (2004: vi).

Table 1: Content of *LALP* as of January 2012.

have survived to date, as they are spread out all over the country in collections, or as single copies in parish and county record offices. Nevertheless, even though *LALP* may only contain a fraction of application letters actually sent at the time when the poor law was operative, the surviving data may be considered of great importance to historical linguists and social historians in that it supplements and may even lead to a revision of what has hitherto been assumed to be a history of the English language by adding data "from below".

3. Challenges in the conversion from letter collection to searchable corpus

In the first instance, we aim to supply data for socio-linguistic research with the *LALP* corpus. Considering that the *Corpus of Early English Correspondence* (*CEEC*; c. 1410-1681) and its supplementary corpora, compiled by the Research Unit for Variation and Change in English (VARIENG) at the University of Helsinki, were specifically designed "to test the applicability of sociolinguistic methods to historical data" (Raumolin-Brunberg/Nevalainen 2007: 148), this group of letter corpora serves as a model for the *LALP* corpus. *LALP* resembles the *CEEC* corpora in that it is a single-genre corpus, i.e. letters, but differs from the latter corpora in that it is restricted to a particular letter subtype, namely the letter of application for poor-relief. In the case of this particular sub-type, the recipients (the parish overseers), despite being different individuals, all have the same function, namely to decide whether the claim for out-relief will be accepted, and if so, in what form the applicants will be helped. In many application letters the parish overseer cannot be identified, as specific names are not mentioned in the address formulae. Extra-linguistic variables can thus only be provided for the senders of the application letters. Due to the applicants' low socio-economic status, it is, in contrast to members of the gentry with a high public profile, difficult to find personal information other than that contained in the application letters. While the sender database of the *CEEC* corpora can contain up to 27 parameters (Raumolin-Brunberg/Nevalainen 2007: 162),[5] the extra-linguistic parameters in the *LALP* corpus are a lot more restricted.

[5] The 27 parameters in the *CEEC* corpora sender database are: 1. Last name, 2. First name, 3. Title, 4. Year of birth, 5. Year of death, 6. First letter, 7. Last letter, 8. Sex, 9. Rank, 10. Father's rank, 11. Social mobility, 12. Place of birth, 13. Main domicile, 14. Migrant, 15. Education, 16. Religion, 17. Number of letters, 18. Number of recipients, 19. Kind of recipients, 20. Number of words, 21. Letter contents, 22. Letter quality, 23. Collection, 24. Career, 25. Migration history, 26. Extra, 27. Complete (Raumolin-Brunberg/Nevalainen 2007: 162).

Before being able to determine extra-linguistic parameters, the question of authenticity needs to be raised. Even though we may assume that the majority of the labouring poor were able to write something by 1800, we cannot be absolutely certain that all the letters are autograph letters. In fact, in some cases it is stated that the application letter for poor-relief was written on behalf of a specific person, a clear indication that we are dealing with a non-autograph letter, that can also be labelled as such. A lot of the letters will not contain this information, however, and may still have been written by somebody else. While one tends to conjecture that the more grammatical letters may have been written by someone other than the applicant, this cannot be proven. At the same time, this lack of knowledge has serious ramifications for the tagging of meta-data, (cf. Nevalainen/Raumolin-Brunberg 1996: 43-45, and Bergs forthc.). To illustrate this point, if we take it that the collection contains an application by a woman aged 40, it may have been the case that this woman was illiterate and therefore asked her neighbour, male and aged 25, to write the letter for her. Not being aware of this, we would tag the letter as being written by a woman, aged 40. These metadata would thus be incorrect and have the effect that the results of a sociolinguistic investigation may be completely skewed. As this problem is unavoidable and irresolvable, the project team concerned with the conversion of the letter collection into a searchable corpus will indicate cases of questionable authenticity in "Extra" in the file header, which contains the sender data. Based on the *CEEC* corpora model, the sender information that can be given in *LALP* is as follows:

1. Last name
2. First name
3. Age (year of birth)
4. Sex
5. Date of letter written
6. Place of current residence
7. Parish of legal settlement
8. Number of letters
9. Number of words
10. Letter contents
11. Extra (authenticity, literacy)

The age of the letter writer can only be given if it is mentioned in the letter itself. As the letters were written by applicants for out-relief, these people have left their parish of legal settlement to look for work elsewhere, their place of current residence. The application for out-relief is addressed to the parish of legal settlement, however. This information not only sheds light on people's migration history, but also on what dialects the applicants may have spoken, based on their parish of legal settlement. While earlier research on the letter collection by Fairman (2007: 275) concludes that "[t]he language of these letters cannot be called dialect. Minimally-schooled English is so similar throughout England that it is possible to consider it as an emerging standard, which the official Standard interrupted", (see also Fairman 2006, and Kortmann/Wagner 2010: 290-291), some reflections of speech contained in the letters may (ideally) shed some light on dialect usage. Having said that, the notion of "parish of legal settlement" requires more explanation. According to seventeenth-century statutes, "everyone should have a single parish of legal settlement in which they were entitled to receive poor relief" (Whyte 2004: 280). He explains further that:

> [s]ettlement rights could be established on the basis of birth, marriage, and, in the nineteenth century, from a father's or even grandfather's parish of settlement. Other mechanisms, such as renting property worth £10 per annum, a year's agricultural service, completing an apprenticeship, paying taxes or serving in a parish office for a year were also grounds for gaining a settlement. People who required poor relief and were living in a parish which was not their parish of settlement could be removed there or, less commonly, be provided with out-relief (ibid.).

The fact that there were many other ways to establish settlement rights other than birth, such as marriage, apprenticeship, and property rental, makes it more complicated to find out where a person originally came from, and thus what original dialect was used. In the case of women, the parish of legal settlement would be determined by marriage. While the parameter "place of legal settlement" may shed some light on a person's origin and also dialect usage, one cannot rely on the fact that the place of legal settlement given in the letters is also a person's birthplace.

Other challenges posed by the *LALP* collection, apart from authenticity and authorship of the letters, are the different levels of literacy and thus the wide range of variant spelling that can be found in the application letters. The letters, which are currently being converted into a computer-readable format,

need to be searchable. As this is not possible with the highly idiosyncratic and therefore unpredictable spelling variants contained in the collection, the spelling needs to be normalized. In order to speed up the normalization process, we have tested a variation detection software named *VARD 2.3*, which can standardize spelling variation, both manually and automatically, in text corpora (see Baron/Rayson 2009).

As this software was initially designed to deal with Early Modern English spelling variation (1500-1700), it can, with the help of some training and complementary tools, successfully normalize spelling variation in the *CEEC* corpora (Baron/Rayson 2009). In contrast to Middle English and Early Modern English data, however, the spelling of words in the *LALP* corpus is less predictable. The problems with the *LALP* data are illustrated in Figure 1.

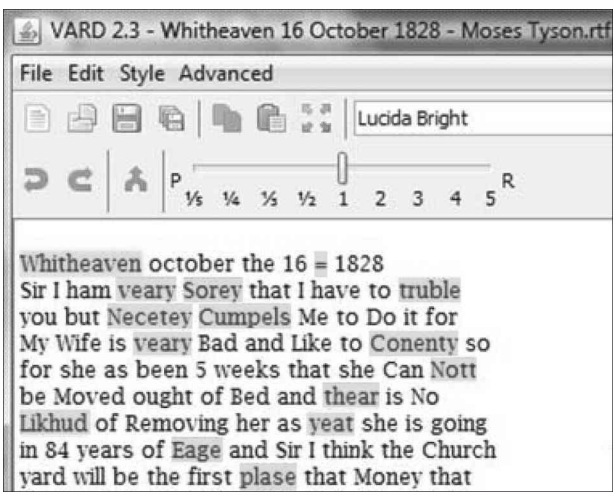


Figure 1: *VARD 2.3* applied to a selected application letter

The words highlighted are those that the software successfully recognizes as variant spelling, e.g. *veary* for *very*, *sorey* for *sorry*, *truble* for *trouble*, *necetey* for *necessity*, *cumpels* for *compels*, *conenty* for *continue*, *thear* for *there*, *likhud* for *likehood*, *yeat* for *yet*, *eage* for *age*, and *plase* for *place*. *VARD 2.3* can be trained so that it recognizes which normalized form the spelling variant corresponds with, and it is thus possible to tag words with normalized forms. Considering that the spelling variants for a particular word can differ from application letter to application letter, and even within one letter, the software needs to be

trained for all possible variants. While this is a straightforward procedure, there are still words in Figure 1 that *VARD 2.3* had difficulties recognizing as non-normalized spelling, e.g. *ham* for *am* in line 3, *as* for *has* in line 5, and *ought* for *out* in line 6. Since the standard lexicon contains the words *ham*, *as* and *ought*, albeit with a different meaning to that in the sample letter, it is difficult for the software to recognize these words as spelling variants. Even though the application of complementary tools such as *DICER* (*Discovery and Investigation of Character Edit Rules*) may improve the software's accuracy, (see Rayson/Baron 2011: 113) with regard to the *LALP* data, we will still need to go through the collection manually in order to ensure that spelling variants have been normalized correctly.

4. Outlook

The *LALP* project is currently preparing a first plain text version of the corpus, which will be completed in June 2012. The technical choices made in the corpus will be outlined in a manual that will be made available at the same time as the first version of the corpus.

The *Letters of Artisans and the Labouring Poor* (*England, c. 1750-1835*) will not only be of great benefit to scholars working on the history of the English language, but also to social historians. Even though the corpus has its disadvantages (cf. Section 3), it provides unique data that allow us to get a more complete picture of what language usage was really like in Late Modern England.

References

Baron, Alistair/Rayson, Paul (2009): Automatic standardisation 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 conference (CL2009), University of Liverpool, UK. http://ucrel.lancs.ac.uk/publications/cl2009/314_FullPaper.pdf

Beal, Joan C./Corrigan, Karen P./Moisl, Hermann L. (eds.) (2007): Creating and digitizing language corpora. Volume 2: Diachronic databases. Houndmills, Basingstoke: Palgrave Macmillan.

Bergs, Alexander (forthc.): Linguistic fingerprints of authors and scribes. In: Auer, Anita/Schreier, Daniel/Watts, Richard J. (eds.): Letter writing and language change. Cambridge: Cambridge University Press.

Biber, Douglas/Finegan, Edward/Atkinson, Dwight/Beck, Ann/Burges, Dennis/Burges, Jene (1994): The design and analysis of the ARCHER corpus: A progress report. In: Kytö, Merja/Rissanen, Matti/Wright, Susan (eds.): Corpora across the centuries. (= Language and Computers. Studies in Practical Linguistics 11). Amsterdam/ Atlanta: Rodopi, 3-6.

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

Caffyn, John (1998): Sussex schools in the eighteenth century: Schooling provision, schoolteachers and scholars. Lewes: Sussex Record Society, No. 81.

Crabbe, George (2007 [1807]): The Parish Register. Teddington: The Echo Library.

Cressy, David (1980): Literacy and social order: Reading and writing in Tudor and Stuart England. Cambridge: Cambridge University Press.

Elspaß, Stephan (2003): A twofold view 'from below'. New perspectives on language histories and historical grammar. In: Elspaß, Stephan/ Langer, Nils/Scharloth, Joachim/Vandenbussche, Wim (eds.): Germanic language histories 'from below' (1700-2000). (= Studia Linguistica Germanica 86). Berlin/New York: Walter de Gruyter, 3-12.

Fairman, Tony (2000): English pauper letters 1800-34 and the English language. In: Barton, David/Hall, Nigel (eds.), Letter writing as a social practice. Amsterdam/ Philadelphia: John Benjamins, 63-82.

Fairman, Tony (2006): Words in English record office documents of the early 1800s. In: Kytö, Merja/Rydén, Mats/Smitterberg, Erik (eds.): Nineteenth-century English: Stability and change. Cambridge: Cambridge University Press, 56-88.

Fairman, Tony (2007): Letters of the English labouring class and the English language, 1800-1834. In: Dossena, Marina/Jones, Charles (eds.): Insights into Late Modern English. Second edition. Bern: Peter Lang, 265-282.

Fairman, Tony (2012): Letters in mechanically-schooled language: Theories and ideologies. In: Dossena, Marina/Del Lungo Camiciotti, Gabriella (eds.): Letter writing in Late Modern Europe. Amsterdam: John Benjamins, 205-228.

Fitzmaurice, Susan (2007): Questions of standardization and representativeness in the development of social networks-based corpora. The story of the network of eighteenth-century English texts. In: Beal et al. (eds.), 49-81.

Kortmann, Bernd/Wagner, Susanne (2010): Changes and continuities in dialect grammar. In: Hickey, Raymond (ed.): Eighteenth-century English. Ideology and change. Cambridge: Cambridge University Press, 269-292.

Lawson, John/Silver, Silver (1973): A Social history of education in England. London: Methuen.

Nelson, Mike (2010): Building a written corpus. What are the basics? In: O'Keeffe, Anne/McCarthy, Michael (eds.): The Routledge handbook of corpus linguistics. London/New York: Routledge, 53-65.

Nevalainen, Terttu/Raumolin-Brunberg, Helena (1996): The corpus of Early English correspondence. In Nevalainen, Terttu/Raumolin-Brunberg, Helena (eds.) Sociolinguistics and language history. Studies based on the corpus of Early English correspondence. Amsterdam/Atlanta: Rodopi, 39-54.

Nevalainen, Terttu/Raumolin-Brunberg, Helena (2003): Historical sociolinguistics: Language change in Tudor and Stuart England. London/New York: Pearson.

Old Bailey Corpus (*OBC*): http://www.oldbaileyonline.org/

Raumolin-Brunberg, Helena/Nevalainen, Terttu (2007): Historical sociolinguistics. The corpus of Early English correspondence. In: Beal et al. (eds.), 148-171.

Rayson, Paul/Baron, Alistair (2011): Automatic error tagging of spelling mistakes in learner corpora. In: Meunier, Fanny/de Cock, Sylvie/Gilquin, Gaëtanelle/Paquot, Magali (eds.): A taste of corpora. In honour of Sylviane Granger. Amsterdam: John Benjamins, 109-126.

Sairio, Anni (2009): Language and letters of the bluestocking network. Sociolinguistic issues in eighteenth-century epistolary English. Helsinki: Société Néophilologique.

Sokoll, Thomas (ed.) (2001): Essex pauper letters 1731-1837. Oxford: Oxford University Press.

Tieken-Boon van Ostade, Ingrid (2011): The Bishop's grammar. Robert Lowth and the rise of prescriptivism. Oxford: Oxford University Press.

van Bergen, Linda/Denison, David (2007): A corpus of late eighteenth-century prose. In: Beal et al. (eds.), 228-46.

Vandenbussche, Wim/Elspaß, Stephan (eds.) (2007): Lower class language use in the 19th century'. In: Multilingua 26 (Special edition), 2-3.

Whyte, Ian (2004): Migration and settlement. In: Williams (ed.), 273-286.

Williams, Chris (ed.) (2004): A Companion to nineteenth-century Britain. Malden, MA: Blackwell.

STEFANIA DEGAETANO-ORTLIEB / HANNAH KERMES /
EKATERINA LAPSHINOVA-KOLTUNSKI / ELKE TEICH

SciTex: a diachronic corpus for analyzing the development of scientific registers

Abstract

In this paper, we report on a project[1] investigating the diachronic development of scientific registers that have emerged in the last thirty years or so as a result of the interdisciplinary contact between selected scientific disciplines and computer science (e.g. computational linguistics or bioinformatics). Our main goal is to gain a better understanding of the principles of register formation in highly specialized scientific domains in this kind of context. For this purpose, we have built a diachronic corpus, the *English Scientific Text Corpus* (*SciTex*). Our theoretical framework is Systemic Functional Linguistics (Halliday 2004) and register/genre theory (Halliday/Hasan 1989; Biber 1988, 1995; Martin 1992). Methodologically, we adopt a variationist approach, looking at lexico-grammatical differences and commonalities between registers under the perspective of recent language change (cf. Mair 2006).

1. Introduction

The investigation of scientific texts is currently a very active research area. Studies are carried out on small text samples or on corpora, ranging from the analysis of single registers (Halliday 1988, O'Halloran 2005) to studies with a wider focus on scientific or academic language (cf., for example, Halliday/Martin 1993, Ventola 1996, Biber 2006 and Hyland 2007). However, an issue that has received little attention so far is the diachronic evolution of scientific registers (see, for example, Halliday 1988 and Banks 2008). In the scientific domain, the pursuit of new knowledge and technological innovation brings about transcendence of the boundaries of established scientific disciplines and the emergence of new, interdisciplinary research fields (seen, for example, in recent years, in bioinformatics, mechatronics and biomechanics). Linguistically, we encounter here a situation of *register contact*, where a newly emerging scientific field draws on the linguistic conventions of two or more established scientific disciplines and possibly develops a new register.

[1] Project 'Registers in contact', funded by Deutsche Forschungsgemeinschaft (DFG) under grant TE-198/2.

The overarching goal of our research is to develop a model of register formation in specialized scientific domains, tracing the major motifs governing the development of a new scientific discipline – *diversification* and *standardization* – in linguistic terms. This involves addressing the following questions:

(1) What are the linguistic features involved in the process of register formation in the scientific domain? Registers are manifested linguistically by particular distributions of lexico-grammatical patterns that are relatively stable in time. The diagnostic for a new register developing would therefore be the observation of redistributions of such patterns. Thus, in analysis, a contrastive approach comparing different scientific registers over time is required.

(2) Which contextual settings are realized by the linguistic features involved in register formation? Register variation is situation-dependent. The canonical view is that situations can be characterized by the parameters of field, tenor and mode of discourse (cf. Halliday 1985, Quirk et al. 1985 and Halliday/Hasan 1989). 'Field' denotes the social action in which participants are engaged in a given situation (e.g. processes and participants); 'tenor' concerns the relationship between participants (e.g. roles and attitudes of participants), and 'mode' is about the symbolic organization of information (information flow, foregrounding and backgrounding of information, etc.). These situational parameters are encoded by particular linguistic subsystems (field: lexis/colligation, tenor: mood and modality, mode: theme-rheme, given-new). It is thus part of the analytical task to interpret the observed linguistic features (and their distributions) in terms of their contextual settings.

The main goal of this paper is to present the particular corpus design (Section 2) and the principal methodology which we adopt in pursuit of our research goals (Section 3). In order to illustrate our approach, we provide selected examples of analysis carried out using the corpus (Section 4). We conclude with a summary (Section 5).

2. *SciTex*: corpus design and processing

To investigate register formation in the scientific domain, we focus on the situation of interdisciplinary contact between computer science and selected other disciplines. The *SciTex* corpus comprises texts from computer science (A-subcorpus), from four interdisciplinary fields (B-subcorpus: computational linguistics, bioinformatics, computer-aided design, microelectronics), and from their respective disciplines of origin (C-subcorpus: linguistics, biology, me-

chanical engineering and electrical engineering) (see Teich/Holtz 2009 and Teich/Fankhauser 2010).

Discipline	Journals (time period)
A-CompSci	Theoretical Computer Science (70s/80s), Journal of the ACM (both), Journal of Computer and System Sciences (both), Journal of Algorithms (2000s)
B1-CompLing	Mechanical Translation (70s/80s), Journal of Computational Linguistics (both), Machine Translation (both), Journal of Natural Language Engineering (2000s)
B2-BioInf	Computers and Biomedical Research (70s/80s), Computers in Biology and Medicine (70s/80s), Bioinformatics (2000s), Journal of Computational Biology (2000s)
B3-CAD	IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (70s/80s), Computers and Industrial Engineering (70s/80s), Advances in Engineering Software (2000s), Advanced Engineering Informatics (2000s)
B4-MicroElec	Microelectronics International (70s/80s), Microelectronic Engineering (70s/80s), IEEE Transactions on VLSI Systems (2000s), International Journal of Embedded Systems (2000s)
C1-Ling	Language (both), Journal of Linguistics (both), Linguistic Inquiry (both), Functions of Language (2000s)
C2-Bio	Nucleic Acid Research (both), Gene (both)
C3-MechEng	Combustion Science and Technology (70s/80s), International Journal of Production Research (70s/80s), Chemical Engineering Science (2000s), International Journal of Heat and Mass Transfer (2000s), Chemical Engineering and Processing (2000s)
C4-ElecEng	International Journal of Electronics (70s/80s), IEEE Transactions on Circuits Theory (70s/80s), Automatica (2000s), Mechatronics (2000s), Control Engineering Practice (2000s)

Table 1: Journals in *SciTex*

SciTex covers two time periods: the 1970s/early 1980s (*SaSciTex*) and the early 2000s (*DaSciTex*).[2] The sources for the corpus are full journal articles; for each discipline at least two different journals were selected and used in both time slices. If journals did not reach back to the 1970s/80s other journal sources were used that matched the academic discipline. Table 1 shows the journals included in *SciTex*. The sources were collected in the form of pdf files and con-

[2] Compare the *Brown* corpus family with the *Brown* (AmE) and *LOB* (BrE) corpora from the early 1960s and the *Frown* (AmE) and *FLOB* corpora (BrE) from the early 1990s for a similar design, cf. Kucera/Francis (1967), Hundt/Sand/Siemund (1999), Hundt/Sand/Skandera (1999).

verted to plain text format. Altogether, *SciTex* contains around 35 million tokens (i.e. approx. 17.5 million per time slice).

A smaller cross-sectional subcorpus of around two million tokens (i.e. one million per time slice) was also created and cleaned of erroneous data produced by the OCR conversion. Also in this subcorpus, formulas (mostly from computer science) and examples (as in linguistics) were tagged to exclude them from linguistic searches. This extensive procedure was important in order to obtain high quality text data at least for a portion of *SciTex*, which can then be employed for detailed analyses that may require further (manual) annotation.

Furthermore, a dedicated processing pipeline (cf. Kermes 2011) was implemented for (1) conversion of the corpus from text files to xml files, while maintaining information about document structure (e.g. paragraphs, sections etc.), (2) tokenization, lemmatization and part-of-speech tagging of the corpus files using the *TreeTagger* (Schmid 1994), (3) transformation of the corpus files into a verticalized text format for segmentation, and (4) encoding of the corpus for query by the *Corpus Query Processor* (CQP; Evert 2005).

3. Methodology: complementary comparisons

The chosen corpus design allows two kinds of perspectives that are necessary for our purposes: the temporal and the disciplinary. From the temporal perspective, we can carry out both *synchronic* (within a time slice) and *diachronic* (across time slices) comparisons. From the disciplinary perspective, we can compare the different disciplines in terms of *register*. To address our overarching research question, concerning register formation in the scientific domain, we obviously need to combine the two perspectives. In doing so, we may want to consider just one triple of A-B-C corpora, zooming in on one particular interdisciplinary field (e.g. computational linguistics), compared to its discipline of origin (e.g. linguistics), and to computer science, in order to detect a trend in that particular interdisciplinary field; or we may look at the interdisciplinary fields as a whole (all B corpora) in the search for a general trend. Furthermore, we may want to compare *SciTex* as a whole to a corpus comprising a number of registers, such as the *BNC* or the *Brown* family (see, for example, Teich/Fankhauser 2010). This may be of interest for comparing diachronic trends in the language as a whole to the development of scientific language (as, for example, in Mair 2006).

In order to detect diachronic trends, we need to determine features that are potentially relevant for the formation of new registers, and that ultimately bring about significant feature redistributions. The theoretical framework of Systemic Functional Linguistics provides a map of lexico-grammatical domains to look into for such features. Lexico-grammatical features typically associated with the contextual variables of field, tenor and mode are:

- field: experiential lexis, collocation/colligation, predicate-argument structure;
- tenor: mood, modality, expressions of stance;
- mode: theme-rheme, given-new.

On the basis of the annotated corpus as described in Section 2, we can then proceed to extract instances of these features. The extraction tool we employ is *CQP*, which allows us to detect feature instances by means of regular expressions, offering several functionalities for the purposes of extraction (e.g. context expansion) and sorting (e.g. counting, grouping of results). This flexibility is very useful when working with linguistic features at various cut-off points of the grammar-lexis cline. The feature frequencies obtained are then evaluated in terms of their discriminatory effects across registers and time slices, using univariate methods (e.g. the chi-square test) on single features, as well as multivariate methods (e.g. principal component analysis, correspondence analysis etc.) on sets of features.

In the following section, we show two examples of analysis using two features associated with field and tenor respectively, and employing univariate evaluation techniques.

4. Sample analyses

4.1 Discourse field: Lexis (most frequent words/keywords)

In the development of a scientific discipline, the creation of a distinctive vocabulary, especially terminology, is a key issue. To gain a first impression, we extract the most frequent nouns from the subcorpora of *SciTex* for both time slices. The most frequent nouns provide a first indication of the topics in a discipline. Table 2 illustrates two triples (A-B1-C1, A-B2-C2) with the five most frequent nouns.

discipline	70s/80s	early 2000s
A-CompSci	*set*	*algorithm*
	time	*time*
	function	*problem*
	proof	*graph*
	algorithm	*set*
B1-CompLing	*word*	*word*
	sentence	*translation*
	rule	*sentence*
	structure	*system*
	system	*example*
C1-Ling	*rule*	*language*
	sentence	*verb*
	form	*case*
	case	*example*
	verb	*word*
B2-BioInf	*system*	*gene*
	computer	*protein*
	time	*method*
	program	*sequence*
	value	*model*
C2-Bio	*DNA*	*gene*
	fragment	*sequence*
	site	*protein*
	gene	*cell*
	plasmid	*DNA*

Table 2: The five most frequent nouns for two triples in both time slices

It can be seen from the table that, diachronically, the most frequent nouns have changed to a different extent for each discipline. For example, in the triple A-B1-C1, all three disciplines have changed their five most frequent nouns to some extent. However, when we look at the interdisciplinary field in this triple, computational linguistics (B1), there is apparently no diachronic change regarding its relation to linguistics (C1) and computer science (A): both in the 1970s/80s and in the early 2000s it leans more towards linguistics than to computer science. Looking at the triple A-B2-C2, a different development is indicated. In the 1970s/80s the interdisciplinary field of bioinformatics (B2) leans more towards computer science (A) in the nouns used most frequently (e.g. *computer*, *time*, *program*), while in the early 2000s, there is a larger overlap with biology (C2) (e.g. *gene*, *sequence*, *protein*).

To further explore these tendencies, we calculate the keyness of the most frequent nouns for each subcorpus, again comparing triples of subcopora. Keyness is calculated by means of log likelihood statistics. The higher the log likelihood value, the more significant is the difference between corpora. A log likelihood of 3.8 or higher indicates a significant difference between two corpora (p-value < 0.05). Positive and negative values indicate which corpus makes more or less use of a given word.

discipline		**nouns**	**keyness in comparison to**	
			C1-Ling	A-CompSci
B1-CompLing	70s/80s	***word***	+ 475.50	+ 1767.27
		sentence	− 3.37	+ 2834.26
		rule	− 2328.87	+ 850.19
		structure	+ 0.90	+ 993.88
		system	+ 639.33	+ 242.12
	early 2000s	***word***	+ 1811.54	+ 8416.12
		translation	+ 5602.33	+ 7785.86
		sentence	+ 1265.74	+ 7921.52
		system	+ 2147.56	+ 2576.71
		example	+ 83.20	+ 2230.67
			C2-Bio	A-CompSci
B2-BioInf	70s/80s	*system*	+ 1711.89	+ 1581.33
		computer	+ 3176.64	+ 3703.04
		time	+ 1138.78	+ 51.52
		program	+ 2931.57	+ 585.25
		value	+ 1106.19	+ 960.09
	early 2000s	*gene*	− 48.20	+ 15216.64
		protein	− 192.48	+ 8429.29
		method	+ 3032.11	+ 5185.33
		sequence	− 305.71	+ 2303.13
		value	+ 2486.16	+ 1623.78

Table 3: Log likelihood for B1-CompLing and B2-BioInf

Table 3 shows the log likelihood values for the comparison of the two triples A-B1-C1 and A-B2-C2 for both time slices.[3] From these values we can observe that in the 1970s/80s computational linguistics (B1) was more similar to linguistics (C1) than to computer science (A): except for the word *rule*, all log likelihood values are smaller for B1 vs. C1 than for B1 vs. A. In the early 2000s the differences from both C1 and A become greater. Regarding bioinformatics

[3] Nouns that are common to two time slices are highlighted by bold face; negative values indicate a less frequent use relative to an interdisciplinary field.

(B2), the differences from computer science (A) also increase over time (larger log likelihood values for B2 vs. A), while the difference from biology (C2) decreases (lower log likelihood values for B2 vs. C2 for all words in B2 except *value*). Of course, these are not meant as general statements about the development of topics in these disciplines as a whole, but about tendencies in our corpus.

4.2 Discourse tenor: Modal verbs

One of the linguistic features relevant for discourse tenor is modality. Here, we discuss an analysis of modal verbs.[4] The overall trend we detect from the quantitative results is that in the early 2000s, the number of modal verbs used is rather stable across disciplines, with 500-1000 occurrences per million words in each discipline. This is in contrast to the 1970s/80s, which exhibit a relatively high variability across disciplines, some using quite a lot of modal verbs (e.g. in linguistics: around 3000 modal verbs per million words) and others rather few (e.g. in computer-aided design: under 500 modal verbs per million words). Overall, we obtain similar results to Mair (2006) who reports a decrease of the modals *shall*, *ought to*, *need to* as well as *must* and *may* in the *Brown* corpus family (see Mair 2006: 327). However, in contrast to the relative stability of the use of *can* as reported by Mair (2006) in English generally, we observe a relative increase of *can* (approx. 10 to 20 %) in our corpus.

To see how the interdisciplinary fields have developed in the given time period, we investigated all A-B-C triples across the two time slices. For this purpose, we applied the following meaning groups as used by Biber et al. (1999: 485):

- permission/possibility/ability: *can*, *cannot*, *could*, *may*, *might*;
- obligation/necessity: *must*, *have to*, *need to*, *ought to*, *should*;
- volition/prediction: *will*, *would*, *shall*.

[4] For other features in the tenor parameter, e.g. evaluative patterns and modal adverbs, see Teich/Degaetano (2011) and Degaetano (2011).

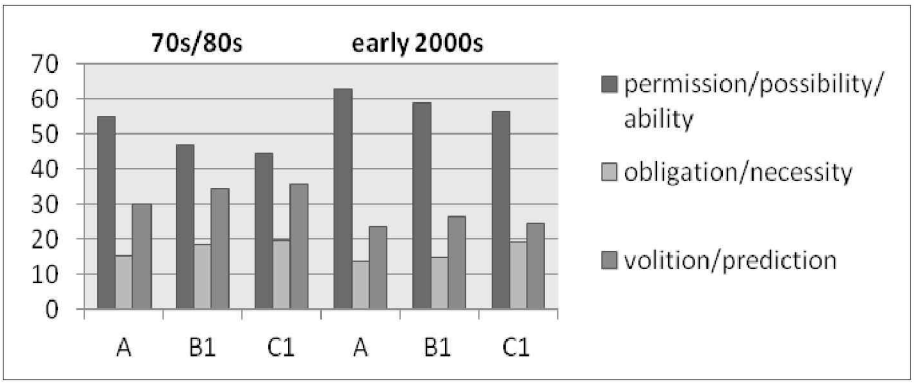


Figure 1: Distribution of modal meanings across the A-B1-C1 triples

Figure 1 shows the A-B1-C1 triple in both time slices in percentages (100% = all modal verbs used). In the 1970s/80s, computational linguistics (B1) seems to be more similar to linguistics (C1), but different from computer science (A), comparing the percentages of the modal meanings (less use of obligation/necessity and volition/prediction in A compared to B1 and C1). In the early 2000s, the picture changes: computational linguistics (B1) seems more similar to computer science (A), while differing from linguistics (C1) in the use of the obligation/necessity group. The diachronic tendency of computational linguistics (B1) being first similar to linguistics (C1) and later on moving towards computer science (A) is confirmed by calculating the p-values using the chi-square test (see Table 4): B1 shows higher significant differences from A than C1 in the 1970s/80s, but lower significant differences from A than C1 in the early 2000s.

Time slice	discipline			p-value
1970s/80s	B1-CompLing	–	A-CompSci	< 2.2e-16
	B1-CompLing	–	C1-Ling	0.0004012
early 2000s	B1-CompLing	–	A-CompSci	8.37e-14
	B1-CompLing	–	C1-Ling	< 2.2e-16

Table 4: p-values for diachronic comparison of the A-B1-C1 triples

Diachronic changes have also been observed for the other interdisciplinary fields: bioinformatics (B2) moved from being similar to computer science (A) to differing from both biology (C2) and computer science (A); computer-aided design (B3) differs from both mechanical engineering (C3) and computer science (A), thus creating its own variation; microelectronics (B4), instead,

remains similar to its discipline of origin (C4: electrical engineering) and differs from computer science (A) in both time slices.

5. Summary and conclusions

We have introduced here a project analyzing the diachronic development of highly specialized scientific registers which have emerged as a result of register contact. In our investigation, we focus on the situation of interdisciplinary contact between selected disciplines and computer science. The prerequisite for our research is an appropriate corpus. We introduced the *SciTex* corpus, compiled from research articles from nine scientific disciplines (Section 2). The *SciTex* corpus enables us to investigate register contact from both the synchronic and the diachronic angles. The framework of Systemic Functional Linguistics and register/genre theory provide the linguistic and contextual categories relevant for the analysis of register variation (Section 3). We then showed two examples of analysis of the corpus, one using a field-related feature, the other using a tenor-related feature (Section 4). The analyses focused on diachronic trends in the four interdisciplinary fields contained in *SciTex*.

So far, we have obtained indications of both of the principal motifs of scientific evolution, diversification and standardization. However, the picture is not uniform across the four interdisciplinary fields investigated: some change from being more similar to their discipline of origin to being more similar to computer science (e.g. modal meanings in computational linguistics); others change in the other direction, away from computer science and towards their discipline of origin (e.g. lexis/keywords in bioinformatics), and some seem to create their own patterns of variation (e.g. modal meanings in computer-aided design). Moreover, the tendencies may differ for different contextual parameters, such as differences in field, but similarities in tenor. Obviously, we need to study more features in order to cover the full spectrum of potential variation. Also, with a larger feature set we will be able to use other, more powerful methods of feature evaluation, such as automatic clustering or classification, which will allow us a more comprehensive and differentiated interpretation. In the analysis of vocabulary, we will explore more advanced methods such as topic models (see Blei forthc.), which promise to get a tighter grip on diachronic topic shifts.

References

Banks, David (2008): The development of scientific writing. Linguistic features and historical context. London: Equinox.

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

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

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

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

Blei, David (forthc.): Introduction to probabilistic topic models. Communications of the ACM.

Degaetano, Stefania (2011): Evaluative options and their choice: modal adjuncts vs. evaluative patterns in academic writing. Paper presented at the International Evaluation Conference (IntEval), Madrid: 6-8 Oct. 2011.

Evert, Stefan (2005): The CQP query language tutorial. IMS, Universität Stuttgart.

Halliday, Michael A.K. (1985): Spoken and written language. Melbourne: Deakin University Press.

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. (2004): Introduction to functional grammar. 3rd edition (with Christian M.I.M. Matthiessen). London: Edward Arnold.

Halliday, Michael A.K./Hasan, Ruqaiya (1989): Language, context and text: a social semiotic perspective. Second edition. Oxford: Oxford University Press.

Halliday, Michael A.K./Martin, James R. (1993): Writing science: literary and discursive power. London/Washington, DC: The Falmer Press.

Hundt, Marianne/Sand, Andrea/Siemund, Rainer (1999): Manual of Iinformation to aAccompany the Freiburg – LOB Corpus of British English ('FLOB'). Freiburg: Englisches Seminar, Albert-Ludwigs-Universität Freiburg. http://khnt.aksis. uib.no/icame/manuals/flob/INDEX.HTM.

Hyland, Ken (2007): Disciplinary discourses: social interactions in academic writing. Ann Arbor: The University of Michigan Press.

Mair, Christian (2006): Twentieth-century English: history, variation and standardization. Cambridge: Cambridge University Press.

Martin, James R. (1992): English text: system and structure. Amsterdam: John Benjamins.

Kermes, Hannah (2011): Automatic corpus creation manual. Institute of Applied Linguistics, Translation and Interpreting. Saarbrücken: Universität des Saarlandes.

O'Halloran, Kay L. (2005): Mathematical discourse: language, symbolism and visual images. London: Continuum.

Quirk, Randolph/Greenbaum, Sidney/Leech, Geoffrey/Svartvik, Jan (1985): A comprehensive grammar of the English language. London: Longman.

Schmid, Helmut (1994): Probabilistic part-of-speech tagging using decision trees. In: Proceedings of international conference on new methods in language processing. Manchester, UK, 44-49.

Teich, Elke/Degaetano, Stefania (2011): The lexico-grammar of stance: an exploratory analysis of scientific texts. In: Dipper, Stefanie/Zinsmeister, Heike (eds.): Beyond semantics: corpus-based investigations of pragmatic and discourse phenomena. In: Bochumer Linguistische Arbeitsberichte 3: 57-66.

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.

Teich, Elke/Holtz, Mõnica (2009): Scientific registers in contact: a methodology and some findings. In: International Journal of Corpus Linguistics 14(4): 524-548.

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.

ANDRÉS ENRIQUE-ARIAS

On the usefulness of using parallel texts in diachronic investigations

Insights from a parallel corpus of Spanish medieval Bible translations[1]

Abstract

This paper addresses a number of methodological, theoretical and practical problems related to corpus-based research in language variation. I show through a number of case studies using data from *Biblia medieval* (a parallel corpus of Old Spanish Bible translations) how this kind of research can profit from parallel texts. To begin with, the perspective afforded by parallel corpus methodology is more open as it is possible to analyze all the forms used to express contents in the source language. Likewise, parallel texts offer direct comparability of concrete examples across different historical periods, as translation equivalents are likely to be inserted in the same or very similar syntactic, semantic and pragmatic contexts of occurrence. Finally, in a parallel corpus it is possible to analyze stylistic variation in a more controlled manner by examining how the same translator selects different linguistic options depending on the genre of each text.

1. Introduction

In this paper I shall present some of the advantages of studying language variation and change using parallel texts as opposed to other data sources. In Section 2 I start with some critical remarks referring to a number of limitations of the two corpora that have been widely used for research on historical Spanish linguistics according to the criteria of comparability and perspective. In Sections 3 and 4, I present glimpses of my own experience with a parallel corpus of Old Spanish translations of the Bible, to show how parallel texts can help overcome some of the limitations of conventional corpora. In Section 5, I draw the necessary methodological conclusions.

[1] This research has benefited from a grant from the Spanish Ministry of Science and Innovation (reference FFI2010-18214) co-financed with FEDER funding as well as from a Collaborative Research Fellowship (2011-2013) from the American Council of Learned Societies.

2. Parallel texts versus conventional corpora

Since the 1990s, tremendous methodological advances have been made in historical linguistics, thanks to the development of computing tools which enable linguists to process and to analyze massive quantities of linguistic data automatically, and the application of variationist methods in the study of language variation and change from a diachronic perspective (Joseph 2008: 182). In the case of Spanish, virtually all the historical investigations of the last decade have made use of either or both of the two large diachronic databases that are freely available online: the *Corpus Diacrónico del Español* (CORDE) or the *Corpus del Español* (CE). CORDE was created in the late 1990s, and was the first large corpus of historical Spanish. It is composed of approximately 250 million words of text, with good representation across the different historical periods. The CE was completed a little later, in 2002. It contains about 100 million words from Old Spanish to the late 1990s. Unlike CORDE, the CE is lemmatized and annotated for part of speech. Both corpora are composed of texts from a variety of genres – poetry, historical writings, prose literature, didactic materials – and have a good representation of the medieval periods (23 and 18 million words in CORDE and CE, respectively). Next to these two corpora, there is a relatively newer one, the *Biblia medieval* corpus (BM), which became available in 2009. Containing over 5 million words, BM is a freely accessible online tool that enables linguists to consult and compare side-by-side the existing medieval Spanish versions of the Bible next to their Hebrew or Latin sources.

CORDE and CE are good examples of what can be considered *conventional corpora*, which are the kinds of corpora that are commonly used in historical investigations. They consist of a computerized database of historical texts from different periods and a search tool to retrieve information from the corpus. In order to access the data, users need to enter a query (i.e. a word or a phrase) and the search application displays all the instances of the search string in the corpus including contextual occurrence and basic background information on the source text, such as title, author and date of composition. In contrast, BM is a *parallel corpus*, that is, a collection of original texts and their translations. In such corpora, the texts are aligned so that it is possible to identify the pairs or sets of sentences, phrases and words in the original text and their corre-

spondences in the other languages.[2] More specifically, BM is composed of the Hebrew bible and the Latin Vulgate, which are the original texts, and their translations into medieval Spanish. Thus, when the user enters a query for any of the parallel versions in the corpus, whether it is the original texts or any of the thirteen Old Spanish versions that it contains, the search application will display all the occurrences of the search string in the relevant version next to the translation equivalents in all the other versions.[3]

The advantages afforded by large conventional corpora like CORDE and CE are obvious, as they allow searching through millions of words of data in a fraction of a second. In the case of the CE, the lemmatization and grammatical annotation allows us to further observe more subtle relationships between different elements (parts of speech, morphological markers, related vocabulary, collocations) and arrive at more nuanced analyses. However, conventional corpora also have certain limitations in areas in which, as I will try to demonstrate in the following pages, a parallel corpus like BM can afford clear advantages.

One such area is *perspective*, understood as how the information is accessed in a corpus. In conventional corpora users enter queries for the forms that are relevant for the research question they are investigating. This means it is necessary to know beforehand, via historical grammars, dictionaries or previous studies, what are the possible expression units for the structure that is being investigated. Two disadvantages emerge from this. One is that, no matter how well we do our research of reference materials, there is always a risk that some relevant form will be overlooked because it has never been studied. Another problem is that, as research with conventional corpora requires searching for explicit markers, it is impossible to identify all the occurrences of linguistic phenomena that may be expressed in a great variety of ways, or even zero-marked. The procedure for accessing the data in conventional corpora may be appropriate when we want to search for closed class elements, or when we know the exhaustive list of possible forms related to the phenomenon to be studied, but is rather inadequate when we are investigating structures that can be expressed with open class elements or for which there is no way of knowing beforehand all the possible expression units.

[2] Parallel corpora have become a key focus of corpus linguistics in the last decade due to their importance as resources for translation and contrastive studies. For a detailed state of the art of the methodological possibilities of parallel and comparable corpora see McEnery/Xiao (2007).

[3] For a detailed description of the corpus refer to the project webpage at www.bibliamedieval.es.

Conventional historical corpora also have limitations in the area of *comparability*. As it is well known, linguistic changes occur after a period of variation in which the original structure and the innovative one coexist. This period is very interesting for the researcher: the examination of the factors that favour the innovative variant will give us the clues to understand the causes, origin, chronology and diffusion channels of the change. Accordingly, quantitative techniques – mainly coocurrence analyses of those linguistic variants that compete in the same contexts of occurrence – have become essential in corpus-based studies of language change. In carrying such analyses historical linguists should be eager to make sure that the empirical basis on which they build their theories is such that it guarantees the highest possible degree of comparability, that is, we have to make sure that the data we draw from texts belonging to different periods are indeed in a relation of equivalence among each other, and thus allow for being compared. But in working with a conventional corpus, this is not always easy.

To begin with there are sample-related problems concerning which texts to compare: is it methodologically sound to compare, let us say, 13th century linguistic data extracted from medieval chronicles with that from 16th century novels? Even though in both cases we are dealing with narrative discourse, these are works with very different textual conventions, and in which the distribution of narration, description and dialogue may be significantly different. That is, corpus developers need to make sure that, for each period represented in the corpus, they are characterizing states of language rather than mere text types.

Likewise, it is rather difficult to identify and to define the contexts of occurrence of linguistic variants, since these are normally conditioned by a complex set of syntactic, semantic and discursive factors. In order to minimize this problem, it would be necessary to locate and to examine a large number of occurrences of the same linguistic structure in versions that were produced at different time periods. Ideally, these occurrences should proceed from texts that have been influenced by the same textual conventions. Conventional corpora, however, are not well-suited for locating such occurrences. An added difficulty is identifying and controlling contextual dimensions that potentially condition variation in older texts (text type and genre, poeticality, orality, dialect, writer demographics, etc.). Historical texts often come down to us devoid of information on author and intended readership, their social and geographic dialects or precise date of composition. In sum, as we cannot control the fac-

tors that condition variation in each individual text, we face the risk of a partial or misleading analysis.

In what follows I would like to look at specific methodological issues involving the use of corpus data to show how using parallel texts can help overcome the limitations of conventional corpora regarding perspective and comparability.

3. Perspective

Linguistic structure is accessed in different ways by historical researchers depending on the material used. As I have already pointed out, when using a conventional corpus, users enter queries for those forms that are supposedly relevant for the research question and obtain textually embedded instances of the form from which its meaning and function can be observed. In this fashion the analysis proceeds from form to function, with the disadvantages that I have already mentioned: poor or insufficient knowledge of the relevant forms will result in an incomplete analysis. In contrast, parallel texts lead the investigator from particular textually embedded contents to form, which, as I will explain, affords a number of advantages.

The first one is the heuristic function of parallel texts, which has no equivalent in other data sources. For instance, let us suppose that we want to study the historical evolution of the linguistic elements used to introduce an exception to a previous statement (i.e. the expressions equivalent to English *except*, *apart from*, *excluding* and the like). If we want to use conventional corpora, first we need to consult reference materials and compile a list of elements that can express this function (i.e. *excepto*, *salvo*, *menos*, *fueras*, etc.), then conduct searches for these words, and finally use the results to examine specific examples in their functional context. Because of the form-to-function perspective in which we are operating, there is no way to know whether the corpus contains other elements that can be used with the same function and in the same contexts. In contrast, in using a parallel corpus like BM we do not need to have an exhaustive list of forms beforehand as the searches in the corpus and the comparisons with the parallel versions will guide us in finding the possible expression units for the structure that is being investigated. In BM we can extract the passages that contain these elements by searching for any equivalent of 'except' in the Latin or Hebrew originals, or in any of the Spanish texts, and then observe the forms that are used in the same context and with the same functions in the parallel versions. In turn, we can search for these forms, wich will yield more

forms and more contexts than can be used for further queries. This perspective, from particular textually embedded contents to form, facilitates the observation of elements that otherwise would have been overlooked.

It is clear, however, that a parallel corpus like BM should never be the only source of information in a diachronic study. Additional sources, such as dictionaries, grammars, studies, and above all, conventional corpora such as CORDE or CE are indispensable sources to make sure that the forms that we discover thanks to the parallel corpus are not just words used only in Bible translation. For instance, in her study of the expression of 'except' in the history of Spanish, Sánchez López (in press) consulted the BM corpus and was able to find, next to numerous instances of the expected *excepto*, *salvo*, etc. 53 occurrences of *salvante*, a form that had not been recorded in previous reference materials. A search in CORDE shows that *salvante* is not a form confined to biblical language (it appears 33 times in 24 texts of different genres, dated between 1380 and 1758).

Another advantage of the function-to-form perspective of the parallel text methodology is that it is possible to search for any way of expressing a linguistic function. This feature of a parallel corpus is helpful in overcoming one of the limitations of conventional corpora: finding examples of linguistic phenomena that may be expressed in a variety of ways or even zero-marked. As a concrete example, consider the expressions that are used in all languages to draw attention to something (English *behold*, *look*, Spanish *he aquí*, *mira*, French *voilà*, Italian *ecco*, etc.). If we want to study how this function is expressed in Old Spanish using a conventional corpus we are restricted to doing searches for those explicit elements that typically could be used in these contexts, such as the discourse markers *he* (and its variants *ahé*, *afé*) or *evás*. In contrast, parallel corpus methodology is much more open as we can search for *any* element used to express these functions: we simply look up all the occurrences of Hebrew *hinneh* or Latin *ecce* in the original and observe how they are translated in the Spanish versions. As can be seen in the translations of Deuteronomy 31:16 below, the corpus allows the analyst to appreciate the wide range of expressions that medieval translators used to convey the meaning and function of this marker:[4]

[4] All passages quoted are from the BM corpus. For a review of the most important issues in regards to dating, description and content of the Old Spanish biblical manuscripts contained in the corpus and for information on the abbreviations used to cite them refer to the website of the project at www.bibliamedieval.es.

(1) Deuteronomy 31:16 (*And the Lord saide unto Moses:* ***behold****, thou shalt sleepe with thy fathers*)

[Hebrew] *wa-yomer YHWH el-moshe* ***hinne****kha shokhev 'im-avotekha*

[Vulgate] *Dixitque Dominus ad Moysen:* ***ecce*** *tu dormies cum patribus tuis*

[Fazienda] *E dixo a Moisén:* ***e*** *tú izrás con tos parientes*

[E8] *Et dixo Dios a Moisén:* ***evás*** *que tú dormirás con tus padres*

[E3] *E dixo Dios a Muisén:* ***cata*** *que tú yacerás con tus parientes*

[E4] *E dixo el Señor a Moisén:* ***hete*** *que dormirás con tus padres*

[E7] *E dixo el Señor a Muisén:* ***ya tú*** *vas a yazer con tus parientes*

[E19] *E dixo Dios a Muisén:* ***aquí*** *tú yacerás con tus parientes*

[Alba] *Dixo el Señor a Moisén:* ***sepas*** *que así como tú yoguieres con tus parientes*

Certainly, this is a real improvement on the use of conventional corpora; here the researcher can observe without limitations what linguistic structures are used to convey the functions expressed by *hinneh* in this context: besides the somewhat expected *he* and *evás* we find a verb of perception (*cata que* 'see that' in E3), deictics of time and space (*ya* 'now', *aquí* 'here' in E7 and E19 respectively), a verb of knowledge (*sepas que* 'know that' in *Alba*) and even zero-marking as in *Fazienda*.

Moreover, if we take full advantage of the parallel nature of the corpus, we can perform more fine-grained analyses. For instance, in a detailed study on this topic (Andrés Enrique-Arias and Laura Camargo (in press)) we have identified two main functions of the equivalents of *hinneh*: (i) to introduce events or objects in the narration, and (ii) to introduce in direct discourse information that is newsworthy for the addressee. Then, using the BM corpus, we were able to compare what expressions are used more often for each one of these two functions: in its function of introducing new information in direct discourse, Old Spanish *ahé* alternates with expressions like *cata que* 'see that' and *sepas que* 'know that', among others, while in its narrative function, it alternates with perception verbs like *ver* 'see' and *fallar* 'find'. In the latter case, zero marking is more likely to occur.

4. Comparability

Direct comparability of concrete examples across different historical periods is a strong point of the parallel text method. While defining contexts of occurrence of linguistic variables in conventional corpora is an arduous task, in parallel texts we have direct access to the evolution of linguistic structures, as translation equivalents are likely to be inserted in the same – or very similar – syntactic, semantic and pragmatic contexts of occurrence. Direct comparability is particularly useful when we are dealing with phenomena that exhibit covariation with a complex set of structural (internal) and contextual (external) factors. Consider for instance the variation in the use of the definite article preceding the possessive marker (*la mi casa* 'the my house') as opposed to possessive alone (*mi casa* 'my house') in Old Spanish, a structure that was optional and whose appearance has been attributed to a considerable number of structural factors, as summarized in (2) (data from Wanner 2005, and my own observations):

(2) Environments that favour article + possessive in Old Spanish[5]

(a) Features of the possessor:
- 1st and 2nd person > third person
- singular > plural

(b) Features of the possessed entity
- inanimate > (animate > terms of kinship)
- body parts > other nouns

(c) Syntactic function of the NP
- subject > object
- bare > with preposition (contra Wanner 2005)

The study of the expression of possession in Old Spanish is further complicated by the fact that the use of the article before possessives is also favoured by stylistic factors (Lapesa 1971 [2000]: 422). Because it is a structure that emphasizes possession, it is used with stylistic functions such as expressivity, solemnity, emphasis or reverence. As a result, when comparing the percentage of article plus possessive in historical texts, it is rather complicated to control for

[5] The vector sign (>) represents that the category to the left is expressed with article plus possessive with greater frequency than the one to the right.

all the possible factors that may be conditioning the variation observed in each text. In contrast, comparisons of this kind are rather straightforward in a parallel corpus like BM: as parallel texts put the discourse contextual factors largely in control, the behaviour of the elements used to express possession can be observed and compared in a focused manner. See for instance the comparison of the different versions of Jeremiah 51:56 in the BM corpus in (3) below. Here we get a good number of occurrences of possessive structures (the Spanish equivalents of 'her mighty men' and 'their bows') embedded in identical syntactic environments:

(3) Jeremiah 51:56 (*and **her mighty men** are taken, every one of **their bows** is broken*)

[Hebrew]	*we-nilkedu **gibboreha** ḥittetah **qashshetotam***
[Vulgate]	*et adprehensi sunt **fortes eius** et emarcuit **arcus eorum***
[E6]	*e son presos **los sos arreziados** e enflaqueció **el arco dellos***
[GE]	*e compresos son **los sos fuertes** e secóse **el so arco***
[E3]	*fueron presos **sus barraganes** quebróse **su ballesta***
[E5]	*serán presos **los sus barraganes** e serán quebrantadas **las sus ballestas***
[BNM]	*serán tomados **sus potentes** e quebrantarse an **sus arcos***
[RAH]	*serán presos **los sus barraganes los sus arcos** serán quebrantados*
[Alba]	*e **los sus barraganes** presos serán e **los sus arcos** serán quebrantados*

In comparing examples like the ones in (3) we can abstract away from the influence of contextual properties and focus instead on the diachronic evolution of structural phenomena.

Another interesting feature of the Bible is that it encompasses texts of varied textual typology: narrative, legislative, lyrical poetry, wisdom literature, epistles, and dialogues. As a result, the BM corpus is particularly suited to explore register variation, as it is possible to examine how the same translator selects language options that are appropriate for each one of the genres represented in the Bible. For instance, the comparison of the distribution of article plus possessive in different textual genres in three medieval Bible translations demonstrates that, for medieval translators, this structure had a definite stylistic value:

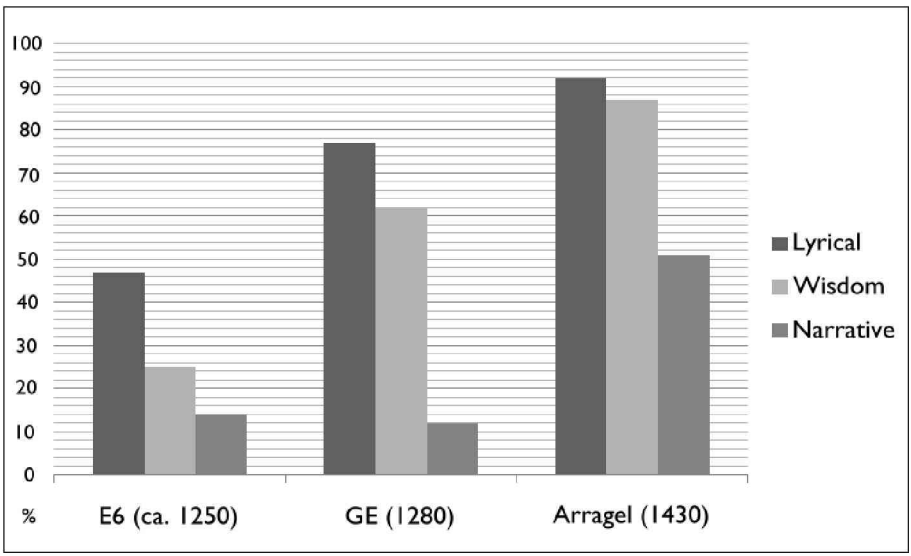


Figure 1: percentage of article + possessive (vs. possessive alone) in three bible translations

As can be seen in Figure 1, the different versions studied fluctuate as to the general frequency of the phenomenon, however they clearly follow a common trend: the construction is most frequent in lyrical passages and least frequent in narrative ones. The data lends support to the view defended by Lapesa (1971 [2000]: 422) that article plus possessive has an expressive value and thus is more used in lyrical passages than in narrative ones.

5. Conclusion

Biblia Medieval opens new perspectives in the historical study of linguistic variation and change in Old Spanish. With the help of the biblical corpus it is possible to apply quantitative and qualitative analyses to the study of variation and change with some clear advantages over conventional corpora. To begin with, the perspective afforded by parallel corpus methodology is more open as it is possible to analyze all the forms used to express contents in the source language. Likewise, parallel texts offer direct comparability of concrete examples across different historical periods, as translation equivalents are likely to be inserted in the same or very similar, syntactic, semantic and pragmatic contexts of occurrence. Finally, in a parallel corpus it is possible to analyze stylistic variation in a more controlled manner by examining how the same translator selects different linguistic options depending on the genre of each text.

While a parallel corpus does not solve all the problems inherent to working with historical texts, it does enable the analyst to observe the historical evolution of structural phenomena while controlling the contextual dimensions that condition variation in older texts.

References

Primary sources

Biblia medieval. Enrique-Arias, Andrés (2009-). http://www.bibliamedieval.es

CORDE. Real Academia Española: Corpus diacrónico del español. http://corpus.rae.es/cordenet.html

Corpus del español. Davies, Mark (2002-). http://www.corpusdelespanol.org

Secondary literature

Enrique-Arias, Andrés/Camargo, Laura (in press): Problemas en torno a la caracterización de un marcador del discurso en español medieval: el caso de he. In: Borreguero Zuloaga, Margarita/Gómez-Jordana Ferary, Sonia (eds.): Marqueurs du discours dans les langues romanes: une approche contrastive. Limoges: Lambert Lucas.

Joseph, Brian (2008): Historical linguistics in 2008. The state of the art. In: van Sterkenburg, Piet (ed.): Unity and Diversity of Languages. Amsterdam/Philadelphia: John Benjamins, 175-188.

Lapesa, Rafael (2000 [1971]): Sobre el artículo ante posesivo en castellano antiguo. In: Cano, Rafael/Echenique, María Teresa (eds.): Estudios de morfosintaxis histórica del español. Madrid: Gredos, 413-435.

McEnery, Tony/Xiao, Richard (2007): Parallel and comparable corpora: The state of play. In: Kawaguchi, Yuji/Takagaki, Toshihiro/Tomimori, Nobuo/Tsuruga, Yoichiro (eds.): Corpus-Based Perspectives in Linguistics. Amsterdam/Philadelphia: John Benjamins, 131-145.

Sánchez López, Cristina (in press): Preposiciones, conjunciones y adverbios derivados de participios. In: Company, Concepción (ed.): Sintaxis histórica de la lengua española. Tercera parte: Adverbios, preposiciones y conjunciones. Relaciones interoracionales. México: Fondo de Cultura Económica/Universidad Nacional Autónoma de México.

Wanner, Dieter (2005): The corpus as a key to diachronic explanation. In: Kabatek, Johannes/Pusch, Claus D./Raible, Wolfgang (eds.): Romance Corpus Linguistics II: Corpora and Diachronic Linguistics. Tübingen: Narr, 31-44.

TOMAŽ ERJAVEC / ALENKA JELOVŠEK

A corpus-based diachronic analysis of Slovene clitics

Abstract

This paper presents a manually annotated corpus of historical Slovene and a study, based on this corpus, of how clitics have changed in the Slovene language over time. The corpus contains 1,000 sampled pages, comprising about 300,000 tokens from over 80 works, spanning the period from the end of the 16th century to the end of the 19th. Each word is manually annotated with its modern day equivalent, lemma and part-of-speech tag. The paper discusses the composition, encoding and availability of the corpus, and then presents a study of word-tokenization mismatches between contemporary and historical Slovene, concentrating on the binding of clitics with their host, and on the variability of clitic orthography in the corpus.

1. Introduction

For empirically-based studies of historical languages, the basic resource needed is a diachronic corpus, which will typically contain proof-read text with links to facsimiles, and manually verified linguistic annotation. In addition to enabling purely linguistic studies, such corpora also facilitate the development of human language technology support for historical language, such as the induction of models for spelling change, lemmatization and tagging, which, in turn, serve to enable better OCR models, and better accessibility of cultural heritage texts in digital libraries. Annotated historical linguistic corpora have already been compiled for a number of languages, for example German (Scheible et al. 2011), and this paper presents a similar attempt for Slovene.

Diachronic corpora typically include hand-validated linguistic annotations of word tokens, consisting of their modern-day equivalent word-form, their modern-day lemma (often referred to as "super-lemma", as it abstracts away from the spelling variability of historical language), and their morphosyntactic tag. This approach has the advantage of making the corpus maximally useful for today's speakers of the language, for example by enabling querying by lemma and having all the word-forms returned, regardless of orthographic variation, but it does present problems where the historical and modern day word forms do not align, when several word-forms in historical language correspond to

one contemporary word-form, or vice-versa. As (word) tokens are the "atoms" of corpora, this mismatch brings with it technical problems in the processing and encoding of corpora, while, on the other hand, it offers the opportunity for a linguistic study of these mismatches.

In this paper we first present the *goo250k* corpus of historical Slovene, describing its sources and content in Section 2, and its annotation and encoding in Section 3. Section 4 then investigates the changes in the concept of the orthographic word in this corpus from a linguistic point of view, with emphasis on the writing of clitics. Finally, Section 5 gives some conclusions and directions for further work.

2. Corpus construction and content

The first stage in the construction of the corpus was the acquisition of high-quality transcriptions, followed by sampling, to arrive at a representative and balanced corpus, which could then be manually annotated. The basis for the reference corpus came from the following sources of proof-read historical texts with facsimiles:

- Successive selected pages from three religious books, from the end of the 16th, 17th and 18th centuries respectively. The scans of the books and proof-read transcriptions were provided by the Scientific Research Centre of the Slovenian Academy of Sciences and Arts. The first two of these books also represent the oldest material in the corpus.
- Complete books from the second half of the 18th and first half of the 19th century. The scans and proof-read transcriptions were provided by NUK, the National and University Library of Slovenia. The books were written in Slovene, and span religious books, plays, fiction and even a cookbook.
- Selected complete issues of one Slovenian newspaper, first published in 1843, and continuing to 1890. The facsimiles and transcriptions were also provided by NUK.
- The *AHLib* digital library (Prunč 2007), containing complete books, mostly from the second half of the 19th century. These are Slovene translations of German books, and span a wide variety of topics, from fiction to textbooks on various subjects.

This collection was the basis for the so called *goo250k* reference corpus. This corpus consists of sampled pages from the text collection, where the sampling procedure aimed at a good coverage of time periods and text types, while taking into account the constraints of the text collection. More weight was, however, given to more recent materials, as the main focus of the corpus is in providing human language technology support for historical language, and the language of the 19th century is still similar enough to the contemporary one for such methods to yield good results, as well as being the most useful, as there are considerably more texts available from the 19th century than from earlier times. Table 1 gives the size of the *goo250k* corpus according to the time periods, and overall, by the number of units (book or newspaper samples), the number of pages (the individual unit of sampling), and the approximate number of tokens. The set size of the corpus was 1,000 pages, which was estimated to be the right size for the manual annotation to be feasible, given the financial and time constraints of the project.

Period	Units	Pages	Tokens
1584	1	8	6,000
1695	1	27	10,000
1751-1800	8	155	27,000
1801-1850	12	206	74,000
1851-1875	36	380	126,000
1876-1900	23	224	51,000
Σ	81	1,000	296,000

Table 1: Corpus size by time period

3. Corpus annotation and representation

The corpus was first automatically annotated, using the *ToTrTaLe*[1] tool, which tokenizes the text, segments it into sentences, transcribes historical words to their contemporary form, tags them with morphosyntactic descriptions and assigns them contemporary lemmas. For tagging and lemmatisation, the tool uses models trained on contemporary Slovene, so the transcription step is not only useful in itself, by making the text more understandable for today's readers, but is also crucial for these two levels of annotation. The transcription is

[1] The tool is described in detail in Erjavec (2011) and is still being actively developed. Unfortunately it cannot be made freely available, as it uses third party components, in particular the *TnT* tagger (Brants 2000) and *Vaam* library (Reffle 2011), which are not open source.

operationalized by the *Vaam* (*Variant Approximate Matching*) finite-state library (Reffle 2011), which uses a lexicon of modern word-forms and a set of transcription patterns of typical spelling changes, which associate historical words to contemporary ones. By inspecting the unannotated corpus, we first developed a set of transcription patterns, and then with the help of the *LeXtractor* editor (Gotscharek et al. 2010) assigned contemporary word-forms to the most frequent (and, typically, unpredictable) words in the collection. With this static lexicon and the transcription patterns we then automatically annotated the *goo250k* corpus.

In the second step the automatically assigned annotations were manually checked and corrected. The annotation editor used was *Cobalt*, developed at the Institute for Dutch Lexicology (INL, Instituut voor Nederlandse Lexicologie), a Web based corpus browser/editor, in which it is possible to load pre-annotated corpora, correct the annotations as well as the transcriptions, and do this in a concordance-oriented view, so identical word-forms can be inspected together. A team of annotators were hired, most of them students involved in previous annotation projects, while the three oldest books were annotated by PhD students of historical Slovene. The *Cobalt* user manual was adapted for Slovene, and additional reference materials (Annotator's Cookbook, FAQ) were written, in tandem with training the annotators on small test corpora. At the manual annotation phase, students corrected mistakes in the transcriptions and annotations, as well as adding glosses to extinct words.

The corpus is encoded according to the *Text Encoding Initiative Guidelines, TEI P5* (TEI consortium (ed.) 2007), with each sampled page as one file, and encoded as a TEI element. The complete corpus, i.e. the XML <TEI> document, is composed of the <teiHeader>, giving extensive meta-data for the corpus, and links to the 1,000 data files.

In Figure 1 we give an example of the annotation for one sentence "*Na to se dolipoklekne.*", containing four words with the intervening whitespace characters and line-breaks, and the full stop punctuation symbol. Each word token is annotated with its normalized form (nform), its modernized form (mform), lemma (lemma) and coarse-grained morphosyntactic corpus tag (ctag). The normalized form is the word-token, decapitalized and with vowel diacritics removed as they are not used in contemporary Slovene.

The third word in the example, "*se*", has a one-to-one mapping with the modern equivalent, which is the default case. The first two words used to be written separately, but are now written together, and the last used to be one word,

but is now written as two. The tokenization is "diplomatic", i.e. it follows the historical texts, and the type attribute on the word gives relations to the modern word forms. The first two words are of multi-word type, and the linguistic annotation is repeated on both parts of the word, while the value of the attribute n identifies the component tokens of the multi-word expression. In other words, both (or, in general, all) parts of the multi-word have the same value of n, in this case mw_752.[2] The last word is "split", i.e. two contemporary words correspond to one historical word. In this case the linguistic annotation is given as a string of underscore-separated items, a somewhat ad hoc, but effective, solution.

```
<s>
  <w type="multiw" nform="na" mform="nato" lemma="nato"
            ctag="Rgp" n="mw_752">Na</w>
  <c> </c>
  <w type="multiw" nform="tu" mform="nato" lemma="nato"
           ctag="Rgp" n="mw_752">tu</w>
  <c> </c>
  <w nform="se" mform="se" lemma="se" ctag="P">se</w>
  <c> </c>
  <lb n="7"/>
  <w type="split" nform="dolipoklekne" mform="doli_poklekne"
           lemma="doli_poklekniti" ctag="Rgp_Vme">dolipoklekne</w>
  <pc ctag=".">.</pc>
</s>
```

Figure 1: Corpus encoding

The *goo250k* corpus, as well as the complete automatically annotated source text collection, is mounted under a Web-based concordancer with CQP (Christ 1994) as its back-end. CQP corpora must be tokenized, and each token can be given arbitrary positional attributes. A rich query language and flexible output methods enable complex analyses over the corpora. In converting the TEI representation into CQP format, we treat multi-words similarly to split words, i.e. as single tokens, so that the query [nform=".*_.*"] gives us all the multiple words (621 nform types/1203 tokens), and the query [mform=".*_.*"] all the

[2] Note that this encoding supports non-contiguous multi-word units, as is necessary with German, for example, or, in certain cases, for Slovene as well.

split words (502 /732). Building on such searches we can analyse changes in the concept of an orthographic word in Slovene over time.

4. Clitic change in Slovene

A cursory examination of diachronic tokenization changes shows that the most salient differences are in the treatment of clitics, an interesting class of words/morphemes in their own right. Clitics are lexical items which can never serve as an independent prosodic domain, because they are prosodically weak and hence unaccented. As such, they must attach to a nonclitic (host) and so become part of the prosodic words to which their hosts belong (Franks/Holloway King 2000: 4). The annotated corpus enables us to discover spelling conventions for specific clitics in a given period (whether they were freestanding or bound[3]), and also study their orthographic variants.

According to Toporišič (2000: 112), clitics in Slovene can be prepositions, particles, conjunctions, personal forms of the auxiliary verb *biti* "to be", and clitical forms of personal pronouns in specific oblique cases. Despite the lack of accent, clitics in contemporary Slovene are typically treated as independent orthographical units, and are written separately from their hosts.

By elimination of the query results not containing at least one clitic, we get a basic division of all possible bound and freestanding clitics in historical Slovene; from the large set of all possible clitics, about 120, only a small set can be bound, and they are identified in Table 2.

Lexical category	Possibly bound clitics	Always freestanding clitics
Preposition	*k, v, z, do, iz, na, ob, po, za*	*pri, pred, čez, med, brez ...*
Particle	*ne, li*	*pa, še, že, pak, naj ...*
Pronoun	*se, mi*	*ga, jih, mu, te, ti ...*
Auxiliary verb *be*	*ste, bi*	*sem, si, je, smo, so ...*
Conjunction	*da, ni*	*in, ki, če, pa ...*

Table 2: Division of clitics to possibly bound and freestanding

Searching by lemmas enables us to investigate orthographic variants of bound clitics. The most variable are the one-letter prepositions *z* "with", *k* "to" and *v* "in", which for a very long time did not have a standardized orthography; the

[3] In this paper the term 'bound clitic' is used for any orthographically bound clitics, that is for clitics written jointly with their host, and not only for special accusative pronouns that are attached to prepositions and cannot be separated from them, for example *zanj* 'for him'.

variability of their spelling also originates from phonological changes due to binding with their host. An example is the palatalized form *ž* of preposition *z* when used with pronouns that start with palatalised [ɲ] (1), or the reduced form *na* of negative *ne* in a word-initial position (2). The writing and pronunciation variants of these four clitics are represented in Table 3.

(1) *Ty ſo hotéli shnym*[4] *govoriti* (published in 1584)
ti so hoteli ž_njim govoriti (transliterated)
ti so hoteli z_njim govoriti (modern)
they AUX want with_him talk
"They wanted to talk to him."

(2) *katiri pak nabo viruvov, bo pogublen* (1777)
kateri pa ne_bo veroval bo pogubljen
who however NEG_AUX believe AUX condemned
"He who will not believe, however, will be condemned."

Clitic	*k* 'to'	*v* 'in'	*z* 'with'	*ne* 'not'
Writing variants	*k', h'*	*u, v, v'*	*s', ſ, sh, z*	*ne, na*
Pronunciation variants	*k, h*	*u, v*	*s, z, ž*	*ne, na*

Table 3: Writing and pronunciation variants of clitics

Clitics most frequently written together with their hosts are the prepositions *k*, *z*, *v* but even these are more often freestanding. The exception is the only 16th century text in the corpus where they are consistently attached to the hosts, but marked with an apostrophe, e.g. *k'njemu* "toward him" for modern *k njemu.* In the 17th century, separate writing of the preposition already prevailed (despite the use of the apostrophe), and bound prepositions appear only sporadically. The trend continued into the 18th century, with the apostrophe gradually disappearing, and in the 19th century, writing the prepositions separately became the rule, creating a norm that still exists. Other one-syllable prepositional clitics were rarely bound. We find some cases of bound prepositions, namely *do* "until/as far as", *iz* "from", *na* "on", *ob* "at", *po* "after" and *za* "for", but of these only *na* is more consistently bound in the only 17th century work in the corpus

[4] In 16th century Slovene the voiced palato-alveolar fricative (IPA: ʒ), which in modern Slovene is written as *ž*, was written *sh*. The palatal nasal in personal pronouns (IPA: ɲ), which in modern Slovene is written as *nj*, was written *n* and followed by *y*, which could either mark the palatalization of the preceding nasal or the long accented vowel *i* (Merše/Jakopin/Novak 1992: 334335).

(3), and even there it is rare compared to the freestanding variant, e.g. *napoſtelo*, "on bed" for contemporary *na posteljo*.

The other lexical category that was sometimes bound is the particle, especially the negative *ne* "not". In the 16th century the proclitic negative was consistently bound with the verb following it. In the 17th century, writing it together with the verb was still prevalent, but in the 18th and 19th century the bound form was becoming rarer, until it disappeared with the progressing standardization of the literary language. The enclitic particle *li* was also sporadically bound. Unlike *ne*, the enclitic *li* was predominantly freestanding in texts regardless of corpus period, with two exceptions: certain works of one 18th century author and several mid-19th century authors, who were also writing it attached to the host, but marked with a hyphen.

The most limited usage was joint writing of cliticized (predominantly reflexive) pronouns and their hosts. It was characteristic only of 17th century writing, and limited to a position directly following the imperative.

(3) *netagotiteſe, inu shnio nepreperajteſe* (1695)
ne_togotite_se in z_njo ne_prepirajte_se
NEG_get_angry_REFL and with_her NEG_argue_REFL
"Do not get angry and do not argue with her."

Forms of the auxiliary verb *biti* "to be" and conjunctions are almost consistently freestanding. The only exception is the sporadic joint writing of two clitics, for example conjunction *da* "that" and auxiliary (*da_ste*, *da_bi*) or conjunction *da* and particle *li* (*da-li*).

We can use the data represented above to make synchronic representations of orthographical properties of clitics for a specific period or author, as shown in Table 4.

Period/author	Bound	Predominantly bound	Predominantly freestanding	Freestanding
16th century	*k, v, z, ne*	–	*Li*	*do, iz, na, ob, za, se*
17th century	–	*ne, se*	*iz, na, ob, za*	*k, v, z, do*
18th century – M. Pohlin	*ne, li*	–	*k, v, z, iz, na, po, za*	*do, ob, se*
18th century – J. Japelj	–	–	*v, z, za, ne*	*k, do, iz, na, ob, se*

Table 4: Orthographical properties of clitics for a specific period or author[5]

[5] Marko Pohlin was a writer from the 18th century who proposed a form of literary language that was founded on spoken language of the central Slovenia and is the author in whose works clitics were most often bound. Jurij Japelj is the main author of the Catholic translation of the Bible (1784–1804). In opposition to Pohlin he used the historically-based literary language that became the foundation for the standardization of Slovene in the 19th century.

As Tables 2 and 4 show, Slovene clitics were predominantly freestanding in all periods. In the 16th century, only some clitics (one-letter prepositions and negation) were consistently orthographically bound. In the 17th and 18th centuries, the norm for writing clitics was unstable, and much variation occurred in the texts, but gradually clitics were more consistently treated as individual orthographic words, thus creating the norm that is used in modern Slovene. Table 3 also shows that, during the unstable period of the literary language, usage was largely dependent on individual spelling conventions of authors, but even those who opted for bound clitics rarely achieved consistency in their writing.

With the help of a list of all bound clitics and their hosts, we can also discover syntactic restrictions for specific clitics, for example the negative particle *ne*. Comparison of the lists of examples with bound and freestanding negative shows that the negative can be bound only when directly preceding a verb, and not when it occurs in front of a particle or adverb.

(4) *ſhivino ne sunej paſti, je ſkodlivo* (1789)
živine ne zunaj pasti, je škodljivo
cattle not outside to_graze, is harmful
"Do not let the cattle graze outside, it is harmful."

(5) *Zhebelle dobru permaſhiti , de v' ſnegi ven neletę* (1789)
Čebele dobro primašiti, da v snegu ven ne lete.
Bees well to_block so in snow out NEG_fly
"Block the bees well so they do not fly out in the snow."

5. Conclusions

This paper presents the *goo250k* corpus of historical Slovene and, on this basis, offers an analysis of clitic change over time. This analysis serves as an example for the possible use of a historical corpus in linguistic research, but is also interesting from a technical standpoint, as it covers the most common cases of tokenization mismatch in historical vs. contemporary Slovene. The analysis shows that free-standing clitics, which are the norm for contemporary Slovene, were typically not bound even in historical Slovene. However, certain classes did have a tendency towards orthographic joining with their phonetic host, depending on syntactic function and placement, and on the individual author.

Acknowledgements

The work presented in this paper was in part supported by the EU FP7-ICT Project IMPACT "Improving Access to Text" and the Google Research Award "Developing Language Models of Historical Slovene".

References

Brants, Thorsten (2000): TnT – A Statistical part-of-speech tagger. In: Proceedings of the sixth Applied Natural Language Processing conference ANLP-2000. Seattle, WA, 224-231.

Christ, Oliver (1994): A modular and flexible architecture for an integrated corpus query system. In: Proceedings of COMPLEX '94: 3rd conference on Computational Lexicography and Text Research, Budapest, Hungary, 23-32.

Erjavec, Tomaž (2011): Automatic linguistic annotation of historical language: ToTrTaLe and XIX century Slovene. In: Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, Portland OR, June 2011. Association for Computational Linguistics, 33-38.

Franks, Steven/Holloway King, Tracy (2000): A handbook of Slavic clitics. New York and Oxford: Oxford University Press.

Gotscharek, Annette/Reife, Ulrich/Ringlstetter, Christoph/Schulz, Klaus/Neumann, Andreas (2010): Towards information retrieval on historical document collections: the role of matching procedures and special lexica. In: International Journal of Document Analysis and Recognition 2010: 1-13,.

Merše, Majda/Jakopin, Franc/Novak, France (1992): Fonološki sistem knjižnega jezika slovenskih protestantov [The phonological system of the Slovene reformation literary language]. In: Slavistična revija 40.4: 321-340.

Prunč, Erich (2007): Deutsch-slowenische/kroatische Übersetzung 1848-1918. Ein Werkstättenbericht. In: Wiener Slavistisches Jahrbuch 53: 63-176.

Reffle, Ulrich (2011): Efficiently generating correction suggestions for garbled tokens of historical language. In: Natural Language Engineering 17. 2: 265-282

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 OR, June 2011. Association for Computational Linguistics, 124-128.

TEI Consortium (ed.) (2007): TEI P5: Guidelines for electronic text encoding and interchange. http://www.tei-c.org/Guidelines/P5/

Toporišič, Jože (2000): Slovenska slovnica [Slovene grammar]. Maribor: Založba Obzorja.

PETER GILLES / EVELYN ZIEGLER

The *Historical Luxembourgish Bilingual Database of Public Notices*[1]

Abstract

Bilingual parallel corpora are increasingly recognised as solid bases for contrastive linguistics, both from a synchronic and diachronic perspective. The *Historical Luxembourgish Bilingual Database of Public Notices* is a diachronic single-genre corpus, comprising French-German parallel texts from the years 1795 to 1920. This paper gives an overview of the text-corpus, specifying the features of the genre 'public notices', and explaining the criteria for text selection. Building on that, the paper details the compilation and presentation of text and image data stored in the corpus. Finally, we describe the technical tools for indexing, searching and managing the text and image data.

1. Introduction

The *Archive of the City of Luxembourg* hosts a large number (around 7,000) of public notices, i.e. printed proclamations presented to the public, between the end of the 18th and the beginning of the 20th century. These large-scale printed public announcements served the city administration during that time as a central means of communicating laws, regulations and organizational matters concerning the public life of the city of Luxembourg. As most of the public notices are written in two languages (French and German) they form an ideal basis[2] for a large parallel corpus, defined as "a source text and its translation into one or more languages" (Aijmer 2008: 276). Parallel corpora guarantee comparability, but also allow the study of monolingual features and developments. However, they are still a desideratum in historical linguistics, as Claridge (2008: 256) points out: "[F]rom a cultural-historical perspective, it would be interesting to link several European languages in a corpus of parallel texts or translations."

[1] This paper was made possible by a project funded by the *Fonds National de la Recherche*, Luxembourg.

[2] We are very grateful to Dr. Evamarie Bange/Archive of the City of Luxembourg, for her permission to use the Affichen-Corpus.

The corpus of the *Historical Luxembourgish Bilingual Database of Public Notices* provides a resource for studies in both contrastive linguistics (e.g. similarities and differences concerning information structure phenomena, lexicography, metaphorical conceptualizations, variation and change of features and constructions in the process of standardization) and in contact linguistics (transfers of lexical, grammatical and orthographic features), as both languages have played (and still play) a long and important role in the history of Luxembourg.

The choice of the genre 'public notice' (see example below) is driven by several facts:

1) This text genre represents a central 'top-down' strategy of information policy carried out by the municipality of Luxembourg. Therefore, these culturally significant documents offer new opportunities for linguistic inquiry into strategies of public language use at the interface of standard language and technical language use.

2) The genre is quite unique and has so far not been considered as a primary research source in linguistics, political science, historical, legal or cultural studies.

3) A single-genre corpus offers the opportunity to study the relation between language and genre, and how a single genre is affected by language change over time (from both macro- and micro-text level approaches).

As mentioned above, the genre 'public notice' formed a central means of top-down communication[3] in the regulatory discourse[4] of the municipality of Luxembourg City during the 18th and 19th centuries. From a socio-cultural perspective, the development of the genre is closely tied to a development that has been referred to as the "linguistic colonization of the public sphere" ("Die Kolonialisierung des öffentlichen Raums durch die Schrift", Auer 2010: 295). Auer means by this that the producers of signs displayed in the public sphere exert agentivity and power, the more so as not every member of society is granted the authority and right to do so. Governmental and administrative institutions considered the emerging public sphere as an arena for displaying and exercising power. One way to do this was to post public notices at the town hall or in other prominent places which provided general public access (e.g.

[3] See Ben-Rafael et al. (2006). For a critical discussion of the concept see Auer (2010).

[4] The concept of "regulatory discourse" goes back to Scollon/Scollon (2003).

church doors). Public notices were used as an instrument of social control, aiming at civil obedience and social order. As Rickards (1973: 7) puts it: "the printed public announcement is essentially an instrument of public control. [...] it represents an extension of the power of the ruler – authority mass-produced." This authoritative gesture is expressed on the text-linguistic level in devices such as royal emblems (e.g. the royal Luxembourg lion emblem), and the use of honorifics (e.g. Grand-Duché du Luxembourg).

2. Development and characteristics of the text genre "public notice"

The history of the genre begins with orders that were proclaimed for the illiterate by a town crier. Rickards (1973: 8) describes this practice as follows: "Summoned by drum or trumpet to the market-place, the people saw the actual document unrolled [...] and heard its oracular pronouncement." In those days the production of a public notice was an individual effort, and documents with the hand-written orders of monarchs are typical examples.

Due to increasing "pedagogicalization" (Mattheier 2003) and popularization of writing and reading, as well as technical innovations such as mass printing and colour typography, town criers were no longer needed, and the public notice rose to general importance as an efficient means of information management, contributing to what is called in modern sociolinguistic terms the "linguistic landscape" of urban spaces (Shohamy et al. 2010). As a consequence, the production of public notices became a process involving a number of people in the composition, revision, error correction, layout design, and printing of the texts.

The genre "public notice" can be described according to both text-external and text-internal characteristics (see Esser 2009). The defining text-external characteristics are as follows:

- Visual mode of communication;
- Written to be viewed at close range, for "targeted perception";
- Temporarily exhibited documents in places where the public is likely to see them;
- Primary text functions/intentions: provision of information about government and municipal activities, and announcement directed at the citizens (instruction, order, prohibition);

- Text producers: mayors of Luxembourg City and/or their secretaries;
- Text recipients: inhabitants of Luxembourg City.

The defining text-internal criteria are:

- Wide range of themes: health, education, election, public affairs, and commerce;
- Mostly bilingual documents written in French and German;
- Use of language for specific purposes (stylistic features, technical lexis) and special notice vocabulary;
- Typographic distinction between French and German through the use of Roman and Gothic (black letter) fonts;
- Variation of layout and typography:
  - alternating horizontal and vertical text vector; changing special notice typography; use of national emblems in accordance with political changes;
  - professionalization of layout and typography over time, e.g. use of display typefaces from the early 1800s;
- Emergence of a global macro-structure; specialization and differentiation of the genre over time;
- Size of text length varies between 700 and 51,000 characters.

Although the names of the writers or those in charge of the texts are generally recorded at the bottom of the documents, nothing is known about the language biographies of the writers or translators. Therefore, the question must remain as to whether there was one single bilingual writer or two, and in which direction the translation went. Only for documents published between 1795 and 1814 do contextualization cues in the German text column make it possible to determine the direction of translation. Among these cues are dates referring to the *French Revolutionary Calendar*,[5] or references to Luxembourg as *Département des Forêts*, which indicate that the direction of translation was from French to German. In fact, from the current perspective, it seems clear that no individual authors are identifiable. Rather, for the production of these texts, several stages of conceptualization have to be assumed, with the involve-

[5] The *French Revolutionary Calendar,* invented during the French Revolution, was in use in France from 1793 until 1805.

ment of several writers, along with the deployment of prefabricated text modules and routines.

3. Compilation of the Corpus

The *Historical Luxembourgish Bilingual Public Notices Database* is a single-genre corpus, consisting of a selection of 2,000 documents out of a total of almost 7,000 stored in the Archive of the City of Luxembourg, and covering the years 1795 to 1920. As the compilation was mainly driven by the intention of allowing a fairly broad scope of research interests, the start and end dates of the corpus are defined, not by linguistic criteria, but by socio-historical landmarks, namely the French annexation of Luxembourg and the end of the First World War. Thus, the start and end dates not only cover a long period (the so called "long 19th century"), but they also link up to the traditional periodization of the history of Luxembourg, including important times of transition.

Nevertheless, sociolinguistic criteria are also taken into account, as the popularization of the text genre 'public notice' in Luxembourg is most closely connected to the effects of the French Revolution, in particular the introduction of democratic ideas, such as public information and participation in decision-making concerning public affairs. Public notices were used to bring about transparency for those who wanted to know more about government and municipal activities. In this vein, public notices were not only an instrument for controlling the public, but also for allowing the public to form their opinions on actions taken by the administration (e.g. estate actions, plans for road building etc.).

The production and publication of the public notices is embedded in a changing socio-historical context. To allow investigation of the relationship between linguistic choices and socio-cultural contexts, the corpus is sub-divided according to the established periodization of the history of Luxembourg. The content of the corpus encompasses four periods based on important landmarks in the history of Luxembourg:

1) LU II: 1795-1814
2) LU III: 1815-1843
3) LU IV, 1: 1844-1890
4) LU IV, 2: 1891-1920

From 1795 to 1814, Luxembourg was under French rule and became part of the *Départements des Forêts*. In 1815, the Congress of Vienna declared Luxembourg a Grand Duchy and allocated it as personal property to William I, King of the Netherlands.

Luxembourg remained under the sovereignty of the Orange-Nassau dynasty until 1890, the year of the death of William III. Between 1815 and 1867, it was also member of the German Confederation (*Deutscher Bund*), and the City of Luxembourg was a confederate fortress with a Prussian garrison. During the Belgian Revolution (1830-1839), Luxembourg was divided and lost half of its territory. The Treaty of London (1839) established Luxembourg's present-day borders and ruled that the western, francophone region should go to Belgium, while the eastern, germanophone region[6] remained, to constitute the Grand-Duchy. In the same year, Luxembourg was granted more independence and became a nation state. With the foundation of the nation state, Luxembourg was confined to the germanophone territory. Nevertheless, French and German continued to be used as both national and administrative languages (see Fehlen 2009, Gilles/Moulin 2003 and Weber 2000).

4. Structure of the corpus

The corpus comprises 2,000 documents. It can be regarded as representative, and sufficient for sociolinguistic research,[7] e.g. for documenting and analyzing language management, language policy, language change and language contact. The corpus allows for typological and genetic approaches, as well as quantitative analysis for less frequent linguistic features and structures. According to standard corpus definitions, it falls in the category of a systematically compiled corpus. The following selection parameters were chosen in compilation:

- Language (only bilingual French-German documents were chosen)
- Topic (health, education, public affairs, commerce)
- Period of time (1795-1814, 1815-1843, 1844-1890, 1891-1920)
- Quantity (approx. 500 documents per period)

As an equal number of texts per period were included, the internal composition of the corpus is balanced, allowing synchronic and diachronic studies as well as inter-period cross-linguistic comparisons.

[6] The concepts of "francophone" and "germanophone" are used to refer to a spectrum of varieties (standard and non-standard) realized in the speech communities. Furthermore, the concept of "germanophone" is intended as a superordinate concept including German and emerging Luxembourgish. It would not be sufficient to conceive of the germanophone part as "German speaking". Finally, it should be pointed out that the distinction between a francophone and a germanophone region is not meant as a clear cut distinction between homogeneous entities but as a distinction that includes cultural diversity.

[7] For a detailed discussion of sociolinguistic research perspectives in corpus linguistics see Romaine (2008) and Mair (2009). For the use of corpus linguistics in the study of language change see Curzan (2009).

In a first step, the 2,000 selected public notices, ranging in size between letter format and DIN A0 (approx. 800 x 1200 mm), and several of them consisting of more than one page, were scanned with a large-scale, high-definition scanner. Information about these image files, along with their metadata (shelf mark, date, title, content, issuing authority, used languages) is stored in a relational database (*MySQL*). Through the web-based frontend, all this information is retrievable through searching facilities. Figure 1 provides an overview of the frontend. In this case the corpus is filtered to show only those public notices of the year 1899.

		1899				
	LU Imp. IV.2_0052	1899 Janvier 26.	Arrêté concernant la police pendant le carnaval		collège échevinal	allemand/fra
	LU Imp. IV.2_0151	1899 Octobre 25.	Kandidatenliste für die Gemeinderathswahl vom Dienstag den 31. Oktober 1899		MOUSEL E., président bureau de vote ppal	allemand
	LU Imp. IV.2_0254	1899 Juillet 15.	Anniversaire de la Naissance de S.A.R. Monseigneur le Grand-Duc	Anniversaire de la Naissance de S.A.R. Monsieur le Grand-Duc		allemand/fra
	LU Imp. IV.2_0261	1899 octobre 20.	Bekanntmachung	Bekanntmachung: Wahl von 7 Mitgliedern für den Gemeinderat		allemand
	LU Imp. IV.2_0293	1899 Janvier 26.	Arrêté concernant la police pendant le carnaval		MOUSEL E., président; FABER J., secrétaire	allemand/fra

Figure 1: Overview of the frontend

When a specific public notice is selected, users get access to the full-sized image, where they can zoom and pan into sections (Figure 2). In technical terms, the image player draws on functionality provided by the Zoomify function of the *OpenLayers* software package.[8] This image viewer is especially useful for navigating in large-size documents with varying font sizes.

Figure 2: Sample section

[8] http://openlayers.org

As well as to the scanned image, the user also has access to the full text of the public notice. All documents were text-digitized by a company specializing in the text-digitization of old documents. Instead of using OCR techniques, which are known to be especially error-prone for old documents, the more traditional way of typing all documents by hand was chosen. In fact, all documents were double-keyed and automatically compared for discrepancies. This technique guarantees extremely low error rates (Büdenbender 2011). In the next step, this raw text data was transformed into XML-format, in accordance with the principles of the *Text Encoding Initiative* (*TEI*), in order to guarantee sustainable data exchange with similar research projects or software. In order to keep as much as possible of the original document structure, the full text contains extensive tagging. This comprises tagging of the different kinds of headings, the switch between font-face and font-type (French text is printed in Roman, while German is printed in Gothic) and the differentiation between sub- and superscript text. As most documents are bilingual, it is of special importance that the different languages (mostly French and German) can be distinguished in the full-text data. Since the XML conversion has not yet been completed, it is not possible to provide definite figures about the size of the corpus. As the average word count for one public notice lies at around 800 words, it can be estimated that the final corpus will contain approx. 2 million running words.

The full text of all public notices is further available for concordance and text retrieval software to facilitate linguistic analyses. The next figure shows the result of a pattern search using the *tlCorpus* programme.[9] This corpus organization and search software utilizes regular expressions, allowing users to find nearly all the desired text patterns. In the example given in Figure 3, '.*ntern' was searched for, matching all words (i.e. strings between spaces) ending in the string 'ntern.' The list with the search results shows the search word in the centre, with an amount of context on the right and on the left (so called KWIC ('key-word-in-context') concordance). Additionally, the name of the document, where the search term has been found, is indicated on the far left.

[9] http://tshwanedje.com/corpus/

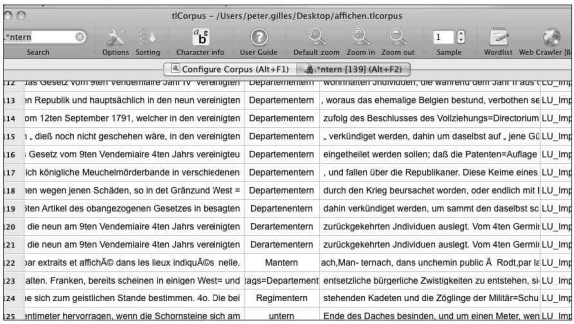


Figure 3: KWIC search

As most of the public notices are bilingual, it is of interest to see how certain elements in one language are replicated in the other. Because of the tagging of the corpus, links between the two languages can be easily and quickly established. This interlinked corpus structure readily allows the finding of equivalent words and structures in the two languages. Due to the fine-grained structure of the corpus, it will be possible to analyze the evolution of certain linguistic features, especially in the domain of language contact, over the timespan of the available data.

5. Conclusion

The aim of this paper was to present the structure of an extensive corpus of a hitherto largely neglected text genre, in studies on language history, namely historical public notices of the city of Luxembourg, which is currently being compiled at the Universities of Luxembourg and Duisburg-Essen. It comprises the period from the end of the 18th to the beginning of the 20th century, and thus covers the pivotal period of establishment of the language of administration, and the development of official communication in the public sphere. Moreover, it is a parallel corpus, as most of the public notices are bilingual (German/French). It can also be accessed as an image database, for the study of the the visual arrangement and layout (e.g. the use of emblems and handwritten additions), and as a searchable full text database, giving the user a flexible analytical tool.

With about 2 million words of running text, this corpus permits the study of several aspects of language use in the administration of the city of Luxembourg. In terms of sociolinguistic research, questions can be addressed concerning language policy and language standardization in the context of multi-

lingualism. Linguistically, the opportunities it offers for studying aspects of language contact are especially promising. For most of the public notices, the direction of translation seems to be from French to German, and the (mostly lexical and syntactical) transfers from French can be studied in situ in the German version.[10]

References

Aijmer, Karin (2008): Parallel and comparable corpora. In: Lüdeling/Kytö (eds.). Vol. 1, 275-292.

Auer, Peter (2010): Sprachliche Landschaften. Die Strukturierung des öffentlichen Raums durch die Schrift. In: Deppermann, Arnulf/Linke, Angelika (eds.): Sprache intermedial. Stimme und Schrift, Bild und Ton. Berlin: de Gruyter, 271-298.

Ben-Rafael, Eliezer/Shohamy, Elana/Amara, Muhammad Hasan/Trumper-Hecht, Nira (2006): Linguistic landscape as symbolic construction of the public space: the case of Israel. In: Gorter, Durk (ed.): Linguistic landscape: a new approach to multilingualism. Clevedon: Multilingual Matters, 7-30.

Büdenbender, Stefan (2011): LexicoLux: EDV-philologische Perspektiven bei der Erstellung eines Wörterbuchnetzes der Grossregion. In: Gilles/Wagner (eds.), 261-275.

Claridge, Claudia (2008): Historical corpora. In: Lüdeling/Kytö (eds.). Vol. 1, 242-259.

Curzan, Anne (2009): Historical corpus linguistics and evidence of language change. In: Lüdeling/Kytö (eds.). Vol. 2, 1091-1108.

Deumert, Ana/Vandenbussche, Wim (eds.) (2003): Germanic standardizations: past to present. Amsterdam: Benjamins.

Esser, Jürgen (2009): Introduction to English text-linguistics. Frankfurt a.M., etc.: Peter Lang.

Fehlen, Fernand (2009): BalaineBis. Une enquête sur un marché linguistique multilingue en profonde mutation. Luxemburgs Sprachenmarkt im Wandel. (= Recherche Etude Documentation 12). Luxembourg: SESOPI Centre Intercommunautaire.

Gilles, Peter/Moulin, Claudine (2003): Luxembourgish. In: Deumert/Vandenbussche (eds.), 303-329.

Gilles, Peter/Wagner, Melanie (eds.) (2011): Linguistische und soziolinguistische Bausteine der Luxemburgistik. (= Mikroglottika. Minority Language Studies 4). Frankfurt a.M., etc.: Peter Lang.

[10] For first analyses of the data see Ziegler (2011).

Lüdeling, Anke/Kytö, Merja (eds.) (2008): Corpus Linguistics. An International Handbook. 2 vols. Berlin: de Gruyter

Mair, Christian (2009): Corpus linguistics meets sociolinguistics: the role of corpus evidence in the study of sociolinguistic variation and change. In: Renouf, Antoinette/Kehoe, Andrew (eds.): Corpus linguistics: refinements and reassessments – proceedings of the 2007 ICAME conference – Stratford-upon-Avon. Amsterdam: Rodopi, 1-26.

Mattheier, Klaus (2003): German. In: Deumert/Vandenbussche (eds.), 211-244.

Rickards, Maurice (1973): The public notice: an illustrated history. New York: Potter.

Romaine, Suzanne (2008): Corpus linguistics and sociolinguistics. In: Lüdeling/Kytö (eds.), Vol. 1, 96-111.

Scollon, Ron/Scollon, Suzie Wong (2003): Discourses in place: language in the material world. New York: Routledge.

Shohamy, Elana/Ben-Rafael, Eliezer/Barni, Monica (eds.) (2010): Linguistic landscape in the city. Bristol: Multilingual Matters.

Weber, Nico (2000): Multilingualism and language policy in Luxembourg. In: Deprez, Kas (ed.): Multilingualism and government: Belgium, Luxembourg, Switzerland, Former Yugoslavia, South Africa. Pretoria: Van Schaik, 82-91.

Ziegler, Evelyn (2011): Sprachenpolitik und Sprachenmanagement in Luxemburg (1795-1920). In: Gilles/Wagner (eds.), 177-203.

BRITTA JUSKA-BACHER / CERSTIN MAHLOW

Phraseological change – a book with seven seals?

Tracing the diachronic development of German proverbs and idioms by a combination of corpus and dictionary analyses

Abstract

Dictionaries and collections of proverbs, idioms, or phrasemes usually provide synchronic information with only little evidence of actual use. While various extensive dictionaries and collections are available for German, a comprehensive description of structural and semantic *changes* of phrasemes over time is still lacking. Our article highlights some issues and challenges, and presents a semi-automatic corpus-based approach for the diachronic investigation of phraseme development. We argue for a combination of dictionary exploration and corpus-based methods, to provide reliable information about the diachronic development of German phrasemes.

1. Introduction: Phraseological change

Phrasemes – the focus here is on idioms and proverbs – are defined by *polylexicality*, relative *stability*, and *idiomaticity* (Burger 2010: 36-42). Diachronic research must consider change with respect to each of these properties. Additionally, due to polylexicality, language change needs to be examined at different levels, i.e. the meaning and structure of each component, as well as phraseme structure and phraseological meaning.

To illustrate change at various levels, let us consider the expression, 'to swim against the current'. Writing in the 16th century, Martin Luther used various forms of what is today a fairly common idiom (Piirainen 2006): *wider den Strom/stram gehen/sein/streben/fechten*[1] ('to go/be/strive/fight against the current'). In this context, the meaning is "to attempt the impossible" (Burger 2007: 97). Today, the most widely used form is *gegen den Strom schwimmen* ('to swim against the current') in the sense of "behaving unlike the majority", as shown by Juska-Bacher (2009: 342). At *component level*, we find different spellings (*stram* vs. *Strom*). At *phraseme level*, two structural differences can be observed: use of prepositions (*wider* vs. *gegen*), and use of verbs (*gehen/sein/streben/fechten* vs.

[1] Luther did not use the verb *schwimmen* ('to swim'). See also Parad (2003: 219).

schwimmen). At the *semantic level*, Parad (2003: 221) notes a weakening, from "hopeless endeavour" to "non-conformist individualism". A hypothesis of diachronic structural and semantic change in phrasemes such as *gegen den Strom schwimmen* can be deduced from a simple juxtaposition of two different empirical data sets. However, to this day, there is no German phraseological dictionary that comprehensively describes such structural and semantic changes. In this article, we first highlight challenges and issues encountered when investigating the diachronic development of German phrasemes. We then present our approach to overcoming some of these issues, by describing the principles, and providing evidence with respect to the example used here.

2. *A book with seven seals*?

At the methodological level, investigations of historical language stages face various fundamental problems: for example, it is impossible to collect empirical data by means of observation, surveys, or experiments. Historical linguists rely on trawling through dictionaries, grammars, and primary sources, usually in printed form. However, sources are limited both in terms of quantity and quality. As Claridge (2008: 247) points out, historical texts "to a large extent reflect the language of the social and educational elite", and "historical corpora can never even remotely capture the full variety of language".

For the study of historical phrasemes, we need to consider idiomaticity and relative stability. Multiword expressions in general can be used both literally and as phrasemes: *Er schwimmt gegen den Strom* ('He swims against the current') may refer to a man who disregards received opinion. Taken literally, however, it describes a man who is actually trying to swim upstream. The lexicographer has to distinguish which meaning is used. This needs to be done manually (cf. Rothkegel 2007), making a full analysis of large corpora of texts extremely time consuming.

Relative stability of phrasemes in the speech community particularly applies to the *current* perspective, while greater lexical, morphosyntactic, and orthographic variability – and hence reduced stability – has been noted for historical examples (Burger/Linke 1998: 747). Inevitably, the linguistic competence of phraseographers examining older language stages will be somewhat impaired, which will affect identification of polylexical units, definition of their obligatory and optional components, pragmatics, and lexicographic descriptions, including decisions on whether examples demonstrate a variant common to

this particular linguistic community, or whether we are dealing with any kind of modification (Burger/Linke 1998: 743). Therefore, the use of tools such as dictionaries and corpora becomes more relevant here than in studies of present-day languages. Metalexicographers have repeatedly criticized the neglect or unsystematic presentation and placement of phrasemes in dictionaries (e.g. Burger 2010: 179-181). This criticism applies even more strongly to historical dictionaries, making the search for phrasemes a time-consuming and often disappointing endeavour (Burger/Linke 1998: 744, Dräger 2009: 33, Dräger 2010: 412f.).

Phraseological dictionaries – contemporary or historical ones – usually cite phrasemes and their meaning at a certain point in time; they thus provide a *synchronic* perspective. Moreover, the sources are rarely, if ever, identified. The state of the art in current phraseography implies that any reliable description of phrasemes should be based on empirical data in the form of corpus analyses (Mellado Blanco 2009: 16). This particularly applies to historical phraseography, because historical lexicographers cannot rely on intuition, data provided by informants, or information found in previous dictionaries (Dräger/Juska-Bacher 2010: 165f).

A fundamental problem that arises when working with corpus analyses is the extremely low frequency of phrasemes (see Colson 2007: 1072); their investigation and documentation require particularly large corpora. Phrasemes occur with even less frequency the further back in time we search (Claridge 2008: 245).

3. Combining corpus and dictionary analyses

The project *OLdPhras – German Proverbs and Idioms in Language Change. Online Dictionary for Diachronic Phraseology*[2] is intended to close a gap in diachronic phraseology for German. The goal is to provide an online dictionary describing the diachronic structural and semantic development of German phrasemes from roughly 1650 to the present. In contrast to Middle and Old High German, plenty of source material – dictionaries as well as texts, some even in large corpora – is available for this period.

[2] "Deutsche Sprichwörter und Redewendungen im Sprachwandel. Online-Lexikon zur diachronischen Phraseologie des Deutschen in neuhochdeutscher Zeit" (http://www.oldphras.net). The project is domiciled at the University of Basel and funded by the Swiss National Science Foundation.

Systematic semi-automatic exploration of a range of mostly synchronic dictionaries, from the 18th to the 21st centuries,[3] enables us to obtain an initial diachronic overview of phrasemes mentioned in older collections, but not in newer ones, and vice versa. So far, corpus-driven methods have not permitted the creation of this kind of phraseme inventory.

We will search historical and current corpora for evidence of actual use in different time periods. Their context will provide references to the citation form as well as restrictions (in order for the citation form to be adapted to certain requirements, see also Moon 2007: 213). This makes it possible to infer pragmatic as well as denotative and connotative aspects of meaning. This will again be a semi-automatic process, as the phraseographer has to evaluate the idiomaticity of each match.

The document-based observation of a phraseme from approximately 1650 to the present allows for a detailed empirical documentation of change. Research literature on the relevant phrasemes will be integrated to provide additional information. In the remainder of this section, we briefly describe our approach using our example of *gegen den Strom schwimmen.*

3.1 Phraseme selection for the online dictionary

For our dictionary we consider only phrasemes containing at least one nominal component. The nominal components of two sets each of two historical and current phraseme collections[4] have been processed automatically to extract phrasemes containing nouns which

1) occur with *high frequency*[5] both in the historical and in the current list of phrasemes, indicating a constant productivity of components: e.g. many words for body parts (somatisms), such as *Hand* ('hand'), *Kopf* ('head'), *Herz* ('heart'), *Auge* ('eye'), *Ohr* ('ear');

2) show a striking *difference in frequency* (change of productivity):
   a) words for animals such as *Affe* ('monkey'), *Laus* ('louse'), or *Kuh* ('cow'), as well as *Narr* ('fool'), *Schnee* ('snow'), or *Feder* ('feather') are more frequent in historical phraseme collections than in contemporary ones;

[3] Among others, Adelung (1793-1801), Campe (1807-1812), Wander (1867-1880), Borchardt (1888), Grimm (1854-1960), Röhrich (2002), and Dudenredaktion (2008).

[4] Historical: Adelung (1793-1801) and Wander (1867–1880); contemporary: Dudenredaktion (2008) and the on-line dictionary *Redensartenindex* (http://www.redensarten-index.de).

[5] We use a threshold of 2%, i.e. if a noun belongs to the top 2% in the frequency list of all nouns of a collection, it is considered frequent in this collection.

b) more frequent in contemporary collections than in historical ones are *Nerv* ('nerve'), *Fall* ('case'), or *Punkt* ('point');

3) are *infrequent*[6] on both lists, i.e. unical components such as *Affenschande* ('beastly shame'; *Affe* = monkey), *Friedenspfeife* ('peace pipe'), *Gnadenbrot* ('charity');

4) are infrequent in contemporary lists, but more frequent in historical lists, like *Krebs* ('crab'/'crayfish'), *Käse* ('cheese'), or *Weib* ('woman')

5) are infrequent in contemporary lists and do not exist in historical lists, like *Fleischwolf* ('meat grinder'), *Brechstange* ('crow bar'), *Sprungbrett* ('diving board'), or *Abstellgleis* ('holding track').

The phrasemes to be investigated in detail will be selected to represent instances of these types. Using several semi-automatic processing steps, we will combine different citation forms (due to various dictionaries and collections using different guidelines) to identify some kind of prototypical form of phraseme and its variants mentioned in various collections. These data will help us search for evidence in corpora by including modifiers, variation in the verbal components, morphological variation, word order, etc.

Example: *to swim against the current*

Comparison of Luther's use of this phraseme with modern occurrences suggests both structural and semantic changes between the 16th and 21st centuries. Various dictionaries were consulted in order to ascertain more detail about these changes, and the results are presented in Table 1. Although the preposition *gegen* appears very early, *wider* is prevalent from the early 17th until the mid-19th centuries. That is when *gegen* takes the lead, and becoming the sole form after the late 19th century,[7] until Röhrich (2002) and Dudenredaktion (2008) list both prepositions as equally acceptable.

The verb *schwimmen* predominates from the early 17th century, even though other verbs occur in the 19th century, and from the middle of the 19th century onwards, dictionaries list *schwimmen* exclusively. Only *Das deutsche Wörterbuch* by Grimm/Grimm (1942; vol. 20) describes structural change in the

[6] A noun is considered infrequent if it appears in only one or two phrasemes within a collection.

[7] This change from *wider* to *gegen* is also observable in the free use of the preposition (Grimm/Grimm 1854-1960). Apart from some exceptions like *für und wider* and fixed combinations like *wider Willen*, *wider* today is characterized as archaic (Grimm/Grimm 1854-1960). Therefore it is even more noteworthy that both forms are listed in Röhrich (2002) and Dudenredaktion (2008).

phraseme. However, the Grimms only document the occurrence of other verbs than *schwimmen* for the 16th and 17th centuries, whilst Wander (1867-1880) gives examples up to the 19th century.

Some of the dictionaries fail to provide any semantic information. Others, as we see in Table 1, indicate a semantic change in the 19th century. Adelung (1793-1801) and Campe (1807-1812) explicitly mention the futility of resistance, whereas Sanders (1859-1865) makes no mention of the notion of resistance at all. It appears again in Wander (1867-1880), with Borchardt (1888) the last to mention it: later collections do not include this element in their explanation. By the 20th century, a weakening to "non-conformist individualism" seems to have become dominant.

1612	Herberger	Wer **wider** den Strom **schwimmt**, muss ersaufen. (He who swims against the current must drown.)
1616	Henisch	Es ist böss **schwimmen gegen** den strom. (It is hard/evil (to be) swimming against the current.)
1793-1801	Adelung	**wider** den Strom **schwimmen** (swim against the current) "wanting to put up resistance (against) overwhelming obstacles"
1807-1812	Campe	**gegen/wider** den Strom **schwimmen** (swim against the current.) "to object, put up resistance where it is futile (to do so)"
1837	Körte	**Wider** den Strom ist schwer zu **schwimmen.** (against the current it is difficult to swim)
1859-1865	Sanders	**gegen (wider)** den Strom **schwimmen** (swim against the current) "to oppose ... the prevalent direction"
1867-1880	Wander	**gegen/wider** den Strom **schwimmen/fahren/gehen/streben** (swim/drive/go/strive against the current) "to prevail against all the odds (or: to overcome all obstacles)"
1888	Borchardt	**gegen** den Strom **schwimmen** (swim against the current) "to defy (sthg) in vain"
1942	Grimm	**wider (gegen)** den Strom **schwimmen** (swim against the current.) "against development, life, general opinion"
1976	Friedrich	**gegen** den Strom **schwimmen** (swim against the current.) "to not conform with generally prevailing opinion and tendencies"
2002	Röhrich	gegen den Strom schwimmen (swim against the current) "to deliberately behave unlike the majority, and to hazard any negative consequences"
2008	Dudenredaktion	**gegen/wider** den Strom **schwimmen** (swim against the current.) "to oppose majority opinion or conventions"

Table 1: Development of phraseme form and meaning since the early 17th century

By consulting various dictionaries, we can adduce possible diachronic semantic changes in the phraseme, *gegen den Strom schwimmen,* but the lack of documentation means that it is not possible to establish a precise chronology for these changes.

3.2 Corpora

The next stage proposed is the search for evidence of the selected phrasemes in texts from 1650 onwards. It is, however, important to remember that any evidence for phrasemes must be interpreted carefully: the absence of a certain phraseme in texts of a given period does not mean that the phraseme was not in use.

In recent years, various diachronic corpora for German have been created and become available, such as the *Deutsches Textarchiv* (*DTA*),[8] with currently 532 full texts from 1650 to 1900; *GerManC*,[9] with 2,000-word samples of texts from 1650 to 1800, and the *Digitale Bibliothek*[10] (*DB*) with roughly 2,700 texts and about 87 million running word forms. Additionally, there are several corpora of 20th century texts, including the *Schweizer Textkorpus*[11] and the *Referenzkorpus der deutschen Sprache des 20. Jahrhunderts.*[12] We will also be able to explore special-purpose corpora, namely *Berg+Text digital*,[13] consisting of the digitized year books of the Swiss Alpine Club since 1864, with 36 million running word forms, and also a subset of the *Collection of Swiss Law Sources*, containing about 4 million running word forms,[14] although these two corpora contain only a relatively small number of written texts.

Fully automatic detection of phrasemes is not possible as yet, which is why lexicographers have to determine idiomaticity manually (Rothkegel 2007). Approaches for finding collocations or phrasemes apply Natural Language Processing (NLP) methods on *well-formed modern* texts (see Fritzinger et al. 2009, and Seretan/Wehrli 2010). However, the orthography in German texts from 1650 onwards is variable, and we have also found variation in inflection, word order, and vocabulary, making vector-based approaches from the field of Information Retrieval (IR) (Salton et al. 1975) more appropriate.

[8] http://www.deutschestextarchiv.de, available under a Creative Commons License.

[9] http://llc.manchester.ac.uk/research/projects/germanc

[10] http://www.textgrid.de/digitale-bibliothek.html, available under a Creative Commons License.

[11] http://www.dwds.ch

[12] http://www.dwds.de/resource/kerncorpus

[13] http://www.textberg.ch

[14] We hope that the inclusion of Swiss German corpora provides some insights into differences between federal and Swiss German phraseology, however, we will not give a systematic overview.

Example: *to swim against the current*

Returning to our example, empirical information can be obtained by consulting the texts in the *Digitale Bibliothek* (*DB*). This corpus contains 96 occurrences of our polylexical unit (including variants), 40 of which show idiomatic usage.

Their distribution across 50-year periods is as follows: 2 prior to 1700, 2 in 1701-1750, 10 in 1751-1800, 7 in 1801-1850, 18 in 1851-1900, and 1 after 1900. This clearly does not represent an even distribution of hits across the study period. Figure 1 shows a comparison of frequencies of occurrence of the prepositions *wider* and *gegen* (bottom), of the verb *schwimmen* vs. other verbs (middle), and of the meaning "oppose majority opinion or conventions" vs. other meanings (top).

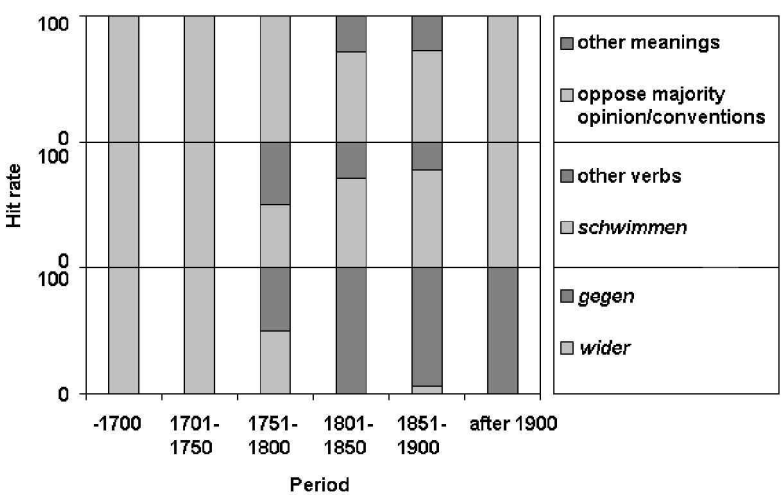


Figure 1: Percentages of occurrences of the prepositions *gegen* and *wider*, of the verb *schwimmen* vs. other verbs, and of the meaning "to oppose majority opinion or conventions" vs. other meanings, in the phraseme *gegen den Strom schwimmen* in *DB*.

According to Figure 1, the transition from the preposition *wider* to *gegen* in our phraseme quite clearly occurs in the second half of the 18th century. However, dictionaries only identify it from the middle of the 19th century, which means that they reflect the change with a certain time lag. However, our findings are in line with Grimm/Grimm (1942), according to whom this transition occurred in the late 18th or early 19th century.

The *DB* documents the verb *schwimmen* in this phraseme during the entire period; it predominates over any other verbs, such as *angehen* ('approach', 'tackle'), *arbeiten* ('work'), *gehen* ('go'/'walk'), *leben* ('live'), *meinen* ('opine'/'think'), *ringen* ('struggle'), *steuern* ('steer'), *wollen* ('want'), which between 1751 and 1900 constitute a proportion of 25-50% of instances, probably indicating that the phraseme is just emerging. In contrast, according to Grimm/Grimm (1942), other verbs only occur in this phraseme in the 16th and 17th centuries.

The earliest documented occurrence of the phraseme dates from 1690, and the most recent one from 1911. As shown in Figure 1, we find the meaning which is still usual today throughout the entire period. It is only between 1801 and 1900 that other meanings can be deduced, i.e. "to attempt the impossible" (2 occurrences), and "to oppose the 'current' of feelings, of custom, of normal procedure" (3 occurrences). Similar to the dictionary data, the hits in this corpus indicate that the 19th century, from which the largest number of documents date, was an important period in the development of this phraseme. However, its weakened meaning would seem to have been quite common even before 1700. These initial results have to be interpreted very carefully, and have to be verified using other corpora as described above.

4. Conclusion

In this article we have described the approach employed in the *OLdPhras* project, demonstrating the advantages of a combination of various methodologies in obtaining information on change in structure and meaning of German phrasemes. While data on the verbal phraseme *gegen den Strom schwimmen* are preliminary, the use of both dictionary and corpus analyses produces a significantly better result than observations of the phraseme at two different moments in time. The use of large corpora promises to yield numerous new examples of actual use, and will therefore significantly improve the description and dating of the development of this kind of phraseme. We hope that the inclusion of systematic dictionary analyses will give a fresh impulse to the analysis of historical corpora.

Acknowledgements

We are grateful to our colleagues Annelies Häcki-Buhofer, Marcel Dräger and Sixta Quaßdorf for collaboration and fruitful discussions. We would also like to thank two anonymous reviewers for helpful comments on an earlier version of this article.

References

Adelung, Johann Christoph (1793-1801): Grammatisch-kritisches Wörterbuch der Hochdeutschen Mundart. Leipzig: Breitkopf & Sohn.

Borchardt, Wilhelm (1888): Die Sprichwörtlichen Redensarten im deutschen Volksmund nach Sinn und Ursprung erläutert. Leipzig: Brockhaus.

Burger, Harald (2010): Phraseologie. Berlin: Erich Schmidt.

Burger, Harald (2007): Semantic aspects of phrasemes. In: Burger/Dobrovol'skij/Kühn/Norrick (eds.), 90-109.

Burger, Harald/Dobrovol'skij, Dmitrij/Kühn, Peter/Norrick, Neal R. (eds.) (2007): Phraseologie. Ein internationales Handbuch der zeitgenössischen Forschung. Berlin/New York: de Gruyter.

Burger, Harald/Linke, Angelika (1998): Historische Phraseologie. In: Besch, Werner/Betten, Anne/Reichmann, Oskar/Sonderegger, Stefan (eds.): Sprachgeschichte: Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung. Berlin/New York: de Gruyter, 743-755.

Campe, Joachim Heinrich (1807-1812): Wörterbuch der deutschen Sprache. Braunschweig: Schulbuchverlag.

Claridge, Claudia (2008): Historical corpora. In: Lüdeling, Anke/Kytö, Merja (eds.): Corpus Linguistics. Vol. 1. Berlin/New York: de Gruyter, 242-259.

Colson, Jean-Pierre (2007): The World Wide Web as a corpus for set phrases. In: Burger/Dobrovol'skij/Kühn/Norrick (eds.), 1071-1077.

Deutsche Literatur von Luther bis Tucholsky (2005): Digitale Bibliothek 125. Berlin: Directmedia.

Dräger, Marcel (2009): Auf der Suche nach historischen Phrasemen – oder: Wörterbücher als Korpora. In: Linguistik online 39: 33-43.

Dräger, Marcel (2010): Phraseologische Nachschlagewerke im Fokus. 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: Gunter Narr, 411-421.

Dräger, Marcel/Juska-Bacher, Britta (2009): Online-Datenerhebungen im Dienste der Phraseographie. In: Ptashnyk, Stefaniya/Hallsteinsdóttir, Erla/Bubenhofer, Noah (eds.): Computergestützte und korpusbasierte Methoden in der Phraseologie, Phraseografie und der Lexikografie. Baltmannsweiler: Schneider Verlag Hohengehren, 162-175.

Dudenredaktion (2008): Redewendungen. Mannheim: Dudenverlag.

Friedrich, Wolf (1976): Moderne deutsche Idiomatik. Systematisches Wörterbuch mit Definitionen und Beispielen. München: Huber.

Fritzinger, Fabienne/Kisselew, Max/Heid, Ulrich/Madsack, Andreas/Schmid, Helmut (2009): Werkzeuge zur Extraktion von signifikanten Wortpaaren als Webservice. In: Hoeppner, Wolfgang (ed.): GSCL-Symposium Sprachtechnologie und eHumanities, 32-43.

Grimm, Jacob/Grimm, Wilhelm (1854-1960): Das deutsche Wörterbuch. Leizig: Hirzel.

Henisch, Georg (1616): Teütsch Sprach und Weissheit: thesaurus linguae et sapientiae Germanicae. Augsburg.

Herberger, Valerius (1612): Hertz Postilla. Leipzig.

Juska-Bacher, Britta (2009): Empirisch-kontrastive Phraseologie. Am Beispiel der Bekanntheit von Bruegels Niederländischen „Sprichwörtern" im Niederländischen, Deutschen und Schwedischen. Baltmannsweiler: Schneider Verlag Hohengehren.

Körte, Wilhelm (1837): Die Sprichwörter und sprichwörtlichen Redensarten der Deutschen. Leipzig: Brockhaus.

Mellado Blanco, Carmen (2009): Einführung. Idiomatische Wörterbücher und Metafraseografie: zwei Realitäten, eine Herausforderung. In: Mellado Blanco, Carmen (ed.): Theorie und Praxis der idiomatischen Wörterbücher. Tübingen: Niemeyer, 1-20.

Moon, Rosamund (2007): Phraseology in general monolingual dictionaries. In: Burger/Dobrovol'skij/Kühn/Norrick (eds.), 909-918.

Parad, Jouko (2003): Biblische Verbphraseme und ihr Verhältnis zum Urtext und zur Lutherbibel. Frankfurt a.M.: Peter Lang.

Piirainen, Elisabeth (2006): Phraseologie in arealen Bezügen: ein Problemaufriss. In: Hallsteinsdóttir, Erla/Farø, Ken (eds.): Neue theoretische und methodische Ansätze in der Phraseologieforschung. New theoretical and methodological approaches to phraseology. (= Linguistik online 27), 195-218. http://www.linguistik-online.de/27_06/piirainen.html.

Röhrich, Lutz (2002): Das große Lexikon der sprichwörtlichen Redensarten. Darmstadt: Wissenschaftliche Buchgesellschaft.

Rothkegel, Annely (2007): Computerlinguistische Aspekte der Phraseme I. In: Burger/Dobrovol'skij/Kühn/Norrick (eds.), 1027-1035.

Salton, Gerard/Wong, Andrew/Yang, Chung-Shu (1975): A vector space model for automatic indexing. In: Communications of the ACM, 18(11): 613-620.

Sanders, Daniel (1859-1865): Wörterbuch der deutschen Sprache. Leipzig: Wigand.

Seretan, Violetta/Wehrli, Eric (2010): Tools for syntactic concordancing. In: Proceedings of the International Multiconference on Computer Science and Information Technology, IEEE, 493-500.

Wander, Karl Friedrich Wilhelm (1867-1880): Deutsches Sprichwörter-Lexikon. Leipzig: Brockhaus.

JOANNA KOPACZYK

Formulaicity in Scots historical corpora and the lexical bundles method[1]

Abstract

This paper draws attention to the newly available corpus resources for the study of Older Scots, and to the application of the lexical bundles method (Biber et al. 1999) in historical specialized discourse. The discussion concentrates on the method adopted from present-day corpus research, which illuminates historical questions which have so far proved unanswerable, e.g. which multi-word elements in text are stable and repetitive. I applied lexical bundles to legal and administrative texts written in Scots, to observe the degree of formulaicity in early specialized discourse. The results of the study show that the Scottish documents contain highly formulaic long lexical bundles (8-grams and 7-grams) when juxtaposed with other specialized discourse texts, such as the Bible. Similarly, shorter bundles helped to identify the impressive degree of formulaicity in comparison to speech-based legal genres, such as trials and depositions (Culpeper/Kytö 2010).

1. An overview of Scots historical corpora

In recent years, corpus linguistics has been moving towards greater inclusiveness of languages other than English. In historical and diachronic studies based on corpora, this trend has also found its echoes. The other Germanic language of the British Isles, Scots, has attracted the attention of corpus compilers as well. Genetically, Scots is a "sister" language to English, stemming from the same Germanic roots, although from a different dialectal background (Old Northumbrian), and developing in different geo-political circumstances (McClure 1994). Until the mid-sixteenth century, Scots clearly enjoyed the status of a separate, national tongue in Scotland, with unrestricted use in major functional domains, from personal correspondence and literary texts, to official royal letters and parliamentary acts. It has been postulated that Scots was on the way towards language standardization (Agutter 1988, Bugaj 2004), which could have been completed, but for the external influence of English in the changing political and social conditions of the Early Modern period.

[1] This research project is supported by the Polish Ministry of Science and Higher Education individual grant no. N N104 014337 (2009-2012).

Nevertheless, the rich textual resources found in Scots both before and after the Union have inspired several corpus projects, summarized by Meurman-Solin (2007). Apart from the multi-genre *Helsinki Corpus of Older Scots*, 1450-1700 (Meurman-Solin (ed.) 1993), the major digital resources for Scots stem from two traditional approaches to this language, namely lexicographic (cf. the Scottish counterpart of the *OED*, the *Dictionary of the Scots Language*, *DSL*), and dialectal. The latter approach has been advanced particularly by the compilation of the *Edinburgh Corpus of Older Scots*, *ECOS* 1380-1500, which serves as the database for the *Linguistic Atlas of Older Scots* (*LAOS*, Williamson (ed.) 2008). These resources have lately been complemented by specialized discourse projects, such as the *Corpus of Scottish Correspondence*, 1542-1708 (Meurman-Solin (ed.) 2003) and the *Corpus of Nineteenth-century Scottish Correspondence* (Dossena (ed.) in prep.). In fact, Williamson's corpus is also an example of a specialized discourse corpus, because it comprises legal and administrative texts from all over Scotland. The *ECOS* corpus is thus uniform in terms of discourse characteristics, and may serve as the database for the enquiry into legal discourse in medieval and early Renaissance Scotland.

2. Specialized corpora of Older Scots: legal and administrative discourse

The *ECOS* corpus is compiled from legal and administrative texts and amounts to some 390,000 words, making it the largest electronic collection of specialized historical Scots available in a single corpus today. To achieve wider chronological coverage, and to complement the materials comprising *ECOS*, relevant samples from the *Helsinki Corpus* (*HCOS*) were added, with care to avoid textual overlaps. The two resources were further conflated with a collection of burgh records from Wigtown in Galloway, in the south-west of Scotland, to achieve a wider geographical coverage. These three resources, combined under the acronym *EdHeW*, guarantee the most comprehensive representation of the earliest instances of public discourse in the Scottish vernacular (1380-1560), amounting altogether to about 580,000 words.

The *EdHeW* corpus has served as the basis for the study of formulaic patterns in early legal and administrative discourse. The development and stabilization of textual patterns, or, in other words, textual standardization, can be traced by extracting repetitive and stable elements from a given collection of texts. In the next section I outline the relevant corpus analysis method (lexical bundles),

recently adapted in historical linguistic research, and discuss its application to the specialized corpus of early Scots legal discourse.

3. Introducing lexical bundles

3.1 Major characteristics

Lexical bundles, also known as *n*-grams, repetitive word strings, word chains (Stubbs/Barth 2003) or word clusters (Scott 1997), are a relatively new tool in corpus linguistics. They can be defined as the "most frequently recurring sequences of words in a register [...] identified using a frequency-driven approach" (Biber 2009: 282). In essence, the method is not geared to answering any specific research question, but rather extracts automatically the most frequent lexical strings, for further inductive interpretations (a corpus-driven approach, Tognini-Bonelli 2001: 84-87). The study by Altenberg (1998) of recurrent patterns in spoken discourse is an early example of how automatic extraction of lexical strings of a given length *n*, or n-grams, can illuminate our understanding of linguistic fixedness. The first major large-scale application of this corpus-driven, frequency-based method for identifying repetitive unchanging elements of discourse was undertaken by Biber et al. (1999) in the *Longman Grammar of Spoken and Written English*. Since then, the lexical bundles method has been propagated by Biber in his work on academic English, and on differences between spoken and written modes of communication (Biber 1997, Biber 2004, Biber/Conrad/Cortes 2003, 2004, Biber/Barbieri 2007; for an overview of other PDE research questions addressed with the lexical bundles methodology, see Kopaczyk 2012).

3.2 Challenges for historical linguistics

In historical linguistics, the only book-length publication so far, where one of the chapters focuses on lexical bundles, is Culpeper/Kytö (2010). The authors extracted lexical bundles from the *Corpus of Early English Dialogues* to explore repetitive strings in historical dialogues, mostly from a functional perspective, looking for orality features in comparison with present-day English material. Lexical bundles have also been used to investigate repetitive patterns in early specialized discourse, along the same lines as in present-day applied research, which involves interpreting the patterns from a structural and functional perspective (Kopaczyk 2013). The degree of formulaicity emerging from this study is discussed in further sections of the present paper.

There are several problems arising in connection with the application of lexical bundles in historical research. The major issues include spelling normalization, the reliability of extraction software, corpus size and cut-off points, as well as digitizing conventions (for further discussion see Ari 2006, Culpeper/ Kytö 2010, Kopaczyk 2012).

Table 1 provides a summary of the methodologies of bundle extraction in selected studies conducted on present-day and historical corpora. The cut-off points mark the frequency above which a given string of lexical items starts counting as a lexical bundle. The decision as to how to set this parameter is largely arbitrary, but will have a crucial influence on bundle counts and formulaicity measurements in each of the studies. I return to this point below.

Author(s) and date	Type of texts in the corpus	(sub)Corpus length (words)	Bundle length	No. of instances at cut-off point
Altenberg (1998)	PDE: spoken	0.5 mln	1- to 8-grams	10
Biber et al. (1999)	PDE: conversation, fiction, press, academic discourse	40 mln	3- to 6-grams	10 / 5 per 1 mln in ≥ 5 texts
Biber et al. (2003)	PDE: conversation, academic discourse	7 mln + 5.3 mln	4- to 5-grams	20 per 1 mln
Biber et al. (2004)	T2K-SWAL spoken and written academic language	2 mln	4-grams	40 per 1 mln
Biber/ Barbieri (2007)	PDE: T2K-SWAL + LSWE subsection: spoken and written academic language	c. 7.9 mln	4-grams	40 per 1 mln
Culpeper/ Kytö (2010)	EModE plays EModE trials	0.22 mln 0.25 mln	3-grams	10
Kopaczyk (2013)	Middle Scots legal texts	0.6 mln	3- to 8-grams	≥10 in ≥10 texts

Table 1: Cut-off points and corpus word counts in selected studies (based on Kopaczyk 2013: 153)

4. Formulaicity in early legal discourse

4.1 Lexical bundles and textual standardization

Formulaicity on the level of text is visible in the recurrent use of identical lexico-syntactic patterns driven by a specific discourse situation. In order to explore this phenomenon in early public discourse in Scotland, a *non-a priori* method should be employed; after all, there is no objective way to establish at the outset which actual constructions would be most fixed and most frequent. This is why the lexical bundles method proves to be the best tool for this particular research question.

Stubbs/Barth (2003) observe that recurrent word chains, or lexical bundles, constitute a "predictable characteristic of different text types" (2003: 62) and that "longer chains discriminate between text types" (2003: 76). In more formulaic types of discourse, longer bundles will indeed be prominent, for example in liturgical language (sermons and the Bible), political speeches and "some kinds of legal texts" (Stubbs/Barth 2003: 78). In legal texts, clarity and completeness are given precedence over stylistic variety, and over a conscious avoidance of structural monotony, which is why their language is even more formulaic than elsewhere.

4.2 Lexical bundles in the EdHeW corpus

4.2.1 Bundle counts

In order to investigate the formulaic patterns in early Scots legal discourse, 3- to 8-grams were extracted from the *EdHeW* corpus (ca. 580,000 words, 1,818 text files; see Table 2 for bundle counts).

	> 5 tokens + > 5 files		≥ 10 tokens + ≥ 10 files		final top 10%	
	all types	all tokens	all types	all tokens	all types	all tokens
3-grams	7,269	166,401	3,535	135,243	354	54,307
4-grams	4,145	80,317	1,913	61,847	191	22,934
5-grams	2,552	43,817	1,142	32,449	114	11,133
6-grams	1,682	26,549	722	18,921	72	6,229
7-grams	1,205	17,388	495	11,837	50	3,841
8-grams	857	11,222	321	7,098	32	2,051

Table 2: Lexical bundle counts in *EdHeW*, cf. Kopaczyk (2013: 154)

Two cut-off points were tried out to establish the lexical bundle threshold: the first one selected the strings which repeated more than 5 times in more than 5 texts (or files) – the 5-5 threshold, and the second one was set at 10 instances in 10 or more texts – the 10-10 threshold. It is clear that the number of types was reduced by more than half with the second threshold (from a 52% drop in 3-grams to 63% in 8-grams), however with the same threshold the number of tokens dropped only by 27% on average (from 19% in 3-grams to 37% in 8-grams). What this change in data counts implies is that the first threshold was too low to extract the most repetitive constructions from the *EdHeW* corpus, and it was only due to raising the cut-off point that the number of available types was restricted and refined, while the number of tokens still pointed to a large membership in a given type. The material was still too abundant to be analyzed qualitatively above the 10-10 threshold, which is why for a detailed structural and functional analysis the top 10% of the types were chosen. That part of the investigation, however, falls beyond the scope of the present paper.

4.2.2 Long bundles as tokens of formulaicity

When it comes to the degree of formulaicity in the *EdHeW* corpus, the first striking feature is the sheer abundance of repetitive lexical strings of all lengths. Even among the 8-grams there are over 300 types of lexical arrangements which repeat in the corpus in an unchanged form more than ten times in ten texts. The most frequent 8-grams account for over 2,000 instances where eight words were arranged in exactly the same order in the texts. Bundles in (1)-(16) constitute the top 10% of the most formulaic types of 8-grams. Out of the thirty-two most formulaic 8-grams, the following sixteen overlapping strings can be built (overlaps are marked with slashes):

(1) *ye / yhere / of / oure lord a thousande four / hundereth / thirty / and*
(2) *ye yhere of god a thousande four hundereth*
(3) *ye / yhere of god a thousande v hundereth / and*
(4) *day / of ye monath of juli ye yhere / of*
(5) *day / of ye monath of may ye yhere / of*
(6) *day / of ye monath of februar ye yhere / of*
(7) *of ye monath of nouember ye yhere of*
(8) *chalans / for ye wrangus haldin fra him of / ane*

(9) *chalans / for ye wrangus haldin fra hir of*

(10) *curt of ye newburch haldin in ye chapel*

(11) *haldin in ye tolbuth of ye samyn be*

(12) *james throu ye mercy of god prior of*

(13) *in witness of ye quhilk thing to thir*

(14) *be / it / kend / til all men be thir / present / letteris / me*

(15) *ye / quhilk / day / ye sutis callit ye curt / effermit / and / absentis / demyt / in*

(16) *balze / in / that tyme and than incontenent ye / said / balze*

The bundles defy rigorous structural categorization as they often span more than one phrase, and often contain fragments of phrases on both ends, e.g. *it kend til all men be this present* in (14). The phrasal incompleteness indicates that there was a fixed core of a given formulaic expression which was employed in varied co-texts (the grammatical and lexical elements which precede and follow the formulaic string). What can be said, however, on the basis of the thirty-two types of 8-grams in the top formulaic range, is that nominal phrases and prepositional phrases recur most. This fact is linked to the informational and referential character of legal, and especially administrative, texts, where the date, the place and the participants have to be specified in the utmost detail, and each entry will inevitably contain the same information. The resulting information package tends to be served in exactly the same wording, which contributes to textual standardization. The formulaic strings stabilize the structure of the text, and simultaneously ground the documents in time, and in the external social context. This is visible, for instance, in the fixed form of the date (1)-(7); note, however, the two competing formulas: 'the year of our lord' (1) and 'the year of god' (2)-(3). Other examples in (8)-(12) refer to the relevant legal charges and their venue, as well as to the authorities involved.

The second prominent structural pattern involves a finite verb or a past participle with some complementation, either by a noun phrase or prepositional phrase, e.g. (11) or (14). This structure typically corresponds to the interpersonal function, and relies on directive and representative speech act verbs such as *be kend* ('to be known' in the sense of "to announce") (14), *call* ('to call') and *affirm* ('confirm') (15). The bundle in (13) may be treated as an indirect commissive speech act, because it refers to committing oneself to witness a legal act. It turns out that the main performative functions associated with legal

discourse emerge in the long formulaic bundles, even though the texts in the corpus do not belong to speech-based genres (cf. Culpeper/Kytö 2010). It is the referential function, however, that remains most prominent in *EdHeW*. Finally, in the last formulaic string the central element, the adverbial *than incontinent* ('then immediately after') acts as a cohesive textual device, providing a narrative link to preceding discourse.

Stubbs/Barth (2003) show that long lexical bundles discriminate between text types. They mention liturgical language as an example of a highly formulaic register, and give examples of repetitive 7-grams: *and the Lord spake unto Moses saying*; *the word of the Lord came unto*; *shall know that I am the Lord*; *and it came to pass in the*; etc. (Stubbs/Barth 2003: 77). There were around 65 7-grams which occurred 20 times or more in the *Authorised Version of the Bible* (1604-1611, ca. 850,000 words), and around 260 7-grams when the threshold was set at 10 instances. The *EdHeW* corpus is shorter (ca. 580,000 words), so a comparable number of the types of 7-grams indicating a formulaic nature of the textual material would have to be around 177, applying the same 10 instances threshold. The bundle counts in Table 1 show that there are, in fact, 495 types of 7-grams above this threshold in the corpus, so it can be asserted with a large degree of confidence that legal and administrative texts in medieval Scotland are more formulaic than the Bible, judging by the frequency of recurrent stable lexical patterns.

4.2.3 Formulaicity indicated by shorter bundles

As yet, there are no studies of long lexical bundles to provide a point of comparative reference for the formulaic strings presented above, or their counts. However, if one compares the *EdHeW* results with the studies of 3- and 4-grams (cf. Table 1), the comparison strongly confirms the structural stability and formulaic nature of early Scots legal discourse. In the grammar by Biber et al. (1999), the adopted threshold for lexical bundles was 10 occurrences per million words in at least 5 texts. This roughly corresponds to the 5-5 cut-off point in *EdHeW*, which comprises 580,000 words. On the basis of their lexical bundle extraction, Biber et al. (1999: 994) found that "3-word bundles occur over 80,000 times per million words in conversation, over 60,000 times per million in academic prose". It means that present-day English conversation is more formulaic, and uses fixed chunks of discourse more frequently than academic prose. Set against these results, medieval legal texts from Scotland come across as extremely formulaic. There are over 160,000 3-grams in *EdHeW*, so the

count per million would be over 270,000 tokens, which is three times as many as the results for present-day conversation.

When confronted with formulaicity measurements provided by Culpeper/Kytö (2010) in their analysis of speech-based early modern genres, the *EdHeW* data look extremely formulaic again. The notion of formulaicity was defined by the authors as follows: "[T]he greater the proportion of the words of a text-type belonging to fewer different lexical bundles the more formulaic that text-type is" (Culpeper/Kytö 2010: 135-136). To calculate the degree of formulaicity, the authors used the following formula: the sum of instances of the top ten bundles, times 3 (because their bundles were 3-grams), divided by the corpus word-count, times 10,000 (Culpeper/Kytö 2010: 137). I applied the same calculations to the *EdHeW* 3-grams, treating the most frequent 3-grams as the "top" ones, just as Culpeper and Kytö did.[2] Table 3 presents the proportions of words used to make up the most frequent 3-grams in speech-based Early Modern English genres, and in medieval and Early Modern Scots legal and administrative texts.

Corpus type	wd-count in 3-grams / 10,000 wds
EModE plays	78.3
EModE fiction	83.0
EModE didactic	95.3
EModE depositions	123.4
EModE trials	138.6
EdHeW	540.5

Table 3: Proportion of corpus word-count used for top 3-grams (per 10,000 words, based on Culpeper/Kytö 2010: 136 and Kopaczyk 2013: 262)

The formulaic nature of *EdHeW* material is striking if compared to other genres, even those belonging to legal discourse, such as depositions and trials. The texts in the *EdHeW* corpus, however, are not speech-based, but rather belong to the formulaic tradition of record-keeping, documentation and promulgation of law. The formulaicity measurement is almost four times higher in *EdHeW* than in early modern trials, and almost seven times higher than in early

[2] Kopaczyk (2013) established formulaicity rankings on the basis of three factors: the token frequency of a bundle, the number of texts in which it appeared, and the relative weight of these two, where frequency was more important than the number of texts.

modern plays. The different character of the records emerges even at first sight, when comparing the frequency counts of the most frequent 3-grams, as well as from their function in respective corpora. In EModE trials, the most frequent 3-gram is *do you know*, an interactive bundle, with normalized frequency of 7.2 per 10,000 words (Culpeper/Kytö 2010: 116). In *EdHeW*, the most frequent 3-gram is *of the said*, a prepositional phrase fragment with a cohesive function, referring to previous discourse, with normalized frequency of 37.5 per 10,000 words (Kopaczyk 2013). In view of these counts and the data in Table 3, the degree of formulaicity in legal records in medieval and early Renaissance Scotland seems truly impressive.

5. Concluding remarks

Automatic extraction of the most frequent, stable chunks of texts allows the investigation of formulaic patterns in texts. The collection of legal documents from medieval and early modern Scotland, compiled on the basis of three available electronic corpora, contains more lexical bundles than any other corpus subjected to lexical bundle extraction and reviewed in the present paper. This observation goes hand in hand with the expected qualities of legal texts, namely a high degree of fixedness and repetition (see, for example, Danet 1980).

I applied the lexical bundles method to a Scots corpus not only because its main part – the *ECOS* – consisted of similar text types, and could be treated as a uniform specialised corpus. I also wanted to draw attention to historical corpora of languages other than English, as there is every reason to reach beyond the most common sets of data. Scots is a language which has an intricate historical relationship with English, while the two legal systems have grown out of different roots, and were subject to different influences and pressures (Walker 2001). It would be interesting to see how the southern neighbour of the day compares in terms of formulaicity in the same text types.

References

Primary sources

Williamson, Keith (ed.) (2008): The Edinburgh Corpus of Older Scots. Edinburgh: University of Edinburgh.

Meurman-Solin, Anneli (ed.) (1993): The Helsinki Corpus of Older Scots. Helsinki: University of Helsinki.

Wigton Burgh Court Book (1512-1534): Unpublished transcript by Alfred Truckell. Unpublished digitized version by Joanna Kopaczyk. Dumfries Archive Centre.

Secondary sources

Agutter, Alex (1990): Restandardisation in Middle Scots. In: Adamson, Sylvia/Law, Vivien/Wright, Susan (eds.): Papers from the 5th International Conference on English Historical Linguistics. Amsterdam: John Benjamins, 1-11.

Altenberg, Bengt (1998): On the phraseology of spoken English: the evidence from recurrent word-combinations. In: Cowie, Anthony Paul (ed.): Phraseology. Oxford: Clarendon Press, 101-122.

Ari, Omer (2006): Review of three software programs designed to identify lexical bundles. In: Language Learning and Technology 10/1: 30-37.

Biber, Douglas (1997): Lexical bundles in spoken and written discourse: what the grammar books don't tell you. In: Gerome, Sally B. (ed.): An update on grammar: how it is learnt – how it is taught (1996 Colloquium proceedings). Paris: TESOL France, 4-8.

Biber, Douglas (2004): Lexical bundles in academic speech and writing. In: Lewandowska-Tomaszczyk, Barbara (ed.): Practical Applications in Language Corpora (PALC 2003). Frankfurt a.M. etc.: Peter Lang, 165-178.

Biber, Douglas (2009): A corpus-driven approach to formulaic language in English: multi-word patterns in speech and writing. In: International Journal of Corpus Linguistics 14/3: 275-311.

Biber, Douglas/Barbieri, Federica (2007): Lexical bundles in university spoken and written registers. In: English for Specific Purposes 26: 263-286.

Biber, Douglas/Conrad, Susan/Cortes, Viviana (2003): Lexical bundles in speech and writing: an initial taxonomy. In: Wilson, Andrew/Rayson, Paul/McEnery, Tony (eds.): Corpus linguistics by the Lune. A Festschrift for Geoffrey Leech. Frankfurt a.M. etc.: Peter Lang, 71-92.

Biber, Douglas/Conrad, Susan/Cortes, Viviana (2004): If you look at... Lexical bundles in university teaching and textbooks. In: Applied Linguistics 25/3: 371-405.

Biber, Douglas et al. (1999): Longman Grammar of Spoken and Written English. London: Longman.

Bugaj [Kopaczyk], Joanna (2004): Middle Scots as an emerging standard and why it did not make it. In: Scottish Language 23: 19-34.

Culpeper, Jonathan/Kytö, Merja (2010): Early Modern English dialogues: spoken interaction as writing. Cambridge: Cambridge University Press.

Danet, Brenda (1980): Language in the legal process. In: Law & Society Review 14/3: 445-564.

Dictionary of the Scots Language [DSL] (2005). Edinburgh: Scottish Language Dictionaries. http://www.dsl.ac.uk/dsl/index.html.

Dossena, Marina (in prep.): Corpus of Nineteenth-century Scottish Correspondence, University of Bergamo.

Kopaczyk, Joanna (2012): Applications of the lexical bundles method in historical corpus research. In: Pęzik, Piotr (ed.): Corpus data across languages and disciplines. Frankfurt a.M.: Peter Lang, 83-95.

Kopaczyk, Joanna (2013): The legal language of Scottish burghs. Standardization and lexical bundles (1380-1560). Oxford: Oxford University Press.

McClure, J. Derrick (1994): English in Scotland. In: Burchfield, Robert (ed.): The Cambridge history of the English language. Vol. 5: English in Britain and overseas: origins and development. Cambridge: Cambridge University Press, 23-93.

Meurman-Solin, Anneli (2007): Manual to the Corpus of Scottish Correspondence. Helsinki: VARIENG. http://www.helsinki.fi/varieng /csc/manual/part1/1_2.html (last access: October 2011).

Scott, Michael (1997): WordSmith Tools Manual. Oxford: Oxford University Press.

Stubbs, Michael/Barth, Isabel (2003): Using recurrent phrases as text-type discriminators: a quantitative method and some findings. In: Functions of Language 10/1: 61-104.

Tognini-Bonelli, Elena (2001): Corpus linguistics at work. Amsterdam: John Benjamins.

Walker, David M. (2001): The Scottish legal system: an introduction to the study of Scots law. 8th edition. Edinburgh: W. Green/Sweet & Maxwell.

Williamson, Keith (ed.) (2008): A linguistic atlas of Older Scots [LAOS]. http://www.lel.ed.ac.uk/ihd/laos1/laos1.html (last access: December 2011).

MELANIE RÖTHLISBERGER / GEROLD SCHNEIDER

Of-genitive versus *s*-genitive

A corpus-based analysis of possessive constructions in 20th-century English

Abstract

This paper examines genitive variation in English, using two methodological approaches. In the manual approach, we extract genitive variants from the parsed subcorpora of the text category J (academic writing) in the *B-Brown* (1931), the *Brown* (1961) and the *Frown* (1991/2) corpora. Focussing on the syntactic parameter, we illustrate how the principle of end-weight gains ground from 1930 to 1990. The automatic approach implements the constraints of the manual approach, confirms the findings of the manual approach and is used to scale to British English. Methodologically, we show how to automatically sift out irrelevant corpus examples whose identification would normally need human intervention – in particular, apparent examples of the two main genitive English constructions which are not in genuine alternation.

1. Introduction

The increase of *s*-genitives (e.g. *my father's house*) at the expense of *of*-genitives (e.g. *the house of my father*) in modern English is a phenomenon that has received increasing attention. Corpus-based works by Altenberg (1982), Jucker (1993), Rosenbach (2002), and Szmrecsany/Hinrichs (2007) have focussed on the interchangeability between the *s*-genitive and the *of*-genitive in areas where both constructions can be chosen. The choice is constrained by a number of conditioning factors: language internal (i.e. syntactic, lexical, phonological, semantic) as well as external (i.e. factors related to processing, economy-related factors, and socio-stylistic factors). This study contributes to previous work by offering a quantitative analysis of three subcorpora of the *Brown*-family, namely *B-Brown* (1931), *Brown* (1961) and *Frown* (1991/2), while focusing on the syntactic parameter only.

We apply two methodological approaches. In a first step, the genitives are extracted from the syntactically parsed subcorpora, and the data manually edited within the context of interchangeability. Manual filtering is a necessary and time-consuming preliminary to an analysis of factors constraining the

choice. An automated procedure would thus be desirable, particularly when dealing with large corpora. Therefore, in a second step, the fine-grained methods from the first step are approximated by an automatic programming-based approach. The aim of the second step is to incorporate previous manual work into an automatic work flow.

This two-fold methodological approach offers unique insights into the possible applicability of manually-applied constraints to computerised automatic searches, and as a consequence, the possible extension of the scope of research to larger amounts of data, to new genres and new varieties.

2. Previous research

Previous research has so far focused on the various parameters that influence the choice of genitive constructions whereby the set of parameters and their relative importance often differs from scholar to scholar (see Szmrecsanyi/Hinrichs 2007: 438). Jucker (1993) counted six factors based on Altenberg (1982): the phonetic, morphological, syntactic, lexical, and relational factors and the degree of formality. Szmrecsanyi/Hinrichs (2007) followed Jucker (1993) and analyzed their data under four major conditioning factors, taking the syntactic and pragmatic levels, as well as communicative aspects and language processing, into account. Rosenbach's (2003) influential work on genitive choice takes only three factors into account: animacy, topicality and possessive relation. She categorically excludes any factors that bias the free choice between the two genitive variants.

The availability of a set of corpora stretching across three time periods and spanning more than half a century offered a unique opportunity to apply the methodological approaches of previous research to a new dataset, and hence give insights into the diachronic changes in genitive choice in American English from 1930 to 1960 to 1990. A previous pilot study with this dataset (Röthlisberger 2009) led to the conclusion that the syntactic factor is one of the most influential parameters in genitive choice, and this has hence been chosen as the focus of this study.

3. Manual Approach

3.1 Data

We used the three corpora *B-Brown*, *Brown* and *Frown* in order to analyse the changes in frequency between 1930 and 1990, and to detect possible differences in change between the periods 1930-1960 and 1960-1990. Within these three corpora we focused on category J (Academic writing). We had to rely on those subcorpora of American English, because the B-Brown corpus is still not completed, and the J and K texts are the only ones fully available in their parsed version.

Surface pattern searches incur a large error rate, since both *s*-genitive forms and *of*-prepositions are ambiguous. The surface form of *s*-genitives is identical to contracted forms of *be* in the 3rd person singular (e.g. *Peter's painting is large* vs. *Peter's painting a house*), and the attachment site of prepositions is ambiguous. In *We accused the man of robbery* the *of*-preposition attaches to the verb, in *the state of emergency of the nation* the second *of*-preposition attaches to *state* and not to *emergency*. In order to minimise the error rate, each text was parsed with the syntactic parser *Pro3Gres* (Schneider 2008). Figure 1 shows an example of the parser output. The *s*-genitive *a layperson's point* is annotated with the dependency label *pos* (possessive), the *of*-genitive point of view with the dependency label *modpp* (modification by PP; this label is used for noun-PP attachment). The possessum is the governor in the relation, the possessor the dependent. Parsing approaches are not error-free either (see Schneider 2008). For the *s*-genitive and *of*-genitive, the accuracy on the 500 sentences *GREVAL* evaluation corpus (Carroll et al. 2003) is as follows: *of*-PP precision=89.6% and recall=85.82%. *s*-genitive precision and recall=97.4%.

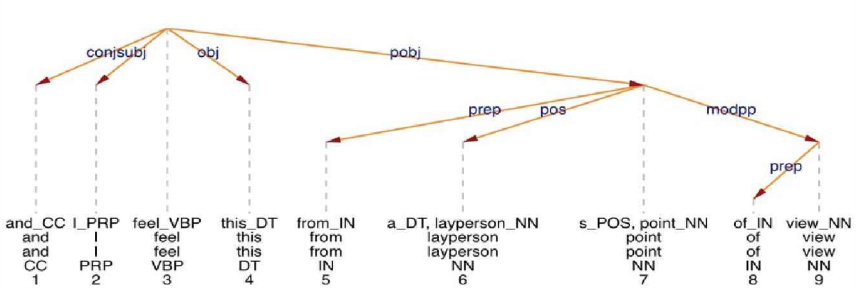


Figure 1: Parser output of a sample sentence.

3.2 Methodology

To analyse the variation between *of*- and *s*-genitive within the three subcorpora, the whole set of genitives was restricted to constructions occurring in so-called 'choice context', which goes back to Labov's principle of accountability of (1969). Rosenbach (2002) defines "choice context" as the linguistic context where both genitive variants can occur in free variation (Rosenbach 2002:40). To extract the relevant interchangeable genitive constructions, this study relies heavily on the methods used by Rosenbach (2002: 28ff), Szmrecsanyi (2007:448), Ljung (1997: 30), Raab-Fischer (1995: 127), and Kreyer (2003: 170-171).

A conversion rule was applied to all genitives found in the dataset: both the original and its alternative genitive construction have to be semantically equivalent and grammatically correct with both possessum and possessor as nouns (thus, for example, excluding genitives with pronominal possessors). The genitives need to have a possessive genitive function and should not appear in an idiomatic expression or conventionalised phrase. The excluded constructions are descriptive genitives, independent genitives, local genitives, post-genitives, nested and group genitives, elliptic genitives, *s*-genitive construction whose possessum is premodified by *own*, and titles of books, films, and works of art that are premodified by their creator's name. In order to bring forth a comparable context, *of*-genitives with a referential device other than a definite element are excluded from the analysis (e.g. *a nest of a bird*) because *s*-genitives are already definite in their nature (Langacker 1995:63). We also exclude almost all *of*-genitives with a possessor that shows a clausal postmodification, due to the conversion rule. Additionally, we exclude measures expressed with *of*-constructions, and *of*-constructions where the possessum modifies the possessor, because such constructions result in an ungrammatical descriptive genitive when converted (e.g. *a king of honour* ≠ *an honour's king*). The data was manually filtered by applying the conversion rule and the set of restrictions as noted above. The three parsed subcorpora were analysed with *SWI-Prolog*, where we programmed a rules file according to the constraints of this study.

3.3 Syntactic factors: theoretical approach

On the level of each constituent, both possessor and possessum can constitute a noun phrase with its governor and modifications (Kreyer 2003: 179). The principle of end-weight and the proximity principle (we do not discuss the lat-

ter here) are two of the most important syntactic factors that affect the individual distribution of pre- and postmodifications on the level of the possessive noun phrase, and on the level of possessor and possessum. This principle implies that longer phrasal constituents tend to follow shorter ones (Szmrecsanyi/Hinrichs 2007: 453). Therefore, a pre-or postmodified possessor-NP should favour *of*-genitive as in 1), whereas a pre-or postmodified possessum-NP favours *s*-genitive as in 2).

1) *the centre* [possessum] *of a guarded heart* [possessor]

2) *my mind's* [possessor] *ability to communicate* [possessum]

3.4 Results

The application of the aforementioned constraints produced the following frequencies of *s*- and *of*-genitives in the three subcorpora:

	s-genitives		*of*-genitives		TOTAL
	N	%	N	%	
B-Brown-J	162	11.0	1306	89.0	1468
Brown-J	179	18.0	814	82.0	993
Frown-J	352	34.4	670	65.6	1022

Table 1: *s*- and *of*-genitives in academic writing in *B-Brown*, *Brown* and *Frown*

	1930-1960	1960-1990
s-genitives	+10.5%	+96.7%
of-genitives	−37.7%	−17.7%

Table 2: Changes in frequency from 1930-1960 and 1960-1990 by genitive type

Table 2 suggests that *of*-genitives decreased to a greater extent in the time period 1930-1960, thus possibly creating a functional and syntactic gap that is filled by the *s*-genitive in the period 1960 to 1990. However, data is too sparse to be able to make a specific claim for such a drag-chain. The differences in frequencies in *B-Brown*, *Brown*, and *Frown* (Table 1) are very highly significant ($df=2$, $p<0.0001$, $\chi^2=2154.06$).

In order to assess the influence of weight and the principle of end-weight, we measure important features related to weight in the following, in particular post- and premodification (Jucker 1993), length and relative length (Szmrecsanyi/Hinrichs 2007).

3.4.1 Syntactic factors: pre- and postmodification

In a first step, the genitive constituents were filtered according to their pre- or postmodifications. Any referential device modifying the possessor in an *of*-construction did not count as modification (e.g. *the motivations of **the** actors*) (cf. Szmrecsanyi/Hinrichs 2007: 453). For methodological reasons, a compound consisting of Noun+Noun was considered to have a modification – namely the first noun. We analysed all genitives according to the pre- and/or postmodifications of their constituents (Table 3).

Type of modification	*s*-gen			*of*-gen		
	1930	1960	1990	1930	1960	1990
No modification	124	97	204	579	279	227
Premodification of possessum	29	53	119	300	185	132
Premodification of possessor	6	17	18	279	190	198
Postmodification of possessum	0	2	0	0	1	0
Postmodification of possessor	0	0	0	0	2	1
Post-and Premod. of possessum	0	0	0	0	0	0
Post- and premod. of possessor	0	0	0	0	0	2
Modification of possessum and possessor	3	10	11	148	157	110
TOTAL	**162**	**179**	**352**	**1306**	**814**	**670**

Table 3: Distribution of modified possessors and possessums by corpus and genitive type

When applying the χ^2-test, the difference between the frequencies of *s*-genitives from 1930 to 1990 is highly significant (df=6, χ^2=31.44, p<0.001). For the *of*-genitive, the χ^2-test at df=6 gives χ^2=56.37, and p=0; the differences from 1930 to 1990 in the choice of *of*-genitives are also highly significant. For a further analysis, we only took those lines into account in which more than half of the numbers are higher than 5, and compared the differences in the corpora between 1930-1960 and 1960-1990, using the log-likelihood test (Table 4).[1]

[1] The critical values to indicate significance are: $p < 0.05$; critical value = 3.84 *; $p < 0.01$; critical value = 6.63 **; $p < 0.001$; critical value = 10.83 ***; $p < 0.0001$; critical value = 15.13 ****

	s-genitives		*of*-genitives	
	1930-1960	1960-1990	1930-1960	1960-1990
No modification	−6.55*	+0.30	−12.82***	−0.02
Premodification of possessum	+4.94*	+0.66	−0.01	−1.58
Premodification of possessor	+4.44*	−3.28	+0.88	+5.39*
Mod. of possessum and possessor	+3.31	−1.73	+21.41****	−1.69

Table 4: The log-likelihood value between the different corpora according to modifications

The significant increase of *s*-genitives with premodified possessums and *of*-genitives with premodified possessors indicates that the principle of end-weight gains ground. Note that this increase occurs for *s*-genitives in the period 1930-1960, and *of*-genitives in the later period. This change could therefore be interpreted as a push-change that starts with the *s*-genitive. The significant increase of *s*-genitives with premodified possessors in the period 1930-1960 runs counter to the concept of end-weight. Further research will be needed in that direction.

3.4.2 Syntactic factors: constituent length

In a second step, we established the boundaries of each genitive construction and calculated the mean possessor and possessum length in orthographic words. Any referential device modifying the possessum or possessor was again not taken into account (Table 5).

Corpora	*s*-genitive		*of*-genitive	
	Mean N1 length	Mean N2 length	Mean N1 length	Mean N2 length
B-Brown-J	1.06	1.23	1.48	1.44
Brown-J	1.18	1.44	1.57	1.52
Frown-J	1.09	1.47	1.65	1.44

Table 5: Mean possessor (N1) and possessum (N2) length in the three corpora

Table 5 illustrates that the mean length of the first constituent in *s*-genitives and *of*-genitives remains fairly stable across the years, while the last constituent in both constructions tends to increase in length. Note that the last constituent in *s*-genitive is N2, while in *of*-genitive it is N1. Again, this points to the influence of end-weight.

A comparison between the lengths of possessor and possessum demonstrates that a difference in length influences the choice of genitive (Table 6).

	s-genitives		*of*-genitives	
	1930-1960	1960-1990	1930-1960	1960-1990
N1>N2	+5.14*	−4.04*	+0.19	+6.22*
N1=N2	−4.30*	+0.02	−1.09	−1.01
N1<N2	+5.18*	+0.56	+1.06	−2.10

Table 6: Log-likelihood test for length of possessor/possessum according to time period

Table 6 indicates that *s*-genitives with a longer first constituent (N1>N2) increase significantly from 1930 to 1960, while the same holds true for *s*-genitives with a longer last constituent (N1<N2). Only the later change follows the principle of end-weight. The first change is either caused by other factors, or may be due to low counts (N=6 is lowest for N1>N2 in 1930, while N=30 is lowest for N1<N2 in 1930). From 1960 to 1990, *s*-genitives with longer first constituents decrease significantly, this time following the principle of end-weight. The principle also seems to hold true for the increase of *of*-genitives with longer last constituents in the time period 1960-1990. The changes across the whole table are highly significant (χ^2 contingency table, df=10, $p < 0.001$).

Overall, the manual approach has shown that the changes in genitive choice tend to follow the principle of end-weight and are generally significant.

4. Automatic approach

We have already automated the syntactic annotation in the manual approach, which constitutes a new method in historical corpora. In this section, we suggest automatic approaches approximating to the manual approach described in Section 3. Automatic approaches have the advantage that they scale, and are consistent and reproducible.

4.1 Methods

We discussed our parsing method in Section 3.1. Only a subset of the Saxon genitives (*s*-genitives) and *of*-PPs are in variation. As the envelope of variation (Labov 1969), which Rosenbach (2003) calls the choice context, is subject to semantic restrictions, its automation is challenging. We now suggest approximations and discuss results in Section 4.2.

4.1.1 Raw counts

Assuming that occurrences of variation and non-variation of the *s*-genitive and *of*-genitive are spread homogeneously across the corpus, raw counts can be used as a coarse measure.

4.1.2 Animacy and proper names

In prototypical *s*-genitives, the possessor is a proper name. Restricting counts to cases where the possessor is a proper name is thus a useful approximation: a large portion of the cases in the variation are covered, and only a few false positives included. Proper names and animacy are related.

4.1.3 Data-driven alternations

The only reliable proof of variation is to test if a token can be in the alternation, in other words that both the original and its alternative genitive construction have the same meaning, which we have tested in the manual approach. As this test relies on semantics and speaker intuition, it cannot be automated easily.

Idioms	*point of view* <≠> **view's point* **(eye) view of bird* <≠> *bird's (eye) view*
Creators	*Spielberg's film* <≠> *?film of Spielberg*
Fixed nominal expressions / Proper names	*Noah's ark* <=> *?Ark of Noah* *Newton's comet* <=> *?Comet of Newton* *Institute of Archaeology* <=> **Archaeology's Institute*
Measures / Quality	*tin of soup* <≠> **soup's tin* *half of (the) century* <≠> **century's half*
Semantic restrictions	*one's recovery* <=> **recovery of one* *God's creation* <≠> *?creation of God*
... many other expressions that are not in the alternation, e.g.:	*image of power* <≠> *?power's image* *concentration of oxygen* <=> *?oxygen's concentration* *faculty of reason* <≠> *?reason's faculty*

Table 7: Examples of automatically excluded alternation candidates

What we can test, however, is whether the alternative form does occur in the corpus. If the two alternatives with the same lexemes are found in the corpus, they constitute a valid alternation.

(LEX) B's A <=> A of B

For example, if both the NP *Peter's friend* and *friend of Peter* are found in the corpus, then they are a valid alternation. There are two differences between such an automatic test and the manual approach. First, the automatic approach is based on performance instead of competence. Second, the requirement of semantic equality cannot be tested, so some false positives will be generated.

4.1.4 Adding semantic classes to overcome sparse data

In practice, there is a third difference. There is typically a serious sparse data problem in that there are relatively few pairs with lexical overlap of both the possessum (governor) and the possessor (dependent). In order to alleviate this problem, we require semantic class overlap instead of lexical overlap.

(SEM.1) B's A <=> class(A) of class(B)

For example, if both the NPs *Peter's friend* and *wife of John* are found in the corpus, they are accepted as valid alternation, because *Peter* and *John* are in the same semantic class, as well as *wife* and *friend*. As lexical class, we use the *WordNet* lexicographer file (Miller 1990). Semantic class overlap shows high correspondence with manual decisions.[2] Classes of automatically excluded pairs are given in Table 7.

One of the restrictions of the manual approach can be automated directly: *of*-genitives with a referential device other than a definite element need to be excluded from the analysis (e.g. *a nest of a bird*). We have added this restriction.

(SEM) B's A <=> class(A) of class(B) AND B is definite

4.2 Results

4.2.1 Raw counts

The raw counts (RAW) are compared in Table 8. The *s*-genitive increases, and *of*-PP seems to decrease, which is in accord with Leech et al. (2009:48) and with our manual data (see Table 2). The increases and decreases are also shown in Table 8. Column 2 gives the ratio (*s*- divided by *of*-), columns 3 and 5 absolute counts, and columns 4 and 6 give percentages.

Corpus	*s*/*of*	*s*-gen #	%	*of*-gen #	%
RAW					
B-Brown-J	0.05	347	4.7%	6998	95.3%
Brown-J	0.06	411	6.1%	6356	93.9%
Frown-J	0.12	716	10.9%	5853	89.1%
1930-60		+18.4%		−9.7%	
1960-90		+74.2%		−7.9%	

Table 8: Raw counts and frequency changes

[2] We did not conduct a formal evaluation of the overlap. Figure 2 gives an indication of the quality.

4.2.2 Animacy and proper names

The proper name counts (PROP) are given in Table 9. Columns 2 and 4 give absolute counts. Columns 3 and 5 give percentages for proper names, showing that proper name genitives are increasingly often realized as *s*-genitives. The readiness for proper name in *of*-PPs is generally low and decreases, as the last two columns show. They give the percentages of genitives with proper names.

Corpus	*s*-gen #	%*s*/ Prop N	*of* #	%*of*/ Prop N	%Prop N/*s*-gen	%Prop N/ *of*-gen
PROP						
B-Brown-J	255	28.7%	634	71.3%	73.5%	9.1%
Brown-J	245	35.4%	447	64.6%	59.6%	7.0%
Frown-J	468	53.7%	404	46.3%	65.4%	6.9%

Table 9: Proper Names counts and percentages

4.2.3 Data-driven alternations

As discussed in Section 4.1.3, the data-driven alternation counts are too low to be reliable or statistically significant, with only between 16 and 35 counts per cell.

4.2.4 Adding semantic classes to overcome sparse data

The counts for semantic class overlap, method SEM.1, are given in Table 10, which can be compared to the results of the manual method given in Table 1. The *of*-genitive overgenerates considerably compared to the manual method.

Corpus	*s*-gen #	% *s*/all	*of* #	% *of*/all
SEM.1				
B-Brown-J	240	6.6%	3 404	93.4%
Brown-J	319	9.5%	3 039	90.5%
Frown-J	564	15.4%	3 101	84.6%

Table 10: Semantic class overlap counts and percentages

The semantic class counts with added indefinite filter, method SEM, are given in Table 11. The *of*-genitive overgenerates less than in SEM.1. The suggested trends are in full agreement with those found by the manual approach.

Corpus	*s*-gen #	% *s*/all	*of* #	% *of*/all
SEM				
B-Brown-J	240	9.8%	2 207	90.2%
Brown-J	319	15.5%	1 737	84.5%
Frown-J	564	25.6%	1 637	74.4%
1930-60		32.9%		−21.3%
1960-90		76.8%		−5.8%

Table 11: Semantic class overlap counts and percentages

4.3 Scaling to British English and other genres

An advantage of automatic approaches is that they scale to other genres, and different, and larger, corpora. We extended our investigation to category K, and to the *LOB* series of corpora (*BLOB* (1931), *LOB* (1961) and *Freiburg LOB* (1991)). We show the results using raw counts (RAW) in Table 12, and semantic class plus indefinite filter (SEM) in Table 13.

Corpus	*s*-gen #	% *s*/all	*of* #	% *of*/all
RAW				
BLOB-J	362	5.0%	6 929	95.0%
LOB-J	425	6.7%	5 897	93.3%
FLOB-J	575	9.0%	5 824	91.0%

Table 12: Raw counts in the *LOB* family

Corpus	*s*-gen #	% *S*/all	*of* #	% *of*/all
SEM				
BLOB-J	243	12.5%	1 706	87.5%
LOB-J	333	15.8%	1 776	84.2%
FLOB-J	372	18.1%	1 687	81.9%
1930-60		37.0%		4.1%
1960-90		11.7%		−5.0%

Table 13: Changes in frequency in the *LOB* family

The trend is similar for the *LOB* family: *s*-genitive increases, and *of*-genitive decreases, although less strongly than in the Brown family, and only relative to the frequency of the *s*-genitive.

4.4 The principle of end-weight

We also investigated the principle of end-weight using the automatic approach SEM.1. We measured the probability of the tokens being pre-modified, which is comparable to the constituent length in Section 3.4.2, Table 5. The results for *Brown* and *LOB* are given in Table 14. Similar results were obtained: the constituent at the end (in bold) has a much higher likelihood of being modified in the more recent corpora, indicating the increased importance of the principle of end-weight.

Corpus	*s*-gen			*of*-gen		
	N	p(DepMod)	**p(GovMod)**	N	**p(DepMod)**	p(GovMod)
B-Brown-J	240	11.7%	**27.1%**	3404	**43.8%**	33.5%
Brown-J	319	13.5%	**32.9%**	3039	**46.0%**	38.4%
Frown-J	563	11.4%	**35.2%**	3101	**47.7%**	37.0%
BLOB-J	243	18.5%	**16.9%**	2407	**39.0%**	31.6%
LOB-J	333	18.0%	**22.8%**	2700	**44.3%**	34.2%
FLOB-J	372	13.4%	**25.0%**	3023	**47.0%**	33.7%

Table 14: End-weight in the *Brown* and *LOB* families

The differences in *Brown* are highly significant (χ^2 contingency, df=2, $p < 0.001$). The differences in *LOB* are significant, but not highly significant (χ^2 contingency, df=2, p=0.039).

5. Discussion

We investigated changes in *s*-Genitives and *of*-Genitives and the principle of end-weight from different perspectives, and observed the same trends in all perspectives. Comparing the absolute counts delivered by the different methods on the *Brown* series in Table 14, it is clear that raw counts overgenerate massively, while the semantic class + indefinite filter counts overgenerate less, as the raw numbers show. Counts for the manual method (MAN) are listed in Table 2, for the raw count method (RAW) in Table 10, and for the semantic class filter (SEM) in Table 11. The percentage increase of *s*-genitives is compared in Figure 2. The comparison shows that SEM is a better approximation to MAN than RAW.

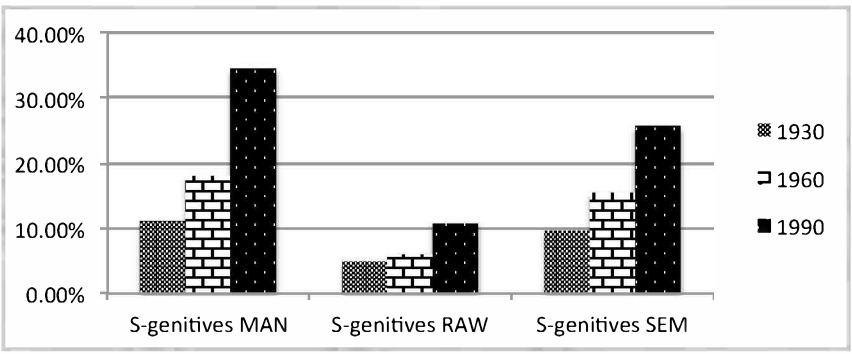


Figure 2: Percentage of Saxon Genitive measured by the different approaches

6. Conclusions

We have shown that the use of the *s*-genitive increased between 1930 and 1990 in both American and British English, while the *of*-Genitive has decreased. The *s*-genitive becomes more restricted to proper names. The differences over the time periods are significant. We have shown that the principle of end-weight has become stronger. We have also presented an approach to the automatic detection of pairs in genitive alternation, which can partly alleviate the workload of the annotator. The manual and automatic approaches are mutually validating. Although automatic approximation overgenerates and delivers a weaker signal, it clearly shows the same trends as the manual approach. In future research, we will conduct a formal evaluation and port the technique to other choice contexts, for example the dative shift.

7. Acknowledgements

We would like to thank two anonymous reviewers for their helpful comments.

References

Altenberg, Bengt (1982): The genitive v. the *of*-construction: a study of syntactic variation in 17th century English. Lund: CWK Gleerup.

Barber, Charles (1964): Linguistic change in Present-Day English. Edinburgh/ London: Oliver & Boyd.

Biber, Douglas (1987): A textual comparison of British and American writing. In: American Speech 62: 99-119.

Carroll, John/Minnen, Guido/Briscoe, Edward (2003): Parser evaluation: using a grammatical relation annotation scheme. In: Abeillé, Anne (ed.): Treebanks: building and using parsed corpora. Kluwer: Dordrecht, 299-316.

Francis, W. Nelson/Kučera, Henry (1982): Frequency analysis of English usage: lexicon and grammar. Boston: Houghton Mifflin Company.

Hawkins, Roger (1981): Towards an account of the possessive constructions. NP's N and the N of NP. In: Journal of Linguistics 17: 247-269.

Jucker, Andreas H. (1993): The genitive versus the *of*-construction in newspaper language. In: Jucker (ed.), 121-136.

Jucker, Andreas H. (ed.) (1993): The Noun phrase in English: its structure and variability. Heidelberg: Winter.

Kreyer, Rolf (2003): Genitive and *of*-construction in modern written English: processability and human involvement. In: International Journal of Corpus Linguistics 8: 169-207.

Langacker, Ronald W. (1995): Possession and possessive construction. In: Taylor, John R./MacLaury, Robert E. (eds.): Language and the cognitive construal of the world. Berlin/New York: Mouton de Gruyter, 51-79.

Labov, William (1969): Contraction, deletion, and inherent variability of the English copula. In: Language 45: 715-762.

Leech, Geoffrey/Smith, Nicolas (2006): Recent grammatical change in written English 1961-1992. Some preliminary findings of a comparison of American with British English. In: Renouf, Antoinette/Kehoe, Andrew (eds.): The changing face of corpus linguistics. Amsterdam/New York: Rodopi, 185-204.

Leech, Geoffrey/Hundt, Marianne/Mair, Christian/Smith, Nicholas (2009): Change in contemporary English: a grammatical study. Cambridge: Cambridge University Press.

Ljung, Magnus (1997): The *s*-genitive and the *of*-construction in different types of English texts. In: Fries, Udo/Müller, Viviane/Schneider, Peter (eds.): From Aelfric to the New York Times. Studies in English Corpus Linguistics. Amsterdam: Rodopi, 21-32.

Mair, Christian (2006): Inflected genitives are spreading in present-day English, but not necessarily to inanimate nouns. In: Mair, Christian/Heuberger, Reinhard (eds.): Corpora and the History of English. Heidelberg: Winter, 235-248.

Miller, George A./Beckwith, Richard/Fellbaum, Christiane/Gross, Derek/Miller, Katherine (1990): Wordnet: an on-line lexical database. In: International Journal of Lexicography 3: 235-244.

Ortmann, Albert (1998): The role of [+/– animate] in inflection. In: Fabri, Ray/Ortmann, Albert/Parodi, Teresa (eds.): Models of inflection. Tübingen: Niemeyer, 60-84.

Osselton, Noel. E. (1988): Thematic genitives. In: Nixon, Graham/Honey, John (eds.): An historic tongue: studies in English linguistics in memory of Barbara Strang. London/New York: Routledge, 138-144.

Raab-Fischer, Roswitha (1995): Löst der Genitiv die *of*-phrase ab? Eine korpusgestützte Studie zum Sprachwandel im heutigen Englisch. In: Zeitschrift für Anglistik und Amerikanistik 43: 123-132.

Röthlisberger, Melanie (2009): *Of*-genitive versus *s*-genitive. A corpus-based analysis of possessive constructions in 20th century American English. Unpublished seminar paper: Universität Zürich.

Rosenbach, Anette (2002): Genitive variation in English. Conceptual factors in synchronic and diachronic studies. Berlin: Mouton de Gruyter.

Rosenbach, Anette (2003): Aspects of iconicity and economy in the choice between the *s*-genitive and the *of*-genitive in English. In: Rohdenburg, Günter/Mondorf, Britta (eds.): Determinants of grammatical variation in English. Berlin/New York: Mouton de Gruyter, 379-411.

Rosenbach, Anette (2005): Animacy versus weight as determinants of grammatical variation in English. In: Language 81: 613-644.

Rosenbach, Anette (2006): On the track of Noun+Noun constructions in Modern English. In: Houswitschka, Christoph/Knappe, Gabriele/Müller, Anja (eds.): Proceedings of the Conference of the German Association of University Teachers of English 27. Trier: Wissenschaftlicher Verlag, 543-557.

Rosenbach, Anette (2007). Exploring constructions on the web: a case study. In: Biewer, Carolin/Hundt, Marianne/Nesselhauf, Nadja (eds.): Corpus linguistics and the Web. Amsterdam/New York: Rodopi, 167-190.

Rosenbach, Anette/Stein, Dieter/Vezzosi, Letizia (2000): On the history of the *s*-genitive. In: Bermudez-Otero, Ricardo/Denison, David/Hogg, Richard M./McCully, Christopher B. (eds.): Generative theory and corpus studies: A dialogue from ICEHL 10. Berlin/New York: Mouton de Gruyter, 185-210.

Schneider, Gerold (2008): Hybrid long-distance functional dependency parsing. Unpublished Dr. Phil. thesis: Philosophische Fakultät, Universität Zürich.

Szmrecsanyi, Benedikt/Hinrichs, Lars (2007): Recent changes in the function and frequency of Standard English genitive constructions: a multivariate analysis of tagged corpora. In: English Language and Linguistics 11: 437-474.

Szmrecsanyi, Benedikt/Hinrichs, Lars (2008): Probabilistic determinants of genitive variation in spoken and written English. A multivariate comparison across time, space, and genres. In: Nevalainen, Terttu/Taavitsainen, Irma/Pahta, Päivi/Kor-

honen, Minna (eds.): The Dynamics of linguistic variation. Corpus evidence on English past and present. Amsterdam/Philadelphia: John Benjamins, 291-309.

The American heritage dictionary of the English language (2000). Fourth edition. Boston: Houghton Mifflin.

Varantola, Krista (1993): Modification of nouns by nouns – bad by definition. In: Jucker (ed.), 69-83.

Wasow, Thomas (1997): Remarks on grammatical weight. In: Language Variation and Change 9: 81-105.

JAVIER RUANO-GARCÍA / MARÍA F. GARCÍA-BERMEJO GINER /
PILAR SÁNCHEZ-GARCÍA

Past tense BE forms in Late Modern Lancashire English

A preliminary corpus-based approach[1]

Abstract

Although the alternation between *was* and *were* has been extensively recorded in modern varieties of British and overseas English, there is comparatively little information about the distribution of *was* and *were* in older varieties of speech. This has been largely due both to the scarcity of old regional material, and the consequent lack of diachronic dialect corpora. In light of this, this paper looks at some of the Lancashire texts included in the *Salamanca Corpus*. It examines the evidence provided by literary representations of the dialect with regard to past tense BE forms. Though largely neglected for linguistic investigation, literary samples of Lancashire English may go some way towards casting light on the forms of BE in the county between 1700 and 1900. Our aim is thus twofold: firstly, to contribute to previous research into past tense BE forms in Lancashire by adding historical data that have not been thus far considered, and secondly, to illustrate the linguistic possibilities of the corpus, arguing that it may serve as a complementary missing link to expand the database of English diachronic dialectology.

1. Introduction

As is true of other regional varieties, our knowledge about early Lancashire speech is characteristically scarce. The linguistic history of English dialects is still distinguished by a great many gaps which render it complex to evaluate the linguistic (dis)continuities between Middle English and modern times. Whilst the increasing availability of textual corpora has enabled successful diachronic research into the history of standard English, variation in regional English dialects remains virtually unexplored. Given these reasons, the *Salamanca Corpus*, which was launched in February 2011, has been conceived as a repository of diachronic dialect material from 1500 to 1950 that might fill some of the *lacunae* still present in the field (see García-Bermejo Giner 2010, 2012).

[1] We wish to thank the University of Salamanca for financial aid to carry out research in Manchester (John Rylands Library, Chetham's Library, Portico Library, Manchester City Library, and local archives) in July 2011.

This paper examines some of the Lancashire texts included in the *Salamanca Corpus*. It builds on previous research into *was/were* alternation in Lancashire and other varieties of British English (see Britain 2007 for a detailed review of the literature), with the aim of casting some light on the forms of BE in Late Modern Lancashire speech. Whilst most scholarly work has thus far concentrated on modern patterns of variation, little research has been conducted into the forms of BE in older speech, Nevalainen (2006) and Kytö/Walker/Grund (2007) being amongst the very few who have approached the subject from a historical perspective. To our knowledge, however, there is no single study attempting to examine the alternation between *was* and *were* in the light of the evidence given by literary representations of regional language. These have been largely, and unjustly, neglected for linguistic purposes, yet samples of dialect literature and literary dialect may go some way towards illuminating the distribution and development of regional features (see Lodge 2010, for example, on the so-called Definite Article Reduction). Given this, our purpose is twofold. Firstly, we aim to shed light on the past forms of BE by means of selected renderings of Lancashire English in the Late Modern period (1700-1900). For this purpose, we shall pay attention to the spread of *r*-forms (e.g. *I were*, *he were*), and to the gradual loss of the verbal *-n* in the past plural (i.e. *they weren* > *they were*). Secondly, we aim to illustrate the linguistic possibilities of the corpus for regional dialect investigation. In doing so, we would hope that this paper may add to the literature on the historical distribution of *was* and *were*, and may prove the linguistic validity of dialect literary documents more generally (see Wales 2010, amongst others).

2. Lancashire data in the *Salamanca Corpus*

As with many other regional varieties, early evidence about Lancashire English is relatively scant. Along with the sparse data which can be gleaned from glossaries, pronunciation dictionaries or survey-books, literary texts are fruitful early sources of information. Documents such as the hitherto unpublished *A Lancashire Tale* (c.1690-1730) and the different editions of John Collier's *A View of the Lancashire Dialect* are outstanding early witnesses to the dialect of the county (Ruano-García 2010, Wagner 1999). Unfortunately, specimens purporting to reproduce the dialect of the area are not particularly abundant during the Early Modern period and the eighteenth century. In fact, Shorrocks (1999a) explains that it was during the nineteenth century that the strongest tradition of dialect literature developed in the county (see further

Ruano-García 2012 for an account of literary and non-literary sources of information).

The earliest localized records which have been compiled in the corpus correspond with the two mentioned above, along with an unpublished anonymous ballad that begins "Robin an's Gonny" (c.1690-1730). So far, we have found relatively few additional documents from the 1700s, but these include some verse dialogues written by John Byrom (1773), Henry Clarke's *The School Candidates* (1788) and Robert Walker's *Plebeian Politics* (1798). For the nineteenth century, retrieval of material for linguistic mining has been made easy by the availability of literary texts in far greater numbers, including works of major Lancashire figures such as Benjamin Brierley (1825-1896) and Edwin Waugh (1817-1890), and of minor writers such as Roper Robinson (1836-1908) and James Bowker. It must be acknowledged, however, that evidence from the first half of the nineteenth century has been hard to find. Shorrocks (1999a: 89-90) argues that

> By 1860, dialect literature was appearing in truly large quantities, and continued to do so for the rest of the century and the first quarter or third of the twentieth century. Its burgeoning between 1850 and 1860 coincided with a marked improvement in material prosperity.

As such, Lancashire texts from the first half of the 1800s are scarce. In particular, only 11 out of the 156 nineteenth-century documents compiled belong to the first half of the century. Interestingly, some of them are cases of dialect literature dated 1800 that are held in the Chetham's Library, Manchester. As can be seen in Table 1, there are certain periods which are underrepresented in the corpus. Our current research aims to repair some of the present textual deficits.

Time span	No. of texts	%
1500-1700	5	2.9
1700-1800	8	4.7
1800-1950	156	92.3
1800-1850	11	
1850-1900	130	
1900-1950	15	
Total	**169**	**99.9**

Table 1: Chronological distribution of Lancashire texts in the corpus (as of October 2011)

3. Past tense BE forms in Late Modern Lancashire English

3.1 Primary data

The selection of data has been made according to three criteria, responding to the need to obtain information which might be taken as representative of language use. Firstly, we have found it necessary to consider cases of dialect literature only, since occurrences of *was* and *were*, as used in the county, are expected to be higher there. Consequently, preference has been given to documents written by natives to Lancashire, in an attempt to avoid vague and impressionistic renderings of the dialect that might provide misleading information. Obviously, natives to Lancashire are expected to have a fairly good knowledge of the dialect. Secondly, only prose texts and dialogues, some of them written in verse, have been considered for scrutiny. This is not meant to suggest that regional drama provides useless data; this kind of textual evidence remains to be investigated for inclusion into the corpus. Finally, it has been our purpose to give a balanced sample of material, but this has not been possible because of the uneven chronological coverage of the corpus. As a result of these criteria, the present analysis concentrates on the period 1700-1900, considering four different subperiods of 50 years each for a clearer analysis of the data. It is worth noting that, whenever possible, we have selected one text per decade, and only a few texts from 1850-1900 have been considered, so as to avoid an excessive bias in the results. These have been chosen in a manner that attempts to avoid idosyncratic traits and individual practices. That is, only works by different authors have been examined. In total, our primary data consist of 14 texts which amount to about 144,000 words. Although this is a relatively small sample, it seems to be statistically significant for the present purpose. Table 2 shows the number of texts and words per (sub)period considered.

Time span	N of texts	N of words
1700-1800	6	29,540
1700-1750	3	10,086
1750-1800	3	19,454
1800-1900	8	113,928
1800-1850	3	7,632
1850-1900	5	106,296
Total	**14**	**143,468**

Table 2: Lancashire material examined: texts and words per (sub)period

3.2 Theoretical background

There exists an important body of studies treating *was/were* variation in British and overseas English, but most of them deal with the modern setting (for instance, Anderwald 2001, Cheshire/Fox 2009, Hollman/Siewierska 2006: 25-26, Pietsch 2005 and Tagliamonte 1998), since historical information is not abundantly available. Needless to say, *was/were* alternation in English is documented as far back as the Middle English (ME) period in view of the large number of alternative forms attested in the record. As regards the North, Pietsch (2005:149-150) explains that *was/were* variation is analogous to the Northern Subject Rule, *was* being used with all subjects except *I*, *you*, *we* and *they* in verb adjacent position. However, this alternation presents different patterns according to the data documented in the traditional dialects covered by the *Survey of English Dialects* (1962-1971) (henceforth *SED*). Drawing on the *SED* findings, Pietsch (2005: 150-151) distinguishes different areas, the first of these being the Central North, covering Cumberland, Northumberland, Westmorland and Durham, where the use of *was* and *were* is similar to that of *is* and *are*, *was* being used throughout the singular, *were* being used throughout the plural. However, *was* is also licensed in plural contexts by the Northern Subject Rule (see Beal 2005: 122). Secondly, he identifies an area including southern Lancashire, southwestern Yorkshire and Derbyshire where *were* is the verb form for both singular and plural (see further Trudgill 2008: 349). Finally, a transitional zone between the above mentioned areas is distinguished, comprising northern Lancashire and the northeastern half of Yorkshire, where *were* is likewise used in the singular, although on a less frequent basis. Clearly, the six northern counties show differences with regard to the use of *was/were*, with the main observations being either alternation or a tendency towards levelling in the paradigm. It therefore seems clear that in Lancashire *were* predominates as the form for both singular and plural, which Shorrocks' (1999b) study of the Bolton area corroborates. Additionally, Beal (2005: 122) sheds some historical light on the picture, by explaining that

> Accounts of the traditional dialects of Yorkshire and Lancashire (Wright 1892; Ellis 1869-1889) suggest that the typical pattern in these areas was one in which *were* occurred with all subjects, singular and plural

From this, one could perhaps assume that *were*-levelling was manifested both in the spread of *r*-forms throughout the paradigm, and in the neutralisation of the singular/plural distinction characteristic of the standard. The question

raised in this paper is whether the past tense of BE in Lancashire has always been distinguished by a preponderance of *r*-forms. We shall examine the contexts in which the different forms of BE appear, focusing also on the gradual loss of verbal *-n*, as the singular/plural distinction has been long preserved in the county.

3.3 Survey and discussion of the data

In her study of Collier's Lancashire dialect, Wagner (1999: 201) claims that "generalisations of the past plural stem to the singular" took place during the Early Modern period. This might have been so, since a summary look at the Linguistic Profiles (LPs) of Lancashire included in the *LALME* suggests that in Late ME *was* and *were* conformed to a very great extent to present-day patterns. In fact, the LPs in which both singular and plural forms are recorded (LPs 6, 21, 23, 25, 113, 154, 167, etc.) show that *s-* and *r-* forms were used for the singular and plural, respectively.[2] However, the documentary *lacunae* existing between late ME and Collier's work render it difficult to corroborate Wagner's contention. Kytö/Walker/Grund (2007) do not record Lancashire data from this time span, but begin their study with the 1700s. Similarly, the *Salamanca Corpus* does little to bridge this gap, for the Early Modern Lancashire material consists of literary dialects written by non-natives to the area. However, taking the evidence supplied by the literary records dated circa 1690-1730, Wagner's statement seems not to be quite accurate. Actually, data extracted from "A Lancashire tale" and "Robin an's Gonny" suggest that *was* was relatively predominant over *were* for the singular in the late seventeenth and early eighteenth centuries, where 14 tokens of *was* against 4 of *were* have been found. As shown below, *was* and *were* are attested with NPs and personal subject pronouns:

(1) *Th' Monn's Cwote wur a Grey* (*A Lancashire Tale*)
['The man's coat was grey']

(2) *he wus down on his back* (*Robin an's Gonny*)

[2] The rest of the LPs record cases of either *was* or *were*. These also have *was* in singular contexts, whilst *were* was used in the plural. The Lancashire data analysed in this paper show different spellings for past tense BE forms. A preference for <u> is attested in the corpus: *wus, wur, wuren, wur'n*, etc. Because of the difficulty of some dialect passages, some of the examples provided in the paper will be given in modern standard English in brackets.

Obviously this does not provide enough evidence to support generalisations of any kind, but as early testimonies to the language of the county, these data should be taken into account.

Table 3 shows, however, that preference for *were* in the singular seems to have been the norm in the 1700s, which concurs with Wagner's argument. Although the cases of *was* could be taken as indicative of *were* not having been fully generalized by then, it is worth noting that the *s*-forms other than those 14 referred to above occur in a single text: Byrom's Lancashire dialogue (1773). An individual bias might explain this.

	1700-1750			**1750-1800**		
	was	*were*		*was*	*were*	
		were	contracted		*were*	contracted
NP	4 (0.4)	31 (3.1)		9 (0.5)	43 (2.2)	
I		10 (0.9)	44 (4.4)	3 (0.15)	6 (0.3)	13 (0.7)
you				1 (0.05)	2 (0.1)	
he, she, it	9 (0.9)	24 (2.4)	13 (1.3)	3 (0.15)	33 (1.7)	16 (0.8)
existential *there*	1 (0.09)	3 (0.3)			16 (0.8)	
relative pn.		4 (0.4)			4 (0.2)	
pronoun		16 (1.6)		2 (0.1)	11 (0.6)	
Total	**14**	**145**		**18**	**144**	
%	**8.8**	**91.2**		**11.1**	**88.9**	

Table 3: Eighteenth-century data: singular contexts[3]

Clearly, there is a strong preponderance of *were* with all singular subjects in the data. Sometimes, adjacent subject pronouns invite contraction, as in (4):

(3) *that wur o meety fawse owd Felly* (*A View of the Lancashire...*)
['that was a very false old fellow']

(4) *hee'r gooink by th' shop dur* (*Plebeian Politics*)
['he was going by the shop door']

As for the nineteenth century, the data recorded in Table 4 show a tendency to use *were* in the singular too. In fact, only *r*-forms have been found, likewise suggesting that *were* was used with (non-)adjacent subject pronouns, NPs, relative pronouns with singular antecedents, etc. By way of illustration:

[3] In this and the ensuing tables, the following information is provided: raw figures, normalized frequencies per 1,000 words in brackets, and total percentages.

(5) *ew wur that* (*A Dialouge between Owd Carder Joan...*)
['I was that']

(6) *Sally were cryin afore hoo'd done* (*The Works of James...*)
['Sally was crying before she had done']

	1800-1850			1850-1900		
	was	*were*		*was*	*were*	
		were	contracted		*were*	contracted
NP		19 (2.5)			226 (2.1)	
I		3 (1.7)	9 (1.2)		209 (1.9)	3 (0.02)
you			1 (0.1)		2 (0.01)	
he, she, it		6 (0.8)	6 (0.8)		304 (2.8)	
existential *there*					59 (0.5)	
relative pn.		1 (0.1)			37 (0.3)	
pronoun		3 (1.7)			45 (0.4)	
Total	**0**	**48**		**0**	**885**	
%	**0**	**100**		**0**	**100**	

Table 4: Nineteenth-century data: singular contexts

The examples extracted from plural contexts show that only *r*-forms were used in the county in the periods considered, irrespective of the type of subject examined. (7) and (8) exemplify *r*-forms with plural subject pronoun, and plural NP, respectively:

(7) *They wurn, its loike, whaint fond o'summut new* (*Miscellaneous...*)
['They were, it seems, very fond of something new']

(8) *Tim an me wor bwoth young then* (*Betty o' Yep's...*)
['Tim and I were both young then']

As these examples, and Tables 5 and 6 below indicate, there is a difference between the periods considered in that the eighteenth-century data manifest a strong preference for forms marked for plurality, whilst the data from the 1800s point in a somewhat different direction.

John Collier commented in the prefatory remarks to the *Miscellaneous Works of Tim Bobbin* (1775: fos. A2-A2v) that one of the most salient features of Lancashire English was the plural *-(e)n* inflection. The verbal *-n* for the present indicative plural dates back to ME, being characteristic of the West Midlands, and has been preserved in some areas of Derbyshire, Cheshire or Staffordshire

(Wright 1905: 435; Upton/Parry/Widdowson 1994: 492). This was likewise used for the past plural of BE in Lancashire or western Yorkshire, still persisting in a relic area of the North-West Midlands (Orton/Sanderson/Widdowson 1978: Maps 21-23). The Lancashire texts analysed support this fact, as forms such as *wuren*, *wurn*, *wur'n*, etc. are attested, especially during the 1700s.

	1700-1750			**1750-1800**		
	weren		*were*	*weren*		*were*
	-n, *-'rn*, *-r'n*	contracted		*-n*, *-'rn*, *-r'n*	contracted	
pl. NP	3 (0.3)			17 (0.9)		1 (0.05)
NP+NP	1 (0.09)			2 (0.1)		1 (0.05)
we						
you	2 (0.2)	1 (0.1)				
they	1 (0.09)	4 (0.4)	1 (0.09)	6 (0.3)	12 (0.6)	1 (0.05)
exist. *there*						7 (0.35)
relative pn.				11 (0.56)		1 (0.05)
pronoun			1 (0.09)			1 (0.05)
Total	**12**		**2**	**48**		**12**
%	**85.7**		**14.3**	**80**		**20**

Table 5: Eighteenth-century data: plural contexts

	1800-1850			**1850-1900**		
	weren		*were*	*weren*		*were*
	-n, *-'rn*, *-r'n*	contracted		*-n*, *-'rn*, *-r'n*	contracted	
pl. NP			2 (0.26)			76 (0.7)
NP+NP			2 (0.26)			15 (0.14)
we				2 (0.01)	5 (0.04)	7 (0.06)
you				1 (0.009)	1 (0.009)	
they				14 (0.1)	33 (0.3)	26 (0.2)
exist. *there*						15 (0.14)
relative pn.			1 (0.13)			21 (0.19)
pronoun						4 (0.03)
Total	**0**		**5**	**56**		**164**
%	**0**		**100**	**25.4**		**74.5**

Table 6: Nineteenth-century data: plural contexts

According to the data, the differences observable in the forms for the plural between the eighteenth and the nineteenth century mostly lie in the gradual decline of verbal *-n* during the 1800s. Although the corpus information from the 1800s suggests that verbal *-n* disappeared from the landscape in the first half of the century and re-emerged in the latter part, it is worth stressing that

only five tokens of past plural BE are attested between 1800 and 1850, which are obviously not sufficient to establish with any certainty what the situation might have been like. Also, it is worth noting that an important number of the forms marked for plurality between 1850 and 1900 are those contracted when preceded by subject pronouns, as in (9). In light of this, it appears that contracted forms may have favoured the preservation of *-n* so as to distinguish the past from the present indicative, as in (10). This seems to have been in decline during the last quarter of the nineteenth century.

(9) *So we'rn marchin' away, but before we'd gwon...* (*Tummus an'...*)
['So we were marching away, but before we had gone'...]

(10) *we're tawkin' abeaut th' dangers o' th' sonds* (*Goblin Tales...*)
['we are talking about the dangers of the sands']

In view of this, it appears likely that, although *r*-forms had been used for both singular and plural since at least the early eighteenth century, Lancashire English began to neutralize the singular/plural distinction in the past tense of BE well into the 1800s.

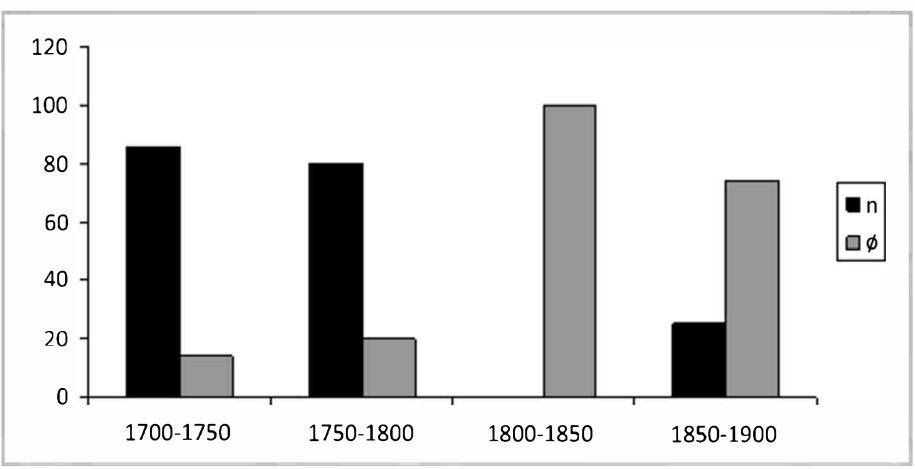


Figure 1: Loss of verbal *-n* in past plural forms: percentages

4. Concluding remarks

This paper has been concerned with shedding some light on the historical distribution of past tense BE forms in Lancashire English. For this purpose, we have analysed some literary renditions of the dialect from the Late Modern

period. The data, though unbalanced at some points, suggest that the dialect was distinguished by a predominance of *r*-forms for both the singular and the plural in the period under investigation. This concurs with modern surveys of the area, which have shown that *were* is the norm for singular and plural subjects. A difference, however, has been detected with modern evidence in that the corpus testifies to the existence of past forms marked for plurality. These, reflecting a typical West Midlands verbal inflection, were apparently more widespread during the 1700s. In fact, the data show a decline of verbal *-n* during the 1800s. Earlier periods did not demonstrate the neutralisation of number distinction shown by modern accounts of BE in Lancashire.

In sum, we hope to have contributed to the history of past tense BE forms in Lancashire English, and to the historical distribution of *was/were* in varieties of English more generally. Although this is an issue which has received a great deal of attention, there are still some aspects which have not been sufficiently considered. The present availability of machine-readable documents in which older varieties of English are reproduced makes it plausible to obtain further diachronic data. There is hope that evidence from other text types and periods which are underrepresented in the corpus will help towards clarifying the full picture, not only in Lancashire.

References

Secondary literature

Anderwald, Liselotte (2001): *Was/Were* variation in non-standard British English today. In: English World-Wide 22: 1-21.

Beal, Joan (2005): English dialects in the North of England. Morphology and syntax. In, Kortmann, Bernd/Schneider, Edgar (eds.): A Handbook of varieties of English. Vol. 2 Morphology and syntax. Berlin: Mouton, 114-141.

Britain, David (2007): Grammatical variation in England. In: Britain, David (ed.), Language in the British Isles. Cambridge: Cambridge University Press, 75-104.

Cheshire, Jenny/Fox, Sue (2009): *Was/were* variation: A perspective from London. In: Language Variation and Change 21: 1-38.

Collier, John (1775): The Miscellaneous Works of Tim Bobbin. Manchester: Printed for the Author.

García-Bermejo Giner, María F. (2010): Towards a history of English literary dialects and dialect literature in the 18th and 19th centuries: The Salamanca Corpus. In: Heselwood/Upton (eds.), 31-41.

García-Bermejo Giner, María F. (2012): The Online Salamanca Corpus of English Dialect Texts. In: Vázquez González (ed.), 67-74.

García-Bermejo Giner, María F./Sánchez-García, Pilar/Ruano-García, Javier (eds.) (2011-): The Salamanca Corpus. A Digital Archive of English Dialect Texts. http://salamancacorpus.usal.es/SC/index.html

Heselwood, Barry/Upton, Clive (eds.): Papers from Methods XIII. Frankfurt a.M.: Peter Lang.

Hollman, Willem/Siewierska, Anna (2006): Corpora and (the need for) other methods in a study of Lancashire dialect. In: Zeitschrift für Anglistik und Amerikanistik 54: 203-216.

Kytö, Merja/Walker, Terry/Grund, Peter (2007): English witness depositions 1560-1760. An electronic edition. In: ICAME Journal 31: 65-86.

Lodge, Ken (2010): Th' interpretation of t' definite article in t' North of England. In: English Language and Linguistics 14: 111-127.

McIntosh, Angus/Samuels, Michael/Benskin, Michael (eds.) (1986): A Linguistic atlas of Late Mediaeval English. Vol. 3: Linguistic profiles. Aberdeen: Aberdeen University Press.

Nevalainen, Terttu (2006): Vernacular universals? The case of plural *was* in Early Modern English. In: Nevalainen, Terttu/Klemola, Juhanni/Laitinen, Mikko (eds.): Types of variation: diachronic, dialectal and typological interfaces. Amsterdam: Benjamins, 351-369.

Orton, Harold/Dieth, Eugene, Halliday, William/Barry, Michael/Tilling, Phillip/Wakelin, Martyn (eds.) (1962-1971): Survey of English dialects (B): the basic material. Leeds: E.J. Arnold & Son Ltd.

Orton, Harold/Sanderson, Stewart/Widdowson, John (eds.) (1978): The linguistic atlas of England. London: Routledge.

Pietsch, Lukas (2005): "Some do and some doesn't". Verbal concord variation in the north of the British Isles. In: Kortmann, Bernd (ed.): A comparative grammar of British English dialects: agreement, gender, relative clauses. Berlin: Mouton, 125-209.

Ruano-García, Javier (2010): I'll tell o how Gilbert Scott sowd is mere Berry: 'A Lancashire tale' as a source for Lancashire speech in the late seventeenth and early eighteenth century. In: Heselwood/Upton (eds.), 53-66.

Ruano-García, Javier (2012): Th' Monn, twoman and t felley: on the definite article in traditional Lancashire English. In: Vázquez González (ed.), 403-431.

Shorrocks, Graham (1999a): Working-class literature in working-class language: the north of England. In: Hoenselaars, Ton/Buning, Marius (eds.): English literature and other languages. Amsterdam: Rodopi, 87-96.

Shorrocks, Graham (1999b): A Grammar of the dialect of the Bolton area. Part II: Morphology and syntax. Frankfurt a.M.: Peter Lang.

Tagliamonte, Sally (1998): *Was/were* variation across the generations: view from the city of York. In: Language Variation and Change 10: 153-191.

Trudgill, Peter (2008): English dialect "default singulars", was versus were, Verner's Law, and Germanic dialects. In: Journal of English Linguistics 36: 341-353.

Upton, Clive/Parry, David/Widdowson, John (eds.) (1994): Survey of English dialects: the dictionary and grammar. London: Routledge.

Vázquez González, Nila (ed.) (2012fc.): Creation and use of historical English corpora in Spain. Newcastle upon Tyne: Cambridge Scholars Publishing.

Wagner, Tamara (1999): John Collier's 'Tummus and Meary'. Distinguishing features of 18th-Century southeast Lancashire dialect – morphology. In: Bulletin of the Modern Language Society, 100: 191-205.

Wales, Katie (2010): Northern English in writing. In: Hickey, Raymond (ed.): Varieties of English in writing. Amsterdam: Benjamins, 61-80.

Wright, Joseph (1905): English dialect grammar. Oxford: Henry Frowde.

Primary literature. Texts from the *Salamanca Corpus* analyzed with the number of words per text

Period 1: 1700-1800

Subperiod 1.1: 1700-1750

Anon. (c.1690-1730): A Lancashire tale. 1321 words.

Anon. (c.1690-1730): Robin an's Gonny.... 322 words.

Collier, John (1748): A View of the Lancashire Dialect. 8,443 words.

Subperiod 1.2: 1750-1800

Byrom, John (1773): Miscellaneous Poems [selection]. 4,783 words.

Clarke, Henry (1788): The School Candidates [selection]. 483 words.

Walker, Robert (1798): Plebeian Politics. 14,188 words.

Period 2: 1800-1900

Subperiod 2.1: 1800-1850

Anon. (1800?): Jone's Ramble from Grinfelt to Owdham. 441 words.

Anon. (1832?): A Dialouge [sic!] between owd Carder... 1,073 words.

Brierley, Benjamin (1850): Gooin' to Cyprus. 6,118 words.

Subperiod 2.2: 1850-1900

Ormerod, Oliver (1856): O full, tru un pertikler okeawnt... 29,636 words.
Lahee, Margaret (1865): Betty o' Yep's Laughable Tale. 8,617 words.
Kershaw, Tom (187?): Tummus an' Meary modernised. 11,149 words
Bowker, James (1883): Goblin Tales of Lancashire. [selection] 29,115 words.
Clegg, James T. (1895): The Works of James Trafford Clegg. [selection] 27,779 words.

OLGA TIMOFEEVA

Anglo-Latin and Old English

A case for integrated bilingual corpus studies of Anglo-Saxon registers

Abstract

This article describes Anglo-Latin and Old English as two codes correlated in Anglo-Saxon England with the same cultural elite. Introducing a taxonomy of Anglo-Saxon registers, it claims that Anglo-Latin material can supplement our knowledge of early Old English lexis. A corpus of *Medieval Latin from Anglo-Saxon Sources* is advocated as a new electronic resource to facilitate bilingual studies in this field.

The interface between Latin and Old English (OE) in the insular period has been traditionally described in terms of language contact, or rather it has been tacitly assumed that such terms or phrases as "Latin influence", "Latin borrowings", etc. can be safely used to describe this situation. Although there seems to be little doubt that these are valid terms, it is striking that what historians of Old English take for granted in their field is clearly at odds with how, say, contact-induced change or bilingualism are understood in contact linguistics (for example in Thomason/Kaufman 1988, or Heine/Kuteva 2005), but see also the discussion of the discrepancy and suggestions for alternative terminology in Timofeeva (2010a,b; 2011). The problematic aspects of the functioning of Latin in Anglo-Saxon England are twofold. On the one hand, there is a controversy (too often ideologically charged in our postcolonial world) over the survival of British Latin – the extent to which it penetrated the various classes of Romanized Celtic society, the upper time limit of its last vestiges, and the geographical distribution of its speakers before and after the Anglo-Saxon settlement (see Jackson 1953: 94-121; Gratwick 1982: 2-6, 69-71; Wollman 1993: 8-15; Wright 2002: 4; Schrijver 2002, 2007 and Tristram 2004: 94-99, etc.). All these make it difficult and often impossible to estimate the circumstances and effects of 'normal' everyday language contact between speakers of British Latin and English.

On the other hand, there is another contact situation between 'high' Latin and Old English among the educated Anglo-Saxons. It is with this second language setting that most studies of the Latin lexical and syntactic influence on OE are

concerned (see selected references in Timofeeva 2006: 48-51; 2010b: 19-22, 78-84, 185). Although they give us valuable insights into development of certain areas of lexis or certain domains of syntax, their authors tend to be somewhat vague as to the sociolinguistic environment of the 'loans' or 'influences' that they discuss, which are generally listed as types of loans (see for example van Gelderen 2006: 93-95). Except for Latin being notoriously (and perhaps too uncritically) referred to as 'the language of the church and administration', we find very little discussion – with the notable exception of Fischer (1992) – of how exactly Latin functioned in Anglo-Saxon society, and whether loans and influences could take place at all in this setting. Moreover, Latin-Old English interaction is typically presented from the point of view of what OE gets from Latin and not of what it gives back. Thus the picture that we have at present lacks both background and dimension. In what follows I would like to suggest that the Latin and English produced by the Anglo-Saxons might be seen as two codes correlated with the same cultural elite. With Latin being the highest among the Anglo-Saxon registers, I defend the idea of integrated bilingual corpus studies of these registers, and introduce my Anglo-Latin corpus project as a first step in this direction.

To begin with, let us briefly consider the interaction between 'high' Latin and Old English from the language-contact position. Three features of the Anglo-Latin bilingualism should be highlighted, namely that it is distant, written, and socially restricted (see Wright 2002: 11-17). Such settings are not universally recognised as legitimate cases of bilingualism (see Thomason/Kaufman 1988: 66-67). Loveday (1996), however, allows for *distant but institutional* bilingualism,[1] in which the speech community as a whole is typically monolingual, and the second-language acquisition is often related to political and cultural dominance. In the OE period, direct contact with native speakers of late Latin-early Romance will have been very rare among the Anglo-Saxons, although within Latin-based institutions (school and church), the intensity of exposure to Latin must have been very high. With a lack of oral exchange with native speakers, written competence in Latin prevails over oral competence. Its acquisition and use are socially restricted to clerical strata, and advanced second-language proficiency is widespread only among the higher secular clergy (i.e. bishops and cathedral priests) and regular clergy (monks and nuns, see

[1] "[T]his kind of contact takes place when the acquisition of a foreign language is not part of community activities, unless in the domain of religion, but is promoted through an institution such as school" (Loveday 1996: 19-20).

Timofeeva 2010a: 1-2, 9-16; 2010b: 8-11). What is also rather unfavourable for the linguistic implications of Latin-Old English language contact is that the size of this bilingual group is well below one per cent of the total population.[2] All this allows us not only to envisage how small the number of people who used Latin was, but also to understand that our knowledge of OE is essentially limited to the language of an extremely small community (see Tristram 2004: 103-105). Since literacy in OE typically presupposes literacy in Latin, that is, any formal schooling is inevitably Latin schooling, in the course of which one can also acquire an ability to read and write OE,[3] it follows that written Latin and written OE are produced and consumed by more or less the same group of people, the professional ecclesiastical minority.[4]

This fact was recognised by philologists at least forty years ago (Bolton 1971) and articulated most eloquently by Lapidge (1993 [1991]: 1-2, n. 1):

> [W]e should always remember that works in Latin and the vernacular were copied together in Anglo-Saxon scriptoria, and were arguably composed together in Anglo-Saxon schools. What is needed, therefore, is an integrated literary history which treats Latin and vernacular production together as two facets of one culture, not as isolated phenomena.

Although a lot has been done to integrate the two literatures,[5] the languages in which they are written continue to be held apart. I would, therefore, like to encourage linguists to consider a possibility of an integrated language history

[2] It is indeed possible to get a rough estimate of how many people knew Latin in the OE period. Given that the clergy is the only group that is likely to be educated in Latin, the estimate of the number of clerics would yield us a figure that would come close to the size of the bilingual group. I have based my calculation on the *Domesday Book* of 1086. The total population in 1086 is estimated to be between 1,100,000 and 2,250,000 people (Russell 1944, 1948; Miller/Hatcher 1978; Hinde 2003). The estimate of the size of the clergy (based on the number of bishoprics, cathedrals, monasteries, and the average number of clerics associated with them) is about 5,500 people (for more details on this calculation, see Timofeeva 2010a: 12-16). Thus, if we divide this figure by the total population, we get between 0.5 and 0.25 per cent, cf. Tristram (2004: 105).

[3] King Alfred's educational plans provided for the reverse acquisition of literacy among free young men in England: the ability to read English first, followed by further instruction in Latin (CPLetWærf 49; cf. Asser, ch. 102), but we do not know whether or how widely this practice extended beyond his palace school (Lapidge 1993 [1991]: 5-12). Ælfric's *Grammar* of c. 1000 is another notable exception (Bullough 1991: 314-317).

[4] Cf. Wormald's conclusions concerning the "restricted literacy" of the Anglo-Saxon period (1977: 113).

[5] See, for example, Pulsiano/Treharne (2001), which brings together articles on Anglo-Latin and Old English literary practices under one title, eloquently phrased *A Companion to Anglo-Saxon Literature*.

which treats Latin and the vernacular together as two facets of one language. Typologically speaking, the two languages of course remain separate, even though examples of various types of code mixing are not too hard to find (see Schendl 2004, Timofeeva 2010a, etc.). My concern, however, is not with typology, but with the taxonomy of registers in Latin-vernacular diglossia. Because both Anglo-Latin and written OE are determined by user characteristics such as religion, class and social power, this diglossia can be best described as *user-oriented*. In Anglo-Saxon England, Latin 'high' (and in due time OE 'high' too) is "superposed acquisitionally and functionally only for a portion of the community" (Britto 1986: 35-53, 331-332) and remains nobody's native language, but one that is only acquired through schooling, and is correlated with its users as the language of the cultural elite.

Let me illustrate the 'one-language' approach with a case study of the notion of 'Latin' in Anglo-Latin and Old English.[6] A diachronic corpus study consisting of two sets of data: Anglo-Latin texts written between the 670s and 800s (based on a selection from *Library of Latin Texts*, Series A in Brepolis databases), and Old English texts written for the most part between the 850s and 1050s (based on a selection from *DOEC*) reveals that the development of vocabulary connected with Latin language and culture shows a clear continuity from Anglo-Latin to OE (see Tables 1 and 2). The main conceptual associations between 'Latin' and 'language', 'literacy', 'education', 'books', 'translation', etc. are first transferred to and formulated in Anglo-Latin from continental Latin, and with the emergence of the vernacular written tradition, they are later re-encoded in OE, with necessary adjustments being made so as to fit these words and phrases to OE morphology.

[6] Described in detail in Timofeeva (forthc.). On language ideologies and attitudes towards 'Latin' in Antiquity, see Fögen (2003); on 'Latin' in the Middle Ages, see Wright (1982, 1991, 2002), Janson (1991), van Uytfanghe (1991), etc. A detailed survey of secondary literature on the term *Latinus* is available in Kramer (1998: 11-57).

called/named/means in Latin	69
in Latin (adverb)	44
translate into Latin	31
Latins as a people	29
Latin language	29
Latin word/book/letter	20
apud Latinos	16
called by the *Latini*	6
'Latin' in context with Romans	4
Latin tradition	4
Latin etymology	3
Latin nouns	2
Latin eloquence	2
correct Latin	1
language of the *Latini*	1
X is Latin	1
Latin authors	1
Latin libraries	1
forest of Latinity	1
Total	**265**

call/mean in Latin	133
write in Latin	16
understand/know Latin	16
Latin books	10
translate from Latin	8
Latins as a people (*Lædenware*)	7
translate into Latin	6
Latin grammar	4
study Latin	3
learned/educated in Latin	2
Latin word	2
speak Latin	1
knowledge of Latin	1
avoid barbarisms in Latin	1
Latin computus	1
mix English and Latin	1
Total	**212**

Table 1 (left): Contexts and collocates of "Latin" and "Latinity" in Anglo-Latin
Table 2 (right): Contexts and collocates of "Latin" and "Latinity" in OE

All the collocations that are present in Anglo-Latin also find their way into OE. Later on, however, as *læden* words are being assimilated in OE, new compounds begin to emerge.[7] In other words these concepts and vocabulary are first adopted by the high written register of the Anglo-Saxons (before the 800s it is Latin by default) and are then infiltrated into their lower written register, OE. Thus, I suggest that the Anglo-Latin data can be used as a supplement primarily to the meagre contents of the OE1 period (dated to before 850 in the *Helsinki Corpus*)[8] and to other periods of OE. Studies based on these two sets

[7] E.g., OE develops three compounds to denote the "Latin-language": *læden-spræc*, *læden-geþeode*, and *læden-gereord*. Two more compounds are *læden-boc* "Latin book" and *boc-læden* "book Latin; written language". The conceptual proximity of 'Latin' and written culture continues to be emphasized in these compounds.

[8] The complete word count for OE1 is 2,190 words. These include a few early charters, *Cædmon's Hymn*, *Bede's Death Song*, the *Ruthwell Cross*, and the *Leiden Riddle* (Kahlas-Tarkka/Kilpiö/Österman 1993: 21-24).

of data can also help us trace the paths of lexical borrowing and assimilation of loans.

What has to be borne in mind, though, is that these studies will continue to describe the two written registers of the educated elite. Tristram (2004) has suggested that the written and spoken English language of the Anglo-Saxon elite was kept comparatively constant throughout the OE period and continued to be cultivated for about two generations following the Norman Conquest. "[T]he vernacular of the bulk of the population" was markedly different from this OE standard. It was, therefore, "the *spoken* language of the formerly repressed low variety" with its substrata of Celtic and Scandinavian that "surfaced after the replacement of the Anglo-Saxon elite by William the Conqueror" and later on gave rise to "a strongly regionalized middle class *written* language" (Tristram 2004: 103-104 – italics in the original). While Tristram's tripartite diglossia model – OE high written, OE high spoken, and OE low spoken – is undoubtedly a valuable contribution to our understanding of the abrupt changes of the early Middle English period, I suggest that a fuller picture may emerge if we envisage the language situation of OE period as still more layered and dynamic:

> **Latin high** is a formal written register, documented between about twelve and twenty times better than the surviving OE (Bolton 1971: 151-152). It was used chiefly by the clergy, whose proficiency in Latin varied greatly depending on time period, possibly location, and, above all, social status.
>
> **OE low 1 → OE high 1** is a formal written register, well documented and used chiefly by the clergy and a few educated laymen. Starting out as a West-Saxon courtly norm of the late ninth century (OE low 1), it gradually developed towards a second written standard (OE high 1), competing with and eventually replacing other existing written norms (Mercian and Northumbrian). This standard continued to be maintained well into the twelfth century.
>
> **OE low 2 → OE high 2** is a less formal spoken variety of the above. It is undocumented and was used, again, by the Anglo-Saxon powerful elite.
>
> **OE low 3** is an informal spoken register, undocumented, used by the lower classes with diverse ethnic/linguistic backgrounds: Celtic, British Latin, and Scandinavian (see Tristram 2004: 103-105).

While spoken OE will largely remain a matter of scholarly speculation, the interfaces between the written registers can be understood more fully if comparative studies of Anglo-Latin and OE (as the one outlined above) are extended to later Anglo-Latin, and set against the context of other vernaculars

and other varieties of Latin. In terms of English historical lexis, there is clearly a lot to be gained from such diachronic multilingual investigations, with concepts connected with local insular culture providing perhaps an obvious point of departure for future studies. Our understanding of Latin-Old English language contact will benefit greatly if the reverse influence, that of first-language OE speakers upon Anglo-Latin, is considered. A sound classification of text-types for both written registers is another important desideratum.

Having outlined the problems and prospects of Latin-Old English linguistic studies, I would like to conclude this paper by introducing a tool that will hopefully help to address both. This tool is a corpus of *Medieval Latin from Anglo-Saxon Sources*. This project was started at the Research Unit for Variation, Contacts and Change in English, University of Helsinki, in 2009, and is presently being continued at the English department of the University of Zurich. Our aim is to compile a corpus of Latin texts from ca. 690-1150 A.D., written by authors with L1 English (or exceptionally L2 English). Ideally the corpus should be compatible with other corpora of medieval Latin from British and continental sources, accessible to and usable by a wide audience of scholars working in medieval history, culture, and language. It will have an appropriate level of metadata and annotation, and provide free and open access to several millions of words.

As an electronic reality today the corpus includes the *Anglo-Latin Minor Poetry sub-corpus* of 60,920 words (as of 31 May 2013), with division into Metrical and Rhythmical parts and division into types of poetry within each part. The files have basic metadata: author, date, place, genre, manuscript, edition, metrical analysis, etc. The prose extension was started in spring 2011 and has grown today to 156,051 words (as of 31 May 2013). Both parts are searchable with WordSmith and will go through an XML conversion in the near future. Apart from the funding institutions mentioned above, the steady progress of the project has been greatly facilitated by the generous support of Michael Lapidge, who donated his collection of Anglo-Latin verse and prose (in manuscripts, photocopies and electronic files) to the corpus, David Howlett, Antonette diPaolo Healey, and Matti Kilpiö, the careful work of Anne Gardner, Alpo Honkapohja, and Sergey Zavyalov, and the proofreading tenacity of five student assistants: Viviane Bergmaier, Lucas Orellano, Irene Rettig, Dominique Stehli, and Eva Stempelova.

References

Bolton, Whitney F. (1971): Pre-Conquest Anglo-Latin: perspectives and prospects. In: Comparative Literature 23: 151-166.

Britto, Francis (1986): Diglossia. A Study of the theory with application to Tamil. Washington, DC: Georgetown University Press.

Bullough, Donald A. (1991): The educational tradition in England from Alfred to Ælfric: teaching *utriusque linguae*. In: Carolingian renewal: sources and heritage. Manchester and New York: Manchester University Press, 297-334.

DOEC = Healey, Antonette di Paolo/Wilkin, John Price/Xiang, Xin (eds.) (2009): The Dictionary of Old English Web Corpus. Toronto: University of Toronto. http://tapor.library.utoronto.ca/doecorpus/.

Fischer, Olga C.M. (1992): Syntactic change and borrowing: the case of the accusative-and-infinitive construction in English. In: Gerritsen, Marinel/Stein, Dieter (eds.): Internal and external factors in syntactic change. (= Trends in Linguistics: Studies and Monographs 61). Berlin/New York: Mouton de Gruyter, 17-88.

Fögen, Thorsten (2003): Forms of language awareness in Antiquity and their significance for Latin linguistics: some theoretical remarks. In: Solin, Heikki/Leiwo, Martti/Halla-aho, Hilla (eds.): Latin vulgaire – latin tardif VI, Actes du VI^{e} colloque international sur le latin vulgaire et tardif. Helsinki, 29 août-2 septembre 2000. Hildesheim: Olms-Weidmann, 29-45.

Gratwick, Adrian S. (1982): Latinitas Britannica: was British Latin archaic? In: Brooks, Nicholas (ed.): Latin and the vernacular languages in Early Medieval Britain. Leicester: Leicester University Press, 1-79.

Heine, Bernd/Kuteva, Tania (2005): Language contact and grammatical change. Cambridge: Cambridge University Press.

Hinde, Andrew (2003): England's Population: a history since the Domesday Survey. London: Hodder Arnold.

Jackson, Kenneth (1953): Language and history in early Britain. A chronological survey of the Brittonic languages first to twelfth century A.D. Edinburgh: Edinburgh University Press.

Janson, Tore (1991): Language change and metalinguistic change: Latin to Romance and other cases. In: Wright (ed.), 19-28.

Kahlas-Tarkka, Leena/Kilpiö, Matti/Österman, Aune (1993): Old English. In: Rissanen, Matti/Kytö, Merja/Palander-Collin, Minna (eds.): Early English in the computer age. Explorations through the Helsinki Corpus (= Topics in English Linguistics 11). Berlin/New York: Mouton de Gruyter, 21-32.

Kramer, Johannes (1998): Die Sprachbezeichnungen *Latinus* und *Romanus* im Lateinischen und Romanischen. Berlin: Erich Schmidt.

Lapidge, Michael (1993 [1991]): Schools, learning and literature in tenth-century England. In: Anglo-Latin Literature 900-1066. London/Rio Grande, OH: Hambledon Press, 1-48.

Library of Latin Texts, Series A. Brepolis Databases. Brepols. http://clt.brepolis.net/llta/Default.aspx.

Loveday, Leo J. (1996): Language contact in Japan: a sociolinguistic history. Oxford: Oxford University Press.

Miller, Edward/Hatcher, John (1978): Medieval England – rural society and economic change 1086-1348 (= Social and Economic History of England). London: Longman.

Russell, Josiah Cox (1944): The clerical population of Medieval England. In: Traditio 2: 177-212.

Russell, Josiah Cox (1948): British medieval population. Albuquerque: University of New Mexico Press.

Schendl, Herbert (2004): 'Hec sunt prata to wassingwellan': Aspects of code-switching in Old English charters. In: VIEWS 13(2): 52-68.

Schrijver, Peter (2002): The rise and fall of British Latin. Evidence from English and Brittonic. In: Filppula, Markku/Klemola, Juhani/Pitkänen, Heli (eds.): The Celtic roots of English (= Studies in Languages 37). Joensuu: University of Joensuu, 87-110.

Schrijver, Peter (2007): What Britons spoke around 400 AD? In: Higham, Nicholas (ed.): Britons in Anglo-Saxon England. Woodbridge: Boydell and Brewer, 165-171.

Thomason, Sarah G./Kaufman, Terrence (1988): Language contact, creolization, and genetic linguistics. Berkeley: University of California Press.

Timofeeva, Olga (2006): Latinskie sintaksicheskie zaimstvovania v drevneangliyskom yazyke [Latin syntactic borrowings in the Old English language]. Sankt Peterburg: Gelikon Plus.

Timofeeva, Olga (2010a): Anglo-Latin bilingualism before 1066: Prospects and limitations. In: Hall, Alaric/Timofeeva, Olga/Kiricsi, Ágnes/Fox, Bethany (eds.): Interfaces between language and culture in medieval England: A Festschrift for Matti Kilpiö (= The Northern World 48). Leiden: Brill, 1-36.

Timofeeva, Olga (2010b): Non-finite constructions in Old English, with special reference to syntactic borrowing from Latin. (= Mémoires de la Société Néophilologique de Helsinki 80). Helsinki: Société Néophilologique.

Timofeeva, Olga (2011): Infinitival complements with the verb *(ge)don* in Old English. Latin influence revisited. In: Leeds Studies in English 42: 93-108.

Timofeeva, Olga (forthcoming): *Of ledenum bocum to engliscum gereorde*: Bilingual communities of practice in Anglo-Saxon England. In: Kopaczyk, Joanna/Jucker, Andreas H. (eds.): Communities of practice in the history of English. (= Pragmatics & Beyond New Series). Amsterdam/Philadelphia: John Benjamins.

Tristram, Hildegard L. C. (2004): Diglossia in Anglo-Saxon England, or: what was spoken Old English like? In: Studia Anglica Posnaniensia 40: 87-110.

van Gelderen, Elly (2006): A history of the English language. Amsterdam/Philadelphia: John Benjamins.

van Uytfanghe, Marc (1991): The consciousness of a linguistic dichotomy (Latin-Romance) in Carolingian Gaul. The contradictions of the sources and of their interpretation. In: Wright (ed.), 114-29.

Wollman, Alfred (1993): Early Latin loan-words in Old English. In: Anglo-Saxon England 22: 1-26.

Wormald, C. P. (1977): The uses of literacy in Anglo-Saxon England and its neighbours. In: The Transactions of the Royal Historical Society, fifth series 27: 95-114.

Wright, Roger (1982): Late Latin and Early Romance in Spain and Carolingian France (= Classical and Medieval Texts, Papers and Monographs 8). Liverpool: Francis Cairns.

Wright, Roger (1991): The conceptual distinction between Latin and Romance: invention or evolution? In: Wright (ed.), 103-113.

Wright, Roger (ed.) (1991): Latin and the Romance languages in the Early Middle Ages. London/New York: Routledge.

Wright, Roger (2002): A Sociophilological study of Late Latin (= Utrecht Studies in Medieval Literacy 10). Turnhout: Brepols.

III. Historical linguistic corpora: Architecture, annotation, and tools

MATHILDE HENNIG

The *Kassel Corpus of Clause Linking*

Abstract

The *Kassel Corpus of Clause Linking* is part of a larger project on the grammar of New High German led by Vilmos Ágel. The project takes as its starting point the assumption that equal consideration of both oral and written language is essential in order to understand developments in grammar during the New High German period. The *Kassel Corpus of Clause Linking* includes four texts from the period 1650-1700 and four texts from 1850-1900, with three texts from each period exemplifying oracy (immediacy) and one text literacy (distance). The corpus was annotated for grammatical features which are relevant for clause linking, such as predicates, subjects and connectors. A major issue in the process of annotation was to identify correlations between single grammatical features and types of clause linking, such as coordination, subordination and ellipsis. In the paper we will explain the principles which formed the basis for compiling and annotating the corpus, and illustrate how correlations between single grammatical features and types of clause linking may be established. We will also provide an example (non-integrative ellipsis) of how the annotated corpus reveals differences between immediacy and distance, as well as showing historical developments.

1. Introduction

The *Kassel Corpus of Clause Linking* is part of a larger project on the grammar of New High German, which is led by Vilmos Ágel at the University of Kassel,[1] and in the following the corpus will be referred to as '*Kajuk*' (for "Kasseler Junktionskorpus".[2] In research into the history of German language, the New High German period (since 1650) has been less thoroughly investigated than earlier periods (see Ágel 2000). Since our knowledge about recent developments in grammar is still rather poor, we do not have a reference grammar covering the whole New High German period. There are grammars of German covering the time from Old High German to Early New High German (Schrodt 2004; Paul 1989; Reichmann/Wegera 1993), and of course there are several grammars of Modern German. But the period between 1650 and the present is

[1] For further information: http://www.uni-kassel.de/fb02/fileadmin/datas/fb02/Institut_für_Germanistik/Fachgebiete/Sprachwissenschaft/Agel/ProjektNhdGramm.pdf.

[2] For further information: http://www.uni-giessen.de/kajuk/index.htm.

not yet covered by reference grammars. Vilmos Ágel intends to close this gap, but there is still a lot of grammatical research to be done before this grammar can be written, and *Kajuk* is being developed as a tool for annotating interesting grammatical features in historical texts.

Kajuk was funded by the *Deutsche Forschungsgemeinschaft* between 2007 and 2009. As it was intended to provide a first approach to the exploration of grammatical phenomena in New High German texts, the project was restricted to one field of grammar, the field of clause linking.

2. Principles of corpus compilation: immediacy and distance

The corpus consists of four texts each from the seventeenth and the nineteenth centuries. Each text consists of 12,000 words, which means that the total corpus comprises nearly 100,000 word tokens. The major principle in compiling the corpus was the consideration of *historical orality*. We shall be using the term 'immediacy' to denote conceptual orality, and the term 'distance' for conceptual literacy, following Koch/Oesterreicher (1985).

How can we consider historical orality in terms of 'immediacy' and 'distance'? These notions were introduced into research on orality by Koch and Oesterreicher in the 1980s. Although they research in the field of the Romance languages, the terms have been widely adopted in German linguistics as well (Hennig 2011). One major reason for the success of this model of immediacy and distance lies in the dissociation of orality and literacy from the medium. In actual fact, the model covers two different notions of orality: the notion of 'medium' and the notion of what Koch and Oesterreicher call the "mode of communication". Whereas "in the medial sense, "oral" (= "phonic") and "written" (= "graphic") are clearly dichotomous" (Koch 1997: 151), the differences in linguistic conception "cover a whole continuous spectrum, ranging from extremely informal oral-type expression to extremely elaborate, formal literate-type language" (Oesterreicher 1997: 193). "It goes without saying that a spontaneous conversation is a more prototypical instance of oral conception than an interview with a politician and that a statute is a more prototypical instance of written conception than an editorial" (Koch 1997: 150). This prototypical understanding of immediacy allows us to transfer the notion to historical texts, even though we do not have tape recordings of historical orality at our disposal. It allows us to look for texts which are fairly close to the pole of immediacy although, strictly speaking, they are written texts.

Vilmos Ágel and I worked further on this issue due to the conviction that a grammar of New High German should cover the whole range of language usage and not only represent the grammar of elites. Therefore, we developed the Koch/Oesterreicher model further by focussing on compiling grammatical features typical of immediacy, and relating them to pragmatic conditions, such as whether the roles of the participants as producers or recipients are fixed or flexible, whether the producers can take their time to plan their utterances or are forced to speak by the presence of the recipients, and so on (Ágel/Hennig 2006). We chose the texts for our corpus by identifying grammatical features of immediacy, and we identify 'texts of immediacy' as texts containing a large number of such features.

		Text	Text type	Dialect area	Degree of immediacy		
					Mikro	Makro	Total
17th cent.	immediacy	Güntzer I	life story	Upper German	28.8	48.3	38.6
		Bauernleben I	chronicle	Central German	26.2	44.4	35.3
		Söldnerleben I	life story	Low German	24.2	62.7	43.4
	distance	Thomasius I			3.3	2.0	2.6
19th cent.	immediacy	Zimmer V	diary	Upper German	14.7	43.2	29.0
		Koralek V	diary	Central German	14.7	63.2	39.0
		Briefwechsel V	private letters	Low German	41.8	36.7	39.3
	distance	Nietzsche V			4.9	3.4	4.1

Table 1: Corpus compilation[3]

[3] In the table, the terms 'Mikro' and 'Makro' represent different approaches on the analysis of the degree of immediacy: Whereas 'Mikro' stands for the analysis of single features of immediacy such as adjacency pairs, repairs, deictic elements, 'Makro' provides us an idea of the overall profile of a text due to syntactic features such as length of sentences, amount of elliptical sentences etc. 'Total' represents the average of the two figures. For further information see Ágel/Hennig (2006).

3. Annotation

In annotating the texts we did not work with standardized annotation tools because of our specific grammatical interests and the special character of the texts of immediacy. Rather we worked with annotation tools which were developed specifically for our purposes by the Centre for Digital Humanities at the University of Trier. All annotations were carried out manually, any kind of automated annotation being impossible, due to the wide range of spellings, and idiosyncratic use of punctuation marks in the texts.

We use the term 'clause linking' to cover all structural means of linking clauses, with semantic relations such as conditional, causal, adversative etc. (Ágel/ Diegelmann 2010). In the field of copulative relations, ellipsis was also considered as a means of clause linking (Hennig 2010). In order to analyze the linking of clauses, we first of all needed to identify the clauses to be linked. As a first step we therefore needed to segment the corpus texts into clauses. Since in the project we consider syntax a means of conveying semantics, we did not take the syntactic category 'clause' as a starting point, but rather the semantic idea of proposition. In the project, we called propositions 'Sachverhaltsdarstellungen', descriptions of facts and circumstances, following the clause-linking theory of Raible (1992) and the linguistic sign model proposed by Bühler (1934). We found it necessary to start from propositions and not from clauses, because not all propositions have the status of syntactic clauses. Propositions can also be realized elliptically, or by other syntactic structures without verbs, as example (1) shows.

(1) Nietzsche

Denn er lebt und leidet mit in diesen Scenen – und doch auch nicht ohne jene flüchtige Empfindung des Scheins

[*because he is living and is also suffering in these scenes – and not also without any feeling of pretence.*]

```
<lb n="16,17">
      <J IR="kaus" EB="prag"><KON>- denn</KON></J>
      <subj real="Pron">er</subj>
      <praed><V ID="Fin"><VV>lebt</VV></V></praed></lb>
<line n="13"/>
```

```
<lb n="17,18,2006,3100">

      <J IR="kop"><KON>und</KON></J>

      <praed real="mikro"><V ID="Fin"><VV>leidet</VV>
      </V></praed>

      mit in diesen Scenen -</lb>

<lb n="18,19,2006,2007,3100">

      <J IR="kop"><KON>und</KON></J>

      <praed real="mikro" type="E" dir="V"><V ID="Fin">
      <VV>leidet</VV></V></praed>

      <J IR="adv"><AP>doch</AP></J>

      auch nicht</lb>

<line n="14"/>

<lb n="2007">

      <J IR="rest"><AD>ohne</AD></J>

      <!--hier line-->jene flüchtige Empfindung des
      Scheins;</lb>
```

Propositions are marked by the Tag `<lb>`, line break. The first proposition in example (1) *denn er lebt* ("because he is living") has the status of a full syntactic clause. The second proposition *und leidet mit in diesen Scenen* ("and is also suffering in these scenes") contains an elliptical pronoun: the pronoun he is not realized. The third proposition *und doch auch nicht* ("and not also") contains an elliptical pronoun and an elliptical verb. Finally, in the last proposition *ohne jene flüchtige Empfindung des Scheins* ("without that transient sense of pretence") the proposition is realized by a prepositional phrase. Since the prepositional phrase contains the nominalization *Empfindung* "sense, feeling", from the verb *empfinden* "feel", it has propositional status.

Assuming that syntax provides structural means for conveying meaning, we then looked for semantic relationships between the propositions, and for the correlating syntactic means of marking these semantic relationships. We found it necessary to differentiate between semantic relationships on the one hand and syntactic means on the other, because semantic relationships can be real-

ized by different syntactic means, or even without any syntactic marking at all. In example 1, the semantic relationships are marked by different types of connectors (tagged by J for *Junktor*), i.e. *denn* "for" and *und* "and" as coordinating conjunctions; *doch* "though" as an adverbial connector, and *ohne* "without" as a prepositional connector. Example 2 provides an example of clause linking without any syntactic marking.

(2) Zimmer

Es sah furchtbar aus die Lokamatife sammt mehrere Wagens lagen zertrimmert da

[*it looked awful, the locomotive and several carriages were destroyed*]

```
<lb n="1002">
      <subj real="Pron">es</subj>
      <praed><V ID="Fin"><VV>sah</VV></V></praed>
      furchtbar
      <praed><VP>aus</VP></praed></lb>
<line n="9"/>
<lb IR="kaus" n="19,1002">
      <subj>die Lokamatife sammt mehrere Wagens</subj>
      <praed><V ID="Fin"><VV>lagen</VV></V></praed>
      <!--hier line-->zertrimmert da,</lb>
```

The second proposition explains why "it" looked awful, and we therefore assumed the existence of a causal relationship which was not marked syntactically.

The major principle of the annotation was to mark any grammatical features which were relevant in reconstructing classes of clause linking. That means that we did not annotate clause linking techniques such as subordination or coordination as such; instead, we annotated grammatical features such as type of connector (i.e. subordinating and coordinating connectors, and adverbial connectors) and the position of the finite verb, which is necessary to reconstruct the clause linking techniques. The grammatical features annotated are:

- propositions (= line breaks)
- semantic relations (= "IR" – *Inhaltsrelationen*)

- relations between propositions (indicated by "house numbers")
- subjects
- predicates (parts of predicates, types of verbs)
- connectors (subordinating and coordinating connectors, adverbial and prepositional connectors)
- correlates
- objects and adverbials (only if necessary for reconstruction of clause linking by ellipsis)

(3) Bauernleben

Obwohl die Schweden und Hesen eines Wesens war, so nam doch ein ider, was er bekomen konnte

[*Although the Swedish and the Hessians were of one type, each one took what he was able to get*]

```
<lb n="24,2172">
      <J IR="konz" norm="obwohl"><SUB>Obwol</SUB></J>
      <subj>die Schweden und Hesen</subj>
      <praed><NGr>eines Wesens</NGr></praed>
      <praed><V ID="Fin"><KV>war,</KV></V></praed></lb>
<lb n="24,2172">
      <KOR>so</KOR>
      <praed><V ID="Fin"><VV>nam</VV></V></praed>
      <J IR="adv"><AP>doch</AP></J>
      <subj>ein ider,</subj></lb><!--hier line-->
<line n="8"/>
<lb>
      <SUB IR="zero">was</SUB>
      <subj real="Pron">er</subj>
      <praed><V ID="Inf"><VV>bekomen</VV></V></praed>
      <praed><V ID="Fin"><MV>konte.</MV></V></praed>
      </lb>
```

Example 3 explains the general principles of the annotation scheme. It consists of three propositions. The first two propositions are linked by a concessive relation, marked by the concessive connector *obwohl* "although", and an adversative relation, marked by the adversative connector *doch* "though". The fact that the two propositions are linked semantically as described is marked by what we called "house numbers", the numbers attributed to the line breaks. Every linked pair of propositions is marked by its own house number. In example (3), the number 24 marks the concessive relationship between the two propositions, and the number 2172 indicates the adversative relation.[4] This was necessary because it is not always the case that semantically related propositions are realized adjacently, and because it can also be the case that more than two propositions are connected. The last proposition is a modifying clause. Modifying and complement clauses were not analyzed in the project because they do not represent semantic relations.

The syntactic technique linking the first two propositions may be called "resumptive subordination", according to König/van der Auwera (1988). This technique can be reconstructed by:

- the annotation of the subordinating connector *obwohl* "although"
- the annotation of the position of the finite at the end of the line break
- the annotation of the correlate *so* "thus" at the beginning of the following line break.

As explained before, the technique is not annotated as such, rather the grammatical features relevant for reconstructing the technique. That makes the annotation into an open tool, which can also be used for reconstructing other grammatical issues besides clause linking.

[4] For further information see http://www.uni-giessen.de/kajuk/dokumentationen/hausnummern.pdf.

4. Usage

Example (4) shows how the annotated corpus can be used for analyzing various grammatical phenomena. It illustrates the phenomenon "integrative vs. non-integrative ellipsis" and the way it was annotated in the project.

(4) Briefwechsel

Und Gott geben mach das Dier bei der Entbindung nichts weiter paßiert und glücklich die sache verleben machst

[*and may God ensure that nothing will happen to you at the childbirth and you will get through this issue happily*]

```
<lb n="169,3046">
      <J IR="kop"><KON>und</KON></J>
      <SUB IR="zero" norm="dass" type="E" dir="V">das</SUB>
      <subj>Gott</subj><!--hier Pagebreak-->
      <!--hier line 1-->
      <praed><V ID="Inf"><VV>geben</VV></V></praed>
      <praed><V ID="Fin"><MV>mach</MV></V></praed></lb>
<lb n="170,3047">
      <SUB IR="zero" norm="dass">das</SUB>
      <obj><obl>Dier</obl></obj>
      bei der Entbindung
      <subj>nichts</subj>
      weiter
      <praed><V ID="Fin"><VV>pa&szlig;iert</VV></V></praed>
      </lb>
<lb n="170,3047">
      <J IR="kop"><KON>und</KON></J>
      <SUB IR="zero" norm="dass" type="E" dir="V">das</SUB>
      <obj type="E" dir="V" change="dat-nom"><obl>Dier
      </obl></obj>
      gl&uuml;cklich
      die sache <!--hier line 2-->
      <praed><V ID="Inf"><VV>verleben</VV></V></praed>
      <praed><V ID="Fin"><MV>machst</MV></V></praed></lb>
```

In the second proposition, the pronoun *dier, you,* (dative case) refers to the person addressed. The pronoun *you* is elliptical in the following proposition. Due to the coordinate status of the ellipsispsis, the reference to the second person singular can be inferred from the previous proposition. But in contrast to regular, integrative ellipsis, the morphological categorization has changed from dative to nominative, and the syntactic function from object to subject. Such instances may be described as "non-integrative ellipsis" (German: *aggregative Koordinationsellipse*, cf. Hennig 2009b).

Table 2 shows the occurrences of non-integrative ellipsis in the corpus texts.

17th century				19th century				Total
immediacy			distance	immediacy			distance	
Güntzer	Bauern-leben	Söldner-leben	Thomasius	Zimmer	Koralek	Brief-wechsel	Nietzsche	
57	65	144	13	11	8	20	1	319
266			13	39			1	319

Table 2: Non-integrative ellipsis

Non-integrative ellipsis is more frequent in the 17th century than in the 19th. In both the 17th and 19th centuries, the extent of non-integrative ellipsis is also greater in texts of immediacy than in texts of distance. Non-integrative ellipsis can thus be shown to be both a historical phenomenon and a phenomenon of immediacy (Hennig 2009a: 156ff.).

5. Prospects

As I have tried to explain, the principle of annotating grammatical features which are relevant for reconstructing clause linking phenomena means that the corpus provides a wide range of grammatical annotations which go beyond the area of clause linking. For example, the corpus may also be used for analyzing syntactic functions, word order, or the structure of the verbal complex. There will soon be open access to the corpus on the *ANNIS* platform for historical corpora.[5]

[5] http://www.uni-giessen.de/kajuk/; http://www.sfb632.uni-potsdam.de/d1/annis/

References

Primary sources

Bauernleben I = (1636-1667 [1998]): Bauernleben im Zeitalter des Dreißigjährigen Krieges. Die Stausebacher Chronik des Caspar Preis 1636-1667. Ed. by Wilhelm A Eckhardt and Helmut Klingelhöfer, with an introduction by Gerhard Menk. (= Beiträge zur Hessischen Geschichte 13). Marburg/Lahn 1998: Trautvetter & Fischer Nachf., 38-69 and 93-101.

Briefwechsel V = (1871-1872 [1999]): "Wenn doch dies Elend ein Ende hätte": ein Briefwechsel aus dem Deutsch-Französischen Krieg 1870/71. Ed. by Isa Schikorsky. (= Selbstzeugnisse der Neuzeit 7). Köln etc. 1999: Böhlau, 99-126.

Güntzer I = Güntzer, Augustin (1657 [2002]): Kleines Biechlin von meinem gantzen Leben. Die Autobiographie eines Elsässer Kannengießers aus dem 17. Jahrhundert. Ed. by Fabian Brändle and Dominik Sieber. (= Selbstzeugnisse der Neuzeit 8). Köln/Weimar 2002: Böhlau, 40v-43v, 54r-63r, 63[a]r-65v, 78r-108r.

Koralek V = Koralek, Ottilie (1889-1890): Lamentatio intermissa I. Tagebucharchiv Emmendingen. Unpublished transcription (Hollmann), 35 and 43-76.

Nietzsche V = Nietzsche, Friedrich (1872 [1999]): Die Geburt der Tragödie. In: Nietzsche, Friedrich: Die Geburt der Tragödie. Unzeitgemäße Betrachtungen. Kritische Studienausgabe. Ed. by Giorgio Colli and Mazzino Montinari. [New edition of KSA Berlin/New York: de Gruyter 1967ff.]. München 1999: Deutscher Taschenbuchverlag, Sections 1-9 (25-67).

Söldnerleben I = (1625-1649 [1993]): Ein Söldnerleben im Dreißigjährigen Krieg. Eine Quelle zur Sozialgeschichte. Ed. by Jan Peters. (= Selbstzeugnisse der Neuzeit 1). Berlin 1993: Akademie-Verlag, 35-111.

Thomasius I = Thomasius, Christian (1696 [1968]): Ausübung der Sittenlehre. (Von der Artzeney wider die unvernünfftige Liebe und der zuvorher nöthigen Erkäntniss Sein Selbst oder Ausübung der Sittenlehre). With a preface by Werner Schneiders. [= Reprint of the edition Halle: Salfeld 1696]. Hildesheim: Olms 1968, 1. Hauptstück (1-36) and 10. Hauptstück (219-257).

Zimmer V = (1861-1864 [2001]): Michael Zimmer's Diary. Ein deutsches Tagebuch aus dem Amerikanischen Bürgerkrieg. Ed. by Jürgen Macha and Andrea Wolf. (= Sprachgeschichte des Deutschen in Nordamerika 1). Frankfurt a.M etc. 2001: Peter Lang, 12-15, 17-23, 25-31, 35-38, 42-49, 57-60, 102-105, 116-117.

Secondary literature

Ágel, Vilmos (2000): Syntax des Neuhochdeutschen bis zur Mitte des 20. Jahrhunderts. In: Besch, Werner/Betten, Anne/Reichmann, Oskar/Sonderegger, Stefan (eds.): Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung. 2nd ed. Vol. 2. (= Handbücher zur Sprach- und Kommunikationswissenschaft 2.2) Berlin/New York: de Gruyter, 1855-1903.

Ágel, Vilmos/Diegelmann, Carmen (2010): Theorie und Praxis der expliziten Junktion. In: Ágel, Vilmos/Hennig, Mathilde (eds.): Nähe und Distanz im Kontext variationslinguistischer Forschung. (= Linguistik – Impulse & Tendenzen). Berlin/New York: de Gruyter, 345-393.

Ágel, Vilmos/Hennig, Mathilde (eds.) (2006): Grammatik aus Nähe und Distanz. Theorie und Praxis am Beispiel von Nähetexten 1650-2000. Tübingen: Niemeyer.

Bühler, Karl (1934/1999): Sprachtheorie: die Darstellungsfunktion der Sprache. (= UTB für Wissenschaft: Uni-Taschenbücher 1159) Stuttgart: Lucius und Lucius.

Hennig, Mathilde (2009a): Nähe und Distanzierung. Verschriftlichung und Reorganisation des Nähebereichs im Neuhochdeutschen. Kassel: Kassel University Press.

Hennig, Mathilde (2009b): Aggregative Koordinationsellipsen im Neuhochdeutschen. In: Ziegler, Arne (ed.), with the collaboration of Christian Braun: Historische Textgrammatik und Historische Syntax des Deutschen. Traditionen, Innovationen, Perspektiven. 2 Vols.. Berlin/New York: de Gruyter, 937-963.

Hennig, Mathilde (2010): Elliptische Junktion in der Syntax des Neuhochdeutschen. In: Schmid, Hans Ulrich (ed.): Perspektiven der germanistischen Sprachgeschichtsforschung. (= Jahrbuch für germanistische Sprachgeschichte 1). Berlin/New York: de Gruyter, 76-103.

Hennig, Mathilde (2011): The notion of immediacy and distance. In: Franco, Mario/Sieberg, Bernd (ed.): Proximidade e Distância. Estudos sobre a Língua e a Cultura. Lisbon: Universidade Católica Editora, 15-32.

Koch, Peter (1997): Orality in literate cultures. In: Pontecorvo, Clotilde (ed.): Writing development: an interdisciplinary view. Amsterdam/Philadelphia: Benjamins, 149-171.

Koch, Peter/Oesterreicher, Wulf (1985): Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. In: Romanistisches Jahrbuch 36: 15-43.

König, Ekkehard/van der Auwera, Johan (1988): Clause integration in German and Dutch conditionals, concessive conditionals, and concessives. In: Haiman, John/Thomson, Sandra (eds.): Clause combining in grammar and discourse. (= Typological studies in Language 18). Amsterdam/Philadelphia: Benjamins, 101-133.

Oesterreicher, Wulf (1997): Types of orality in text. In: Baker, Egbert/Kahane, Ahuvia (eds.): Written voices, spoken signs: tradition, performance and the epic text. Cambridge, MA/London: Harvard University Press, 190-214.

Paul, Hermann (1989): Mittelhochdeutsche Grammatik. 23. ed., revised by Peter Wiehl and Siegfried Grosse. (= Sammlung kurzer Grammatiken germanischer Dialekte A. Hauptreihe 2). Tübingen: Niemeyer.

Raible, Wolfgang (1992): Junktion. Eine Dimension der Sprache und ihre Realisierungsformen zwischen Aggregation und Integration. (= Sitzungsberichte der Heidelberger Akademie der Wissenschaften, Philosophisch-historische Klasse). Heidelberg: Winter.

Reichmann, Oskar/Wegera, Klaus-Peter (eds.) (1993): Frühneuhochdeutsche Grammatik. (= Sammlung kurzer Grammatiken germanischer Dialekte A. Hauptreihe 12). Tübingen: Niemeyer.

Schrodt, Richard (2004): Althochdeutsche Grammatik II: Syntax. (= Sammlung kurzer Grammatiken germanischer Dialekte A. Hauptreihe 5,2). Tübingen: Niemeyer.

BRYAN JURISH / MARKO DROTSCHMANN / HENRIETTE AST

Constructing a canonicalized corpus of historical German by text alignment

Abstract

Historical text presents numerous challenges for contemporary natural language processing techniques. In particular, the absence of consistent orthographic conventions in historical text presents difficulties for any system requiring reference to a static lexicon indexed by orthographic form. Canonicalization approaches seek to address these issues by assigning an extant equivalent to each word of the input text and deferring application analysis to these canonical forms. Quantitative evaluation of canonicalization techniques in terms of precision and recall requires reference to a ground-truth corpus in which the canonical form for each corpus token has been manually verified, but such manually annotated corpora are difficult to come by and in general both costly and time-consuming to create. In this paper, we describe a method for bootstrapping a ground-truth canonicalization corpus with minimal manual annotation effort by means of automatic alignment of historical texts with current editions of the same texts, coupled with a two-phase manual review process.

1. Introduction

Virtually all conventional text-based natural language processing techniques 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. Unconventional 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 et al. 2005; Gotscharek et al. 2009a, b; Reffle et al. 2009; Jurish 2010, 2011) 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.

A quantitative evaluation of any canonicalization technique in terms of the information retrieval notions of precision and recall requires reference to both *retrieval* and *relevance* relations over corpus target items (tokens or types). In general, a retrieval relation can be defined for any elementary canonicalization function f as the equivalence kernel $\sim_f = f \circ f^{-1} = \{(x, y) : f(x) = f(y)\}$;

for a given query, all and only those items are retrieved which share a canonical form with that query. The relevance relation however should be independent of the canonicalization technique chosen, in order to ensure comparability of evaluation results for different canonicalization methods. Further, relevance should be determined as far as possible by manual inspection in order to avoid confounding evaluation results and to ensure that problematic phenomena such as lexical ambiguity are adequately accounted for. Clearly, these desiderata would be fulfilled by a ground-truth corpus of historical text in which the canonical form for each corpus token has been manually verified, thus providing both canonicalization input data (the original corpus text) as well as a relevance relation (the equivalence kernel for the manually determined canonical forms), but such manually annotated corpora are difficult to come by and in general both costly and time-consuming to create.

In this paper, we describe a method for constructing such a ground-truth canonicalized corpus with minimal manual annotation effort using automatic text alignment coupled with a two-phase manual review process. The core intuitions underlying our approach can be summarized as follows:

(1) When they exist, contemporary editions of historical texts already incorporate the desired relevance relation; and

(2) since language change is a comparatively slow process, only a small subset of the relevance relation can be expected to consist of "interesting" non-identity pairs.

Intuition (1) suggests that we can extract the desired relevance relation by aligning a historical text with a contemporary edition of the same text, while intuition (2) can be used to guide the alignment process by attempting to maximize the number of identity alignments.

2. Construction

2.1 Sources

We applied our construction to a prototype corpus of 13 volumes of historical German text published between 1780 and 1880 (Table 1).[1] The text of the historical editions was drawn from the *Deutsches Textarchiv* ("German Text Archive"; Geyken/Klein 2010),[2] encoded according to the Text Encoding Initiative (TEI) P5 Guidelines.[3] Contemporary editions of the selected volumes were provided by the online libraries *Project Gutenberg*[4] and *Zeno*.[5]

N	Text
12405	C. Brentano: *Geschichte vom braven Kasperl und dem schönen Annerl*. Berlin: Vereinsbuchhandlung, 1838.
1865	W. Busch: *Max und Moritz*. München: Braun & Schneider, 1865.
14490	J. W. von Goethe: *Iphigenie auf Tauris*. Leipzig: Göschen, 1787.
42970	J. W. von Goethe: *Wilhelm Meisters Lehrjahre*. Bd. 1. Berlin: Unger, 1795.
43933	J. W. von Goethe: *Wilhelm Meisters Lehrjahre*. Bd. 2. Berlin: Unger, 1795.
45255	J. W. von Goethe: *Wilhelm Meisters Lehrjahre*. Bd. 3. Berlin: Unger, 1795.
63215	J. W. von Goethe: *Wilhelm Meisters Lehrjahre*. Bd. 4. Berlin: Unger, 1796.
24771	J. W. von Goethe: *Torquato Tasso*. Leipzig: Göschen, 1790.
3164	I. Kant: „Beantwortung der Frage: Was ist Aufklärung?" In: *Berlinische Monatsschrift*, 1784, H. 12, S. 481-494.
5925	G. E. Lessing: *Die Erziehung des Menschengeschlechts*. Berlin: Voss, 1780.
30922	F. Schiller: *Kabale und Liebe*. Mannheim: Schwan, 1784.
50697	J. Spyri: *Heidi's Lehr- und Wanderjahre*. Gotha: Perthes, 1880.
9702	T. Storm: *Immensee*. Berlin: Duncker, 1852.

Table 1: Historical source texts used to construct the prototype corpus

[1] The 13-volume prototype corpus represents only a small portion of an ongoing corpus construction project using the methods described here. We use the 13-volume corpus as an example throughout this article because it represents the most thoroughly annotated subset of the corpus at the time of writing.

[2] http://deutschestextarchiv.de

[3] http://www.tei-c.org/Guidelines/P5/

[4] http://www.gutenberg.org

[5] http://www.zeno.org

The raw historical corpus was heuristically tokenized into 417,249 tokens of 30,101 distinct surface types in 20,872 sentences. Of these, 349,541 tokens (84%) of 28,146 distinct surface types (94%) contained only alphabetic and hyphenation characters and were thus considered "word-like".

2.2 Text Alignment

The first phase of the construction process is the heuristic alignment of a historical source text with a contemporary edition of the same text. The contemporary edition or "target text" is assumed to adhere to contemporary orthographic conventions, and the purpose of the alignment phase is to extract a significant portion of the canonicalization relevance relation exhibited by the editorial changes in the target text. Effectively, the alignment of source and target texts should bootstrap a relevance relation based on the linguistic competence of the human editor(s) responsible for the contemporary edition.

Input to the alignment phase were pairs of files representing the raw historical source and contemporary target editions of each corpus text. The optimal alignment itself was computed by GNU `diff` (Hunt/McIlroy 1976; MacKenzie et al. 2003) under the '`--minimal`' switch,[6] which returns an alignment based on the longest common subsequence (LCS) for the given source and target texts. Since `diff` aligns its argument files line-by-line, both source and target texts were heuristically tokenized into a one-word-per-line format before alignment in order to abstract over differences in formatting. Additionally, since `diff` aligns input lines exclusively on the basis of surface string identity, the conservative transliteration function from Jurish (2010) was applied to the historical text to help account for extinct graphemes.[7]

The initial alignment returned by `diff` is a sequence of *hunks*, where each hunk is either:

- an *identity hunk*, a string of adjacent (transliterated) tokens occurring in both source and target texts;[8]

[6] The GNU `diff` manual glosses this option as "Try hard to find a smaller set of changes".

[7] In particular, the transliterator was responsible for mapping the historical long 'ſ' to a conventional round 's', as well as superscript 'e' to the conventional *Umlaut* diacritic '◌̈', as in the transliteration *Abſtände* ↦ *Abstände* ("distances").

[8] Strictly speaking, `diff` does not output identity hunks at all. The location and content of identity hunks can however easily be reconstructed from the line addresses associated with the adjacent non-identity hunks.

- a *deletion hunk*, a string of tokens occurring only in the source;
- an *insertion hunk*, a string of tokens occurring only in the target; or
- a *change hunk*, equivalent to a simultaneous deletion and insertion with no intervening identity hunk.

Insertion hunks were ignored during the alignment phase. Identity hunks were marked as valid canonicalizations and copied verbatim to the aligned output file, in accordance with intuition (2) from Section 1.[9] For each token in a deletion hunk, a corresponding token with an empty canonical form was included in the output file, where it was marked as unaligned and therefore requiring further manual attention.

In order to extract potential canonicalizations beyond those for which strict identity of (transliterated) string forms applies, each change hunk was inspected more closely. First, each change hunk was tested for identity modulo token boundaries in order to accommodate common concatenative morphological phenomena such as exhibited by the canonicalizations *zwei und vierzig* ↦ *zweiundvierzig* ("forty-two") or *allzuweit* ↦ *allzu weit* ("all too far"). Change hunks which were entirely accounted for by identity of concatenated transliterated forms were flagged as such and accepted with the corresponding substring identities into the output file.[10]

Each remaining change hunk was passed to an additional fine-grained alignment subroutine using the Wagner-Fischer (1974) algorithm for computing string edit-distance (Levenshtein 1966) to align the deletion and insertion portions of the change hunk on the character level. The resulting character-wise alignment was used to determine the most likely target word in the insertion portion for each source word in the deletion portion, using a scoring function based on the empirical probability of a (case-insensitive) match operation per source token character. Word alignments thus extracted from change hunks

[9] Violations of intuition (1) arising e.g. from use of non-standard orthography in both source and target texts will therefore result in spurious identity canonicalizations during this phase. Minor violations of intuition (2) such as might arise from a highly deviant source text will only increase the required manual annotation effort, while major violations of intuition (2) stemming e.g. from major grammatical discrepancies between source and target texts will cause the construction as given to fail, since an adequate treatment of these would require more sophisticated alignment techniques than a simple LCS-based method can provide.

[10] Such treatment is justified to the extent one assumes (as we do) that despite diachronic changes in word boundary placement, the historical forms remain compositionally grammatical.

were copied as candidate canonicalizations to the output file, but flagged as non-identity alignments in need of further attention.

The final output of the text alignment phase for each pair of source and target files was a single XML file containing one token for each token of the source text. Each output token was assigned attributes for both the original source string (before transliteration) and the aligned target word string (if any), in addition to the administrative flags described above.

2.3 Manual Annotation

The automatic text alignment procedure discovered candidate canonicalizations for over 98% of word-like input tokens. Of these, over 77% were literal identity pairs and over 94% were identical after transliteration. Even accepting the validity of the transliterated-identity alignments,[11] we are still left with 23,205 word-like tokens requiring human attention. While this represents a substantial reduction in required manual annotation effort with respect to the full 349,541-word corpus, the situation can be further improved by splitting the manual annotation process into *type-wise* and *token-wise* phases.

2.3.1 Type-wise Confirmation

Natural language text is known to obey Heaps' Law (Heaps 1978; Baeza-Yates/Navarro 2000), a correlate of the more widely known Zipf rank-frequency correlation (Zipf 1949; van Leijenhorst/van der Weide 2005; Lü et al. 2010). The former empirical law states that there is a log-linear correlation between vocabulary size in types and corpus size in tokens. In the current context, Heaps' Law implies that a comparatively small number of alignment word-pair types can be expected to account for a large portion of the candidate tokens discovered by the alignment phase. Moreover, an incremental corpus construction process can be expected to encounter ever fewer novel candidate alignment types as the number of aligned tokens increases.

The next step toward minimizing the manual annotation effort required by our corpus construction is therefore a type-wise manual confirmation phase. In this phase, a human annotator is presented with a series of (*source* ↦ *target*) word-pair types representing candidate canonicalizations discovered by

[11] Note that automatic alignment with a contemporary text should serve to minimize any bias introduced by the transliteration function, since the contemporary target text provides independent evidence for any transliterations which are accepted by this heuristic.

the alignment phase, and is asked to decide for each presented type whether or not the given *target* word is to be considered a valid equivalent contemporary form for the given *source*. Each alignment type is presented at most once,[12] and the annotator's decisions are saved to a persistent database and re-used for each newly aligned text, so that the effort required for type-wise confirmation decreases as the corpus grows.

Since each decision regarding the validity of an alignment type is final, achieving our goal of a high-quality output corpus suitable for use as a ground-truth relevance relation means that great care must be taken to ensure that the decisions made at this stage are based on conservative criteria. As an example, consider the canonicalization candidate (*über* ↦ *aber:* "over" ↦ "but"): the heuristics used by the text alignment phase can easily suggest the alignment of these two types by virtue of their common string suffix *-ber*, but given the high frequencies of the closed-class words involved, the potential for spurious alignments of the corresponding types is very great indeed.

For this reason, type-wise annotators were instructed to accept only those proposed alignments of which they were certain. Additional guidelines given to the type-wise annotators included the instructions:

(1) In general, accept changes in letter case and common historical allographs; e.g. accept any of the source forms *Bei, Bey, bei,* or *bey* for the target word *bei* ("by").

(2) Reject alignments involving a change in lexical root, part-of-speech, or morphosyntactic features; e.g. (*das* ↦ *dass:* "the" ↦ "that"), (*Ewigkeiten* ↦ *Ewigkeit:* "eternities" ↦ "eternity").

(3) Reject alignments of suspected graphical origin such as printing-, OCR-, or transcription errors; e.g. (*Gerechtigkelt* ↦ *Gerechtigkeit:* "justice"), (*zuükhalten* ↦ *zurückhalten:* "hold back").

(4) Reject alignments in which the proposed target is itself archaic or extinct; e.g. (*danach* ↦ *darnach:* "afterwards"), (*Licht* ↦ *Lichte:* "light") – the respective inverse alignments would however be acceptable.

(5) Reject alignments whose source components are surface-identical to non-equivalent contemporary words. This criterion applies chiefly

[12] Identity alignments, identity-of-transliteration alignments, and unaligned source words are not presented at this stage.

to ambiguities involving the archaic dative *-e* suffix and contemporary plurals; e.g. (*Orte* ↦ *Ort:* "place(s)"), (*Lande* ↦ *Land:* "land(s)").

(6) Reject alignments of proper names which involve any graphematic changes beyond transliteration of extinct characters, e.g. (*Franciska* ↦ *Franziska*) and (*Oehi* ↦ *Öhi*), but (*Göthe* ↦ *Goethe*) is allowed.

For the prototype corpus described in Section 2.1, the 23,205 unconfirmed token alignments were reduced to a set of 7,166 alignment pair types of which only 5,780 elements representing 17,839 tokens corresponded to successful alignments arising from change hunks whose source and target components were not surface-identical modulo transliteration. Of these, 4,483 alignment types (77%) representing 16,083 tokens (90%) were accepted in the type-wise confirmation phase, thus eliminating over 69% of the remaining uncanonicalized tokens by manually inspecting less than one quarter of the available unconfirmed items.

The annotation effort required for type-wise confirmation was estimated by explicitly measuring the time needed for confirmation of a random sample of 100 corpus types. Annotation of the sample proceeded at an average confirmation rate of 3.95 seconds per pair, corresponding to a projected total annotation time of about 6.3 hours for the entire corpus. In terms of the original input corpus size, the type-wise confirmation phase proceeded at an estimated rate of over 15 words per second, so the corpus construction up to and including the type-wise confirmation phase does indeed display a very high throughput.

2.3.2 Token-wise review

Although the combination of automatic text alignment and type-wise manual confirmation is able to provide canonicalizations for the vast majority of input tokens (ca. 98%) with only very little manual annotation effort, a small fraction of input tokens do remain unaccounted for by these techniques. These as-yet uncanonicalized words however are likely to be of particular interest for diachronic corpus-based studies since they include those canonicalization patterns which cannot be reduced to simple string identities or common "run-of-the-mill" allography relations, as well as those which involve ambiguities with valid contemporary forms. In order to achieve a more accurate model of the canonicalization relevance relation, we therefore introduced an additional

manual review phase for direct annotation of canonical cognates for as-yet uncanonicalized word-like tokens in sentential context.

Not all of the uncanonicalized tokens returned by the type-wise confirmation phase represent "interesting" non-trivial canonicalization patterns, however. In particular, editorial changes to the original text involving front or back matter, marginalia, speaker designations or stage directions were purged from the corpus by means of a simple XPath filter. Later investigations showed that in some cases – especially in verse collections – chunks of source text spanning multiple pages failed to be automatically aligned at all, usually due to heavy editorial intervention (re-ordering) in the contemporary edition. An additional filter was developed to heuristically detect and remove such unaligned chunks from the corpus using a moving window of n=3 sentences and a minimal alignment threshold of p=75%. It was also noted that the change-hunk-internal heuristic scoring function used in the text alignment phase often failed for short closed-class words such as *der* ("the"), *und* ("and"), or *nicht* ("not"), causing an inordinate inflation of uncanonicalized tokens due to these words' high frequencies. For this reason, a lexicon of 213 high-frequency closed-class items and appropriate canonicalizations was created and applied to the uncanonicalized portion of the corpus.

After pruning and application of the closed-class exception lexicon, the corpus contained a total 405,150 tokens of which 341,798 (84%) were "word-like". Of these, only 3,476 (1.1%) were uncanonicalized. The pruning and closed-class lexicon heuristics together eliminated over half of the remaining uncanonicalized tokens by discarding a mere 2.2% of word-like corpus material as "uninteresting". The pruned corpus was separated into blocks of roughly ten pages which were then randomly sorted and concatenated into a single corpus file for token-wise annotation.

Token-wise annotation itself was performed using a dedicated graphical interface in conjunction with the character-level text-to-image coordinate mapping used by the *Deutsches Textarchiv* online corpus search utility. The annotator was presented with each as-yet uncanonicalized token together with its immediate sentential context in document order, and was asked to assign each such token a lexically equivalent extant cognate. If an automatically discovered alignment was present for the token, it was presented as the default canonical form. The annotator was also asked to provide additional administrative data for each canonicalization if and when appropriate, specifically:

- Whether the token presented is in fact a valid token, or whether it instead represents an error on the part of the heuristic tokenizer.
- Whether the sentence containing the token presented is in fact a valid sentence-like unit, or whether it represents a tokenization error.
- Which of a set of eight pre-defined coarse-grained lexical classes the current token is to be considered an instance of. The set of lexical classes from which the annotator could choose were:

  LEX: a "normal" lexical word; this was the default class assigned if no other class was explicitly chosen.

  JOIN: used together with sentence-level attributes to indicate a string of multiple source tokens to be canonicalized into a single target token. The annotator was additionally asked to map the individual source tokens to compositionally plausible contemporary equivalents where possible.

  SPLIT: used together with an auxiliary target attribute to indicate a single source token to be canonicalized into multiple target tokens. The annotator was additionally asked to map the source token to a single compositionally plausible (e.g. hyphenated) target token where possible.

  FM: foreign-language material.

  GONE: an extinct lexeme without any contemporary cognate.

  GRAPH: an error of graphical origin.

  NE: a proper name, e.g. a person or place name.

  BUG: an encoding error in the source corpus.

Canonical cognates were determined by direct etymological relation of the source root in addition to matching morphosyntactic features. Proper names were canonicalized in accordance with guideline (6) from Section 2.3.1. Otherwise, proper names, extinct lexemes, and foreign-language material were treated as their own canonical cognates. Problematic tokens were explicitly marked as such and later subjected to review by an expert.

For efficient annotation of (potentially ambiguous) medium- and high-frequency words, the interface supported batch-level edit operations with optional user selection of target tokens based on a fixed-width context window. As

additional visual aids, the annotator was presented with colour-coded "traffic light" status frames for the current source and target forms which indicated whether or not the corresponding word was known to the high-coverage TAGH morphology for contemporary German (Geyken/Hanneforth 2006), and whether or not it satisfied a set of morphological security heuristics (Jurish 2011: A.4). Finally, each edit operation was logged together with its timestamp and the annotator's user-name to a local history list in order to provide basic revision control functionality.

Of the 3,746 uncanonicalized tokens passed into the token-wise review phase, 3,263 (87%) were directly assigned canonical cognates by the original annotator, and the remaining 483 (13%) were flagged and subjected to expert review. 43 word-like tokens and 102 sentences were marked as tokenization errors. Since only complete sentences containing no invalid tokens were included in the final output corpus, tokenization errors resulted in the elimination of 2,827 word-like tokens (<1%) from the corpus. The distribution of the lexical classes assigned to the annotated tokens is given in Table 2.

Class	N	% Edited
LEX	2684	59.22 %
NE	874	19.29 %
JOIN	792	17.48 %
GRAPH	101	2.23 %
SPLIT	72	1.59 %
BUG	40	0.88 %
GONE	8	0.18 %
FM	1	0.02 %

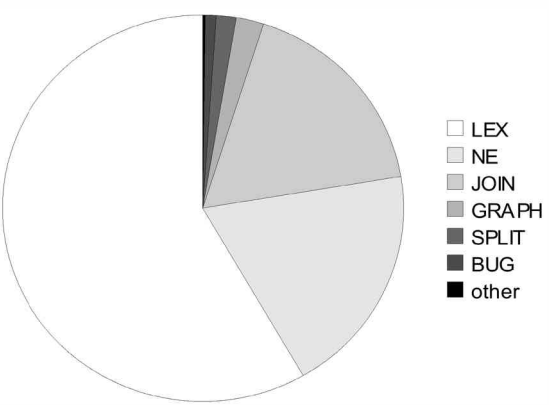


Table 2: Distribution of word classes assigned during token-wise review

Annotation effort was estimated using the intervals between timestamps associated with each manual edit operation. Edit intervals of less than 1 second or greater than 30 minutes were ignored for purposes of the computation. The original annotator applied 5,253 edit operations[13] in editing sessions totaling

[13] Multiple edit operations were applied to some tokens, and 786 tokens were edited which had already been canonicalized in the type-wise alignment phase. The latter cases were for the most part batch operations which set administrative flags.

55.9 hours. Expert review involved 964 edit operations in sessions totaling 11.7 hours. The manual annotation effort for the token-wise review phase was therefore 67.6 hours, and the total manual annotation effort for the entire corpus was only 74 hours, roughly 2 full-time work weeks. This corresponds to an average throughput of about 1.3 words per second for the whole prototype corpus from start to finish.

3. Conclusion

We have presented a method for constructing a ground-truth corpus of canonicalized historical text with minimal manual annotation effort using automatic text alignment techniques coupled with a two-phase manual review process. Automatic text alignment with a contemporary edition provided an efficient means of discovering non-trivial historical spelling variants, and allowed the subsequent manual review process to draw on the linguistic intuitions of the contemporary edition's editor(s). Manual review was divided into a conservative type-wise confirmation phase and a subsequent token annotation phase in order to leverage the logarithmic growth of vocabulary size for natural language text conforming to Heaps' Law. We estimated an annotation rate of approximately 1.3 words per second for a fully annotated corpus of 13 volumes of 18th-19th century German text.

The 13-volume corpus described above constitutes only the initial portion of an ongoing corpus construction project. We are currently working on incrementally extending the canonicalized corpus using the methods described here based on the historical texts from the *Deutsches Textarchiv*. At the time of writing, an additional 116 volumes containing 5,843,664 tokens in 286,091 sentences have been automatically aligned and passed through the type-wise confirmation phase, requiring manual annotation of an additional 58,644 alignment pair types. Of these, 3,730,781 tokens in 177,390 sentences have also passed through the initial token-wise annotation phase and are awaiting expert review.

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 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.

References

Baeza-Yates, Ricardo/Navarro, Gonzalo (2000): Block-addressing indices for approximate text retrieval. In: Journal of the American Society for Information Science (JASIS) 51(1): 69-82.

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.

Geyken, Alexander/Klein, Wolfgang (2010): Deutsches Textarchiv. In: Jahrbuch 2009 der Berlin-Brandenburgischen Akademie der Wissenschaften. Berlin: Akademie Verlag, 320-323.

Gotscharek, Annette/Reffle, Ulrich/Ringlstetter, Christoph/Schulz, Klaus U. (2009a): On lexical resources for digitization of historical documents. In: Proceedings of DocEng '09. New York: ACM, 193-200.

Gotscharek, Annette/Neumann, Andreas/Reffle, Ulrich/Ringlstetter, Christoph/Schulz, Klaus U. (2009b): 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.

Heaps, Harold S. (1978): Information Retrieval: Computational and Theoretical Aspects. Orlando: Academic Press.

Hunt, James W./McIlroy, M. Douglas (1976): An algorithm for differential file comparison. (= Computing Science Technical Report 41). Murray Hill, NJ: Bell Laboratories.

Jurish, Bryan (2010): More than words: Using token context to improve canonicalization of historical German. In: Journal for Language Technology and Computational Linguistics 25(1): 23-40.

Jurish, Bryan (2011): Finite-state canonicalization techniques for historical German. PhD thesis, Universität Potsdam.

Levenshtein, Vladimir I. (1966): Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet Physics Doklady 10: 707-710.

Lü, Linyuan/Zhang, Zi-Ke/Zhou, Tao (2010): Zipf's Law leads to Heaps' Law: Analyzing their relation in finite-size systems. In: *PLoS ONE*, 5(12): e14139.

MacKenzie, David/Eggert, Paul/Stallman, Richard (2003): Comparing and merging files with GNU Diff and Patch. Bristol: Network Theory Ltd.

Rayson, Paul/Archer, Dawn/Smith, Nicholas (2005): VARD versus Word: A comparison of the UCREL variant detector and modern spell checkers on English historical corpora. In: Proceedings of the Corpus Linguistics 2005 conference, Birmingham, July 14-17.

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: 165-174.

van Leijenhorst, Dirk C./van der Weide, Theo P. (2005): A formal derivation of Heaps' Law. In: Information Sciences 170(2-4): 263-272.

van Rijsbergen, Cornelis J. (1979): Information retrieval. Newton, MA: Butterworth-Heinemann.

Wagner, Robert A./Fischer, Michael J. (1974): The string-to-string correction problem. In: Journal of the ACM 21(1):168-173.

Zipf, George K. (1949): Human behaviour and the principle of least-effort. Cambridge, MA: Addison-Wesley.

SONJA LINDE / ROLAND MITTMANN

Old German reference corpus: digitizing the knowledge of the 19th century

Automated pre-annotation using digitized historical glossaries

Abstract

The project *Referenzkorpus Altdeutsch* ('Old German Reference Corpus') aims to establish a deeply-annotated text corpus of all extant Old German texts. In order to minimize manual work, an important target was to automate the retrieval of as much data as possible from existing sources. Whilst the texts themselves were already available in a digital form, the annotation data could to a large extent be found within a set of glossaries associated with each text. After digitizing these, the information contained in them could be automatically and semi-automatically linked to the texts. Subsequent manual editing focuses on any remaining gaps and misattributions, rejecting inapplicable alternatives and adjusting details to the annotation standards of the project. Throughout the process, various problems have been encountered that require special attention to find particular solutions.

1. Introduction

Creating a linguistically annotated corpus of texts in a historical language is a task that seems to imply a huge amount of manual annotation work. For every single word, the required additional data purportedly has to be collected or abstracted from grammars and glossaries, keeping those involved unnecessarily busy looking things up, and spending their time dealing with similar or even identical cases again and again. Thankfully, modern technical facilities allow us to digitize the secondary resources needed, and to automate both the gathering of information and its assembly into a reasonably searchable data structure. Thus, using the precise information in the authoritative glossaries and grammars for the Old German texts, in combination with the technical developments of the last few decades, the DFG-funded research project *Referenzkorpus Altdeutsch*[1] ('Old German Reference Corpus') aims to produce a deeply-annotated corpus of all preserved texts from the oldest stages of German (Old High German and Old Saxon), which date from ca. 750 to 1050 CE.

[1] http://www.deutschdiachrondigital.de

The glossaries, many of which are themselves over a hundred years old, preserve detailed knowledge of the scholars of their era, and constitute a rich field of information waiting to be exploited systematically.

Comprising a total of 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 wholly in an Old German language, which are mainly incantations. The translations are mainly of religious literature, prayers and hymns, but some also relay the writings of ancient authors and scientific writings. The largest coherent subcorpora are the Old High German works of Notker Labeo and Otfrid of Weissenburg (*Evangelienbuch*), an Old High German translation of the gospel harmony of Tatian the Assyrian (*Diatessaron*) and the Old Saxon gospel harmony now known as the *Heliand*. Edited versions of all texts exist in print; they have been digitized by the *TITUS*[2] project.

The following sections describe the approach used in gathering the data needed – the texts themselves and the glossaries – and in combining this information in order to create a deeply-annotated corpus. The paper then focuses on the limitations of the automatic data linking, and the remaining manual work. A final short section deals with the specific problem of creating a standardized version of the corpus.

2. Digitization and automated pre-annotation

Every text in the corpus is to be included in three ways: 1) a close transcription of the manuscript, 2) a scholarly edited version and 3) a standardized version. Thus, the corpus comprises a near approximation to a diplomatic variant, a second variant reflecting the most appropriate printed edition, and a third variant representing standardized morphology and orthography, in accordance with the forms given in Splett (1993). For instance, a word form from the beginning of the *Sangaller Credo* transmitted as *almah/ticum* in the manuscript[3] and given as *almahticun* in the printed edition will thus be complemented by a standardized form *alamahtīgun*, created from the standard lemma *alamahtīg* (Splett 1993: 582) plus the corresponding inflectional ending *-un* (Braune 2004: 226).[4] Each of the three text versions is to be included twice –

[2] Thesaurus of Indo-European Text and Language Materials, http://titus.uni-frankfurt.de

[3] cf. http://titus.uni-frankfurt.de/texte/etcs/germ/ahd/klahddkm/klahd005.htm

[4] As Splett (1993) uses Upper German word forms, the ending *-un* instead of Middle German *-on* was chosen, cf. Braune (2004: 207).

once with word tokens as the smallest unit, and once subdivided into single, aligned letters to enable phonological and morphological research (cf. also Figure 4).

The printed edition version is imported from the *TITUS* website,[5] together with indications of the language of every word form (Old High German, Old Saxon or Latin, with a few exceptions), and information about the subdivision of the text. This information provides the basic structure for the data. In addition, the corpus will contain standardized lemmata, with their translations, for each word token, as well as information on the morphological features of the word form and its lemma, on phrase types and, if applicable, on rhyme schemes.

```
- <entry>
    <lem>got</lem>
    <pos>st. m.</pos>
    <trlat>deus (dominus)</trlat>
  - <case>
      <form>nom.</form>
    - <inst>
        <rec>1, 1</rec>
        <rec>4, 14</rec>
        <rec>5, 9</rec>
        <rec>13, 14</rec>
        <rec>21, 7 (3)</rec>
        <rec>etc.</rec>
      - <rem>
          <com>zus. 28 mal</com>
        </rem>
      </inst>
    - <inst>
        <expr>got Abrahames (Isakes)</expr>
        <rec>127, 4</rec>
      </inst>
```

gomman - barn *st. n. männliches Kind, masculinum: nom. sg.* 7, 2.
gomo *sw. m. im Compos.* brûti-gomo.
got *st. m. deus (dominus): nom.* 1, 1. 4, 14. 5, 9. 13, 14. 21, 7 (3) *etc.* (*zus.* 28 *mal*). got Abrahames (Isakes) 127, 4. got totero 127, 4. truhtin got Israhelo (unser) 4, 14. 128, 2. *voc.* got 118, 2. 3. got min 207, 2 (2). min got 233, 7. *gen.* gotes 82, 9. 90, 4. 126, 3. 244, 2; *vgl.* 4, 18.

Figure 1: Detail from Sievers (1892), XML (l.) and original (r.) versions

The whole corpus is covered by a range of existing glossaries,[6] each of which covers a section of the corpus, and the structure of these needs to be closely scrutinized. Typically, these glossaries give the lemmata together with their translations and at least some of the morphological features specific to each lemma (see Figure 1). The entries for lemmata that cover several parts of speech, or possess a number of semantic nuances, are subdivided according to these aspects. The entries are completed by a list of all attestations, sorted according to the specific morphological features of each attestation, and followed by a reference to their location within the text. In a few instances, mainly in the

5 http://titus.uni-frankfurt.de/texte/texte2.htm#ahd and #asachs

6 Heffner (1961), Hench (1890), Hench (1893), Kelle (1881), Sehrt (1955), Sehrt (1966), Sievers (1874), Sievers (1892) and Wadstein (1899).

case of very frequent lemmata, only a selection of attestations is listed. The glossaries were digitized into an XML format. This had to be completed manually, as current OCR programs have difficulty recognizing older fonts.

The required data are retrieved and stored in a file containing all attestations, with reference to their location, as well as the corresponding lemma and the morphological features pertaining to the particular lemma and word form in context (see Figure 2).[7] The file also gives additional grammatical information not contained in the glossaries, for example a more exact indication of the inflectional class. The information for this is obtained manually from grammars of the relevant language.[8] Within the file, all part-of-speech and inflectional information is transferred into the standard of the *Deutsch-Diachron-Digital-Tagset* (*DDDTS*), a tagset developed by the project specifically for use with the historical stages of German and Latin, and built on the basis of the *Stuttgart-Tübingen-Tagset* (*STTS*)[9] for modern (i.e. New High) German. As we aim to list the correct record for every lemma, the fact that most glossaries only give the records within their context presents a further challenge. Every word of the given phrase has to be compared to the lemma – and, in case of suppletion, to its other stems – to determine the correct correspondence.

```
Lem | Lem2 | Lem3 | PoS | Flex | Form | Expr | Expr2 | Rec
       Lemma   DDDTS   Lemmabezug     Belegbezug     Flexion

&ecirc;uuo | ēuuo | euuo | m. |  | acc. sg. | ęuun | ęuun | 37, 17
       NA      NA      n_Masc         n_Masc          Sg_Acc
```

Figure 2: Title and sample line from glossary data file on *Isidor* (cf. Hench 1893: 138)

In the case of the Old High German texts, the various forms of each lemma are given in a unified form corresponding to the entry in Splett (1993), which covers the whole Old High German lexicon using a standardized orthography. Automatically generated lists of lemmata from all the Old High German glossaries listed are expanded by giving the form and the translation found in Splett (cf. Figure 3). This is done by hand because of copyright restrictions. However, we deviate from Splett's practice in that ‹e› unaffected by umlaut is marked as ‹ë›, and fricative ‹z› as ‹ʒ›, in order to separate these pairs of pho-

[7] However, not all dictionaries give record-specific morphological features, cf. e.g. Kelle (1881).

[8] Braune (2004) and Gallée (1993).

[9] Schiller et al. (1999).

nemes according to the orthography used in, for instance, Braune (2004). This task can also only be automated to a limited extent and therefore has to be completed manually. For Old Saxon, Sehrt (1966) serves as a standard, since Tiefenbach (2010) was published too late to be taken into account.

```
uuazar wa33ar 'Wasser, Gewässer, Meer'
```

Figure 3: Sample line from lemma concordance file on the *Monsee Fragments* (cf. Hench 1890: 205 and Splett 1993: 1073)

A subsequent program then links the pre-processed glossary data file and the lemma concordance file to the *TITUS* text. This program matches every word of the text with the records in the glossary data file. It then goes on to retrieve the corresponding dataset, or a range of datasets, if it is not possible to allocate it unambiguously. For manual processing, it proved easier to discard partial datasets (consisting of lemma, translation and morphological features) which had been incorrectly allocated to a record, rather than to enter the correct partial dataset by hand. Figure 4 shows a pre-annotated dataset before discarding one alternative interpretation. To avoid cases where the appropriate reference to the place of the word is missing, and an unmanageable number of allocations would be given, a manually edited list of possible allocations is provided for those records. For instance, if an adverb can function as a separable prefix to a number of verbs, and the program gives a range of lemmata with this adverb as a prefix, then the manual list is searched. The lemmata are then replaced by the standardized ones and their translations from Splett (1993), by means of the concordance files.

The texts and the annotation information obtained automatically are divided into more manageable sections, and transferred into the software *ELAN*[10], developed by the Max Planck Institute for Psycholinguistics at Nijmegen, the Netherlands. This software makes manual adaptation of multilevel annotations possible without more extensive technical knowledge.

[10] http://www.lat-mpi.eu/tools/elan

Reference Text Words	that			
Reference Text Letters	t	h	a	t
Lemma	that	the		
Translation	daß	dieser, diese, dieses, jener, jene, jenes, der, die, das		
PoS Lemma	KO	DD		
PoS Record	KOUS	DDA		
Inflectional Information		Neut_Sg_Nom, Acc		
Chapter	I			
Line	86			

Figure 4: Sample word pre-annotation from the *Heliand* in *ELAN* format (simplified schematic representation)[11]

3. Limitations of information gained from grammars

Some types of information which also need to be included in the corpus can only with difficulty be added automatically, as they are not provided in the glossaries. Perhaps the most obvious such problem is the question regarding the position of adjectives and adpositions. Like *STTS*, *DDDTS* requires it to be clearly indicated whether these elements precede or follow the noun which they qualify, yet most glossaries give no information on this. In order to predict this, it would be necessary to check the text against the annotation data before their combination. This would result in an unmanageable increase in complexity for computational purposes, and must therefore be done manually. The *DDDTS* information on adjectives and adpositions remains underspecified in the pre-annotation. For instance, this is the case with the expression *durft mihhil* in verse 18 of the *Muspilli*: *mihhil* is pre-annotated as an adjective (cf. Heffner 1961: 106), but its position after the noun has to be indicated manually.

Information about the inflection of adjectives (i.e. strong/pronominal, weak/nasal or endingless) constitutes a rather borderline case. This indication is rarely provided in the glossaries and can be gathered from the grammars, but as all spelling variants have to be taken into consideration for this task, and as

[11] *DDDTS* abbreviations:
- KOUS: Konjunktion, unterordnend mit Satz (conjunction, subordinating with clause)
- DDA: Determinativ, definit, artikelartig (determiner, definite, article-like)

it is usually quite straightforward to attribute the inflectional forms correctly when number, case and gender are already given, this has only been done in a few exemplary cases. The form *hêlagas* in line 50 of the *Heliand*, for instance, is pre-annotated according to the information *gen. sg. masc.* in the glossary (Sehrt 1966: 243) as *Pos_Masc_Sg_Gen_*, and only *st* has to be added at the end of the given information to indicate a strong/pronominal inflection.

On the other hand, the strong verb classes are an example of information which can be provided without excessive difficulty. No indication of these is customarily given in the glossaries, but the grammars give clear rules on assigning them, which can easily be formalized. Provided that their spelling is not too unusual, strong verbs of class 2b, for instance, can be expressed as all verbs containing *eo* or *io*, followed by *d, t, s, z, h* or *hh*, in turn followed by *an* and a word end (cf. Braune 2004: 279).

4. Problems in the automatic attribution of word forms to lemmata

That the dictionaries were written for human readers and not for automated processing is reflected by the fact that they often provide contexts without highlighting the word forms in question, and occasionally even give phrases that do not contain the lemma in question at all. One such example occurs in Kelle's *Otfrid* glossary (Kelle 1881: 17). The lemma *ango* 'anxious' appears only once within Otfrid's works, and the remaining fourteen case forms given in the *Otfrid* glossary represent the author's use of adverbs of manner in general and in different syntactic environments. They are preceded by an extensive comment explaining this. However, in automated digitization, the examples that do not contain the lemma concerned cannot be discarded automatically. Thus, it is possible for the program to interpret the word *in*, or sometimes the word *unsen*, as an occurrence of *ango*, since less attention is paid to vowels due to *Ablaut* alternation in strong verbs. This results in *unsen* (the dative plural of the possessive pronoun *unsēr* ('our')) being wrongly attributed to the lemma *ango*, for example in Chapter 6, verse 65 of the fifth book of Otfrid's *Evangelienbuch*, rather than to its correct lemma (*unsēr*), since this occurrence of *unsen* is not listed under its correct lemma.

A very frequent case of ambiguous lemmatization leading to multiple attribution is the case of Old Saxon *that*. This can be the nominative or accusative singular of the neuter demonstrative pronoun *the*, which at the time was only

developing into a definite article, or it can be the conjunction *that*. Any instance of *that* not listed in the glossary with its specific reference will hence lead to its being attributed to both possible lemmata (cf. Figure 4). One of these can be discarded in *ELAN* by only three mouse clicks.

The attribution of annotation information fails completely if the word is not listed in the glossary. This is the case for the word *forstuotun* ('they understood') in sentence 7 of chapter 104 in the Old High German *Tatian*, which Sievers (1892: 432) does not list under the lemma *furstantan* 'understand'. As the form is irregular (the standard would be *forstuontun*), no other third person singular form of the past of *furstantan* could be found when searching for this record. In this case, lemma, translation and morphological information have to be supplied manually. This example is an instance of the tendency of some glossaries to cover all semantic nuances rather than all different spellings.

5. Manual Annotation

After the data from the glossaries has been incorporated, the *ELAN* files must be adjusted manually. The automatically inserted lemmata are verified with respect to the specific context in which they occur, and the same applies to part-of-speech tagging and morphological annotation.

Furthermore, information on clause types and rhyme schemes has to be inserted completely by hand, and the manuscript version of the texts (see Section 2) also has to be copied from digital manuscript photographs. However, if a glossary comprises not only the word tokens of the printed edition, but also the forms used in the manuscript, these can be automatically inserted into the *ELAN* file, reducing the manual work to proofreading. This is the case for the *Heliand*: the glossary by Sehrt (1966) provides the word forms of the four (of six) manuscripts that were known by the time of its publication, so all these word forms can be added automatically.

In the same way as automatic pre-annotation, manual annotation gives rise to various challenges and problems which are outlined here by way of example. Inflectional classes of lemmata and records do not always coincide. In many cases, the inflectional class of a lemma may vary, a fact that has to be indicated in its annotation. As the lemma *ërda* "earth", for instance, varies between an *ō*-stem ('strong') and an *n*-stem ('weak') inflection (see for example Sievers 1892: 318), both lemma and records are pre-annotated as *o,n_Fem*. However,

the annotation of the records has to be adjusted in each case, with a genitive singular *erdun* for *n_Fem*, and with an accusative singular *erda* for *o_Fem*. A nominative singular *erda* is ambiguous and may thus remain unchanged.

The pre-annotation fails with fixed terms that have to be manually converted into multi-word lexemes. The adverbial expression or conjunction *aftar thiu* 'thereafter', 'after' + clause, for instance, is pre-annotated as a sequence of the preposition or adverb *aftar* ('back', 'behind', 'after', 'thereafter') and the neuter instrumental singular of the demonstrative pronoun *dër*, so the annotation has to be adapted to the specific context.

In late Old High German and Old Saxon texts, the original inflectional classes can often no longer be distinguished. Therefore, inflectional forms are marked according to a synchronic view of the texts. To give an example, the lemma *sunta* 'sin' originally inflects as a *jō*-stem, (see Braune 2004: 197). However, Heffner (1961: 145) lists forms lacking the semivowel (e.g. *sundono*, *sundon* in the genitive plural) as well as forms where it is preserved (e.g. *suntiono*, *sundeno*). The former records are accordingly annotated as *ō*-stems, interpreting this phenomenon as a change of inflectional class.

In some cases, lemmata synchronically represent a part of speech other than their primary one and should consequently be treated as such, but sometimes the older use may subsist. The lemma *filu*, for instance, is pre-annotated as an adverb ('very', 'much') (cf. Splett 1993: 231), and in most cases this analysis perfectly fits its interpretation as an adverb of degree. However, it does not only trace back to a *u*-stem noun *filu* ('multitude)', but is in some cases still used as such. Verse 69 of chapter 16 of the third book of Otfrid's *Evangelienbuch* reads: *Fílu thero liuto giloubta in drúhtinan tho* 'Then, many [of the] people believed in the Lord'. Here, *Fílu* is the subject of the clause, having a genitive plural attribute *thero liuto* and agreeing with the singular verb *giloubta*. Therefore, in this case, the pre-annotation has to be changed from an adverb to a noun.

6. Creating standard representations of the words

Once all manual annotation work is completed, the standard version of the texts (see Section 2) can be automatically generated on the basis of the lemmata and the disambiguated and completed morphological information, using a digitized version of the grammar of the relevant stage of the language. To create the standardized word form *alamahtīgun*, mentioned at the beginning

of Section 2, the part-of-speech information on the form *almahticun* from the printed edition of the *Sangaller Credo*, obtained from the glossary as *a s m* (cf. Heffner 1961: 6) and automatically replaced by *Pos_Masc_Sg_Acc_*, has to be amended manually to give *Pos_Masc_Sg_Acc_st*. This completed declaration triggers the addition of *un* at the end of an adjective lemma, in this case the standard dictionary lemma *alamahtīg* (cf. Splett 1993: 582), replacing the glossary lemma *almahtig*.

Where the grammars give several alternatives as correct, the one selected is that which is oldest, or corresponds best to the overall structure of the language. However, there are cases where there are no explicit rules for the selection of variant forms, as the conditioning is purely lexical. This is the case, for example, for the comparative suffix of adjectives, which may appear as *-ōr-* or *-ir-*, with the latter also triggering umlaut. Here, the choice is made according to the spelling of the particular attested form. Furthermore, as geminates are written with a single letter at the end of words, lists have to be provided of words whose final letter is doubled in inflectional forms.The automatically generated standard word forms will require further manual proofreading, which will certainly give rise to further problems.

7. Conclusion

Within the range of projects aimed at compiling historical text corpora, the *Referenzkorpus Altdeutsch* is distinguished by the extent to which it automates existing data. Not only are the texts themselves digitized on the basis of existing sources, but the digitization of 19th century glossaries, and the exploitation of their data in the annotation of medieval texts, constitute a time-saving innovation that helps focus on the specific problems of annotation, preventing progress being slowed down by the need for consideration of straightforward cases. Texts with automated pre-annotation have to be scrutinized even more carefully than those which have not been through a process of pre-annotation, so that misattributions are not overlooked. The not inconsiderable effort involved in preparing the automation also has to be taken into account. Nonetheless, the approach outlined here allows for an effective and efficient creation of large historical text corpora.

As the project has not yet been finished, and the largest subcorpus, the Old High German *Notker*, comprising roughly two thirds of the whole corpus, is still outstanding, the methods developed still have chance to prove themselves with even larger quantities of text.

References

Primary Literature[12]

Heliand. In: Taeger, Burkhard (ed.) (1984): Heliand und Genesis. 9th edition. Tübingen: Niemeyer.

Isidor. In: Eggers, Hans (ed.) (1964): Der althochdeutsche Isidor. Tübingen: Niemeyer.

Monseer Fragmente. In: Hench, George Allison (ed.) (1890): The Monsee Fragments. Straßburg: Trübner.

Muspilli. In: Steinmeyer, Elias (ed.) (1916): Die kleineren althochdeutschen Denkmäler. Berlin: Weidmann, 66-81.

Otfrid. In: Erdmann, Oskar (ed.) (1973): Otfrids Evangelienbuch. 6th edition, revised by Ludwig Wolff. Tübingen: Niemeyer.

Sangaller Credo. In: Steinmeyer, Elias (ed.) (1916): Die kleineren althochdeutschen Denkmäler. Berlin: Weidmann, 27f.

Tatian. In: Sievers, Eduard (ed.) (1892): Tatian. Lateinisch und althochdeutsch mit ausführlichem Glossar. 2nd edition. Paderborn: Schöningh.

Secondary Literature

Braune, Wilhelm (2004): Althochdeutsche Grammatik. Band I: Laut- und Formenlehre. 15th edition, ed. Ingo Reiffenstein. Tübingen: Niemeyer.

Gallée, Johan Hendrik (1993): Altsächsische Grammatik. 3rd edition, ed. Heinrich Tiefenbach. 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.

Schiller, Anne et al. (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS. (Großes und kleines Tagset). http://www.sfb441.uni-tuebingen.de/a5/codii/info-stts-en.xhtml

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

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

[12] All texts are accessible via http://titus.uni-frankfurt.de/texte/texte2.htm#ahd and #asachs resp.

Sievers, Eduard (1874): Die Murbacher Hymnen. Halle/Saale: 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.

Tiefenbach, Heinrich (2010): Altsächsisches Wörterbuch. A concise Old Saxon dictionary. Berlin/New York: de Gruyter.

Wadstein, Elis (1899): Kleinere altsächsische Sprachdenkmäler. Norden/Leipzig: Soltau.

BARBARA McGILLIVRAY / ADAM KILGARRIFF

Tools for historical corpus research, and a corpus of Latin

Abstract

We present *LatinISE*, a Latin corpus for the *Sketch Engine*. *LatinISE* consists of Latin works comprising a total of 13 million words, covering the time span from the 2nd century BC to the 21st century AD. *LatinISE* is provided with rich metadata mark-up, including author, title, genre, era, date and century, as well as book, section, paragraph and line of verses. We have automatically annotated *LatinISE* with lemma and part-of-speech information, enabling users to search the corpus with a number of criteria, ranging from lemma, part-of-speech, context, to subcorpora defined chronologically or by genre.

We also illustrate word sketches, one-page summaries of a word's corpus-based collocational behaviour. Our future plan is to produce word sketches for Latin words by adding richer morphological and syntactic annotation to the corpus.

1. Introduction

Latin is the language of the first electronic corpus, the *Index Thomisticus*, compiled by Father Roberto Busa between the late 1940s and the 1970s, and typically considered to have marked the beginning of linguistic computing (Lüdeling/Zeldes 2007). Since those times, corpus linguistics and computational linguistics have developed into mature disciplines, and a number of modern languages have been provided with large annotated corpora and computational tools. In particular, sophisticated corpus query systems have been created that allow linguists to carry out advanced searches on corpora. Despite its promising start, Latin, like other dead languages, has partially been left behind in this process, especially with regard to the availability of large corpora with rich syntactic and semantic information and of advanced corpus query systems.

This paper focuses on the *Sketch Engine*, a leading corpus query tool that is widely used in a number of lexicographic and corpus research projects (Kilgarriff et al. 2004). We present the functions and resources we have added to the *Sketch Engine* in order to meet the needs of historical corpus research. In particular, we illustrate *LatinISE*, the first historical corpus included in the

Sketch Engine. *LatinISE* is a 13-million word Latin corpus whose texts range from the 2nd century BC to the beginning of the 21st century AD. It has been automatically annotated with state-of-the-art Natural Language Processing (NLP) tools: a lemmatizer and a part-of-speech (POS) tagger.

The rest of the paper is organized as follows: Section 2 gives the background on existing corpora for Latin and shows the motivation for the *LatinISE* project; Section 3 illustrates the *Sketch Engine* and its main functionalities; Section 4 describes how we have collected and automatically annotated *LatinISE* and, finally, Section 5 concludes by outlining future research.

2. Latin corpora

Latin can be considered a less-resourced language from a computational point of view, if we compare it with modern languages like English. However, for a dead language, the range of available corpora for Latin is quite large.

Over the past years several projects have dealt with digitizing the immense amount of texts produced throughout the history of the Latin language. These projects have created a large number of digital editions, which can be browsed and searched through *ad hoc* search engines. Thanks to such tools, philologists, linguists and literary scholars can look up occurrences of single word forms, or sequences of word forms, to extract their contexts in the texts (concordances).

Some Latin digital editions have been designed for philologists, and therefore contain rich information on the tradition of the texts, one example being *Musisque deoque* (http://www.mqdq.it), comprising a collection of Latin works by poets from the archaic to the modern era. Other projects have opted to include just one particular edition of each work, rather than displaying complete philological information, for example the *Library of Latin Texts* (*CTLO* 2010), which contains more than 50 million words, is available on CD-ROM and is searchable by lexical form and chronological era. Another private collection of Latin texts is the *Bibliotheca Teubneriana Latina* (*CETEDOC* 1999), containing 10 million words and published as a CD-ROM. Among the open-access digital collections worth mentioning is the *Perseus Digital Library* (Bamman/Crane 2008: http://www.perseus.tufts.edu), consisting of 10 million words. Thanks to the morphological analyzer *Morpheus*, the *Perseus Digital Library* is searchable by word forms and lemmas. Around 53,000 words belonging to several classi-

cal works collected in the *Perseus Digital Library* have been morphologically and syntactically annotated in the *Latin Dependency Treebank* (Bamman/Crane 2006).

The *Index Thomisticus*, consisting of 11 million lemmatized words and collecting Thomas Aquinas' *opera omnia*, is still among the largest existing Latin corpora, and it is now available online (Busa 1974-1980: http://www.corpusthomisticum.org). A morphologically and syntactically annotated portion of this corpus is available as the *Index Thomisticus Treebank* (Passarotti 2007: http://itreebank.marginalia.it).

A third treebank for Latin was created as part of the *PROIEL* (*Pragmatic Resources of Old Indo-European Languages*) project (http://www.hf.uio.no/ifikk/english/research/projects/proiel/) and contains around 90,000 words, mainly from the Jerome's translation of the Bible (Haug/Jøndal 2008).

As attested by this brief and non-exhaustive overview, a considerable number of Latin corpora are available to linguists nowadays. However, the searches that are possible on them are limited by their annotation. The majority of these collections contain raw text, so the search options are limited to word forms, or in some cases lemmas, while more advanced searches are only possible in the three treebanks, which are very limited in size (between 53,000 and 120,000 tokens).

The aim of our project was precisely to fill this gap and build a large, richly annotated corpus of Latin, covering an extensive time span. To do that, we decided to apply state-of-the-art NLP tools, which allow fast and consistent automatic annotation. In addition, we wanted to make the corpus searchable through a flexible and sophisticated corpus query tool: the *Sketch Engine*.

3. Sketch Engine

Since its launch in 2003, the *Sketch Engine* has been in use in several dictionary projects, and its value for lexicography is illustrated in Kilgarriff/Tugwell (2001), among others. One of its advantages is that it is provided with a wide range of corpora, and is able to handle large amounts of data (the largest corpus to date contains 8 billion words). The web interface allows the user to upload their own corpus or to build it automatically from the web. In addition, the *Sketch Engine* provides highly developed search options on the corpora, which makes it an ideal tool for dictionary making. These options include

word form, lemma, phrase and CQL (Contextual Query Language) search, as well as filters on contexts of a target word, such as the size of the left/right context, the lemmas and parts-of-speech of the words in the context. The output of such searches is a set of concordances, with customizable view and sorting settings.

In addition to these advanced concordance features, the *Sketch Engine* provides *word sketches* , its distinctive feature. Word sketches are one-page automatic corpus-based accounts of a word's grammatical and collocational behaviour.

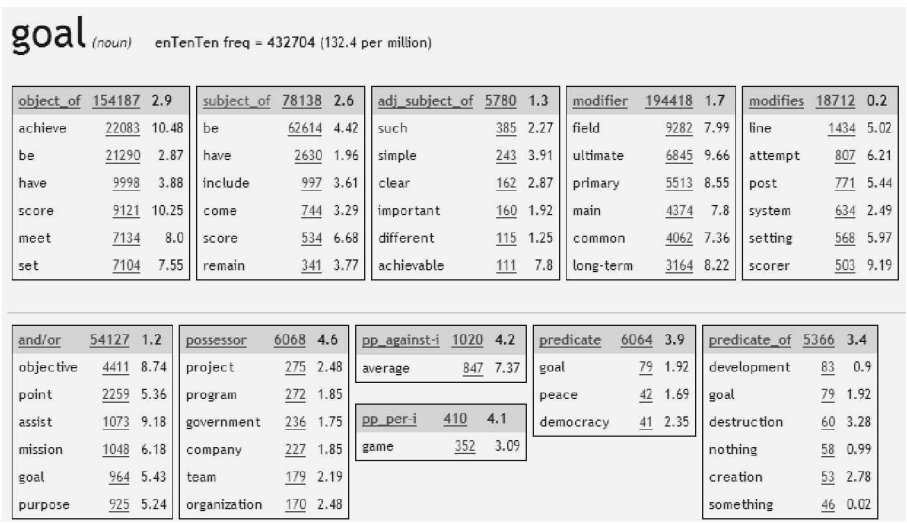


Figure 1: Example of word sketch for the noun *goal* in the *enTenTen* corpus.

Figure 1 shows the word sketch for the noun *goal* in the *enTenTen* corpus for English, containing over 3 billion tokens. The word sketch is organized by grammatical relation and is produced from a syntactically parsed corpus. Each section of the word sketch shows which words stand in a particular grammatical relation with the target word *goal*. For example, the section "object_of " contains verbs whose syntactic object in the corpus is *goal*. Each such collocate is shown with its corpus frequency and salience.

This example gives an idea of the potentialities of word sketches for corpus-based linguistic studies on words' behaviour. Along the same lines as word sketches, *sketch differences* show the differences in the corpus behaviour of two target words, for example by highlighting which collocates are shared by the two words, and which ones are specific to only one of them.

No large historical corpus has been provided with such a rich range of search options so far, and our project aimed to make Latin the first dead language to be included in the Sketch Engine.

4. *LatinISE*: a Latin corpus in the *Sketch Engine*

In this section we discuss the project phases, from explaining how we collected the texts (Section 4.1), to describing the metadata and subcorpora (Section 4.2), illustrating the morphological annotation (Section 4.3) and POS tagging (Section 4.4), and finally exemplifying how the corpus can be searched and displayed (Section 4.5).

4.1 Collecting the texts

The first phase of our project consisted of the collection of the texts. These were assembled from three online digital libraries: *LacusCurtius* (http://penelope.uchicago.edu/Thayer/I/Roman/home.html, by Bill Thayer), *IntraText* (http://www.intratext.com), and *Musisque Deoque* (http://www.mqdq.it). These digital libraries contain texts from standard editions, and cover a wide time span, as well as a variety of genres. In this respect, they were ideal for our purposes of creating a large and wide-ranging corpus for Latin.

The texts had to be converted from HTML format into the verticalized format required by the *Sketch Engine*. While converting the HTML files, special care was devoted to keeping the metadata mark-up specifying authors, title, books, sections, paragraphs and lines (for poetry). In the verticalized text each line corresponds to a token, a punctuation mark or a tag, and looks like this:

```
<character name="Th">
<line>
praemia
si
cessant
<g/>
,
```

The `<g/>` tag always precedes punctuation marks and has the effect of suppressing space characters between two tokens.

4.2 Metadata and subcorpora

In a historical corpus, especially a diachronic one, rich metadata annotation is essential, given the specifically literary and/or diachronic interest of the users. All three digital libraries provide the texts with metadata information, which was therefore extremely helpful. The metadata were also used to automatically eliminate duplicates of the same texts, an important task in automatic corpus building.

Our metadata cover author, title of the work, genre (prose or poetry), era, date of the work (when available), and century. The oldest text in our corpus is the *Senatus consulta de Baccanalibus* (186 BC), and the most recent one is *Dominus Iesus* (2000), by the Vatican Congregation for the Doctrine of the Faith. Below we show an example of how the metadata information is encoded in the corpus for the first text from *LacusCurtius* (LC):

```
<doc id="LC" n=1 author="uncertain"
title="Einsiedeln Eclogues" genre="poetry"
era="Romana, Postclassica" date="cent. 1 AD"
century="cent. 1 AD">
```

Our classification in eras follows the one adopted in *IntraText* and includes *Romana Antiqua* (VII-II cent. BC), *Romana Classica* (I cent. BC), *Romana Postclassica* (I-VI cent. AD), *Mediaevalis* (VII-XIV cent. AD), and *Nova* (XV-XXI cent. AD).

The *Sketch Engine* allows the corpus builder to define subcorpora according to specific metadata features. For example, the prose subcorpus has 9,935,401 tokens,[1] while the poetry subcorpus has 3,818,603 tokens.

4.3 Morphological annotation

In order to annotate the corpus, we used state-of-the-art NLP tools. Automatic methods are less accurate than manual, but are far faster and cheaper, and automatic annotation can be easily updated as the input corpus increases or changes.

We aimed to enrich the texts with lemmas and POS tags. For the lemmatization phase, we used the *PROIEL* project's morphological analyser developed by Dag Haug's team; for those word forms that were not recognized by this ana-

[1] In the *Sketch Engine* a *token* is a word or a punctuation mark.

lyser, we used *Quick Latin* (http://www.quicklatin.com/). The input to the analysers was the verticalized text; for example the output of the phrase *sumant exordia fasces* 'let the fasces open the year' looked like this:

```
> sumant
sumo<verb><3><pl><present><subjunctive><active>
> exordia
exordium<noun><n><pl><acc>
exordium<noun><n><pl><nom>
exordium<noun><n><pl><voc>
> fasces
no result for fasces
```

For each word form the analyser gave all possible analyses, with lemma and POS, as well as other morphological tags (gender, number, case, mood, person and voice).

4.4 POS tagging

Once the possible analyses of each token were available, the next question was how to disambiguate these analyses to find the right one. In particular, we focussed on obtaining the most likely lemma and POS for each token in context adopting a machine-learning approach.

Machine-learning POS taggers work on the assumption that if we train a model on some annotated text (training set), it will learn patterns of regularities and will thus be able to tag unseen text.

Lemmatized and morpho-syntactically annotated data for a total of over 242,000 tokens are available from the three Latin treebanks we introduced in Section 2: the *Index Thomisticus Treebank*, the *Latin Dependency Treebank* and the PROIEL project's Latin treebank. Therefore, we opted to use those data to train *TreeTagger* (Schmid 1995), a language-independent POS tagger developed by Helmut Schmid at the University of Stuttgart.

The input to *TreeTagger* was the output from the morphological analyser, with lemma and POS. Based on the contexts each token occurred in, *TreeTagger* learned what POS was the most likely among all those possible. We then assigned the token to that POS and its corresponding lemma.

The output of the annotation was added to the verticalized text, so that the first column contained the word form, the second one its POS, and the third one its lemma. For example, the sentence *praemia si cessant,* 'if the prizes are lacking', uttered by the character Thamyra in the *Einsiedeln Eclogues* (1st century A. D.), is represented in the corpus as follows (ADJ is for adjectives, C conjunctions, N nouns, V verbs):

```
<character name="Th">
<line>
praemia   N     praemium
si        C     si
cessant   V     cesso
<g/>
.
```

4.5 Searching *LatinISE*

The annotation provided in *LatinISE* allows the user to search for a lemma by its POS. For example, the Latin word *cum* can be a preposition ('with') or a conjunction ('when', 'because'). The user can choose to restrict the search to one POS, or to view both the lemma and the POS ('C' and 'PRE') in the concordances. In the latter case the output would look like Figure 2.

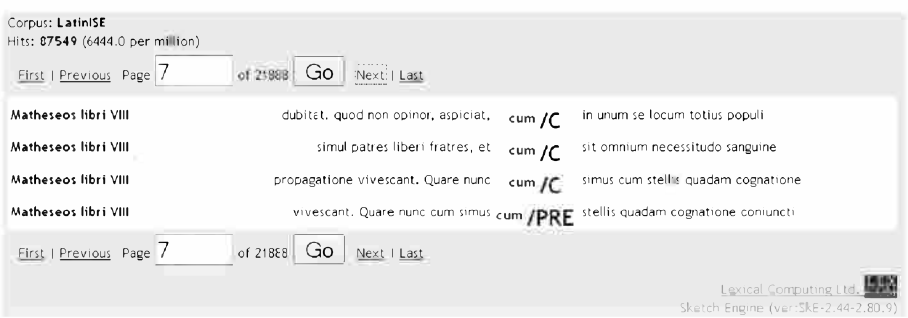


Figure 2: Concordances for *cum* 'with; when, because' in *LatinISE*.

A wide range of possibilities are offered by the view options, where the user can display different metadata information (title of the work in Figure 2), the size of the context, the order of the concordance lines by context, and so on. In addition to simple search on word forms, lemmas, and phrases, it is possible to specify the left/right context of a word by the lemma, POS and number of tokens in its context. This allows the user to extract syntactic constructions like *dico/puto/credo* 'believe', 'think'+*quod* 'that' (Figure 3), and get an overview of the distribution of these constructions in the corpus.

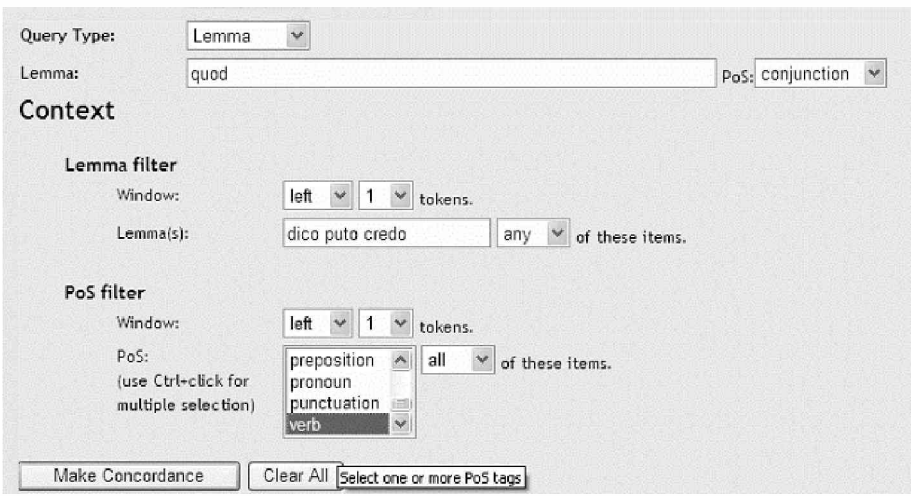


Figure 3: Context-dependent concordance search for the conjunction *quod* 'that' followed by forms of the verbs *dico*, *puto* or *credo* 'think', 'believe'.

5. Conclusion and future research

We have presented *LatinISE*, a 13-million token corpus for Latin. *LatinISE* is the first historical corpus included in the *Sketch Engine*, and was automatically lemmatized and POS tagged using state-of-the-art NLP tools. The texts contained in *LatinISE* cover a time span of 22 centuries, from Early Latin to the beginning of our century. Its rich metadata and linguistic annotation make it possible to carry out diachronic studies on various aspects of the Latin lexicon.

We plan to enrich the annotation with morphological tags (case, number, gender, mood, voice, person) and, ultimately, syntactic relations. This would allow us to produce word sketches, showing the collocational behaviour of Latin lemmas over time.

References

Bamman, David/Crane, Gregory (2006): The design and use of a Latin dependency treebank. In: Proceedings of the Fifth International Workshop on Treebanks and Linguistic Theories. Prague: Institute of Formal and Applied Linguistics, 67-78.

Busa, Roberto (1974-1980): Index Thomisticus. Stuttgart-Bad Cannstatt: Frommann-Holzboog.

Crane, Gregory/Chavez, Robert F. (2001): Drudgery and deep thought: designing digital libraries for the Humanities. In: Communications of the ACM, 44(5): 34-40.

Centre Traditio Litterarum Occidentalium (CTLO) (2010): Library of Latin Texts CLCLT-6. Turnhout: Brepols.

CETEDOC (1999): Bibliotecha Teubneriana Latina. Turnhout: Brepols.

Lüdeling, Anke/Zeldes, Amir (2007): Three views on corpora: corpus linguistics, literary computing, and computational linguistics. In: Jahrbuch für Computerphilologie, 9: 149-178. http://computerphilologie.tu-darmstadt.de/jg07/luedzeldes.html.

Haug, Dag Trygve Truslew/Jøndal, Marius Larsen (2008): Creating a parallel treebank of the old Indo-European Bible translations. In: Calzolari, Nicoletta/Choukri, Khalid/Maegaard, Bente/Mariani, Joseph/Odjik, Jan/Piperidis, Stelios/Tapias, Daniel (eds.): Proceedings of Language Technologies for Cultural Heritage Workshop (LREC 2008). Paris: European Language Resources Association, 27-34.

Kilgarriff, Adam/Rychly, Pavel/Smrz, Pavel/Tugwell, David (2004): The Sketch Engine. In: Williams, Geoffrey/Vessier, Sandra (eds.): Proceedings of the Eleventh Euralex International Congress. Lorient: Université de Bretagne-Sud, 105-116.

Kilgarriff, Adam/Tugwell, David (2001): WORD SKETCH. Extraction and display of significant collocations for lexicography. In: Proceedings of the ACL workshop on COLLOCATION: Computational extraction, analysis and exploitation. Toulouse, 32-38.

Passarotti, Marco (2007), Verso il Lessico Tomistico Biculturale. La treebank dell'Index Thomisticus. In: Petrilli, Raffaella/Femia, Diego (eds.): Il filo del discorso: intrecci testuali, articolazioni linguistiche, composizioni logiche. Atti del XIII Congresso Nazionale della Società di Filosofia del Linguaggio. Roma: Aracne, 187-205.

Schmid, Helmut (1995): Improvements in part-of-speech tagging with an application to German. In: Proceedings of the ACL SIGDAT-Workshop. Dublin, 47-50.

ALEXANDER MEHLER / SILKE SCHWANDT /
RÜDIGER GLEIM / ALEXANDRA ERNST

Inducing linguistic networks from historical corpora

Towards a new method in historical semantics

Abstract

In this paper, we experiment with exploring linguistic networks as a new method in historical semantics. Our starting point is a long-term historical corpus (i.e., the *Patrologia Latina*) which we analyse regarding the conceptual stability of a key concept in medieval literature (i.e., *virtus*). Most analyses in historical semantics explore small data sets by focusing on narrow contexts of lexical usages, but we propose a more comprehensive method based on lexical networks that represent the underlying documents as a whole. We demonstrate both the topological stability of document-based lexical networks and their usefulness in providing empirical evidence in historical semantics.

1. Historical semantics

The view a researcher has of language can vary widely depending on the field of research he or she is coming from. The following section deals with the view of the historian, working in the field of humanities, whose questions, opinions, and interpretations may differ from those of corpus linguists. Nevertheless, *historical semantics* has quite a long tradition in historical research. So far corpus methodology has mostly been applied to language corpora of non-historical languages (see Steinmetz 1993), and such research projects have dealt with rather short periods of time and comparatively small corpora.

This chapter presents a modified methodology for a corpus-based approach to the analysis of language patterns over a longer period of time. It will do so in three steps, dealing firstly with language, secondly with the relationship between language and society, and thirdly, it will introduce a way of representing language as linguistic networks.

1.1 Concerning language

Broadly speaking, sense and meaning are attributed to single words or multiple word units. Within a given system, these attributions are *situational*, *ambiguous*, *uncertain* and *variable* (Barwise/Perry 1983; Rieger 1995; Geeraerts 1997). Sense-attributions are, for example, situational in that they occur in specific situations of word-use. As a result of such usages in ever new situations, the connection between sign vehicle and meaning does not become fixed, and has to be actualised by means of repetition (Peirce 1993). Without repeated usages it falls out of existence and can be replaced by other sense-attributions. Sense-attributions can also be ambiguous, as there is frequently more than one possible meaning for a given sign.[1]

These characteristics establish the variability of sense-attributions. They change from situation to situation and over time, i.e. historically (Geeraerts 1997). Of course, such changes do not render the sense-attributions arbitrary. Every mechanism of making sense has to be plausible and depends on successful communication. It therefore depends on the situations in which it is used. Hence, sense-attributions can be viewed as indicators of the social acceptance of meaning, that is, as indicators of social processes (Halliday 1977).

1.2 Concerning society

Historical semantics as a methodology applied in historical research focuses on the relationship between language and society (Jussen et al. 2007). Language as a sign system reflects the society or culture it is being used in. Thus language change can be viewed as a measure of societal change.

There are several traditions in historiography that dealt with language and the analysis of meaning. One of the most important traditions in Germany, the *Begriffsgeschichte* ('history of concepts'), started with the works of Reinhart Koselleck in the 1970s. The original idea was to trace back in time the meaning and use of politically and socially relevant concepts central to our understanding of society. Together with a large group of his colleagues, Koselleck assembled hundreds of articles on concepts like the state, the nation, revolution, and freedom (see Brunner et al. 1972-1997). Over more than 30 years, the *Begriffsgeschichte* developed the following main arguments:

[1] See the literature cited above for thorough analyses of these and related properties of the meaning relation.

1) when language changes the images and notions of society, the latter can be grasped through the former;
2) key aspects of meaning can be identified by collecting and analysing the antonyms of words manifesting a certain concept;
3) semantics need to be repeatable; that is, there has to be a minimum consent about the meaning of words in order for them to be understandable (see Koselleck et al. 2006).

Apart from historiography and the linguistic sciences, there is another concept of semantics dealing with language and society that fits into the reflections presented above. It was developed by the German sociologist Niklas Luhmann. He talks about "cultivated semantics" and the plausibility of sense (see Luhmann 1980-1995: vol. 1). His setting is that a sign vehicle (e.g. on the lexical level) connects with several possible sense-attributions over time, and can always be connected to new ones. These processes are contingent, but once they are part of a successful communication – once they have made sense – they are registered and put into the pool of "cultivated semantics". In that way, society is able to use them again. Every sense-attribution is a snapshot in time. There are indicators as to how plausible a certain sense-attribution is at a certain point in time, or in a certain historical situation. In addition, this plausibility is dependent on society. Society has to make sense of things, of words, of circumstances, and only if this is successful can it be repeated. If it can be repeated, it enters into "cultivated semantics".

All these theories concentrate on moments of change. The method presented in this paper also focuses on change, its visualization and its interpretation.

1.3 Concerning networks

Following the framework of the foregoing paragraphs, sense and meaning are constituted by means of situational sense-attributions. For the analysis of these attribution processes it is essential to look at the co-occurrences of single words or multi word units in question. Doing that it is possible to isolate language patterns and trace them within a diachronic corpus. These patterns, or situational word-nets, represent semantic snapshots in time. In order to pinpoint moments of change and interpret them properly, one needs to compare the snapshots with each other. One way of visualizing the relation of co-occurrences to the search-term and to each other is to plot them as a network (see Mehler et al. 2011 and Section 3 of this chapter). The comparison of the word

usage networks and their situational circumstances raises the question of whether the changes observed can be accounted for by means of our analytical method.

The paper is organized as follows: in Section 2, we outline the (still ongoing) preprocessing of the corpus that underlies our study (i.e., the *Patrologia Latina*) as a prerequisite of network analysis in historical semantics. In Section 3, we exemplify our method by means of lexical networks on the level of word forms, and introduce the notion of sonar-word-induced networks as input to historical semantics. Section 4.1 provides a first interpretation of such networks. Finally, the paper concludes and gives a prospect on future work in Section 6.

2. Preprocessing the *Patrologia Latina*

Our experiments are based on the *Patrologia Latina* (PL) (Migne 1844–1855). The PL is a corpus of ecclesiastical documents that stem from between the 4th century and the beginning of the 13th century. It reveals several stages of the development of Latin in the direction of Early Romance languages, on various levels of linguistic resolution. The PL has been digitized and edited by Chadwyck-Healey. This edition, which is encoded in SGML, and which does not provide annotations below the level of paragraphs, is not directly suited for corpus analyses. Therefore, we initially mapped the PL onto the document model of the *Text Encoding Initiative* (TEI P5) (TEI Consortium 2010) by correcting several annotation errors in the source edition, and by adding a range of annotation layers. Amongst others, this includes the detection of sentence boundaries, tokenization and lemmatization.

Texts	8,508
Sentences	4,990,602
Tokens	106,515,458
Types	1,077,932
Lemmas	782,453

Table 1: Some statistics of the *Patrologia Latina* (PL) based on our latest preprocessing. Differences from previous publications (e.g. Mehler et al. 2011) are due to a complete renewal of the underlying preprocessor (including sentence recognition, handling of embedded sentences, and lemmatization).

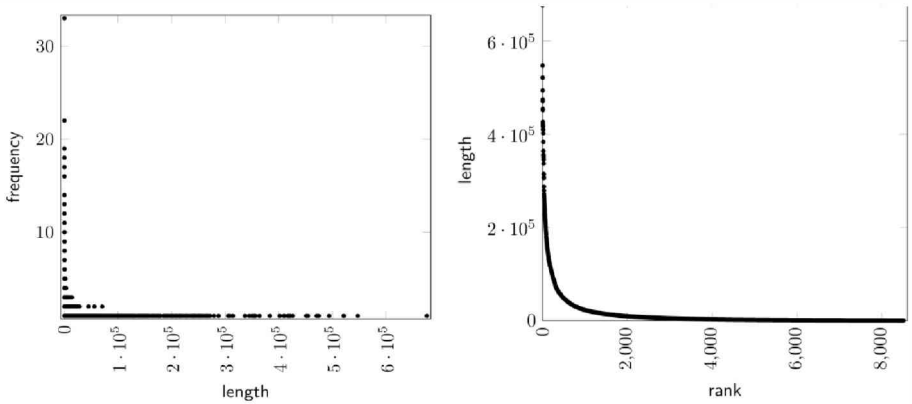


Figure 1: Left: length-frequency distribution of documents in the *Patrologia Latina*, displaying the number of documents (*y*-axis) as a function of the length of the documents (*x*-axis). Length is measured by the number of tokens by disregarding sentences in note- and foreign-tags. Right: the corresponding rank length distribution starting with the longest document in descending order.

According to our latest preprocessing, the *Patrologia Latina* consists of 8,505 texts, which may contain notes in languages other than Latin, such as Greek or French. Table 1 gives an overview of the corpus with a focus on Latin content only. This table shows some differences from previously published figures (Mehler et al. 2011), the result of our ongoing work on improving, for example, sentence recognition, lemmatization and foreign word detection (see Sukhareva et al. 2012) in the PL. Figure 1 shows the length-frequency distribution of the PL (where length is measured by the number of Latin tokens), while Figure 2 shows the rank distribution of the type-token ratio (in descending order for the overall set of 8,508 documents.

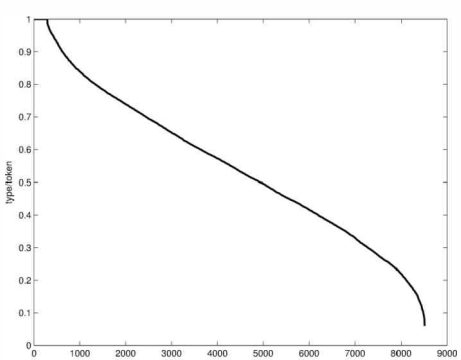


Figure 2: Rank distribution of the type-token ratio of the documents in the PL.

The Corpus can be queried and analysed using the *Historical Semantics Corpus Manager* (Jussen et al. 2007), which is part of the *eHumanities Desktop* (Gleim et al. 2009).[2]

3. Inducing lexical networks

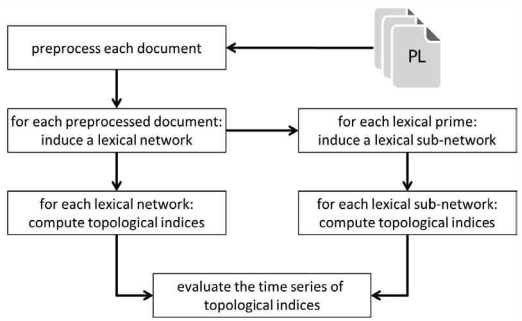


Figure 3: Sequence of operations of mapping a historical long-term corpus like the PL onto a time series of graph invariants of lexical networks.

In Mehler et al. (2011), a procedure has been introduced that maps streams of input documents onto time series of lexical networks. In Mehler et al. (2010a), this procedure has been formalized with the aid of graph theory. Mehler et al. use k-layer graphs to formally represent the time-aligned networking on the level of lexemes, sentences and word forms.[3] The underlying procedure of this process of graph induction is summarized in Figure 3. It includes the following steps:

1) Starting from a historical corpus whose preprocessing occurs according to Section 2, each input document is mapped onto a time line. This can be done by mapping the documents onto the century of their publication, onto the date of their publication (if known), or onto the end date of the productive period of the corresponding author(s).
2) Then, for each document a separate lexical (or sentential – Mehler et al. 2010a) network is induced whose nodes are linked whenever the corresponding words co-occur in a sentence of that document. As a result, edges between lexical nodes are weighted by the frequency of their sentence-based co-occurrences. Note that any further processing of the edges'

[2] See http://www.hucompute.org/ressourcen for further information on these tools.

[3] In Mehler et al. (2010b), this model has been used to model alignment in verbal communication according to the time scale of dialogues.

weights can be done by means of a co-occurrence measure (Heyer et al. 2006; Evert 2008). In what follows, we will not vary this parameter.

3) In addition to what has been done in Mehler et al. (2010a) and Mehler et al. (2011), we induce subgraphs of lexical networks using lexical primes that manifest key concepts in historical semantics. In Section 4.1, this is exemplified by *virtus* in the work of John of Salisbury and of St. Augustine. Subnetworks of this sort, which are henceforth called *sonar-word induced lexical networks*, are spanned by a prime word's neighbourhood in the lexical network of the underlying document.
4) The next step is to compute so called graph invariants (Diestel 2005) for each lexical network and its sonar-induced sub-networks. Informally speaking, a graph invariant is a characteristic of a graph (i.e. network) that is preserved by isomorphic graphs. The idea behind computing such invariants is to gain insights into the laws of linguistic networking. Mehler et al. (2010a), for example, show a remarkably stable pattern of lexical networking that holds both for a present day language and for a historical language. Note that this law-like behaviour is not based on a simple frequency distribution (as in the case of Zipf's law, cf. Zipf 1972), but on the structure of the networks under consideration.
5) Finally, the resulting time series of graph invariants are input to investigate the law-like behaviour of linguistic networking over time.

In this paper, we complement this procedure by adding a further level of investigation. In contrast to earlier research, we induce networks of *word forms* by focusing on their order (i.e. the number of vertices) as the independent variable. There are two reasons for doing this:

1) First, the variance observed in Mehler et al. (2010a) may be partly due to the variance of the order of the networks. In order to investigate this potential dependency, we look at the values of certain graph invariants as a function of order.
2) Secondly, lemmatizing historical corpora manifested by a language in flux (such as Latin) represents an enormous amount of work. Thus, the question arises whether word form networks that circumvent this effort are as expressive as lemma networks.

Accordingly, in this paper, we analyze word form networks. More specifically, we aim to test whether word form networks show law-like behaviour as previously mentioned for lemma networks. Extending our previous studies, we se-

lect a subset of 2,696 medieval documents of the Patrologia Latina, containing at least 1,000 word forms,[4] concentrating on Medieval Latin in order to neutralize the effects of language change (from Classical Latin to Late Latin). Additionally, we want to reduce the effect of small document sizes that possibly correlates with the divergence of the types of texts collected by the PL (see Section 5). The complete PL contains 8,508 documents whose size ranges from a few tokens to more than 600,000 tokens (see Figure 1 for the corresponding length-frequency distribution). We make a cut on the left side of this distribution, and therefore disregard documents that are too small in terms of their number of tokens.

[4] The *Patrologia Latina* is not so much a structured corpus as a rather arbitrary collection of texts. Chadwyck-Healey's digital edition of these texts contains more than 8,000 documents. These include documents which do not represent Medieval Latin texts. Therefore, the collection had to be examined and was reduced according to the following criteria:

1. The digital documents of the Chadwyck-Healey edition were originally annotated according to the period of time in which they were supposedly written. Our revised collection contains mainly those documents marked as "medieval". We also included some documents marked as "uncertain" because the authorship is unknown. In these cases we relied on evidence within the documents. The documents of our collection date from the 1st to the 13th century AD; most have been annotated with the century of their origin.
2. All secondary documents (indices, commentaries, and notes written by the early modern editors of the Patrologia Latina, etc.) have been removed from our collection.
3. We have also removed from the collection all documents which only contain references to texts in other volumes of the original edition of the Patrologia Latina.

All these criteria are listed within our set of annotations, which also address the genre of the documents (see Figure 7).

4. Experiment

4.1 Network model

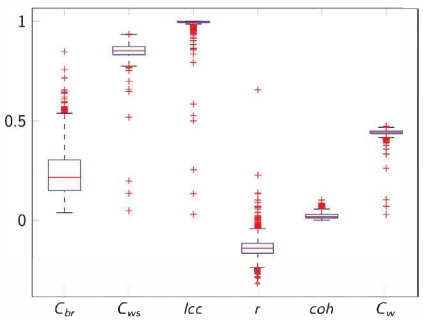


Figure 4: Boxplots of six graph invariants in a corpus of 2,696 networks (documents) of at least 1,000 vertices (word forms): C_{br} (cluster coefficient of Bollobás/Riordan 2003), C_{ws} (cluster coefficient of Watts/Strogatz 1998), lcc (fraction of vertices of a largest connected component), r (assortativity), coh (cohesion), C_w (weighted cluster coefficient).

In this Section we show an experiment based on the subcorpus of the PL described above. We provide an analysis based on graph invariants of all documents in this subcorpus. More specifically, we examine six invariants:

1) As a simple graph invariant, we consider the cohesion $coh(G) = \frac{\sum_{v \in V} d_G(v)}{|V|^2 - |V|} \in [0,1]$ of a graph G as the fraction of the number of co-occurrences observed in relation to the overall number of possible co-occurrences ($d_G(v)$ is the degree of vertex v; in the present case this is the number of its collocates). For sparse networks, the cohesion tends to converge to 0 for growing numbers of vertices. Therefore, we also consider the following invariant as a measure of connectedness.

2) $lcc(G) = \frac{|LCC(G)|}{|V(G)|} \in [0,1]$ is the fraction of vertices that belong to a largest connected component of G in relation to its order $|V(G)|$ (number of vertices). Together with cohesion, this provides an expressive network model. The reason is that small-world networks (cf. Watts/Strogatz 1998) are typically sparse networks with a giant component that dominates all other components in terms of its order. In such a network, nearly every pair of vertices is connected by a path of, in the present case, lexical association.

3) Small-world networks typically exhibit a high rate of complete graphs of order 3 (so called 'triangles'). In order to test this hypothesis for word form networks, we consider the probability by which words that are syntagmatically associated to the same word form tend to be associated on their own. This is done with the help of the cluster coefficient of Watts/Strogatz (1998): $C_{WS}(G) = \frac{1}{n}\sum_{i=1}^{n} c_{WS}(v_i) = \frac{1}{n}\sum_{i=1}^{n} \frac{adj(v_i)}{\binom{d_G(v_i)}{2}} \epsilon[0,1]$ where $adj(v_i)$ is the number of edges ending at neighbours of v_i.

4) A variant of the latter model has been proposed by Bollobás/Riordan (2003), who consider weights of vertices as a function of the size of their neighbourhood: the larger this neighbourhood, the more important the vertex: $C_{br}(G) = \frac{\sum_{v\epsilon V}\binom{d_G(v)}{2} c_{WS}(v)}{\sum_{v\epsilon V}\binom{d_G(v)}{2}} \epsilon[0,1]$.

5) Further, we compute the weighted variant C_W of the cluster coefficient C_{WS} of Watts/Strogatz according to Serrano et al. (2006), in order to gain insights into effects of weighting edges according to the frequency of associations.

6) Finally, we compute the assortativity of the networks by a correlation coefficient $r\epsilon[-1,1]$ of the degrees of adjacent vertices according to Newman/Park (2003).

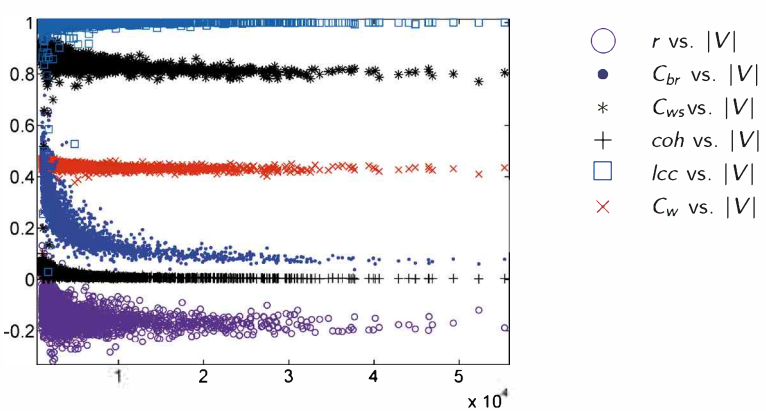


Figure 5: Distributions of the values of six graph invariants (see Figure 4) each as a function of graph order (i.e. number of vertices in the underlying graph).

As can be seen in Figure 5, when ordering the values of these invariants (y-axis) as a function of the networks' order (x-axis), we observe rather stable, law-like behaviour (see also Figure 4 for the corresponding boxplots). We get values of *lcc* according to which the giant component covers nearly all words in the underlying network. This is not a simple result of the fact that sentences rarely consist of a single token. Rather, the way words are used within the documents results in the connectedness of nearly all their lexical constituents. As mentioned above, cohesion of document networks tends to be zero for large documents. Thus, at least, we can say that cohesion as observed here does not contradict a law-like networking of lexical constituents. Things look different, however, if we consider the cluster coefficient C_{br} in contrast to C_{ws} and its weighted counterpart C_w: whereas the latter two coefficients show a remarkably low variance in conjunction with a linear dependency on the order of the networks, it is the former that obviously exhibits nonlinear behaviour as a function of order. Thus, we may say that the larger the network, the smaller its clustering in terms of C_{br}. Moreover, as the values of C_{br} fall much below those of C_{ws}, we get the information that in word form networks, clustering tends to be a matter of less frequently linked vertices, that is, of words that tend to enter syntagmatic associations infrequently. Conversely speaking, frequently linked words tend to be linked to those that are infrequently linked. This observation is at least not contradicted by the value distribution of r, which shows disassortative mixing of dissimilarly linked word forms.

According to this analysis, word form networks as induced here seem to exhibit law-like networking behaviour that is independent of their order, at least from the point of view of the invariants considered so far.[5] These findings are in line with what we have observed in the case of lemma networks (Mehler et al. 2010a). They hint at a stable pattern of lexical networking from the point of view of its topological structure: independently of the content of the underlying documents, their topics are manifested in a way that leads to structurally similar networks. The question arises whether there is still enough semantic variance to discriminate among these networks so that historical semantics can rely on this data structure. This question is addressed in the next section by example of two sonar-word-induced networks starting from the same lexical prime.

[5] The elaboration of a statistical test for findings of this sort is a task for the future.

4.2 Historical interpretation of lexical networks

The examples for the interpretation given in this section draw their data from two medieval Latin political treatises: *De civitate Dei* by St. Augustine (427 AD) (Figure 6, right) and the *Policraticus* of John of Salisbury (1159 AD) (Figure 6, left). The central term to which these word-induced networks are referring is *virtus*, one of the most frequently used terms in Medieval Latin literature. The spectrum of its possible meanings is similar to that of its English equivalent *virtue*. The exact meaning is determined by the situation in which the term is used.

The graphs in Figure 6 show *virtus* and its co-occurrences connected by grey lines. The additional information given is the connection between these co-occurring terms amongst themselves. The width of the lines represents the frequency of the connection. As a whole, these graphs offer the possibility to discern the language patterns around *virtus*; they hint at the contexts in which *virtus* is being used, thus suggesting certain discursive functions of the term.[6]

At first glance the graphs themselves look slightly different. For *De civitate Dei* there are far more thick lines connecting single terms than in the graph for the *Policraticus*. This suggests that the semantic field surrounding *virtus* has a higher density in the language use of St. Augustine than in that of John of Salisbury. Although there are quite a few distinct themed clusters, their interconnection is higher with St. Augustine than with John. Interestingly, most of the thick lines in the graph representing the *Policraticus* show a frequent combination between the Latin words *vera*, *Dei*, and *vero*. The same words are shown as being closely related in the other graph as well. There seems to be a traditional word use that has been stable over several hundred years.

The notion of *vera virtus* is one of the central elements in the apologetic argumentation of St. Augustine. He opposes the philosophically influenced *virtus* of the ancient world with the "true" *virtus* of Christians. This also explains the close connection to the use of *Deus* (*De civitate Dei*)/*Dei* (*Policraticus*). In the graph representing *De civitate Dei* there is a far more frequent connection between this cluster and another one consisting of *imperium*, *gloriam*, *honorem*, *finem*, and *opus* than can be shown for the *Policraticus*. This has to do with the social context in which the two treatises were written. In Late Antiquity, St.

[6] The graphs show only word forms. Thus they show only the co-occurrences of *virtus* in the nominative singular.

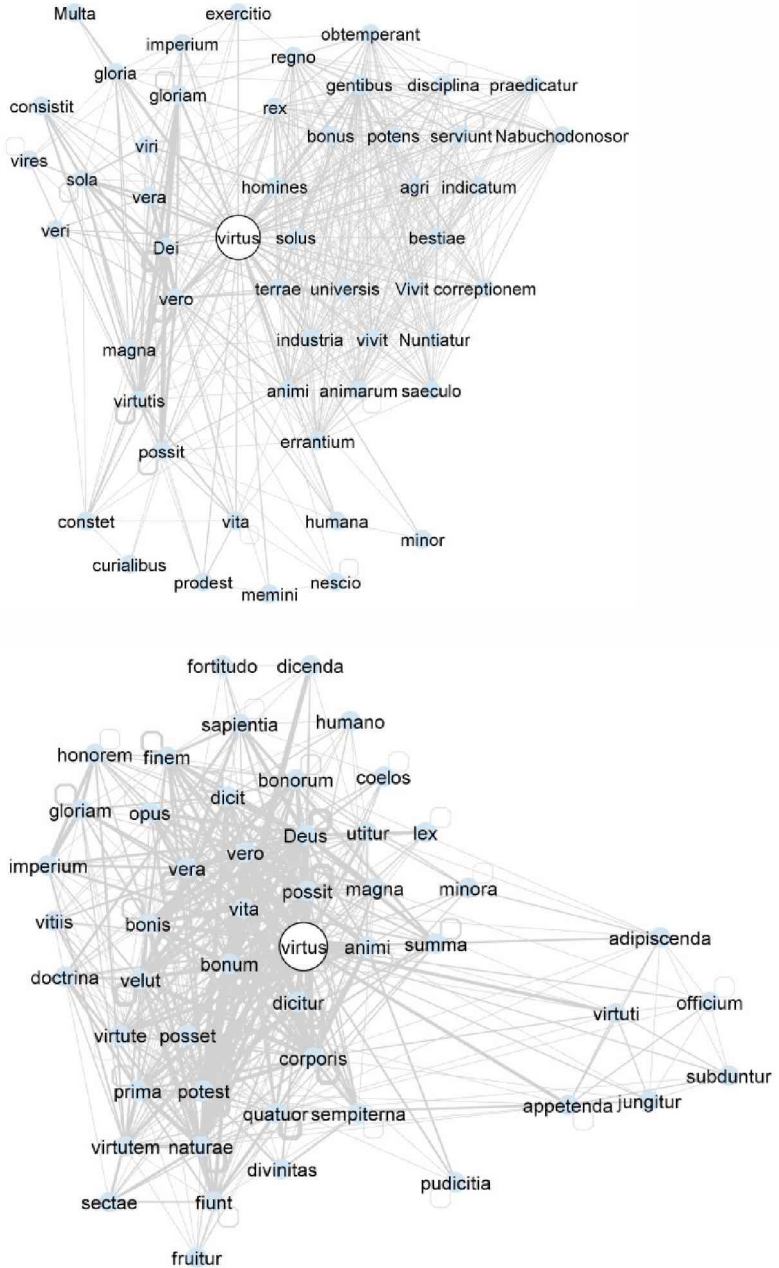


Figure 6: Top: sonar-word-induced network using *virtus* as a prime in John of Salisbury's *Policraticus* (Migne 1844–1855). Bottom: sonar-word-induced network using *virtus* as a prime in St. Augustine's *De civitate Dei* (Migne 1844–1855).

Augustine wrote for an audience which was still highly influenced by Graeco-Roman ideals of ethics and society. The quoted cluster represents the Roman ideal of public life that should amount to the achievement of command and leadership (*imperium*), honour (*honorem*), and glory (*gloria*) through personal deeds (*opus*). For the High Middle Ages, the time of John of Salisbury, this context is no longer of importance, either politically or socially. Nevertheless, the terms of this cluster can also be found in his *Policraticus* which, again, suggests some sort of continuity within the semantics of *virtus*. This continuity also shows that the individual aspects of meaning once connected to *virtus* can be easily used again. It is the social contexts, and with them the logic of argumentation, that change. This example shows how word-induced networks can help to single out argumentative clusters, similarities and differences of word use as well as their correlating trends in (medieval) thought. Moreover, the semantic change as manifested on the lexical level takes place in the context of lexical networks whose macroscopic structure is remarkably stable. That is, latent semantic changes are microscopic processes that take place against the background of law-like lexical networking on the level of texts as a whole.

5. Intertextuality in Medieval Latin

For further research it will be important to improve control over the sets of networks being compared, since the examples above seem to be rather randomly chosen. Aiming at sustainable diachronic interpretations of language change, the corpora serving as data sources should be formed more carefully. Therefore, one of the next steps within the development of our method is to assign the texts of the *Patrologia Latina* to a typology of text genres as suggested in Figure 7, that is, to sources of typological intertextuality. Once the grid is completed and tested, it can be applied to any other text offering its integration into the given corpus. Using a grid like this for corpus-building makes it possible to concentrate, for example, on the analysis of legal, polemic, or paraenetic language. Comparing the relations between such languages will lead to an analysis of historical discourses. Further, it may help in explaining the variance that we still observe – though to a minor degree – in Figure 5, which has also previously been identified in the case of lemma networks (Mehler et al. 2010a).

Other forms of (referential) intertextuality that are very frequent in Latin medieval texts are quotations and references, which are not marked as such in most cases, unlike today. They serve as labels of authority, especially those coming from the Bible, or from other highly authoritative texts such as of the church-fathers. With the help of lexical networks, it may be possible to find these references (whether marked or not) and relate them to the texts they originate from. This would enable research on proliferation, reception, and perception of ideas and arguments in those texts. Another research possibility along these lines is the reconstruction of texts known to medieval authors (who did not usually have the originals but were using compilations instead).

6. Conclusion

In this paper, we analyzed word form networks in historical semantics. Our aim was to induce lexical networks as a representation format that captures more information than lists of keywords in context. Amongst others, this relates to the networking of lexical items, as well as to the strength and clustering of this networking. Starting from the *Patrologia Latina* as a long-term historical corpus, we shed light on the law-like networking of lexical units based on the framework of complex network theory. Our findings are threefold:

1) First, we observed a remarkable structural stability in lexical networks on the level of complete texts, in line with recent observations in the example of lemma networks.
2) Secondly, we exemplified the interpretation of sonar-word induced subgraphs of word form networks as a further representation format in historical semantics. By means of this method, we exemplified a change in the lexical context of the prime *virtus* that hints at a change in the underlying processes of sense attributions. However, starting with the same word, we also described a remarkably stable conceptualization across different authorships that are separated by a relatively long period of time.
3) This dual observation is a first hint of a relationship between microscopic processes of lexical semantic change on the one hand, and macroscopic processes of text structure formation on the other. One consequence of this finding is that we need far more *fine-grained* network analysis methods that operate on *very small* networks, in order to shed light on such microscopic processes of semantic change.

Sonar-word induced networks seem to be expressive enough to enable such analyses in historical semantics, in terms of studying semantic change and related processes. Future work will focus on a systematic evaluation of the expressiveness of sonar-word induced networks in this area of research. We plan to perform a time series analysis, in which we will compute the similarities and dissimilarities of such networks as indicators of semantic change. For this task, we need to further develop the apparatus of k-layer networks. In other words, we need to clarify the degree to which networking, for example, on the lexical level, is indeed law-like over time. This, in turn, requires that we make a thorough text-typological analysis of the *Patrologia Latina* and related corpora, in order to keep control of genre-based effects on networking.

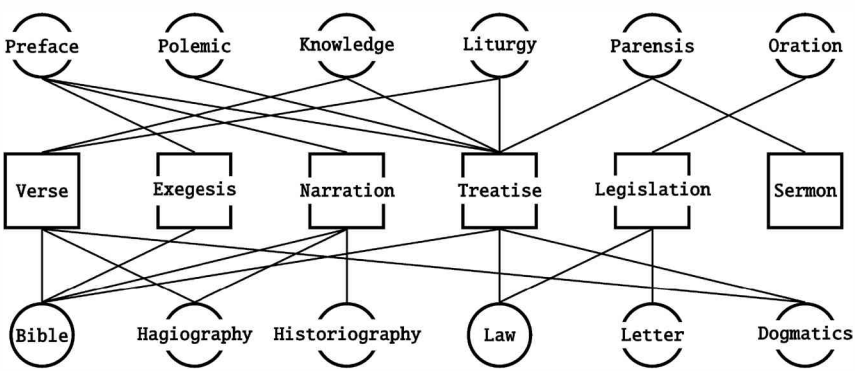


Figure 7: Outline of text types (middle row) in the PL and the sources of their features.

References

Barwise, Jon/Perry, John (1983): Situations and attitudes. MIT Press: Cambridge MA.

Bollobás, Béla/Riordan, Oliver M. (2003): Mathematical results on scale-free random graphs. In: Bornholdt, Stefan/Schuster, Heinz G. (eds.): Handbook of graphs and networks: from the Genome to the Internet. Weinheim: Wiley-VCH, 1-34.

Brunner, Otto/Conze, Werner/Koselleck, Reinhart (eds.) (1972-1997): Geschichtliche Grundbegriffe. Historisches Lexikon zur politisch-sozialen Sprache in Deutschland. Stuttgart: Klett-Cotta.

Diestel, Reinhard (2005): Graph theory. Heidelberg: Springer.

Evert, Stefan (2008): Corpora and collocations. In: Lüdeling, Anke/Kytö, Merja (eds.): Corpus linguistics: an international handbook of the science of language and society. Belin/New York: Mouton de Gruyter, 1212-1248.

Geeraerts, Dirk (1997): Diachronic prototype semantics. Oxford: Clarendon Press.

Gleim, Rüdiger/Waltinger, Ulli/Ernst, Alexandra/Mehler, Alexander/Esch, Dietmar/Feith, Tobias (2009): The eHumanities Desktop – an online system for corpus management and analysis in support of computing in the humanities. In: Proceedings of the Demonstrations Session of the 12th Conference of the European Chapter of the Association for Computational Linguistics EACL. Athens, 30 March - 3 April 2009.

Halliday, Michael A. K.(1977): Text as semiotic choice in social context. In: van Dijk, Teun A./Petöfi, Jånos S. (eds.): Grammars and descriptions. Berlin/New York: de Gruyter, 176-225.

Heyer, Gerhard/Quasthoff, Uwe/Wittig, Thomas (2006): Text mining: Wissensrohstoff Text. Herdecke: W3L.

Jussen, Bernhard/Mehler, Alexander/Ernst, Alexandra (2007): A corpus management system for historical semantics. In: Sprache und Datenverarbeitung 31(1-2): 81-89.

Koselleck, Reinhart/Spree, Ulrike/Steinmetz, Willibald/Dutt, Carsten (eds.) (2006): Begriffsgeschichten. Studien zur Semantik und Pragmatik der politischen und sozialen Sprache. Frankfurt a.M.: Suhrkamp.

Luhmann, Niklas (1980-1995): Gesellschaftsstruktur und Semantik. Studien zur Wissenssoziologie der modernen Gesellschaft. 4 vols. Frankfurt a.M.: Suhrkamp.

Mehler, Alexander/Gleim, Rüdiger/Waltinger, Ulli/Diewald, Nils (2010): Time series of linguistic networks by example of the Patrologia Latina. In: Fähnrich, Klaus-Peter/Franczyk, Bogdan (eds.): Proceedings of INFORMATIK 2010: Service Science, 27 September - 1 October 2010, Leipzig. (= Lecture Notes in Informatics 2), 609-616.

Mehler, Alexander/Lücking, Andy/Weiß, Petra (2010): A network model of interpersonal alignment. In: Entropy 12(6): 1440-1483. doi: 10.3390/e12061440.

Mehler, Alexander/Diewald, Nils/Waltinger, Ulli/Gleim, Rüdiger/Esch, Dietmar/Job, Barbara/Küchelmann, Thomas/Pustylnikov, Olga/Blanchard, Philippe (2011): Evolution of Romance language in written communication: network analysis of late Latin and early Romance corpora. In: Leonardo 44(3): 244-245.

Migne, Jacques-Paul (ed.) (1844-1855): Patrologiae cursus completus: Series Latina, vols. 1-221. Cambridge: Chadwyck-Healey.

Newman, Mark E. J./Park, Juyong (2003): Why social networks are different from other types of networks. In: Physical Review E: 68:036122.

Peirce, Charles Sanders (1993): Semiotische Schriften 1906-1913. Vol. 3. Frankfurt a.M.: Suhrkamp.

Rieger, Burghard (1995): Situation semantics and computational linguistics: towards informational ecology. In: Kornwachs, Klaus/Jacoby, Konstantin (eds.): Information: new questions to a multidisciplinary concept. Berlin: Akademie-Verlag, 285-315.

Serrano, M. Ángeles/Boguñá, Marian/Pastor-Satorras, Romualdo (2006): Correlations in weighted networks. In: Physical Review E: 74:055101.

Steinmetz, Willibald (1993): Das Sagbare und das Machbare. Zum Wandel politischer Handlungsspielräume. England 1780-1867. (= Sprache und Geschichte 21). Stuttgart: Klett-Cotta.

Sukhareva, Maria/Islam, Zahurul/Hoenen, Armin/Mehler, Alexander (2012): A three-step model of language detection in multilingual ancient texts. In: Journal for Language Technology and Computational Linguistics 26(2): 167-179. http://www.jlcl.org/2011_Heft2/8.pdf.

TEI Consortium (ed.) (2010): TEI P5. Guidelines for electronic text encoding and interchange. TEI Consortium. http://www.tei-c.org/Guidelines/P5/.

Watts, Duncan J./Strogatz, Steven H. (1998): Collective dynamics of 'small-world' networks. In: Nature 393: 440-442.

Zipf, George K. (1972): Human behavior and the principle of least effort: an introduction to human ecology. New York: Hafner Publishing Company.

ACHIM STEIN / SOPHIE PRÉVOST

Syntactic annotation of medieval texts

The *Syntactic Reference Corpus of Medieval French* (*SRCMF*)

Abstract

This article presents the *Syntactic Reference Corpus of Medieval French* (*SRCMF*). The corpus is composed of texts taken from the two major Old French corpora, the *Base de Français Médiéval* and the *Nouveau Corpus d'Amsterdam*. This contribution describes some of the core principles of the annotation model, which is based on dependency grammar, as well as the annotation procedure and representation formats.

1. Introducing the *SRCMF*

The project *SRCMF*[1] builds a syntactic dependency annotation on top of the two principal Old French (henceforth "OF") corpora: the *Base de Français Médiéval* (*BFM*,[2] Guillot et al. 2007) and the *Nouveau Corpus d'Amsterdam* (*NCA*,[3] Stein et al. 2006, Stein/Kunstmann 2007). The annotation principles rely on the concept of dependency (close to the models of Tesnière 1965 and Polguère/Mel'čuk 2009), and sentences are described as a hierarchy of connected words rather than a tree of immediate constituents. One reason for choosing such a model is that dependency is more appropriate to give an account of a language with a relatively free word order such as OF, which does not have the rigid SVO order of Modern French. It is less constrained with respect to topicalization (e.g. of objects or adverbials, often resulting in verb-second structures) and adjacency conditions (e.g. of heads and modifiers or of auxiliaries and main verbs). The second reason is the desire to introduce as few theoretical assumptions as possible. Thus, for example, the *SRCMF* grammar does not postulate a default word order, and consequently does not need to represent movement using traces. This distinguishes the *SRCMF* project from

[1] Funded by the *Agence nationale de la recherche* (*ANR*) and the *Deutsche Forschungsgemeinschaft* (*DFG*), 1.3.2009-29.2.2012. For more information see the SRCMF wiki on https://listes.cru.fr/wiki/srcmf.

[2] *BFM – Base de Français Médiéval* [online version]. Lyon: UMR ICAR/ENS-LSH, 2005, http://bfm.ens-lsh.fr.

[3] *NCA – Nouveau Corpus d'Amsterdam*, Stuttgart: Institut für Linguistik/Romanistik, 2006, http://www.uni-stuttgart.de/lingrom/stein/corpus.

the first major syntactic resource for medieval French, the corpus *Modéliser le changement: les voies du français* (*MCVF*),[4] which contains, for Old and Middle French (until 1500), about 72,000 annotated sentences with *PENN*-style constituent structure annotation (Matineau 2008, 2009). A third reason is that the goal of the *SRCMF* project is to provide a reference corpus not only for syntactic research, but also for the training of dependency parsers.

2. The *SRCMF* grammar model

2.1 General principles

A word is represented by a node that depends (as a *dependent*) on its *governor* (we also use the term 'head'). The inflected verb is the topmost governor. Each dependency relation is labelled with its function. Following the specifications of the *NotaBene* annotation tool (Mazziotta 2010a, 2010b), *SRCMF* uses a class hierarchy for syntactic structures and functions. The structures and functions and their abbreviations ('tags') are listed in Table 1, where structures are distinguished by '[S]'. Each dependency relation is expressed by the triple 'governor-function-dependent'.

Tag	Function	Tag	Function
Apst	apostrophe	NgPrt	negative particle
AtObj	attribute of object	NMax [S]	non-maximum structure
AtSj	attribute of subject	NSnt [S]	non-sentence
Aux	auxiliation	Obj	object
AuxA	active auxiliation	Regim	oblique
AuxP	passive auxiliation	Rfc	reflexive clitic
Circ	adjunct	Rfx	reflexive pronoun
Insrt	comment clause	RelC	coordinating relator
Cmpl	complement	RelNC	non-coordinating relator
GpCoo [S]	coordinated group	SjImp	impersonal subject
Coo [S]	coordination	SjPer	personal subject
Intj	interjection	Snt [S]	sentence
ModA	attached modifier	VFin [S]	finite verb
ModD	detached modifier	VInf [S]	infinitival verb
Ng	negation	VPar [S]	participle verb

Table 1: tagset of *SRCMF* syntactic categories

[4] The *MCVF* corpus is freely available on http://www.voies.uottawa.ca and on CD-ROM.

The *SRCMF* model does not use null elements (empty nodes or traces). This is avoided by encoding the linear surface order of words without assuming movement of any kind. Discontinuous structures, which occur very frequently in free word order languages like OF, are connected by the dependency relations alone, thus accepting crossing branches in the representation. However, the model uses duplicated forms in some special cases. In the relative clause (1), the relative pronoun *qui* is a non-coordinating relator (RelNC) whose duplicate is a subject (SjPer). This allows the user to retrieve the complete argument structure of verbs regardless of clause type.[5]

(1) *Souffrance si est semblable a esmeraude* ***qui*** *toz jorz est vert.*
„Sufferance such is like an emerald which all day is green."
(*Queste del Saint Graal* v. 17-18)

In (2), the contracted form *nes* (*ne+les*) is a negation (Ng); its duplicate is an object (Obj):

(2) *sovent dit qu' or veut morir s' il* ***nes*** *ocit.*
„often says that now wants die if he not+them kills"
(*Tristan* de Béroul v. 1985-1986)

Duplicated forms are linked by a special type of relation, different from the dependency relation.

2.2 Governing nodes and functional elements

The selection of the governing node is crucial for a dependency annotation. Whereas some dependency models prefer functional nodes as heads (thus coming closer to generative approaches), the *SRCMF* model prefers the main lexical node: each structure is headed by the lexical head (verb, noun, adjective, adverb). According to the principles of dependency grammar, each main clause must contain a finite verb (VFin) as the top node of the structure. This means that coordinated main clauses as in (3) are analysed as two separate clauses, governed by *monte* and *part*.

(3) (*Et li reis monte*) (*et se part de la cort*)
„and the king mounts and refl. leaves from the court"
(*Tristan* de Beroul, v. 121)

[5] Again, this approach is different from the Turin University Treebank, where a trace-filler system accounts for discontinuous structures, and where slash categories are used for nodes which combine more than one function (e.g. subject and verb in causative constructions; see Bosco 2004: 152ff.).

The fact that lexical heads are generally preferred over functional heads as top nodes of a structure is an important feature which also distinguishes the *SRCMF* model from some other dependency annotations, like *TUT*. In our example sentence (4) the main clause is governed by the inflected verb (i.e. the first inflected element of the verb complex, here *a*). This verb immediately dominates the verb of the subordinate clause (*entra*). The functional category (e.g. the conjunction *que*) depends on the verb. Similarly, prepositional phrases are headed by the noun; the preposition (*entre*) depends on the noun (*cuises*).

(4) *Elle a juré [...] qu' entre ses cuises nus n' entra*
„She has sworn that between her thighs no one not entered"
(*Tristan* de Beroul, v. 121)

The dependency of functional elements is shown in (5), where the governing nodes are printed in bold and the functional categories are underlined.

(5) (**VFin *a*** (SjPer *elle*) (AuxA *juré*) (**Obj *entra*** (RelNC *qu'*) (SjPer *nus*) (Ng *n'*) (**Circ *cuises*** (RelNC *entre*) (Det *ses*))))

The structure in (5) also shows that in complex verb forms the finite verb (auxiliary or modal) dominates the non-finite verb (participle or infinitive): thus, *juré* depends on *a* at the same level as the subject *elle*.

One reason for preferring lexical governors is that functional categories are often absent in Medieval French (genitives without preposition, nouns without determiner, relative clauses without relative the pronoun etc.).

3. Annotation

3.1 The annotation procedure

Due to the limited size of the OF corpora (about 3 million words in each corpus, *BFM* and *NCA*, with a considerable number of shared texts), the *SRCMF* project adopted a manual annotation procedure during the three-year funding period, in order to provide resources which are as reliable as possible.

NotaBene is a tool for manual syntactic annotation (Mazziotta, 2010b).[6] It makes it possible to create and modify the syntactic annotation by means of a graphic interface. It allows the user to manipulate tree structures, to add free comments to any node of the structure, as well as to search and list them. Script-based semi-automatic correction is also provided, and text-specific or

[6] *NotaBene* is open-source and freely available on http://sourceforge.net/projects/NotaBene/.

user-specific annotations can be created by simple modification of labels. *NotaBene* can compare two versions of the same text and highlight the differences in the annotations. RDF graphs are used ("resource description format"; see Bechhofer et al. 2004) for the internal representation of the annotation, and dependency relations (i.e. governor-function-dependent triples) are expressed by RDF triples which form a directed graph. The RDF data is encoded in a W3C-defined XML format which can easily be converted. Although *NotaBene* can be freely adapted to other annotation tasks, a number of its functions are closely linked with the workflow of the *SRCMF* project (Figure 1).

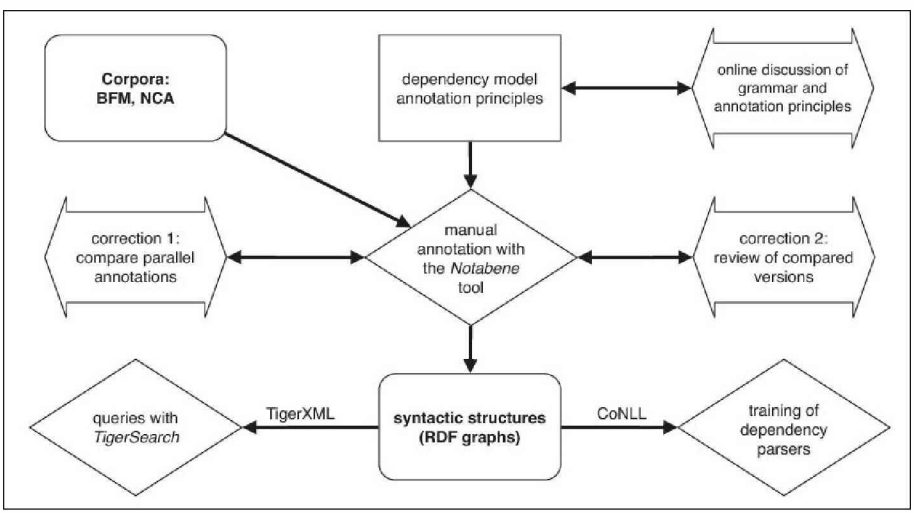


Figure 1: Annotation workflow of the *SRCMF* project

The manual annotation procedure has been designed to attain a high level of accuracy by means of redundancy. At the first level ("manual annotation", in Figure 1), two annotators produce two separate analyses of a text. At the next level ("correction 1"), they compare their analyses in order to eleminate annotation errors. In the next step ("correction 2"), two correctors compare and review both versions using the comparison function of the *NotaBene* tool, decide about cases of syntactic ambiguity, and produce the final version. This step is also executed using *NotaBene*, and the final result is therefore encoded in RDF graphs, and will be published in that format, which contains the complete information of the syntactic analysis.

3.2 Distribution formats and queries

The last two steps shown in Figure 1 are not part of the annotation procedure proper, but they exemplify the formats which can be derived from the RDF graphs. Currently, *NotaBene* can convert RDF into dot (*GraphViz*) format to visualize graph images, as well as into the two application-oriented formats TigerXML and CoNLL.

TigerXML has been specified for the *TigerSearch* query software (IMS, Stuttgart; Lezius 2002) and has been chosen because *TigerSearch* provides a user-friendly environment for syntactic queries, either as a stand-alone application[7] or as a plugin for the TXM platform.[8] Since TigerXML was conceived for the representation of constituent graphs (where words have to be terminal nodes), some modifications were necessary. TigerXML is being developed further in the *tiger2* project, one of whose goals consists in representing both constituency and dependency analyses simultaneously in the same graph.[9]

The other export format is the standard tabular format used in dependency parsing, as defined by the Conference on Computational Natural Language Learning (in the CoNLL 2009 shared task). One of the goals of the manual annotation is to provide a reliable gold-standard for the training of dependency parsers. Promising tests were made with the *mate-tools* (Bohnet 2010; Björkelund et al. 2010): unlike other graph-based dependency parsers, the *mate* parser implements a "maximum spanning tree" which not only considers the nodes depending directly on a given node, but also the grand-children and sibling nodes.

Due to this technique, *mate* is well suited for the *SRCMF* grammar model: as explained in Section 2.2, our grammar is verb-centered, i.e. the verb is the top node of main clauses as well as of subordinate clauses, and functional categories are dependent on the lexical ones. For the automatic analysis however, functional categories provide important information. Consider the example given in (4): for a dependency parser without 'maximum spanning tree', sub-

[7] For Windows, Mac and various versions of Unix, see http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/oldindex.shtml.

[8] TXM was developed in the project *Textométrie* at the École Normale Supérieure of Lyon, see Heiden et al. (2010).

[9] TigerXML is currently being elaborated in the *tiger2* project. One of its goals consists in representing both constituency and dependency analyses simultaneously in the same graph. For more information see http://korpling.german.hu-berlin.de/tiger2/.

ordination would be a mere verb-verb dependency (*a juré – entra*). The *mate* parser, however, looks further ahead to the functional category (*qu'*) and – judging by these very first tests – performs quite well even for complex structures like coordinations or subordinate clauses of this kind. In the unabbreviated version of the sentence (6), the coordinated predicates (*a juré* "has sworn" and *et mis en vo* "and put in oath") as well as the subordinate clause (*qu'entre...*) were analyzed correctly, although the parser had been trained on only 3,000 manually annotated sentences. The parser output shown in Figure 2 shows that only *nus* was erroneously analyzed, as an attributive adjective (ModA) instead of an indefinite subject pronoun.

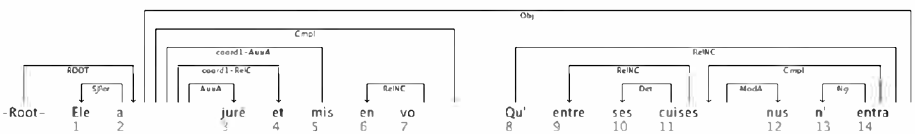


Figure 2: Automatic dependency annotation with the *mate* parser

4. Conclusions and perspectives

The *SRCMF* project is work in progress, and the manual annotation of the *BFM* and *NCA* corpora will be pursued even after the end of the funding period. The results will be published from 2012 on. The first tests with dependency parsers like *mate* have encouraged us to conclude that the combination of manually annotated training corpora and automatic parsing could be an interesting perspective for the continuation of the project.

References

Bechhofer, Sean/van Harmelen, Frank/Hendler, Jim/Horrocks, Ian/McGuinness, Deborah L.F./Patel-Schneider Peter/Andrea Stein, Lynn (2004): OWL Web Ontology Language Reference. W3C Recommendation 10 February 2004. http://www.w3.org/TR/owl-ref/.

Bjorkelund, Anders/Bohnet, Bernd/Hafdell, Love/Nugues, Pierre (2010): A high-performance syntactic and semantic dependency parser. In: Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations. Stroudsburg, PA: Association for Computational Linguistics, 33-36.

Bohnet, Bernd (2010): Top accuracy and fast dependency parsing is not a contradiction. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010). Beijing: Coling 2010 Organizing Committee, 89-97.

Bosco, Cristina (2004): A grammatical relation system for Treebank annotation. Unpublished PhD Thesis: Università degli Studi di Torino.

Guillot, Céline/Marchello-Nizia, Christiane/Lavrentiev, Alexeij (2007): La Base de Français Médiéval (BFM): états et perspectives. In: Kunstmann/Stein (eds.), 56-65.

Heiden, Serge/Magué, Jean-Philippe/Pincemin, Bénédicte (2010): TXM. Une plateforme logicielle open-source pour la textométrie – conception et développement. In: Bolasco, Sergio/Chiari, Isabella/Giuliano, Luca (eds.): Statistical analysis of textual data. Proceedings of 10th International Conference JADT 2010, Rome, 9-11 June 2010.

Kunstmann, Pierre/Stein, Achim (2007): Le Nouveau Corpus d'Amsterdam. In: Kunstmann/Stein (eds.), 9-27.

Kunstmann, Pierre/Stein, Achim (eds.) (2007): Le Nouveau Corpus d'Amsterdam. Actes de l'atelier de Lauterbad, 23-26 février 2006. Stuttgart: Steiner.

Lezius, Wolfgang (2002): Ein Suchwerkzeug für syntaktisch annotierte Textkorpora. Stuttgart: Institut für Maschinelle Sprachverarbeitung (IMS).

Martineau, France (2008): Un Corpus pour l'analyse de la variation et du changement linguistique. In: Corpus 7. Constitution et exploitation des corpus d'ancien et de moyen français, 135-155.

Martineau, France (2009): Le corpus MCVF. Modéliser le changement: les voies du français. Ottawa: Université d'Ottawa.

Mazziotta, Nicolas (2010a): Logiciel NotaBene pour l'annotation linguistique. Annotations et conceptualisations multiples. In: Recherches qualitatives. Hors-série 9, 83-94.

Mazziotta, Nicolas (2010b): Building the *Syntactic Reference Corpus of Medieval French* using *NotaBene RDF annotation tool*. In: Proceedings of the 4th Linguistic Annotation Workshop (LAW IV). http://www.aclweb.org/anthology-new/W/W10/W10-1820.pdf.

Polguère, Alain/Mel'čuk, Igor (2009): Dependency in linguistic description. Amsterdam/Philadelphia: Benjamins.

Prévost, Sophie (2003): Détachement et topicalisation: des niveaux d'analyse différents. In: Cahiers de praxématique 40, 97-126.

Stein, Achim et al. (2006): Nouveau Corpus d'Amsterdam. Corpus informatique de textes littéraires d'ancien français (ca 1150-1350), établi par Anthonij Dees (Amsterdam 1987), remanié par Achim Stein, Pierre Kunstmann et Martin-D. Gleßgen. Stuttgart: Institut für Linguistik/Romanistik.

Tesnière, Lucien (1965): Éléments de syntaxe structurale. Paris: Klincksieck.