Diachronic Corpora, Genre, and Language Change

EDITED BY

Richard J. Whitt

Studies in Corpus Linguistics 85

JOHN BENJAMINS PUBLISHING COMPANY

Diachronic Corpora, Genre, and Language Change

Studies in Corpus Linguistics (SCL)

ISSN 1388-0373

SCL focuses on the use of corpora throughout language study, the development of a quantitative approach to linguistics, the design and use of new tools for processing language texts, and the theoretical implications of a data-rich discipline.

For an overview of all books published in this series, please see
http://benjamins.com/catalog/scl

Volume 85

Diachronic Corpora, Genre, and Language Change
Edited by Richard J. Whitt

Diachronic Corpora, Genre, and Language Change

Edited by

Richard J. Whitt
University of Nottingham

John Benjamins Publishing Company
Amsterdam / Philadelphia

™ The paper used in this publication meets the minimum requirements of the American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Cover design: Françoise Berserik
Cover illustration from original painting *Random Order*
by Lorenzo Pezzatini, Florence, 1996.

DOI 10.1075/scl.85

Cataloging-in-Publication Data available from Library of Congress:
LCCN 2018027602 (PRINT) / 2018050031 (E-BOOK)

ISBN 978 90 272 0148 5 (HB)
ISBN 978 90 272 6350 6 (E-BOOK)

John Benjamins Publishing Company · https://benjamins.com

Table of contents

Part I: Methods in diachronic corpus linguistics

Part II: Genre and diachronic corpora

Part III: Genre-based analyses of linguistic phenomena

Preface and acknowledgments

Papers in this volume represent a selection of the presentations given at the *Diachronic Corpora, Genre, and Language Change* conference, which was held at the University of Nottingham on April 8–9, 2016, and featured forty presentations, being attended by seventy-five delegates representing sixteen countries. The impetus for this conference was the completion of my own project on "Evidentiality and Genre in the Histories of English and German", which was funded generously by the University of Nottingham's Nottingham Research Fellowship Scheme. This project was a corpus-based investigation into the connection between the use of evidential markers and genre variation in the histories of English and German (from the early modern period onwards) and employed a number of multi-genre and specialised diachronic corpora. After initial studies of multi-genre corpora failed to turn up any substantial connection between the development of evidential markers and genre – although relevant findings concerning grammaticalisation processes were discovered – attention was concentrated on scientific, particularly medical, discourse because this was one domain where a major epistemological shift (from Scholasticisim to Empiricism) concentrated on differing traditions valuing different sources of knowledge. This project also resulted in the creation and release of a small specialised corpus of early modern German texts on midwifery and gynaecology to enable comparison with extant resources for English.[1] During the course of this project, it was also noted that although there has been great interest recently in the development and application of diachronic corpora, there have been few if any concentrated efforts to focus discussion on the role genre[2] plays in language change and how diachronic corpora – most, if not all, having been created with genre factored into corpus architecture – can help us better understand this role. Particularly, what role can multi-genre diachronic corpora study play in the of phenomena of language change, including recent change, versus the role of single-genre diachronic

[1] *The Nottingham Corpus of Early Modern German Midwifery and Women's Medicine (ca. 1500–1700)*, available at the Oxford Text Archive. <http://ota.ox.ac.uk/desc/2562> (17 July 2017).

[2] Genre was broadly conceived as situated language usage for purposes of the conference, as different delegates proceeded with varying understandings of the concept, see introductory chapter.

corpora in studying change within a certain genre, including recent change? What computational and statistical methods can or should be applied to the analysis of data from multi- and single-genre diachronic corpora? Project reports of multi- and single-genre diachronic corpora currently in development were also deemed crucial to any examination of the state-of-the-art. Finally, a desideratum of the conference was to have as many languages as possible represented, and overall, a dozen languages (Arabic, Dutch, English, French, High German, Low German, Greek, Old Norse, Polish, Sardinian, Spanish, and Turkish) were discussed, with seven of these (Arabic, English, French, High German, Low German, Greek, and Spanish) being represented in the current volume.

I am grateful to The University of Nottingham for allowing me to pursue my project on historical genre studies and corpus linguistics, which included a generous amount of research support able to subsidise this conference. In particular, I would like to thank Nicola McLelland for her support of my application and further mentoring once I arrived on campus. She also provided great advice on conference organisation. Thanks are also due to Michaela Mahlberg, who officially opened the conference and gave a warm welcome to guests. I also owe a debt of gratitude to Nottingham Conferences, who arranged for a suitable conference space and provided excellent catering. Also, the staff of the Chatsworth House in Derbyshire enabled a stimulating post-conference excursion. The conference assistants who made sure things ran smoothly throughout the event deserve no less recognition: Jacqueline Cordell, Caitlin Sayce, Fen Gao and her daughter, Gloria. Finally, I owe my biggest debt of gratitude to my wife, Errolinda, who provided perennial support throughout the project and assisted during the conference as well.

Richard J. Whitt
Nottingham
April 2018

Using diachronic corpora to understand the connection between genre and language change

Richard J. Whitt
The University of Nottingham

1. Introduction

The emergence of diachronic corpus linguistics from early pioneering projects like the *Bonn Corpus of Early New High German* in the 1970s through major endeavors in English like the Helsinki and ARCHER corpora has allowed us unprecedented access to data in the study of past stages of language and linguistic change. Major areas of interest that have emerged in this field include the study of genre (broadly conceived) as a major factor affecting language variation and change, as well as the notion of genre as a locus of change. In addition, there has been great interest in how specific genres have (or have not) changed over time. The goal of the present volume is to present a state-of-the-art overview, drawing from a number of languages, as to where this relatively young field of research is at more than three decades after its inception.

2. What is genre?

Pinning down a precise definition of genre is nothing short of daring to find one's way through a "terminological maze" (Moessner 2001). Besides *genre*, terms such as *text type* and *register* appear interchangeably throughout linguistic literature, and the matter is further complicated when one attempts to draw comparisons between languages and the disparate use of terminology which various scholarly traditions behind these languages employ. There appears to be broad agreement in the literature that both linguistic and extralinguistic factors must be considered to arrive at a satisfactory working definition of "genre" (Miller 1984; Hyon 1996; Moessner 2001; Biber & Conrad 2009; Tardy & Swales 2014), that is, sociocultural considerations must be taken alongside linguistic ones. This includes issues

https://doi.org/10.1075/scl.85.01whi

related to social contexts and practices, i.e. the broader communicative context; the notion of genre can be seen as a form of social action, or as Miller (1984: 159) puts it, "typified rhetorical actions based in recurrent situations". But beyond this broad notion of social context, there is little unity in approach: Biber and Conrad (2009) distinguish genre from register, noting that the latter is what is actually based in the larger communicative context, whereas the former deals with conventional textual, rhetorical, and linguistic structures of specific texts. Hyon (1996) provides an overview of varying approaches to genre, noting that the tradition of New Rhetoric focus more on language external variables and ethnography (695–696), whereas linguistic approaches such as Hallidayan Systemic Functional Linguistics (696–697) focus more on the linguistic realisations of social processes (cf. Halliday 1978; Eggins & Martin 1997). The tradition of English for Specific Purposes (ESP) focuses on both formal properties and social contexts as part of genre, being particularly interested in structural move analysis, i.e. the study of a text's global organisational patterns (Hyon 1996: 695; cf. Swales 1990). Literary scholar and language philosopher Mikhail Bakhtin (1986) sees genre as a combination of thematic content, style, and compositional structure, which themselves are "inseparably linked to the *whole* [emphasis Bakhtin's] of the utterance and are equally determined by the specific nature of the particular sphere of communication" (1986: 60).

Adding the historical dimension to genre studies – or adding the genre dimension to historical studies – only complicates the maze further. This is nevertheless a necessary complication, for as Bakhtin puts it, "historical changes in language styles are inseparably linked to changes in speech genres" (1986: 65; cf. Diller 2001: 3). Therefore any examination of how genre changes have affected language change (or vice versa) will have to consider extralinguistic factors (Jucker & Taavitsainen 2013: 149–151). Swales' notion of discourse communities (1990) can be helpful here, as it can illuminate the precise social relationships that exist between language users/text producers on the one hand and text recipients on the other (Jucker & Taavitsainen 2013: 151–153; cf. Jauss 1979 on the notion of "horizons of expectation"). However, as Diller (2001: 20ff.) notes, this can be difficult to pin down in diachronic terms because historical linguists tend to merely "use" the notion of genre rather than define or problematise it. Perhaps, but as we will see in Section 3, the compilers of diachronic corpora have been quite aware of the challenges posed by imposing genre labels in text classification procedures, and although not all approaches have proceeded along the same lines, it is worth noting there are more similarities than differences in how many final categorisation schemata of diachronic corpora turned out.

3. Diachronic corpora: Challenges in design, compilation, and use

Efforts to compile diachronic corpora began not long after the advent of corpus linguistics in the 1960s.[1] In the mid-1970s, for example, Klaus-Peter Wegera and his associates compiled the *Bonn Corpus of Early New High German* (*Bonner Frühneuhochdeutschkorpus*),[2] covering the years 1350 through 1699 and containing a number of *Textgattungen* 'text types': academic/scientific, chronicles, literary, devotional, and theological (Wegera 2013; see also papers in Lenders & Wegera 1982 and the corpus web-site, fn. 2). The most widely known of the early diachronic corpora, though, is no doubt the *Helsinki Corpus of English Texts*,[3] which was compiled by Matti Rissanen and his team at the University of Helsinki in the 1980s and released in 1991; it covers the history of English from Old English through the early modern period (ca. 730–1710) and contains a number of genres, including legal texts, handbooks (for Old and Middle English), Bible translations, philosophy, fiction, biography, correspondence, and several others (Rissanen 2009; Nevalainen 2013; see also the project web-site, fn. 3). This was followed shortly thereafter by another substantial diachronic corpus of English, ARCHER (*A Representative Corpus of Historical English Registers*),[4] compiled by Douglas Biber and colleagues in the early 1990s. This covers the years 1650 to 1999 and contains twelve genres: advertising, drama, fiction, sermons, journals, legal, medicine, news, early prose, science, letters, and diaries (Biber et al. 1993; Yáñez-Bouza 2011; see also links in fn. 4). All of these early diachronic corpora have undergone further modifications, and numerous other diachronic corpora (multi-genre and specialised) have been created in the meantime, but as can be seen here, genre classification is one of the key features of each of these corpora.

There are several challenges faced by compilers and users of diachronic corpora that are inherent when dealing with historical materials (Rissanen 1989, 2009: 64–66; cf. Nevalainen 2013: 38). For one, there is the risk that corpus data will be taken to be the language itself (known as the "God's truth fallacy"), rather than simply a sampling thereof. This is also a problem for corpus data on present-day language varieties, but it poses a greater challenge for historical linguistics

[1] Johansson (2009) provides a good historical overview of the early days of corpus linguistics.

[2] <https://korpora.zim.uni-duisburg-essen.de/Fnhd/> (11 July 2017).

[3] <http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/> (10 July 2017).

[4] <http://www.projects.alc.manchester.ac.uk/archer/>; see also <http://www.helsinki.fi/varieng/CoRD/corpora/ARCHER/updated%20version/background.html> (11 July 2017).

because fewer materials are available for past stages of a language, and the further back one goes, the more scarce materials become. This ties in with Labov's (1992) notion of "bad data" and the challenge to make corpus data representative of the language variety (or period) it is supposed to exemplify (see also Claridge 2009: 246–249; Curzan 2009: 1094; Wegera 2013: 64–67; Durrell 2015). The difficulties in avoiding this fallacy will obviously vary depending on which period(s) and genre(s) one is interested in. Another issue is the "mystery of vanishing evidence", whereby low-frequency phenomena become more difficult to track with smaller data sets; and most historical corpora are considered quite "small" when compared to the likes of present-day corpora such as the *British National Corpus* (BNC), which contains roughly 100 million words (Claridge 2009: 245; Wegera 2013: 56–59). Indeed, the size of one's data set determines to what degree quantitative results can be subject to statistical testing and generalisability (Hilpert & Gries 2016). Finally, "the philologist's dilemma" connects meaningful and correct conclusions from the data to a good command of the language (and texts) in question, which is at issue when one is dealing with a large number of texts and/or covering a broad time span. Again, the nature of one's own research interests dictates how many (or few) precautions must be taken when investigating genre-related diachronic phenomena.

Another challenge (and benefit) to users and compilers of diachronic corpora is the ability to represent language change both "from above" and "from below" (Labov 1994: 78; cf. Rissanen 2009: 57); that is, not only can or should genres representing the upper echelons of society (language "from above") – such as literary or academic texts – be considered, but so too should the genres representing the language of "normal" or non-elite people (language "from below") – diaries, private letters, etc. – be considered as drivers of language change. Closely related to this are the notions of immediacy and distance in language (Koch & Oesterreicher 1985, 2007, 2012; see also Elspaß 2014): some genres are more representative of face-to-face, spoken, interpersonal ("immediate") communication while other genres are of a more impersonal, written, and formal ("distant") nature, and a continuum exists between these two poles. Biber's (1986, 1988) multi-dimensional approach to variation makes a similar contrast. Such notional distinctions were clearly at work in the selection of genres to be included in both the Helsinki and ARCHER Corpora, as well as more specialised corpora such as the *Corpus of English Dialogues*[5] (Culpeper & Kytö 2010) or the *Corpus of Cape Dutch Correspondence* (Deumert 2004).

[5] <http://www.engelska.uu.se/forskning/engelska-spraket/elektroniska-resurser/a-corpus> (12 July 2017).

There are also a number of technical challenges facing compilers of diachronic corpora. The most obvious is getting texts into machine-readable form, and where optical character recognition (OCR) software proves unfeasible (as it often does), manual inputting – a timely and costly affair – is required, although advances are being made with OCR and ancillary technologies (see, for example, Springmann & Lüdeling 2017). Once texts are in machine-readable form, the degree of spelling variation one finds in texts that pre-date standardisation efforts, or from writers who knowingly or unknowingly deviate from any extant orthographic prescriptions, can pose a significant challenge for tagging and annotation efforts. For English, this has been alleviated to some degree by the Variant Detector (VARD) software programme,[6] developed at the University of Lancaster, which can be trained to automatically normalise the spelling in early modern English texts (Baron & Rayson 2008), thus enabling more accurate forms of automatic tagging (Rayson et al. 2007). Where such programmes do not exist, compilers can apply more bootstrapping approaches involving varying degrees of manual training/annotation and automatic tagging (Kroch & Taylor 2000; Kroch et al. 2004; Dipper 2011; Rögnvaldsson et al. 2011; Walkden 2016); compilers can also train their tagging or annotation software on a small sample of their corpus (a "gold standard" corpus), and subsequently apply the trained tagger to the entire corpus (Sánchez-Marco et al. 2011; Scheible et al. 2011, 2012). The majority of these efforts have focused on morphosyntactic tagging, although there have been efforts to tag (diachronic) corpora for semantic and pragmatic information as well (Rayson & Stevenson 2009; Archer et al. 2009; Archer 2014).

A brief word should also be said about the notions of "corpus-based" vs. "corpus-driven" approaches to linguistics (Tognini-Bonelli 2001: 65ff.; McEnery & Hardie 2012: 5–6, 147ff.) and their connection to diachronic language study related to genre. Corpus-based approaches to language employ corpora to test, validate, refute, or refine an extant theory of linguistic phenomena in the literature, whereas corpus-driven approaches – at least in their most extreme form (see, for example, Teubert 2005) – view corpus data as an end in and of themselves: the corpus itself has theoretical status and can serve as the basis of drawing conclusions and theorising about language. To do such with diachronic data would clearly make one guilty of engaging in the "God's truth fallacy" as discussed by Rissanen (1989, 2009) and Nevalainen (2013), especially when looking at genre factors, which require linguistic data to be viewed in light of language external considerations such as writer-addressee relations, discourse communities (Swales 1990), and "horizons of expectations" (Jauss 1979). In practice this is rarely if ever

[6] <http://ucrel.lancs.ac.uk/vard/about/> (12 July 2017).

done among corpus linguists studying present-day varieties (McEnery & Hardie 2012: 150–152), and the most "corpus-driven" approach one finds among historical linguists is better known as the "bottom-up" method (Pahta & Taavitsainen 2010: 563; cf. Grund 2012, Whitt 2016a, b): a small sub-section of the corpus is read closely to determine what features might best serve for a larger corpus search, and this is often used in tandem with the "top-down" method, whereby items to be searched for have been determined beforehand (such as what has already been discussed in the relevant literature).

4. Some diachronic corpora

English historical linguistics is without question in the lead when it comes to scope and availability of multi- and single-genre diachronic corpora. The Helsinki and ARCHER corpora, discussed above, were among the first resources available in the field of diachronic corpus linguistics. And in the meantime, a number of larger and/or more specialised corpora have become available. For example, *The Corpus of Historical American English (COHA)*[7] covers the years 1810 to the early 2000s and contains 400 million words (Davies 2012), whereas Helsinki contains ca. 1.5 million words and ARCHER contains just over 3 million words (see Nevalainen 2013 for a discussion of the development of the field of English diachronic corpus linguistics). Besides the *Corpus of English Dialogues* mentioned earlier, there are a number of other specialised diachronic corpora available: the *Electronic Text Edition of Depositions 1560–1760 (ETED)* corpus (Kytö et al. 2011); the *Corpus of Early English Correspondence (CEEC)*[8] in its several versions (Nurmi 1998; see also Auer et al. 2015); a number of corpora devoted to the history of medical writing in English since the medieval period (Taavitsainen et al. 2005; Taavitsainen et al. 2010; Taavitsainen et al. 2014; see also Taavitsainen & Pahta 2004, 2011);[9] the *Coruña Corpus of English Scientific Writing* (Moskowich & Crespo García 2007; Moskowich & Parapar López 2008; see also Moskowich & Crespo García 2012, Moskowich et al. 2016), which contains several corpora covering scientific writing in the history of English;[10] the *Hansard Corpus*, containing British parliamentary

[7] <http://corpus.byu.edu/coha/> (14 July 2017).

[8] <http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/index.html> (14 July 2017).

[9] <http://www.helsinki.fi/varieng/CoRD/corpora/CEEM/> (14 July 2017).

[10] <http://www.udc.es/grupos/muste/corunacorpus/index.html> (14 July 2017).

speeches from 1803 to 2005;[11] and there are numerous others (Kytö 2010 provides an excellent overview of resources compiled up to 2010; see also http://www.helsinki.fi/varieng/CoRD/corpora/index.html, 14 July 2017).

Some multi- and single-genre diachronic corpora exist for languages other than English. Besides the *Bonn Corpus of Early New High German* mentioned above, the *German Manchester Corpus (GerManC)* provides a representative sample of German from 1650 to 1800 and contains a number of texts reflecting both "immediate" (drama, letters, newspapers, sermons) and "distant" (legal, humanistic, narrative, scientific) genres (Durrell et al. 2012; cf. Koch & Oesterreicher 2012).[12] There is also the more specialised *Nottingham Corpus of Early Modern German Midwifery and Women's Medicine, ca. 1500–1700 (GeMi)* (Whitt 2016c), which provides a sampling of the earliest vernacular texts devoted to midwifery and gynaecology to be printed in German. *The Compilation Corpus of Historical Dutch* provides a sampling of Dutch-language chancellery and narrative texts dating from the medieval period to 2000 (Coussé 2010).[13] And at Brigham Young University, Mark Davies has compiled both the *Corpus del Español* and the *Corpus do Português*, which contain millions of words representing several genres in both Spanish and Portuguese dating back to the thirteenth and fourteenth centuries, respectively.[14] A number of other multi- and single-genre diachronic corpora on languages other than English will be discussed in the present volume (see papers by Atwell, Enrique-Arias, Farasyn et al., Niehaus & Elspaß, and Trips & Stein).

There are also a number of massive on-line text collections which, although not corpora in the strictest sense of the word because they are not balanced for time period or genre (see Wegera 2013: 55–59), can still provide resources for the study of genre and language change, as well as for the creation of *ad hoc* diachronic corpora. For English, the Text Creation Partnership (TCP) offers *Early English Books Online (EEBO)*, which contains thousands of texts dating from the fifteenth century through the seventeenth century, the *Eighteenth Century Collection Online (ECCO)*, and *Evans Early American Imprints*.[15] Texts in EEBO formed the basis of the *Corpus of Early Modern English Medical Texts (EMEMT)* (Taavitsainen & Pahta 2010), and as McEnery and Baker (2016) have shown, even EEBO itself can

[11] <http://www.hansard-corpus.org/intro.asp> (14 July 2017).

[12] <http://www.alc.manchester.ac.uk/modern-languages/research/german-studies/germanc/> (17 July 2017).

[13] <http://www.eviecousse.be/compilationcorpus.htm> (17 July 2017).

[14] <http://corpus.byu.edu/neh2015.asp> (17 July 2017).

[15] <http://www.textcreationpartnership.org/> (17 July 2017).

be used successfully for corpus-based sociohistorical investigations. The *Deutsches Textarchiv* ('German Text Archive') contains a number of texts from the seventeenth to the early twentieth century and can also be used for corpus linguistic investigations (Geyken & Gloning 2015; cf. Thomas & Wiegand 2015).[16] Finally, the *EuroDocs* site interfaces with a number of other sites and provides links to various historically-oriented text collections from over four dozen languages.[17] The above is hardly an exhaustive list of resources, but it does provide an overview of what kinds of text collections are available for use.

5. The present volume

This volume seeks to complement other collections of papers on diachronic corpus linguistics (Bennett et al. 2013; Gippert & Gehrke 2015) by focusing exclusively on issues related to the role of genre in the creation and analysis of historical corpora. It is divided into three sections. The focus of Part I is on "Methods in diachronic corpus linguistics" and presents papers whose primary focus is to highlight new or pertinent methods in the compilation and use of diachronic corpora. It begins with "'From above', 'from below' and regionally balanced: Towards a new corpus of nineteenth-century German" by Konstantin Niehaus and Stephan Elspaß, who discuss the importance of accounting for regional and register variation when compiling a historical corpus, particularly of an under-studied period of a language such as nineteenth-century German. They demonstrate the significance of including both genres "from below" and "from above" in their corpus through case studies on both syntactic (*Ausklammerung* 'exbraciation') and morphological (diminutives, plural forms) phenomena. Bryan Jurish, in "Diachronic collocations and genre", presents the open-source software tool "DiaCollo" and explains how it can be deployed in analyses of diachronically-oriented studies of collocations and their connection to genre. The section concludes with Eric Atwell's paper on "Classical and modern Arabic corpora: Genre and language change", in which the rich linguistic annotation surrounding corpora of classical Arabic – based on the religious texts of the Quran and Hadith – can be applied to NLP research on modern-day Arabic corpora which feature substantially different genres, such as Twitter tweets and Amazon customer reviews.

[16] <http://www.deutschestextarchiv.de/> (17 July 2017).

[17] <https://eudocs.lib.byu.edu/index.php/Main_Page> (17 July 2017). Note, however, that many of the resources available here are simply digital facsimiles, which means additional work would be necessary to make them machine readable.

Part II, “Genre and diachronic corpora”, homes in on how the study of change in a single domain of usage can be facilitated by diachronic corpus linguistics. Irma Taavitsainen opens this section with “Scholastic genre scripts in English medical writing 1375–1800”, in which she discusses the notion of “genre scripts” and how the “top” levels of medical writing created during the period of medieval Scholasticism (such as commentaries) enjoyed afterlives long after the scholastic tradition had come to an end. In “Academic writing as a locus of grammatical change: The development of phrasal complexity features”, Bethany Gray and Douglas Biber use a number of corpora to argue against the long-held belief that academic writing, being a more “conservative” register, is resistant to change. In fact, they argue that academic writing is a hotbed of change when it comes to phrasal complexity features and that much internal variation exists among different disciplines of writing.

Part III (“Genre-based analyses of linguistic phenomena”) contains a variety of papers using multi-genre corpora to see the connection between genre and particular phenomena of language change. Georgia Fragaki and Dionysis Goutsos’ contribution on “The importance of genre in the Greek diglossia of the 20th century: A diachronic corpus study of recent language change” begins the discussion; here, the authors show how genre has been a driver in recent (twentieth-century) change in Greek, particularly in the area of diglossia where variant forms, both “high” and “low”, for certain items like prepositions exist, and such usage depends on generic context. The next contribution comes from Florian Haas, whose examination of the second-person impersonal pronoun *you* in “‘You can’t control a thing like that’: Genres and changes in Modern English human impersonal pronouns” reveals that this pronoun has recently gained in frequency and enjoyed some degree of functional specialisation, and that although genre is certainly pertinent to these changes, other cross-linguistic factors may be at work as well. Change in English is also in focus in Ole Schützler’s “Concessive conjunctions in written American English: Diachronic and genre-related changes in frequency and semantics”, in which the author shows that semantic changes affecting certain conjunctions and prepositions can occur across a number of genres at the same time. Karolina Rudnicka takes up the issue of sentence length in “Variation of sentence length across time and genre: Influence on syntactic usage in English” and discusses the role genre has played in changes related to sentence length in written late modern English. The role of genre in medieval English-French contact-induced change is discussed in Carola Trips and Achim Stein’s contribution, “A comparison of multi-genre and single-genre corpora in the context of contact-induced change”. The authors draw data from a number of multi- and single-genre diachronic corpora to highlight how change in use of the recipient passive in English is likely to be due to contact with French and that genre-specific usage – particularly letter writing – can highlight the nature of such contact-induced change. The benefit of using parallel corpora in studying language change is discussed by Andrés Enrique-Arias in "Some methodological issues in the corpus-based study of morphosyntactic variation: The case of Old Spanish possessives". Here, Enrique-Arias discusses how ART+POSS constructions vary across certain biblical genres and translations into Old Spanish and how a parallel corpus can be used not only to track morphosyntactic change across time, but also to account for generic factors contributing to such variation and change. Melissa Farasyn, George Walkden, Sheila Watts, and Anne Breitbarth argue in "The interplay between genre variation and syntax in a historical Low German corpus" that genre must be considered a factor when accounting for syntactic change, as their study of null referential subjects, resumptive pronouns, relative particles, and coordination gaps in Middle Low German demonstrates that the inclusion or exclusion of the genre variable can produce different results. Finally, Luise Kempf concludes the section, and the volume, with her discussion of "Genre influence on word formation (change): A case study of German adjectival derivation", in which newspaper and scientific texts are identified as particularly innovative in mustering change in the domain of German derivational morphology.

6. Reflection

The papers in this volume certainly give credence to Bakhtin's thesis that "historical changes in language styles are inseparably linked to changes in speech genres" (1986: 65), and diachronic corpus linguistics has played a major role bringing genre to the forefront of studies on language change. Even so, the notion of genre itself remains elusive after nearly three decades of the discipline, and the papers in this volume admittedly reflect the multiplicity of understandings that still exist; the terminology ranges from "genre" to "register" to "text type" in the contributions found here. This may be inevitable if we are to take Mair's (2009: 1122–1123) plea for methodological pluralism in diachronic corpus linguistics to heart, given how many other research traditions appropriate the notion of "genre" in one form or another. That said, Diller's (2001: 20ff.) critique remains valid that historical linguists (vis-à-vis diachronic corpus linguists) often use the term without critically defining or problematising it, and the best way to remedy this in the future is to develop a better cross-linguistic understanding of situated language usage – i.e., are the similarities and disparities that exist between different "genres" of a similar or differing nature in different languages, and do the changes that occur over time in said genres proceed in a similar or different fashion across the different languages? Only one of the papers in the present volume is cross-linguistic in nature (the contribution by Trips and Stein – although Enrique-Arias does consider the

Hebrew and Latin source material of the Old Spanish Bible translations as well), and the problem here is that not enough corpus resources exist to enable comparative study. As mentioned earlier, a plethora of excellent resources exist for studying the history of the English language, but this is not necessarily the case for other languages. Maybe when such resources are finally available will we be able to arrive at a more satisfactory understanding of genre and its role in language change.

References

Archer, Dawn. 2014. Historical pragmatics: Evidence from the Old Bailey. *Transactions of the Philological Society* 112(2): 259–277. https://doi.org/10.1111/1467-968X.12011

Archer, Dawn, Culpeper, Jonathan & Davies, Matthew. 2009. Pragmatic annotation. In *Corpus Linguistics: An International Handbook*, Vol. 1, Anke Lüdeling & Merja Kytö (eds), 613–642. Berlin: De Gruyter.

Auer, Anita, Schreier, Daniel & Watts, Richard J. 2015. *Letter Writing and Language Change*. Cambridge: CUP. https://doi.org/10.1017/CBO9781139088275

Bakhtin, Mikhail. 1986. The problem of speech genres. In *Speech Genres and Other Late Essays by M. M. Bakhtin*, Caryl Emerson & Michael Holquist (eds), Vern W. McGee (trans.), 60–102. Austin TX: University of Texas Press.

Baron, Alistair & Rayson, Paul. 2008. VARD 2: A tool for dealing with spelling variation in historical corpora. Proceedings of the *Postgraduate Conference in Corpus Linguistics, Aston University*, Birmingham, UK, 22 May 2008. <http://acorn.aston.ac.uk/conf_proceedings.html> (12 July 2017).

Bennett, Paul, Durrell, Martin, Scheible, Silke & Whitt, Richard J. 2013. *New Methods in Historical Corpora*. Tübingen: Narr.

Biber, Douglas. 1986. Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language* 62(2): 384–414. https://doi.org/10.2307/414678

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511621024

Biber, Douglas, Finegan, Edward & Atkinson, Dwight. 1993. ARCHER and its challenges: Compiling and exploring a representative corpus of historical English registers. In *English Language Corpora. Design, Analysis and Exploitation*, Jan Aarts, Pieter de Haan & Nelleke Oostdijk (eds), 1–13. Amsterdam: Rodopi.

Biber, Douglas & Conrad, Susan. 2009. *Register, Genre, and Style*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511814358

Claridge, Claudia. 2009. Historical corpora. In *Corpus Linguistics: An International Handbook*, Vol. 1, Anke Lüdeling & Merja Kytö (eds), 242–259. Berlin: De Gruyter.

Coussé, Evie. 2010. Een digitaal compilatiecorpus historisch Nederlands. *Lexikos* 20: 123–142. https://doi.org/10.4314/lex.v20i1.62688

Culpeper, Jonathan & Kytö, Merja. 2010. *Early Modern English Dialogues: Spoken Interaction as Writing*. Cambridge: CUP.

Curzan, Anne. 2009. Historical corpus linguistics and evidence of language change. In *Corpus Linguistics: An International Handbook*, Vol. 2, Anke Lüdeling & Merja Kytö (eds), 1091–1109. Berlin: De Gruyter. https://doi.org/10.1515/9783110213881.2.1091

Davies, Mark. 2012. Expanding horizons in historical linguistics with the 400 million word Corpus of Historical American English. *Corpora* 7: 121–157. https://doi.org/10.3366/cor.2012.0024

Deumert, Ana. 2004. *Language Standardization and Language Change: The Dynamics of Cape Dutch* [Impact: Studies in Language and Society 19]. Amsterdam: John Benjamins. https://doi.org/10.1075/impact.19

Diller, Hans-Jürgen. 2001. Genre in linguistic and related discourses. In *Towards a History of English as a History of Genres*, Hans-Jürgen Diller & Manfred Görlach (eds), 3–43. Heidelberg: C. Winter.

Dipper, Stefanie. 2011. Morphological and part-of-speech tagging of historical language data: A comparison. *Journal for Language Technology and Computational Linguistics* 26(2): 25–37.

Durrell, Martin. 2015. 'Representativeness', 'bad data', and legitimate expectations: What can an electronic historical corpus tell us what we didn't actually know already (and how)? In Gippert & Gehrke (eds), 13–33.

Durrell, Martin, Bennett, Paul, Scheible, Silke, Whitt, Richard J. & Ensslin, Astrid. 2012. *GerManC Corpus: A Representative, Multi-Genre Corpus of Early Modern German, 1650–1800*. Oxford: Oxford Text Archive. <http://ota.ahds.ac.uk/headers/2544.xml> (17 July 2017).

Eggins, Suzanne & Martin, James R. 1997. Genres and registers of discourse. In *Discourse as Structure and Process*, Teun A. van Dijk (ed.), 230–256. London: Sage.

Elspaß, Stephan. 2014. The use of private letters and diaries in sociolinguistic investigation. In *The Handbook of Historical Sociolinguistics*, Juan M. Hernándex-Campoy & J. Camilo Conde-Silvestre (eds), 156–169. Chichester: Wiley Blackwell.

Geyken, Alexander & Gloning, Thomas. 2015. A living text archive of 15th–19th-century German: Corpus strategies, technology, organization. In Gippert & Gehrke (eds), 165–179.

Gippert, Jost & Gehrke, Ralf. 2015. *Historical Corpora: Challenges and Perspectives*. Tübingen: Narr.

Grund, Peter J. 2012. The nature of knowledge: Evidence and evidentiality in the witness depositions from the Salem witch trials. *American Speech* 87(1): 7–38. https://doi.org/10.1215/00031283-1599941

Halliday, Michael A. K. 1978. *Language as a Social Semiotic: The Social Interpretation of Language and Meaning*. London: Edward Arnold.

Hilpert, Martin & Gries, Stefan Th. 2016. Quantitative approaches to diachronic corpus linguistics. In *The Cambridge Handbook of English Historical Linguistics*, Merja Kytö & Päivi Pahta (eds), 36–53. Cambridge: CUP. https://doi.org/10.1017/CBO9781139600231.003

Hyon, Sunny. 1996. Genre in three traditions: Implications for ESL. *TESOL Quarterly* 30(4): 693–722. https://doi.org/10.2307/3587930

Jauss, Hans Robert. 1979. The alterity and modernity of medieval literature. *New Literary History* 10(2): 181–229. https://doi.org/10.2307/468759

Johansson, Stig. 2009. Some aspects of the development of corpus linguistics in the 1970s and 1980s. In *Corpus Linguistics: An International Handbook*, Vol. 1, Anke Lüdeling & Merja Kytö (eds), 33–53. Berlin: De Gruyter.

Jucker, Andreas H. & Taavitsainen, Irma. 2013. *English Historical Pragmatics*. Edinburgh: EUP.

Koch, Peter & Oesterreicher, Wulf. 1985. Sprache der Nähe – Sprache der Distanz: Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch* 36: 15–43.

Koch, Peter & Oesterreicher, Wulf. 2007. Schriftlichkeit und kommunkative Distanz. *Zeitschrift für Germanistische Linguistik* 35: 346–375. https://doi.org/10.1515/zgl.2007.024

Koch, Peter & Oesterreicher, Wulf. 2012. Language of immediacy – language of distance: Orality and literacy from the perspective of language theory and linguistic history. In *Communicative Spaces. Variation, Contact, and Change. Papers in Honour of Ursula Schaefer*, Claudia Lange, Beatrix Weber & Göran Wolf (eds), 441–473. Frankfurt: Peter Lang.

Kroch, Anthony & Taylor, Ann. 2000. *The Penn-Helsinki Parsed Corpus of Middle English (PPCME2)*, 2nd edn. Department of Linguistics, University of Pennsylvania. <https://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-4/index.html> (12 July 2017).

Kroch, Anthony, Santorini, Beatrice & Delfs, Lauren. 2004. *The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME)*. Department of Linguistics, University of Pennsylvania. <http://www.ling.upenn.edu/histcorpora/PPCEME-RELEASE-3/> (12 July 2017).

Kytö, Merja. 2010. Data in historical pragmatics. In *The Handbook of Historical Pragmatics*, Andreas H. Jucker & Irma Taavitsainen (eds), 33–67. Berlin: De Gruyter.

Kytö, Merja, Grund, Peter J. & Walker, Terry. 2011. *Testifying to Language and Life in Early Modern England*. Amsterdam: John Benjamins. https://doi.org/10.1075/z.162

Labov, William. 1992. Some principles of linguistic methodology. *Language in Society* 1: 97–120. https://doi.org/10.1017/S0047404500006576

Labov, William. 1994. *Principles of Linguistic Change, Vol. 1: Internal Factors*. Oxford: Blackwell.

Lenders, Winfried & Wegera, Klaus-Peter. 1982. *Maschinelle Auswertung sprachhistorischer Quellen: Ein Bericht zur computerunterstützten Analyse der Flexionsmorphologie des Frühneuhochdeutschen*. Tübingen: Niemeyer.

Mair, Christian. 2009. Corpora and the study of recent change in language. In *Corpus Linguistics: An International Handbook*, Vol. 2, Anke Lüdeling & Merja Kytö (eds), 1109–1125. Berlin: De Gruyter. https://doi.org/10.1515/9783110213881.2.1109

McEnery, Tony & Hardie, Andrew. 2012. *Corpus Linguistics*. Cambridge: CUP. https://doi.org/10.1093/oxfordhb/9780199276349.013.0024

McEnery, Tony & Baker, Helen. 2016. *Corpus Linguistics and 17th-Century Prostitution: Computational Linguistics and History*. London: Bloomsbury.

Miller, Carolyn. 1984. Genre as social action. *The Quarterly Journal of Speech* 70: 151–167. https://doi.org/10.1080/00335638409383686

Moessner, Lilo. 2001. Genre, text type, style, register: A terminological maze? *European Journal of English Studies* 5(2): 131–138. https://doi.org/10.1076/ejes.5.2.131.7312

Moskowich, Isabel & Crespo García, Begoña. 2007. Presenting the Coruña Corpus: A collection of samples for the historical study of English scientific writing. In *'Of Varying Language and Opposing Creed': New Insights into Late Modern English*, Javier Pérez Guerra, Dolores González-Álvarez, Jorge L. Bueno-Alonso & Esperanza Rama-Martínez (eds), 341–357. Bern: Peter Lang.

Moskowich, Isabel & Crespo García, Begoña. 2012. *Astronomy 'playne and simple': The Writing of Science between 1700 and 1900*. Amsterdam: John Benjamins.

Moskowich, Isabel, Camiña Rioboo, Gonzalo, Lareo, Inés & Crespo García, Begoña. 2016. *'The Conditioned and the Unconditioned': Late Modern English Texts on Philosophy*. Amsterdam: John Benjamins.

Moskowich, Isabel & Paraper López, Javier. 2008. Writing science, compiling science: The Coruña Corpus of English Scientific Writing. In *Proceedings from the 31st AEDEAN Conference*, María Jesús Lorenzo Modia (ed.), 531–544. A Coruña: Universidade da Coruña.

Nevalainen, Terttu. 2013. English historical corpora in transition: From new tools to legacy corpora? In Bennett et al. (eds), 37–53. Tübingen: Narr.

Nurmi, Arja. 1998. *Manual for the Corpus of Early English Correspondence Sampler (CEECS)*. Department of English, The University of Helsinki. <http://clu.uni.no/icame/manuals/CEECS/INDEX.HTM> (14 July 2017).

Pahta, Päivi & Taavitsainen, Irma. 2010. Scientific discourse. In *The Handbook of Historical Pragmatics*, Andreas H. Jucker & Irma Taavitsainen (eds), 549–586. Berlin: De Gruyter.

Rayson, Paul, Archer, Dawn, Baron, Alistair & Smith, Nick. 2007. Tagging historical corpora – The problem of spelling variation. Proceedings of *Digital Historical Corpora, Dagstuhl-Seminar 06491*, International Conference and Research Center for Computer Science, Schloss Dagstuhl, Wadern, Germany, 3rd–8th December 2006. <http://drops.dagstuhl.de/portals/index.php?semnr=06491> (12 July 2017).

Rayson, Paul & Stevenson, Mark. 2009. Sense and semantic tagging. In *Corpus Linguistics: An International Handbook*, Vol. 1, Anke Lüdeling & Merja Kytö (eds), 564–579. Berlin: De Gruyter.

Rissanen, Matti. 1989. Three problems connected with the use of diachronic corpora. *ICAME Journal* 13: 16–19.

Rissanen, Matti. 2009. Corpus linguistics and historical linguistics. In *Corpus Linguistics: An International Handbook*, Vol. 1, Anke Lüdeling & Merja Kytö (eds), 53–68. Berlin: De Gruyter.

Rögnvaldsson, Eiríkur, Ingason, Anton Karl, Sigurðsson, Einar Freyr & Wallenberg, Joel. 2011. Creating a dual-purpose treebank. *Journal for Language Technology and Computational Linguistics* 26(2): 141–152.

Sánchez-Marco, Cristina, Boleda, Gemma & Padró, Lluís. 2011. Extending the tool, or how to annotate historical language varieties. Proceedings of the *ACL-HLT 2011 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2011)*. Portland, Oregon. June 24, 2011. 1–9. <http://dl.acm.org/citation.cfm?id=2107636&picked=prox> (12 July 2017).

Scheible, Silke, Whitt, Richard J., Durrell, Martin & Bennett, Paul. 2011. A gold standard corpus of early modern German. Proceedings of the *5th LAW Workshop*. Portland, Oregon, June 23–24, 2011, 124–128. <http://dl.acm.org/citation.cfm?id=2018966&picked=prox> (12 July 2017).

Scheible, Silke, Whitt, Richard J., Durrell, Martin & Bennett, Paul. 2012. GATE to GerManC: A GATE-based annotation pipeline for historical German. Proceedings of the *8th International Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul, Turkey, May 21–27, 2012, 3611–3617. <http://www.lrec-conf.org/proceedings/lrec2012/summaries/978.html> (12 July 2017).

Springmann, Uwe & Lüdeling, Anke. 2017. OCR of historical printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus. *Digital Humanities Quarterly* 11(2). <http://www.digitalhumanities.org/dhq/vol/11/2/000288/000288.html> (12 July 2017).

Swales, John M. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge: CUP.

Taavitsainen, Irma & Pahta, Päivi. 2004. *Medical and Scientific Writing in Late Medieval English*. Cambridge: CUP.

Taavitsainen, Irma & Pahta, Päivi. 2010. *Early Modern English Medical Texts*. Amsterdam: John Benjamins. https://doi.org/10.1075/z.160

Taavitsainen, Irma & Pahta, Päivi. 2011. *Medical Writing in Early Modern English Medical Texts*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511921193

Taavitsainen, Irma, Pahta, Päivi & Mäkinen, Martti. 2005. *Middle English Medical Texts*. CD-ROM. Amsterdam: John Benjamins.

Taavitsainen, Irma, Hiltunen, Turo, Lehto, Anu, Marttila, Villa, Pahta, Päivi, Ratia, Maura, Suhr, Carla & Tyrkkö, Jukka. 2014. *Late Modern English Medical Texts 1700–1800*: A corpus for analysing eighteenth-century medical English. *ICAME Journal* 38: 137–153. https://doi.org/10.2478/icame-2014-0007

Tardy, Christine M. & Swales, John M. 2014. Genre analysis. In *Pragmatics of Discourse*, Klaus P. Schneider & Anne Barron (eds), 165–187. Berlin: De Gruyter. https://doi.org/10.1515/9783110214406-007

Teubert, Wolfgang. 2005. My version of corpus linguistics. *International Journal of Corpus Linguistics* 10(1): 1–13. https://doi.org/10.1075/ijcl.10.1.01teu

Thomas, Christian, & Wiegand, Frank. 2015. Making great work even better: Appraisal and digital curation of widely dispersed electronic textual resources (c. 15th–19th centuries) in CLARIN-D. In Gippert & Gehrke (eds), 181–196.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work* [Studies in Corpus Linguistics 6]. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.6

Walkden, George. 2016. The HeliPaD: A parsed corpus of Old Saxon. *International Journal of Corpus Linguistics* 21(4): 559–571. https://doi.org/10.1075/ijcl.21.4.05wal

Wegera, Klaus-Peter. 2013. Language data exploitation: Design and analysis of historical language corpora. In Bennett et al. (eds), 55–73. Tübingen: Narr.

Whitt, Richard J. 2016a. Evidentiality in early modern English medical treatises (1500–1700). *Journal of Historical Sociolinguistics* 2(2): 235–263. https://doi.org/10.1515/jhsl-2016-0014

Whitt, Richard J. 2016b. Using corpora to track changing thought styles: Evidentiality, epistemology, and early modern English and German scientific discourse. *Kalbotyra* 69: 265–291.

Whitt, Richard J. 2016c. *The Nottingham Corpus of Early Modern German Midwifery and Women's Medicine (ca. 1500–1700)*. Oxford: Oxford Text Archive. <http://ota.ox.ac.uk/desc/2562> (17 July 2017).

Yáñez-Bouza, Nuria. 2011. ARCHER: Past and present (1990–2010). *ICAME Journal* 35: 205–236.

PART I

Methods in diachronic corpus linguistics

‘From above’, ‘from below’, and regionally balanced

Towards a new corpus of nineteenth-century German

Konstantin Niehaus & Stephan Elspaß
University of Innsbruck / University of Salzburg

In this chapter, we report on an ongoing project on a corpus of German in the long nineteenth century. Similar to other historical corpora of German (and unlike existing corpora for the nineteenth century), the *Nineteenth-Century German Corpus* will focus on regional variation. In addition, it will account for an increasing demand for register variation, e.g. by considering formal as well as informal ‘oral’ texts. Thus, text genres ‘from above’, such as regional newspaper texts, and ‘from below’, such as private letters, are incorporated. Three case studies of variation in nineteenth-century grammar will demonstrate the benefits of such a corpus design for the study of language continuity and language change.

1. Introduction

The paper presents the conceptual outlines of a corpus project on a crucial period of Modern German, i.e. the *Nineteenth-Gentury German Corpus* (NiCe German), and some first results of a pilot study. The first section will explain the motivation for this project, i.e. the lack of corpora of nineteenth-century German, which pose an obstacle to the investigation of ongoing standardisation process in this century and of the roots of variation in contemporary German. The second section will outline the conceptual and methodological foundations of the corpus project. Here, we will resort to the concept of ‘language of distance’ vs. ‘language of immediacy’ (Koch & Oesterreicher 2012) and draw on the corpus design of the GerManC corpus. In the third section, we will present three case studies – *Ausklammerung* (‘exbraciation’), diminutive suffixes, and noun plurals in German. The last section will provide a summary and a brief conclusion.

https://doi.org/10.1075/scl.85.02nie

2. Motivation for a (new) corpus of nineteenth-century German

What motivated us to embark on a new corpus project is the persistence of a myth in the historiography of German. The myth is that a unified written German language has been in existence from the end of the eighteenth century. The following quotation from Bach's *Geschichte der deutschen Sprache* ('history of the German language') is one of many representations of this myth.

> Es waren nicht die Grammatiker, die schließlich die sprachl. Einigung des dt. Volkes vollenden konnten. Unvergleichlich mehr als ihre Bemühungen um eine Regelung wirkte das mitreißende Vorbild der großen Dichter und Schriftsteller am Ende des 18. Jh.'s. Erst als die Aufklärung die konfessionellen Gegensätze abgeschwächt hatte und Mittel- und Norddeutschland mit Klopstock, […] Schiller und Goethe die Heimat eines allen Gebildeten der Nation verehrungswürdigen vaterländ. Schrifttums geworden waren, errang die von ihnen gebrauchte dt. Gemeinsprache allgemeine Geltung.
>
> It was not the grammarians who were eventually able to accomplish the linguistic unification of the German people. Incomparably more than their efforts to regulate the language was the infectious example of the great poets and writers at the end of the eighteenth century. Only after the Enlightenment had weakened confessional differences, and central and northern Germany – with Klopstock, […] Schiller and Goethe – had become home to the patriotic literature which was venerated by all educated people of the nation, did the Common Language used by them gain general prestige. (Bach 1965: 352; our translation, KN, SE)

Similar accounts are given in other textbooks on the history of the German language as well as in the research literature (e.g. Besch 1993: 136). This myth has been and is still nurtured by several language ideologies and their outcomes, such as

- a teleological view of language history, in which the emergence of a virtually 'unified' standard variety (cf. Watts's 2011: 291 'funnel view model' of the history of a language) and the establishment of language norms are considered major achievements in the history of a language;
- a powerful standard language ideology (cf. Milroy & Milroy 1985), which ranks standard varieties higher than other varieties of the same language, thus raising the standard variety to the linguistic 'measure of all things';
- and a 'language history from above' perspective, which is characterised by a bias in favour of printed language and a focus on the language production of experienced, mostly male writers from the social elites in a language history (cf. Elspaß 2007: 3–4).

In the case of German, this view has led to the belief that written German was and has remained a standardised language for the last 200 years. In classroom practice, it has resulted, for instance, in futile attempts of generations of students to understand original (and largely unannotated) texts from the time of the classical period in German literature. The consequence of this myth for the historiography of German is that few efforts have been made to investigate language variation and change in the nineteenth century. In addition, (hand-written) texts by non-elite writers in this century of mass literacy, mass migration and the rise of mass media have long been ignored.

Thus, there are at least three good reasons for compiling a corpus of nineteenth-century German.

1. There are corpora of earlier language periods of German, and since 2012, there is a (Middle) New High German corpus for the period 1650 to 1800, the *German Manchester Corpus*[1] (GerManC, as part of the 'Deutsch Diachron Digital' project, cf. <http://www.alc.manchester.ac.uk/modern-languages/research/german-studies/germanc/> (30 June 2017), but no corpus on nineteenth-century German (see Section 3.1 below).
2. In view of empirical evidence and in accordance with the more recent research literature, we are convinced that a written Standard German was not achieved before 1900 (Elspaß 2002, 2005; cf. also Stedje 2007: 172). Moreover, until well into the twentieth century, spoken German was still dominated by local dialects, and in the twentieth century in many regions such dialects were gradually replaced by emerging regional standard and non-standard varieties of German. Thus, the nineteenth and twentieth centuries can be regarded as a highly dynamic period in the history of German.
3. The nineteenth and early twentieth centuries are of eminent importance for present-day German, as the foundations of Standard German were laid in this period. The linguistic dynamics in these two centuries, which we just referred to, left their mark on present-day variation in both standard (cf. Elspaß & Kleiner in print) and non-standard varieties of German.

3. Methodology: Towards a new corpus of nineteenth-century German

3.1 Existing corpora of nineteenth-century German and their limits for variational analysis

The two major historical text corpora for modern German are the *Deutsches Referenzkorpus* (DeReKo) ('German Reference Corpus') and the *Deutsches Textarchiv*

[1] In other words, only ten years ago we did not even have a historical corpus of New High German, that is, from the mid seventeeth century onwards.

(DTA) ('German Text Archive'). For various reasons, their usefulness for variational analysis is rather limited.

- The 40 billion word DeReKo has a small 70 million word 'virtual' subcorpus of historical texts ("Archiv HIST"). A list on the website of the *Institut für Deutsche Sprache* (http://www1.ids-mannheim.de/lexik/abgeschlosseneprojekte/historischeskorpus/historisches-korpus.html, 27.01.2017) only reveals that it consists of sub-corpora (of different sizes) from different text genres. It includes only a relatively small proportion of sub-corpora from the nineteenth century, and these sub-corpora consist mostly of fictional texts, some newspaper texts (see below), articles from encyclopedias, Grimms' tales, and a selection of the complete works of Marx and Engels. These texts represent printed texts exclusively, and they are not balanced for register or region.
- The DTA is intended "to provide a basic stock of German language texts from the period 1600 to 1900, which encompasses various disciplines and genres" (our translation, KN, SE, 27.01.2017).[2] It consists of approximately 145 million word forms, and 1,019 out of a total of 2,610 fictional, academic and 'functional' texts (*Gebrauchsliteratur*) are from the nineteenth century (27.01.2017). Again, only printed texts are represented. The DTA core corpus is balanced for time (208, 492, and 650 texts for the seventeenth, eighteenth, and nineteenth centuries respectively) and 'text category' (504 texts from fiction, 239 'functional' texts, and 628 academic texts; 11.10.2016), but there is no identifiable balance of text genres and regions. The texts of the DTA have been included in the *Digitales Wörterbuch der Deutschen Sprache* (DWDS) ('Digital Dictionary of the German Language'.[3]

Considering the heavily increasing number of newspapers and their circulation in the 'long' nineteenth century, the neglect of newspaper corpora in current corpus projects is noteworthy. This is true for English (cf. Percy 2012: 192), but particularly for German. "Archiv HIST" in the DeReKo contains sub-corpora which comprise almost 60 million words from fiction and philosophy, but only 4.1 million from historical newspapers and historical magazines (in the *Mannheimer Korpus historischer Zeitungen und Zeitschriften (*khzm), 'Mannheim Corpus of Historical Newspapers and Magazines', cf. Pfefferkorn & Fankhauser 2014: 42). The *khzm* is

[2] "Das Deutsche Textarchiv stellt einen disziplinen- und gattungsübergreifenden Grundbestand deutschsprachiger Texte aus dem Zeitraum von ca. 1600 bis 1900 bereit." <http://www.deutsches-textarchiv.de/> (27 January 2017).

[3] Cf. <www.dwds.de> (27.01.2017).

balanced neither for time periods nor for region.[4] A considerable part of the corpus consists of newspapers and magazines from the eighteenth century rather than the nineteenth century, from randomly selected time periods, with issues from particular newspapers from particular regions "representing" these (short) periods.

To sum up, whereas other corpus projects of historical German such as the GerManC present selections of texts balanced for time periods, regions and genres, a corresponding corpus for the nineteenth century is still lacking.

3.2 A new corpus: *The Corpus of Nineteenth-Century German* (NiCe German Corpus)

With the *Corpus of Nineteenth-Century German* (NiCe German Corpus), we seek to provide the missing link between existing corpora of historical German and present-day German, thus facilitating the investigation of language variation and change for the modern period. In compiling this corpus our particular interest is directed towards register variation and regional variation.

We will account for *register variation* by using texts "from above" and texts "from below", and for *regional variation* by building a balanced corpus of such texts from different regions. By the notion of texts "from above" and "from below", we mean texts from prototypically formal and informal registers and, more generally, we refer to the concept of "language of distance" vs "language of immediacy" according to the well-known model of Koch and Oesterreicher (2012). As prototypical formal texts (representing the "language of distance"), we include printed newspaper texts – to be more precise, texts with regional news from nineteenth-century papers. In general, prototypically informal texts (representing the "language of immediacy") are represented by private conversations. As there are no spoken data available for the nineteenth century (and earlier centuries), we take private letters – as "the 'next best thing' to authentic spoken language" (Nevalainen & Raumolin-Brunberg 2012: 32), and we focus on letters written by members of the lower ranks of society. Such hand-written texts, which were not written to be printed, are known to be least affected by epistolary norms of style and other "traditions of a 'writing of distance'" (Elspaß 2012: 158). A particularly large collection of nineteenth-century letters of "ordinary" people was compiled as part of the extensive correspondence of emigrants to the U. S. and their relatives and friends (cf. Elspaß 2005: 55–72 for details).

[4] Cf. <http://repos.ids-mannheim.de/fedora/objects/clarin-ids:mkhz1.00000/datastreams/CMDI/content> (27 January 2017)

Ideally, a corpus of nineteenth-century German would account for all available text genres. For our pilot corpus, we concentrated first on text genres which are maximally different from each other in the degree of formality and orality respectively, and secondly on texts from different regions of the German-speaking countries. We expect that such a corpus design will render the maximum range of possible variation and thus be most rewarding for a study of language change linked to register and regional variation.

Texts from these two main genres, newspaper texts and private letters, are available from every region of the German-speaking countries in the nineteenth century and to such an extent that it is possible to compile a corpus containing texts from both genres from the five major historical dialect regions (as in GerManC). The regional division of our corpus texts is based on the concept of pluriareality rather than pluricentricity, which implies that we do not divide the corpus along national borders, but rather according to these historical dialect regions. In view of the fact that for the most part of the nineteenth century the German-speaking countries were not yet distinct nation states, the pluriareal model appears to be the more adequate concept.

Currently, our letter corpus contains approximately 465,000 word forms. The newspaper corpus is much smaller, amounting presently to approximately 107,000 words.[5] Our plan is, of course, to work on an extension of the newspaper corpus, with a focus on texts from different newspaper sub-genres and with an enhanced regional balance.

4. Case studies

In this section, we present three grammatical case studies, one from syntax (*Ausklammerung* 'exbraciation'), one from derivational morphology (variants of diminutives) and one from inflectional morphophonology (variants of noun plural forms in the case of *Wagen* vs. *Wägen* 'cars, carriages'). We chose these three examples, firstly, to account for the potential range of grammatical features displaying variation in nineteenth-century German, and, secondly, for practical reasons: all three cases provide sufficient instances of variants in our still relatively small newspaper subcorpus. All the case studies provide absolute figures as well as relative figures in order to compensate for the varying size of our sub-corpora and also with respect to the small absolute numbers that may occur. 'Relative' means

[5] For more details on the corpus design cf. Elspaß & Niehaus (2014: 51–52) and Elspaß (2015a: 396–399).

that the absolute frequency is measured to 10,000 words per sub-corpus. At this stage, we can only present descriptive statistics without recourse to tests of statistical significance.

4.1 *Ausklammerung*

Ausklammerung ('exbraciation') is a grammatical construction identified in terms of the model of syntactic fields (see below). It is commonly assumed that *Ausklammerung* is a feature more frequent in spoken colloquial German. However, this depends very much on the grammatical elements involved. Thus, *Ausklammerung* of numerous constituents and also of adverbials is more likely to appear in colloquial language, whereas in written German such types of *Ausklammerung* are much harder to find. In general, the use of *Ausklammerung* in written German seems to depend more on stylistic preferences, e.g. the 'stylisation' of colloquial German in writing (cf. Niehaus 2016: 170–171). These characteristics make *Ausklammerung* an ideal candidate for a genre-sensitive analysis.

The concept of *Ausklammerung* rests on the assumption that German[6] clauses and sentences can be subdivided into positional fields (in German: *Stellungsfelder*) which are determined by the two parts of a verbal bracket (*Satzklammer*): initial position (*Vorfeld*), central position after the first part of the bracket (*Mittelfeld*) and position after the closing bracket (*Nachfeld*), all of which may be filled with certain syntactic elements (constituents). In the prototypical German clause with a verbal bracket, the main clause statement, most of its syntactic elements (constituents) are sandwiched between the two parts of the verbal bracket. A bracket may be formed by [finite verb + infinite verb/verbal cluster/verbal particles] in main clauses or by [subordinator + verb or verbal cluster] in subordinate clauses. Table 1 shows one example of a bracket in a main clause and the corresponding *Ausklammerung* of the constituent *im Club* ('at the club').

Table 1. Example of *Ausklammerung* in German

Initial position	Left verbal bracket	Central position	Right verbal bracket	Post verbal position
Sie *She*	hat *has*	ihn gestern **im Club** *him yesterday at the club*	gesehen. *seen*	
Sie *She*	hat *has*	ihn gestern *him yesterday*	gesehen *seen*	**im Club**. *at the club*

[6] For *Ausklammerung* in Dutch cf. Kooij (2009: 122–123).

As the example illustrates, *Ausklammerung* means moving a constituent from its position between the two parts of the bracket to the postverbal position. The clause given in this example is a rather common case, in that it is often prepositional phrases which are placed after the closing verbal bracket, though other phrases and even clauses (e.g. relative clauses) may appear as *Ausklammerungen* in certain contexts (cf. Niehaus 2016: 176). The theory of syntactic fields was originally modelled on examples from standard written language (this is why such theories have been characterised as 'scripticistic' by Ágel 2003: 4–11), and the *Ausklammerung* is to date one of the most complex and controversial phenomena in this theory (cf. also Ágel 2000: 1873).

In view of the different concepts on *Ausklammerung* in the research literature (cf. Niehaus 2016: 117–122), we considered only those constructions that correspond to *Ausklammerung* in its narrowest sense for our case study. Thus, we only examined prepositional and noun phrases as *Ausklammerung*, and we excluded clauses and other constructions (such as comparatives with *als* 'than' and *wie* 'as …as'), which are usually placed in the postverbal position and considered grammaticalised (cf. Eisenberg 2004: 402). Table 2 provides an overview of the results of our corpus study.

Table 2. Instances of phrasal *Ausklammerung* in the two sub-corpora of the NiCe German Corpus

Region	**Newspaper corpus(absolute / relative per 10,000 words)**	**Private letter corpus(absolute / relative per 10,000 words)**
North German (NG)	14 / **8.3**	37 / **2.9**
West Central German (WCG)	26 / **41.2**	35 / **3.3**
East Central German (ECG)	8 / **6.0**	11 / **5.5**
West Upper German (WUG)	36 / **19.6**	24 / **3.0**
East Upper German (EUG)	10 / **4.8**	23 / **2.3**
Total	94 / **12.4**	130 / **3.4**

Firstly, the results challenge the widespread view that *Ausklammerung* is used more often in colloquial German than in written formal German, i.e. that it is a typical feature of orality (cf. – among others – Zifonun, Hoffmann & Strecker 1997: 1657, 1660; Weinrich 2003: 83; Hentschel & Weydt 2013: 390). Rather, our findings suggest that in the nineteenth century *Ausklammerung* was employed more frequently in formal texts. In relative numbers (again measured to 10,000 words per respective sub-corpus), the rate of *Ausklammerung* in the newspaper sub-corpus is much higher than in the letter sub-corpus. *Ausklammerung* of prepositional phrases constitutes the vast majority of instances in both the newspaper sub-corpus (75 instances, 79.8%) and the private letter corpus (111 instances, 85.3%). Adverbials account for about half of these prepositional phrases (newspapers sub-corpus: 53 instances = 56.4%, private letter sub-corpus: 62 instances, 47.7%). *Ausklammerung* of prepositional objects[7] in the private letter sub-corpus (47 instances, 36.1%) is used more frequently than in the newspaper sub-corpus (10 instances, 9.4%), while *Ausklammerung* of subjects occurs more often in the newspaper sub-corpus (13 instances, 13.8%) than in the private letter sub-corpus (3 instances, 2.3%), mainly "in order not to overstretch the verbal bracket"[8] (cf. Durrell 2011: 466). All in all, both sub-corpora – not only the private letters – show a considerable amount of variation across syntactic forms and functions.

Somewhat surprising is also the second result, in that the newspaper subcorpus shows regional variation to a much greater extent than the private letter sub-corpus. In the newspaper sub-corpus, *Ausklammerung* occurs particularly frequently in the West Central German and the West Upper German area. The regional news texts from these areas show a stronger tendency towards *Ausklammerung* than the private letters.[9] This is unlikely to be the result of a few authors' idiosyncratic usage; the tendency is stable if we add supra-regional news texts to our newspaper sub-corpus (cf. the results in Niehaus 2016: 169). Various explanations for the emergence of *Ausklammerung* have been offered in the research literature. One explanation is that already at the end of the eighteenth century, authors of moral weeklies as well as *Sturm und Drang* writers rejected the strict compliance with the verbal bracket in chancery German and used *Ausklammerung* as a stylistic device to imply 'naturalness' or 'natural' language usage rather than 'artificial'

[7] This includes prepositional direction complements (cf. Durrell 2011: 380) as they are part of the verb's valency and as there is empirical diachronic evidence that *Ausklammerung* of these complements may occur if auxiliary and past participle form the verbal bracket (cf. Niehaus 2016: 168). This leaves rather rare cases where *Ausklammerung* of direction complements cannot occur – cases with often disputable verbal brackets of phrasal verbs (cf. Niehaus 2016: 168).

[8] E.g. *daß dem 1. Senat zugeteilt werden* ***die Sachen der Kammern für Handelssachen von Mainz, Worms und Offenbach und die Patente für Schutzsachen*** 'that the 1st senate is supplied with **the affairs of the chambers of commerce of Mainz, Worms and Offenbach as well as the patents for protection details**' (Darmstädter Zeitung, 03.01.1910; emphasis ours, KN, SE).

[9] Due to the small numbers, a differentiation of syntactic forms and functions is hardly valid. But Niehaus (2016: 169–170) observes the same areal preference in newspaper agency texts, especially with prepositional phrases (regardless of their respective syntactic functions).

or 'latinised' German (Admoni 1990: 215).[10] Another explanation given in the research literature is language contact. International correspondents and regional writers alike could have had regular language contact with French and English (languages in which postverbal elements such as prepositional phrases occur relatively often) and then borrowed such syntactic structures (cf. Moser 1982: 336), which eventually resulted in (higher numbers of) *Ausklammerung* in German. Von Polenz (2013: 342 [referring to Blackall 1966]) suggests that eighteenth-century German editors of moral weeklies and other periodical essays intentionally borrowed postverbal syntactic structures from English in order to create a supposedly "easily" comprehensible style. Whatever the reason for the regional distribution in the nineteenth-century newspaper corpus, it is noteworthy that *Ausklammerung* has continued to vary regionally up to the present, although with a slightly different areal pattern in a present-day Standard German corpus (with a much higher frequency of *Ausklammerung* in East Central texts, in particular, cf. Fuchs-Richter 2018a).

4.2 Diminutive *-chen/-gen/-lein*

The diminutive of a noun in German may be formed using various suffixes. In nineteenth-century written German, three main suffixes were used: *-chen*, *-gen* (a spelling variant of *-chen*) and *-lein*, all triggering *Umlaut* of the stressed syllable where possible, so that, for example, the diminutive of *Haus* 'house' could be *Häuschen*, *Häusgen*, or *Häuslein*. In present-day Standard German only *-chen* and *-lein* are possible, with *-chen* being the more widely used variant, and *-lein* chiefly occurring with nouns ending in *-(n)g* and *-ch*.

Our results shows register as well as regional variation (cf. Table 3).

In terms of register, the use of *-gen* is restricted to the letter corpus. Here, it only occurs in the Central and West Upper German areas, where it was dominant in informal as well as formal registers in the eighteenth century. Not restricted to register, but clearly to region is *-lein*. In our nineteenth-century corpora, *-lein* only occurs in texts from the Upper German regions, where it was also used in dialects (cf. Seebold 1983). Moreover, it was a feature of the so-called *oberdeutsche Literatursprache* ('Upper German literary language') of the eighteenth century.[11] The

[10] In contradiction to this, Behaghel (1932: 134) regards the very style of *Sturm und Drang* as highly artificial.

[11] Cf. the quote from the grammar by Aichinger (1754: 143): "*Gen*, oder wie Herr Gottsched will, *chen*, ist bey den Sachsen; *lein* bei den Oberteutschen gebräuchlicher" ('*-gen*, or *-chen*, as Mr Gottsched will have it, is used by the Saxons; *-lein* is more popular with the Upper Germans') [our translation, KN, SE].

variants *-gen* and *-lein* appear to have been in gradual decline in printed German from the end of the eighteenth century, the areal patterns in the corpus data suggest continuities in regional use from the eighteenth to the nineteenth century (cf. Durrell, Ensslin & Bennett 2008: 270–271; Elspaß 2015a: 400–403). In the case of *-gen*, the decrease in use can most certainly be attributed to the effect of prescriptivist judgements. Eighteenth century grammarians like Gottsched and Adelung had declared *-gen* to be the "wrong" form (cf. Elspaß 2015a: 402–403), and this appears to have had an immediate effect on printed German, but not on private correspondence. Interestingly, while there is no instance of *-lein* in the late eighteenth century newspapers of the GerManC (cf. Durrell, Ensslin & Bennett 2008: 270–271), we find it being used occasionally in the nineteenth century. However, its absence in GerManC may simply be due to the relatively small size of the sub-corpora.

Table 3. Diminutive suffixes in the two sub-corpora of the NiCe German Corpus[12]

Region	**Newspaper corpus (absolute / relative per 10,000 words)**			**Private letter corpus (absolute / relative per 10,000 words)**		
	-chen	*-gen*	*-lein*	*-chen*	*-gen*	*-lein*
North German (NG)	4 / **2.0**	–	–	5 / **4.8**	–	–
West Central German (WCG)	–	–	–	1 / **1.1**	2 / **2.2**	–
East Central German (ECG)	4 / **3.6**	–	–	6 / **6.1**	–	–
West Upper German (WUG)	4 / **1.7**	–	1 / **0.4**	19 / **17.6**	2 / **1.8**	1 / **0.9**
East Upper German (EUG)	8 / **3.5**	–	1 / **0.4**	16 / **8.0**	1 / **0.6**	1 / **0.6**
Total	22 / **2.9**	–	2 / **0.3**	42 / 7.5	5 / **0.9**	2 / **0.3**

4.3 Noun plural forms with or without Umlaut (*Wägen*/*Wagen*)

The plural of *der Wagen* ('the car, carriage') has two variants, both without a plural suffix, but one with Umlaut (*die Wägen*) and one without Umlaut (*die Wagen*, and thus indistinguishable from most singular forms). Although the morphologically opaque variant, *Wagen* is more widespread in both present-day colloquial and

[12] Hits for *-chen* without lexicalised *Mädchen* ('girl'), hits for *-lein* without lexicalised *Fräulein* ('Miss').

Standard German. Again, our corpus study provides noteworthy results with respect to register and regional variation (cf. Table 4).

Table 4. Plural *Wagen/Wägen* ('cars, carriages') in the two sub-corpora of the NiCe German Corpus

	Newspaper corpus (absolute / relative per 10,000 words)		**Private letter corpus (absolute / relative per 10,000 words)**	
	Wagen	***Wägen***	***Wagen***	***Wägen***
North German (NG)	1 / **0.6**	–	9 / **0.7**	1 / **0.1**
West Central German (WCG)	–	–	7 / **0.7**	–
East Central German (ECG)	1 / **0.8**	–	2 / **1.0**	1 / **0.5**
West Upper German (WUG)	–	–	6 / **0.8**	8 / **1.0**
East Upper German (EUG)	1 / **0.5**	1 / **0.5**	6 / **0.5**	9 / **0.7**
Total	3 / **0.4**	1 / **0.13**	30 / **0.74**	19 / **0.46**

From a regional perspective, it is evident that *Wägen* was mainly used in the southern parts of the German speaking countries (cf. Table 4). This continues to be the case in present-day German, as can be seen from a map of the *Atlas zur deutschen Alltagssprache (*AdA) ('Atlas of colloquial German', cf. Elspaß & Möller 2003 et seqq., cf. Figure 1) and from (relative) figures for the distribution of the two plural forms in the present-day newspaper corpus of the *Variantengrammatik* ('Variational Grammar of Standard German') project[13] (cf. Figure 2).

Other nouns showing a similar regional distribution of variants with or without Umlaut include *Kragen* ('collar'), *Hochwasser* ('flood'), and *Polster* ('cushion, padding'). As for register variation in the NiCe German Corpus, the use of *Wägen* is almost exclusively restricted to the informal register, i.e. the private letter subcorpus (cf. Table 4). There is, however, one instance of *Wägen* in an East Upper German regional news text (*die neue Zukunfts-Straßenlocomotive mit den Tramway****wägen***, Innsbrucker Nachrichten, 20.02.1884).

In contrast to the case of diminutive suffix variants, the figures on noun plural forms with or without Umlaut provide evidence to support the assumption that the influence of prescriptive grammars on actual language use can on occasion

[13] Cf. Dürscheid & Elspaß (2015) for details on the *Variantengrammatik* project.

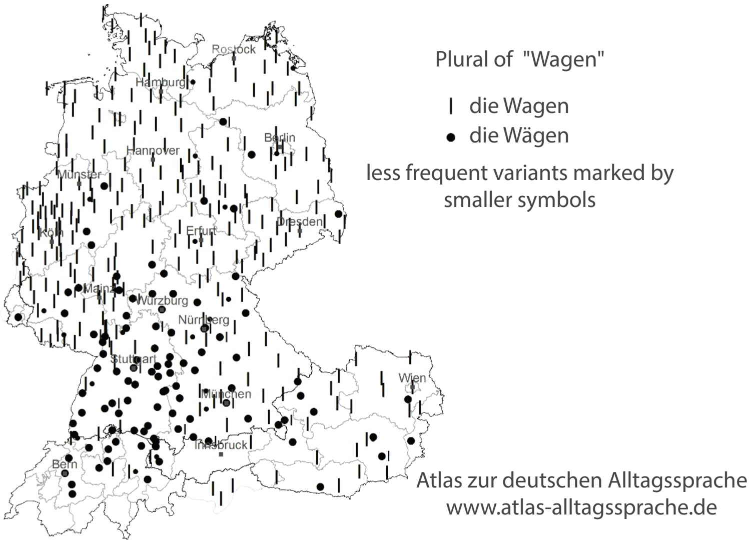


Figure 1. Regional distribution of plural variants *Wagen/Wägen* ('cars, carriages') in colloquial German, according to the Atlas zur deutschen Alltagssprache (AdA) (Elspaß & Möller 2003 et seqq.)

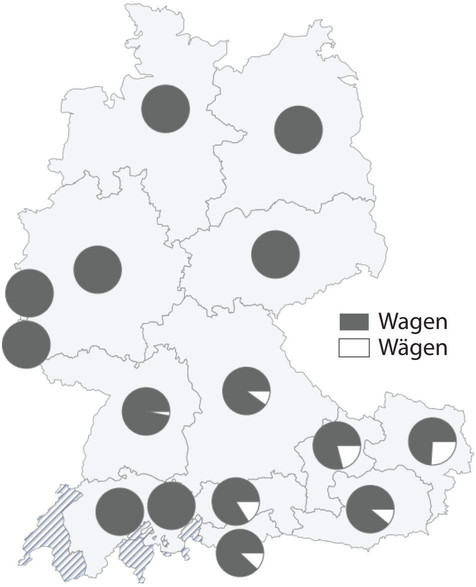


Figure 2. Regional distribution of plural variants *Wagen/Wägen* ('cars, carriages') in Standard German, according to newspaper corpus of the Variantengrammatik ('Variational Grammar of Standard German', cf. Rimensberger forthcoming)

be limited. While such grammars have classified *Wägen* as less refined[14] (cf. Adelung 1811: 1335) or consider it as “non-standard” (cf. Duden-Zweifelsfälle 2016: 1007), in present-day newspapers from the East Upper German areas, *Wägen* is employed so frequently that the *Variantengrammatik* describes it as a “standard” variant in these regions (cf. Figure 2 and Rimensberger forthcoming).

4.4 Other features and future research

In this section, we give a brief account of other constructions which display register and/or regional variation in the corpus, and we cast a look at other ‘candidates’ which appear to be worth a closer investigation with respect to such variation.

In Elspaß and Niehaus (2014) we examined a particular type of serialisation in three-verb clusters,[15] discontinuous pronominal adverbs[16] and the lexical variants *Samstag* vs. *Sonnabend* (‘Saturday’). Whereas the first two (grammatical) features did not show any regional variation in the newspaper sub-corpus, they very much did so in the letter sub-corpus, showing a clear north-south divide. This is a characteristic distribution that continues into present-day Standard German, as is confirmed by data from Negele (2012: 241–244) for the variation of pronominal adverbs, and the *Variantengrammatik* (Niehaus 2018b) for the variation in verbal clusters. The lexical variation of *Samstag* and *Sonnabend* ‘Saturday’ shows a similar

[14] Adelung writes: “… doch ist Wagen in den edlern, und Wägen in den gemeinen Sprecharten [des Hochdeutschen] am üblichsten” (‘*Wagen* is the most common variant in the more noble forms of speech of High German, whereas *Wägen* is most common in general speech’).

[15]

E.g. …	*dass*	*er*	*das*	*Spiel*	***hat***^{1}	*sehen*^{3}	*können*^{2}
…	that	he	the	match	has$^{1\text{–}fin}$	watch$^{3\text{–}inf}$	can$^{2\text{–}inf}$
vs.							
…	*dass*	*er*	*das*	*Spiel*	*sehen*^{3}	***hat***^{1}	*können*^{2}
…	that	he	the	match	watch$^{3\text{-}inf}$	has$^{1\text{-}fin}$	can$^{2\text{-}inf}$
	‘that he has been able to watch the match’						

[16]

E.g. ***damit***	*habe*	*ich*	*nichts*		*zu*	*tun*	
this-(PRO)-with	have	I	nothing		to	do	
vs.							
da	*habe*	*ich*	*nichts*	***mit***	*zu*	*tun*	
this-(PRO)	have	I	nothing	with	to	do	
vs.							
da	*habe*	*ich*	*nichts*	***damit***	*zu*	*tun*	
this-(PRO)	have	I	nothing	this-(PRO)-with	to	do	
‘I have nothing to do with this’							

regional distribution (cf. Elspaß & Niehaus 2014: 60–61). However, there is no evidence of register variation in our data. Both the letter corpus and the newspaper corpus displayed a dominant use of *Samstag* in the south (and some parts of the north-west) and of *Sonnabend* in the north and east. The results for this lexical variable thus constitute an instance of what may be called 'variational continuity' in both 'texts from above' and 'texts from below'.

Further 'candidates' for a study of register and regional variation in the *Nineteenth-Century German Corpus* include grammatical features which have been the subject of recent work on diachronic and present-day variation, e.g.

- *-s* vs. *-es* suffixes in the genitive singular of masculine and neuter nouns, e.g. *des Jahrs/Jahres* ('of the year'), *des Vertrags/Vertrages* ('of the contract') (cf. Szczepaniak 2010; Fehringer 2011),
- (the absence of) the *-e* suffix in the dative singular of masculine and neuter nouns, e.g. *dem Jahr(e)*, *dem Vertrag(e)* (Elspaß 2005: 348–354; Durrell, Ensslin & Bennett 2008: 267–268; von Polenz 2013: 277–281; Durrell 2016: 223–229),
- the use of genitive vs. dative case with prepositions like *wegen* ('because of'), *trotz* ('in spite of') or *während* ('during') (cf. Elspaß 2005: 320–325; Sato 2016),
- adverbial genitives vs. prepositional phrases indicating a time of day, *des morgens* vs. *am Morgen*, *in der Früh* ('in the morning') (cf. Elspaß 2005: 336–339),
- constructions of progressive and habitual aspect consisting of a prepositional phrase with *am/im/beim* + nominalised infinitive and *to be* (e.g. *Er ist am/im/beim Arbeiten*. 'He is working.', cf. Van Pottelberge 2004; Elspaß 2010: 134–135),
- *nicht brauchen* + infinitive as a modal verb used with vs. without the infinitive particle *zu* (e.g. *Er braucht nicht (zu) kommen*. 'He doesn't need/have to come.') (cf. Elspaß 2015b: 48–49),
- *würde* + infinitive construction (cf. Durrell & Whitt 2016), *täte* + infinitive construction vs. synthetic conjunctive forms (e.g. *würde/täte kommen* vs. *käme* 'would come').

Apart from grammatical features, all kinds of lexical variables are suitable for an investigation of register and regional variation in the NiCe German Corpus. As for regional variation, any variable which displays such variation in present-day German (as documented, for instance, in the *Variantenwörterbuch* (VWB), cf. Ammon, Bickel & Lenz et al. 2016 or the *Atlas zur deutschen Alltagssprache (*AdA), cf. Elspaß & Möller 2003 et seqq.) and which are sufficiently frequent could be searched for in the corpus. First observations in the NiCe German Corpus reveal that variants which are nowadays identified as "Austrian standard German" (cf. Ammon, Bickel & Lenz et al. 2016) were used in newspapers and letters from regions in the south-west and the west of the German-speaking countries, e.g. *Jänner* for *Januar* 'January' (*Gülich und bergische wöchentliche Nachrichten*, Düsseldorf, 07.01.1800; letter by August Schnurr from Renchen near Freiburg, 03.08.1863) and *retour* for *zurück* 'back' (*Schwäbischer Merkur*, Stuttgart, 05.06.1852). Even a quick GoogleBooks search (restricted to "the nineteenth century" and pages in German) confirms a larger regional distribution of both words.[17] It would be interesting to see when (and how) such words came to be "Austrian".

These observations lead us to suppose that many lexical variants which are nowadays considered national variants of German used to have a wider distribution in the nineteenth century, and it would be interesting to explore in which cases and under what sociolinguistic conditions a regional variant became a national variant.

If in future research such grammatical and lexical features could be investigated on the basis of larger digitised diachronic corpora, we would not only be able to assess language change with more sensitivity to genre and variation, but also to provide an answer to when and how today's "patterns" of national and regional variation in (Standard) German came about. In particular, Peter von Polenz's concept of increasing "monocentricity" in late nineteenth-century Standard German versus a tendency to "pluricentricity" in late twentieth century Standard German (cf. von Polenz 1989: 15) could be investigated more thoroughly.

5. Summary and conclusion

We hope to have demonstrated the usefulness of our approach to compile a new diachronic corpus of nineteenth-century German sensitive to both the role of text genres and to the role of areal variation in language change. The lack of corpora of nineteenth-century German which are balanced for register and region motivated us to start a new corpus project, the *Corpus of Nineteenth-Century German* (NiCe German Corpus). In an attempt to use "texts from above" and "texts from below"

[17] <https://www.google.at/search?q=J%C3%A4nner&ie=utf-8&oe=utf-8&client=firefox-b&gfe_rd=cr&ei=s0qUWL73DrDi8AeRiZ7oAg#q=J%C3%A4nner&tbs=cdr:1,cd_min:1800,cd_max:1899,lr:lang_1de&tbm=bks&lr=lang_de> (27 January 2017) and <https://www.google.at/search?q=retour&ie=utf-8&oe=utf-8&client=firefox-b&gfe_rd=cr&ei=nUqUWK-YAbDi8AeRiZ7oAg#q=retour&tbs=cdr:1,cd_min:1800,cd_max:1899,lr:lang_1de&tbm=bks&lr=lang_de> (27 January 2017).

and thus capture the opposite poles of the continuum between "language of distance" and "language of immediacy" (cf. Koch & Oesterreicher 2012), we initially focus on newspaper texts as representatives of printed, formal text genres on the one hand and on private letters as representatives of handwritten, informal text genres on the other.

Moreover, our corpus design takes account of the fact that the German language – like most other European languages – is and has always been characterised by marked regional differences. Likewise, regional variation has always been a potentially influential factor in language change. Thus, following the example of other modern projects of diachronic corpora of German, such as GerManC, the NiCe German Corpus is not only divided into (two) text genres, but also into subcorpora from five major dialect regions.

We then carried out case studies which gave evidence of regional and genre-specific variation. Patterns of regional variation in the nineteenth century appear as pluriareal variation rather than pluricentric variation – pluricentric in the sense of diatopic (standard) variation following political borders (which is a fairly recent and rather controversial concept, cf. Auer 2013). Genre-specific variation becomes evident in terms of quantity (as far as our small corpus allows for any conclusions), meaning that newspaper texts generally have less variation than private letters (with the exception of certain cases of *Ausklammerung*). Finally, we presented perspectives for future research on grammatical and lexical variation in nineteenth-century German. Particularly in a view "from below", there is more variation to be discovered than has met the eye of "scripticistic" historical research on the history of modern German.

Finally, we hope to have illustrated that – at least for German – variation in the nineteenth century is crucial to any research on contemporary variation, varieties and genres. The making of Standard German can be traced back in actual use by means of diachronic corpora, not only in prescriptivist metalinguistic comments. To give a full account of what actually happened in the standardisation process, we need larger and more sophisticated digital corpora. This holds true for German as well as for other standardised European languages. Historical corpus linguistics, we argue, does not only allow us to analyse past periods of one language, but can also help us to reach a deeper understanding of present-day variation.

Acknowledgement

The authors would like to thank Richard J. Whitt and two anonymous reviewers for their helpful comments.

References

Adelung, Johann Christoph. 1811. *Grammatisch-kritisches Wörterbuch der hochdeutschen Mundart, mit beständiger Vergleichung der übrigen Mundarten, besonders aber der Oberdeutschen, Vol. IV: Seb–Z*. Wien: Bauer.

Admoni, Wladimir. 1990. *Historische Syntax des Deutschen*. Tübingen: Niemeyer.

Ágel, Vilmos. 2000. Syntax des Neuhochdeutschen bis zur Mitte des 20. Jahrhunderts. In *Sprachgeschichte – ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung* [Handbücher zur Sprach- und Kommunikationswissenschaft 2.2], 2nd edn, Werner Besch, Anne Betten, Oskar Reichmann & Stefan Sonderegger (eds), 1855–1903. Berlin: De Gruyter.

Ágel, Vilmos. 2003. Prinzipien der Grammatik. In *Neue historische Grammatiken. Zum Stand der Grammatikschreibung historischer Sprachstufen des Deutschen und anderer Sprachen* [Reihe Germanistische Linguistik 243], Anja Lobenstein-Reichmann & Oskar Reichmann (eds), 1–46. Tübingen: Niemeyer. https://doi.org/10.1515/9783110913194.1

Aichinger, Carl Friedrich. 1754. *Versuch einer teutschen Sprachlehre, anfänglich nur zu eignem Gebrauche unternommen, endlich aber, um den Gelehrten zu fernerer Untersuchung Anlaß zu geben, ans Licht gestellt*. Leipzig: Kraus.

Ammon, Ulrich, Bickel, Hans & Lenz, Alexandra N. (eds). 2016. *Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen*, 2. völlig neu bearbeitete und erweiterte Auflage. Berlin: De Gruyter. https://doi.org/10.1515/9783110245448

Auer, Peter. 2013. Enregistering pluricentric German. In *Pluricentricity. Language Variation and Sociocognitive Dimensions*, Augusto Soares da Silva (ed.), 19–48. Berlin: De Gruyter. https://doi.org/10.1515/9783110303643.19

Bach, Adolf. 1965. *Geschichte der deutschen Sprache*, 8th edn. Heidelberg: Quelle & Meyer.

Behaghel, Otto. 1932. *Deutsche Syntax. Eine geschichtliche Darstellung, Vol. IV: Wortstellung. Periodenbau*. Heidelberg: Winter.

Besch, Werner. 1993. Regionalität – Überregionalität. Sprachlicher Wandel zu Beginn der Neuzeit. Mit 9 Karten. *Rheinische Vierteljahrsblätter* 57: 114–136.

Duden-Zweifelsfälle. 2016. = *Duden. Das Wörterbuch der sprachlichen Zweifelsfälle. Richtiges und gutes Deutsch* [Duden 9], 8th edn, Mathilde Hennig (ed). Berlin: Dudenverlag.

Durrell, Martin. 2011. *Hammer's German Grammar and Usage*, 5th edn. London: Routledge.

Durrell, Martin. 2016. Textsortenspezifische und regionale Unterschiede bei der Standardisierung der deutschen Sprache. In *PerspektivWechsel oder: Die Wiederentdeckung der Philologie, Band 1: Sprachdaten und Grundlagenforschung in der Historischen Linguistik*, Sarah Kwekkeboom & Sandra Waldenberger (eds), 211–231. Berlin: Schmidt.

Durrell, Martin & Whitt, Richard J. 2016. The development of the *würde* + infinitive construction in Early Modern German (1650–1800). *Beiträge zur Geschichte der deutschen Sprache und Literatur* 138(3): 325–364. https://doi.org/10.1515/bgsl-2016-0028

Durrell, Martin, Ensslin, Astrid & Bennett, Paul. 2008. Zeitungen und Sprachausgleich im 17. und 18. Jahrhundert. Special issue *Der Schreiber als Dolmetsch. Sprachliche Umsetzungstechniken beim binnensprachlichen Texttransfer im Mittelalter und früher Neuzeit*, Werner Besch & Thomas Klein (eds). *Zeitschrift für deutsche Philologie* 127: 263–279.

Dürscheid, Christa & Elspaß, Stephan. 2015. Variantengrammatik des Standarddeutschen. In *Regionale Variation des Deutschen. Projekte und Perspektiven*, Roland Kehrein, Alfred Lameli & Stefan Rabanus (eds), 563–584. Berlin: De Gruyter. https://doi.org/10.1515/9783110363449-024

Eisenberg, Peter. 2004. *Grundriß der deutschen Grammatik, Vol. 2: Der Satz*, 2nd rev. edn. Stuttgart: Metzler. https://doi.org/10.1007/978-3-476-03763-3

Elspaß, Stephan. 2002. Standard German in the 19th-century? (Counter-) Evidence from the private correspondence of 'ordinary people'. In *Standardization. Studies from the Germanic Languages* [*Amsterdam Studies in the Theory and History of Linguistic Science, Series IV: Current Issues in Linguistic Theory 235*], Andrew R. Linn & Nicola McLelland (eds), 43–65. Amsterdam: John Benjamins. https://doi.org/10.1075/cilt.235.05els

Elspaß, Stephan. 2005. *Sprachgeschichte von unten. Untersuchungen zum geschriebenen Alltagsdeutsch im 19. Jahrhundert* [Reihe Germanistische Linguistik 263]. Tübingen: Niemeyer. https://doi.org/10.1515/9783110910568

Elspaß, Stephan. 2007. A twofold view 'from below': New perspectives on language histories and language historiographies. In *Germanic Language Histories 'From Below' (1700–2000)* [Studia Linguistica Germanica 86], Stephan Elspaß, Nils Langer, Joachim Scharloth & Wim Vandenbussche (eds), 3–9. Berlin: De Gruyter. https://doi.org/10.1515/9783110925463.3

Elspaß, Stephan. 2010. Regional standard variation in and out of grammarians' focus. In *Grammar between Norm and Variation* [VarioLingua 40], Alexandra N. Lenz & Albrecht Plewnia (eds), 127–144. Frankfurt: Peter Lang.

Elspaß, Stephan. 2012. The Use of Private Letters and Diaries in Sociolinguistic Investigation. In *The Handbook of Historical Sociolinguistics*, Juan Manuel Hernández-Campoy & Juan Camilo Conde-Silvestre (eds), 156–169. Chichester: Wiley-Blackwell. https://doi.org/10.1002/9781118257227.ch9

Elspaß, Stephan. 2015a. Grammatischer Wandel im (Mittel-) Neuhochdeutschen – von oben und von unten. Perspektiven einer Historischen Soziolinguistik des Deutschen. *Zeitschrift für Germanistische Linguistik* 43(3): 387–420. https://doi.org/10.1515/zgl-2015-0022

Elspaß, Stephan. 2015b. Private letters as a source for an alternative history of Late Modern German. In *Letter Writing and Language Change* [Studies in English Language Series], Anita Auer, Daniel Schreier & Richard J. Watts (eds), 35–52. Cambridge: CUP. https://doi.org/10.1017/CBO9781139088275.004

Elspaß, Stephan & Kleiner, Stefan. in print. Forschungsergebnisse zur arealen Variation im Standarddeutschen. In *Language and Space, Vol. 4: Deutsch* [Handbücher zur Sprach- und Kommunikationswissenschaft 30(4)], Joachim Herrgen & Jürgen Erich Schmidt (eds). Berlin: De Gruyter.

Elspaß, Stephan & Möller, Robert. 2003 et seqq. *Atlas zur deutschen Alltagssprache (AdA)*. <www.atlas-alltagssprache.de> (27 January 2017).

Elspaß, Stephan & Niehaus, Konstantin. 2014. The standardization of a modern pluriareal language. Concepts and corpus designs for German and beyond. *Orð og tunga* 16: 47–67.

Fehringer, Carol. 2011. Allomorphy in the German genitive. A paradigmatic account. *Zeitschrift für Germanistische Linguistik* 39: 90–112. https://doi.org/10.1515/zgl.2011.005

Fuchs-Richter, Werner. 2018a. Ausklammerung. In *Variantengrammatik des Standarddeutschen. Ein Online-Nachschlagewerk. Verfasst von einem Autorenteam unter der Leitung von Christa Dürscheid, Stephan Elspaß und Arne Ziegler*. <http://mediawiki.ids-mannheim.de/VarGra/index.php/Ausklammerung> (5 July 2018).

Fuchs-Richter, Werner. 2018b. Wortstellung im Verbalkomplex. In *Variantengrammatik des Standarddeutschen. Ein Online-Nachschlagewerk*, Christa Dürscheid, Stephan Elspaß & Arne Ziegler (eds). <http://mediawiki.ids-mannheim.de/VarGra/index.php/Wortstellung_im_Verbalkomplex> (5 July 2018).

Hentschel, Elke & Weydt, Harald. 2013. *Handbuch der deutschen Grammatik*, 4th rev. edn. Berlin: De Gruyter. https://doi.org/10.1515/9783110312973

Koch, Peter & Oesterreicher, Wulf. 2012. Language of immediacy – Language of distance: Orality and literacy from the perspective of language theory and linguistic history. In *Communicative Spaces. Variation, Contact, and Change. Papers in Honour of Ursula Schaefer*, Claudia Lange, Beatrix Weber & Göran Wolf (eds), 441–473. Frankfurt: Peter Lang.

Kooij, Jan G. 2009. Dutch. In *The World's Major Languages*, 2nd edn, Bernard Comrie (ed.), 110–124. London: Routledge.

Milroy, James & Milroy, Lesley. 1985. *Authority in Language. Investigating Language Prescription and Standardisation*. London/New York: Routledge and Kegan Paul.

Moser, Hugo. 1982. Regionale Varianten der deutschen Standardsprache. *Wirkendes Wort* 32: 327–339.

Negele, Michaela. 2012. *Varianten der Pronominaladverbien im Neuhochdeutschen. Grammatische und soziolinguistische Untersuchungen* [Studia Linguistica Germanica 108]. Berlin: De Gruyter. https://doi.org/10.1515/9783110273281

Nevalainen, Terttu & Raumolin-Brunberg, Helena. 2012. Historical sociolinguistics: Origins, motivations, and paradigms. In *The Handbook of Historical Sociolinguistics*, Juan Hernández-Campoy & Juan Camilo Conde-Silvestre (eds), 22–59. Chicester: Wiley-Blackwell. https://doi.org/10.1002/9781118257227.ch2

Niehaus, Konstantin. 2016. *Wortstellungsvarianten im Schriftdeutschen. Über Kontinuitäten und Diskontinuitäten in neuhochdeutscher Syntax* [Germanistische Bibliothek 58]. Heidelberg: Winter.

van Pottelberge, Jeroen. 2004. *Der am-Progressiv. Struktur und parallele Entwicklungen in den kontinentalgermanischen Sprachen* [Tübinger Beiträge zur Linguistik 478]. Tübingen: Narr.

Percy, Carol. 2012. Early advertising and newspapers as sources of sociolinguistic investigation. In *The Handbook of Historical Sociolinguistics*, Juan Hernández-Campoy & Juan Camilo Conde-Silvestre (eds), 191–210. Chichester: Wiley-Blackwell. https://doi.org/10.1002/9781118257227.ch11

Pfefferkorn, Oliver & Fankhauser, Peter. 2014. On the role of newspapers in disseminating foreign words in German. In *Proceedings of the Ninth Conference on International Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), 42–45. <https://ids-pub.bsz-bw.de/files/2605/Pfefferkorn_Fankhauser_On+the+role+of+historical+newspaper_2014.pdf> (27 January 2017).

von Polenz, Peter. 1989. Das 19. Jahrhundert als sprachgeschichtliches Periodisierungsproblem. In *Voraussetzungen und Grundlagen der Gegenwartssprache*, Dieter Cherubim & Klaus J. Mattheier (eds), 11–30. Berlin: De Gruyter. https://doi.org/10.1515/9783110852905.11

von Polenz, Peter. 2013. *Deutsche Sprachgeschichte vom Spätmittelalter bis zur Gegenwart, Vol. II: 17. und 18. Jahrhundert*, 2nd edn, Claudine Moulin & Dominic Harion (eds), Berlin: De Gruyter.

Rimensberger, Bettina. Forthcoming. Pluralbildung mit / ohne Umlaut. In *Variantengrammatik des Standarddeutschen. Ein Online-Nachschlagewerk*, Christa Dürscheid, Stephan Elspaß & Arne Ziegler (eds). <http://mediawiki.ids-mannheim.de/VarGra/index.php/Pluralbildung_mit_/_ohne_Umlaut> (will be available in December 2018 at the latest).

Sato, Megumi. 2016. Soziopragmatische Überlegungen zur Kasusrektion bei *wegen* in inszeniert mündlichen Texten des 18. und 19. Jahrhunderts. *Sprachwissenschaft* 41(3–4), 403–420.

Seebold, Elmar. 1983. Diminutivformen in den deutschen Dialekten. In *Dialektologie. Ein Handbuch zur deutschen und allgemeinen Dialektforschung* [Handbücher zur Sprach- und Kommunikationswissenschaft 1(2)], Werner Besch, Ulrich Knoop, Wolfgang Putschke & Herbert Ernst Wiegand (eds), 1250–1255. Berlin: De Gruyter.

Stedje, Astrid. 2007. *Deutsche Sprache gestern und heute*, 6th edn., rev. by Astrid Stedje & Heinz-Peter Prell. Paderborn: Fink.

Szczepaniak, Renata. 2010. *Während des Flug(e)s/des Ausflug(e)s?* German short and long genitive endings between norm and variation. In *Grammar between Norm and Variation* [VariOLingua 40], Alexandra N. Lenz & Albrecht Plewnia (eds), 103–126. Frankfurt: Peter Lang.

Watts, Richard J. 2011. *Language Myths and the History of English*. New York, Oxford: Oxford University Press.

Weinrich, Harald. 2003. *Textgrammatik der deutschen Sprache*, 2nd rev. edn. Hildesheim: Olms.

Zifonun, Gisela, Hoffmann, Ludger & Strecker, Bruno. 1997. *Grammatik der deutschen Sprache*, Vol. 2. Berlin: De Gruyter. https://doi.org/10.1515/9783110872163

Diachronic collocations, genre, and DiaCollo

Bryan Jurish
Berlin-Brandenburgische Akademie der Wissenschaften

This chapter presents the formal basis for diachronic collocation profiling as implemented in the open-source software tool "DiaCollo" and sketches some potential applications to multi-genre diachronic corpora. Explicitly developed for the efficient extraction, comparison, and interactive visualization of collocations from a diachronic text corpus, DiaCollo is suitable for processing collocation pairs whose association strength depends on extralinguistic features such as the date of occurrence or text genre. By tracking changes in a word's typical collocates over time, DiaCollo can help to provide a clearer picture of diachronic changes in the word's usage, especially those related to semantic shift or discourse environment. Use of the flexible DDC search engine[1] back-end allows user queries to make explicit reference to genre and other document-level metadata, thus allowing e.g. independent genre-local profiles or cross-genre comparisons. In addition to traditional static tabular display formats, a web-service plugin also offers a number of intuitive interactive online visualizations for diachronic profile data for immediate inspection.

1. Introduction

DiaCollo is an open-source software tool for automatic *collocation profiling* (Church and Hanks 1990; Evert 2005) in diachronic corpora such as the *Deutsches Textarchiv*[2] (Geyken 2013) or the Corpus of Historical American English[3] (Davies 2012) which allows users to choose the projected collocate attributes and the granularity of the diachronic axis on a per-query basis (Jurish 2015; Jurish et al. 2016). Unlike conventional collocation extractors such as DWDS Wortprofil (Didakowski and Geyken 2013) or Sketch Engine (Kilgarriff and Tugwell 2002), DiaCollo is suitable for extraction and analysis of diachronic collocation data, i.e. collocation pairs whose association strength depends on the

[1] "DWDS/Dialing Concordance", http://sourceforge.net/projects/ddc-concordance

[2] http://www.deutschestextarchiv.de

[3] http://corpus.byu.edu/coha

https://doi.org/10.1075/scl.85.03jur

date of their occurrence and/or other extralinguistic features such as author or genre. By tracking changes in a word's typical collocates over time or corpus subset and applying J. R. Firth's famous principle that "you shall know a word by the company it keeps" (Firth 1957), DiaCollo can help to provide a clearer picture of associated changes in the word's usage.

Developed in the context of the European Union CLARIN project[4] to aid historians in their analysis of the changes in discourse topics associated with selected terms as manifested by changes in those terms' context distributions, DiaCollo has been successfully applied to both mid-sized and large corpus archives, including the *Deutsches Textarchiv* (1600–1900, ca. 3.2K documents, 197M tokens) and a large heterogeneous newspaper corpus[5] (1946–2015, ca. 11M documents, 4.4G tokens). A modular web-service plugin provides access to the corpus hits for any returned collocation pair whenever the DiaCollo instance is associated with an independent DDC search engine, allowing users to proceed from the abstract, "distant" summary of a collocation profile to a more detailed "close" reading of the relevant corpus content (Moretti 2013).

The remainder of this chapter is organized as follows: after a brief discussion of previous work on both synchronic and diachronic collocation profiling in Section 2, the formal basis for diachronic collocation profiling as implemented in DiaCollo is presented in Section 3. Section 4 demonstrates some of DiaCollo's capabilities for genre-sensitive profiling by means of two simple examples. Finally, Section 5 summarizes the contribution and its implications for future work.

2. Related work

A great deal of previous work on collocation discovery with respect to synchronic corpora can be found in the literature, including early work by Church and Hanks (1990), the textbook presentation by Manning and Schütze (1999), and more recent discussions by Evert (2005, 2008) and Rychlý (2008). The techniques and association measures developed for collocation discovery in synchronic corpora are well understood and have gained a wide degree of acceptance in the corpus linguistic community. Traditional approaches treat the source corpus as a homogeneous whole however: no provision is made for changes to a word's collocation

[4] http://www.clarin.eu

[5] The *Alle Zeitungen* ('all newspapers') corpus hosted by the *Digitales Wörterbuch der deutschen Sprache* project ('digital dictionary of the German language', DWDS), cf. Geyken et al. (2017).

behavior over time or corpus subset. Since co-occurrence frequencies are traditionally collected only for pairs of (surface) strings, the potential influence of occurrence date or extra-linguistic environment is irrevocably lost.

A number of diachronic corpus studies have made use of traditional collocation measures and other distributional features to track semantic change over time. Typically, such approaches begin by manually partitioning the corpus into "epochs" or "slices" by document date and proceed by performing an independent synchronic collocation analysis of each epoch, the results of which are then manually collated for interpretation with respect to a specific research question. Baker et al. (2008) for example partition their diachronic corpus into 10 annual sub-corpora, Sagi et al. (2009) use 5 corpus epochs each covering roughly 100 years, Gulordava and Baroni (2011) focus on 2 distinct epochs of 1 decade each, and Kim et al. (2014) manually partition their data into 160 epochs of 1 year each. Scharloth et al. (2013) exploit document metadata to implicitly partition a weekly newspaper corpus into ca. 3400 micro-epochs of one issue each, using a post-processing phase to categorize target terms and estimate frequencies as moving averages over a fixed-width temporal window, and Kilgarriff et al. (2015) describe a system for neologism detection using collocation analysis in any diachronic corpus admitting at least 3 distinct epochs. The vast discrepancies in number and size of corpus epochs is not arbitrary however. On the contrary, Gabrielatos et al. (2012) argue that the selection of temporal granularity must be dependent on the research question, with due consideration given to corpus size and expected frequency of the target item(s).[6] This criterion was one of the fundamental design goals of DiaCollo.

In many cases, diachronic corpus studies employ distributional-semantic vector-similarity measures rather than traditional collocation profiles. Wang and McCallum (2006) introduced the "Topics over Time" variant of latent Dirichlet allocation, which models a continuous time variable jointly with document-level word co-occurrences as mediated by a finite set of opaque topics. Sagi et al. (2009) employ latent semantic analysis to create a compressed vector-space model with respect to co-occurrence with the 2000 most frequent content-bearing collocates,

[6] As an anonymous reviewer pointed out, Gries and Hilpert (2008) made essentially the same observation four years earlier, going on to propose a method for dynamically partitioning an input corpus into epochs of potentially non-uniform size. Their technique crucially relies on a pair of high-level functional parameters: a similarity measure and an amalgamation rule. Since instantiation of these parameters is in turn highly dependent on a particular user's research question, it is not immediately apparent how such a technique could be meaningfully implemented in the context of a user-agnostic diachronic profiling framework such as DiaCollo aims to provide.

and Kim et al. (2014) induce a series of vector-space models using neural networks as described by Mikolov et al. (2013). Relying as they do on compressed vector-space models defined in terms of opaque "topics" or "latent" dimensions, these approaches can provide at best a coarse approximation of a word's collocation behavior. While such approximations can be useful, their success crucially depends on the choice of compile-time parameters such as number of topics, which severely limits their applicability for generic diachronic collocation discovery.

3. Implementation

DiaCollo is implemented as a Perl library, and provides both a command-line interface as well as a modular RESTful web service plugin (Fielding 2000) with a form-based user interface for evaluation of runtime database queries and interactive visualization of query results. The remainder of this section presents the formal model of diachronic collocation profiles as implemented in DiaCollo. Section 3.1 provides a top-down sketch of DiaCollo's runtime functionality on the basis of a simple example using the form-based web interface. Section 3.2 describes the underlying corpus data model required for flexible attribute selection and on-the-fly diachronic partitioning. Acquisition of raw diachronic co-occurrence frequency distributions from suitably encoded corpus data is described in Section 3.3. Section 3.4 describes the online computation of association scores and the subsequent construction of a pruned diachronic profile from a raw co-occurrence frequency distribution, and Section 3.5 extends these techniques to provide direct diachronic comparisons of independent operand profiles.

3.1 Overview

Users typically interact with a DiaCollo corpus index through the web-service interface provided by the `DiaColloDB::WWW` module, an annotated screenshot of which is included here as Figure 1. In the simplest case, a user must provide at least a '`query`' parameter (q) indicating the collocant(s) for which a diachronic profile is to be computed; for the example in Figure 1, the selected collocant is the lemma *Revolution*. To acquire such a profile, the DiaCollo process must first identify all corpus lexemes matching the query and extract co-occurrence frequencies for each candidate collocate. In doing so, the corpus is implicitly partitioned into epochs of the size specified by the optional user parameter '`slice`' (E). Since not all lexical attributes may be relevant to a particular query, only those collocate attributes specified by the optional '`groupby`' (G) parameter are projected onto the result set. The diachronic interval to be queried can be restricted by means of the ‘date’ parameter, and the admissible values of indexed lexical attributes can be specified by means of optional restriction clauses in the ‘groupby’ parameter (*H*). In Figure 1, collocates are grouped jointly by their ‘Lemma’ and ‘Pos’ (part-of-speech) attributes, the diachronic interval is restricted to 1600–1899, and only common noun collocates (Pos=NN) are considered.[7]

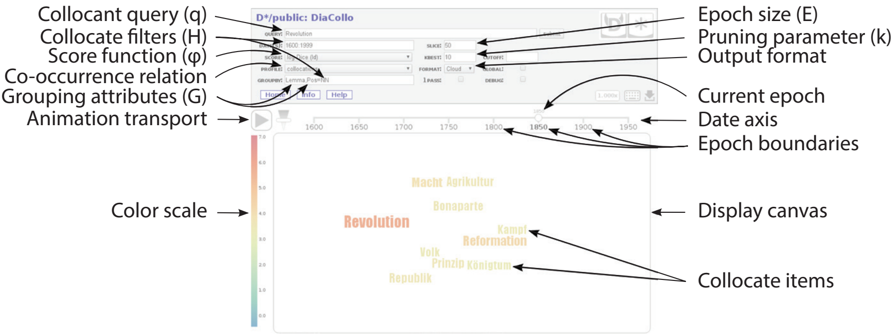

QUERY: Revolution
DATE(S): 1600:1999
SLICE: 50
SCORE: log Dice (ld)
KBEST: 10
CUTOFF:
PROFILE: collocations
FORMAT: Cloud
GLOBAL: □
GROUPBY: Lemma,Pos=NN
1PASS: □
DEBUG: □

Figure 1. Annotated screenshot of the web-based DiaCollo user interface displaying a dynamic tag-cloud visualization of the 10 best common noun collocates per 50-year epoch for the collocant *Revolution* over the interval 1600–1899. Variable names in parentheses are those used by the formal description in Sections 3.2–3.5

The precise meaning of “co-occurrence” depends on both the collocant query and the DiaCollo profiling relation specified by the ‘profile’ parameter (Section 3.3) – each co-occurrence relation supported by DiaCollo returns an epoch-labeled raw frequency profile for the requested collocant(s). Since raw frequency alone is often not a good indicator of association strength (Evert 2008), each candidate collocate is assigned a scalar association score by means of the quaternary operation specified by the ‘score’ (*φ*) parameter. Since the user is not typically interested in an exhaustive list of all candidate collocates, DiaCollo prunes the results to the highest-scoring candidates in each epoch, retaining only up to ‘kbest’ (*k*) collocates per epoch.

DiaCollo’s web interface supports a number of different output formats for displaying diachronic profile data, specified by the ‘format’ parameter. Figure 1 shows a dynamic tag-cloud visualization using the D3.js library,[8] which maps collocate items’ association strength magnitude to font size and color according to

[7] The corpus in question uses the Stuttgart-Tübingen tagset (Schiller et al. 1995) to annotate part-of-speech tags. DiaCollo itself is language- and tagset-agnostic, but requires that all annotations to be indexed are present in the input corpus at index compilation time.

[8] http://d3js.org/

a dynamically computed scale.[9] A horizontal diachronic axis acts as a slider, allowing users to "drag" a handle between individual epochs. For convenient inspection of diachronic development, this visualization mode also offers an animation transport in the form of a play/pause button and a playback speed slider: "playing" the animation causes the display canvas to interpolate smoothly between the discrete epochs represented by the underlying profile data, causing the display collocate items to change size and color as their (interpolated) association scores change, and to fade in or out as they enter or leave the set of k-best collocates for the current epoch, respectively. Detailed information on collocate items themselves as well as hyperlinks to close approximations of the underlying corpus hits are available by a pop-up dialogue invoked by clicking on a item in the main display canvas.

The rest of this section will be concerned with the precise formal characterization of the corpus data model underlying DiaCollo's diachronic collocation profiles and of the compile- and run-time computations required for their acquisition. For further information on the web interface, its capabilities and limitations, and concrete usage examples, the interested reader is referred to the introductory tutorial[10] (in German) and the DiaCollo user documentation.[11]

3.2 Corpus data

In order to track changes in a word's collocation behavior over time, we must first ensure that each corpus token is associated at least with the date of its occurrence. To this end, DiaCollo treats corpus tokens as n-tuples of potentially salient attribute values – including the occurrence date – rather than simple atomic strings. In the simplest case, a corpus need only provide surface strings $s_w \in \Sigma^*$ and associated dates $y_w \in \mathbb{N}$ as non-negative integers, encoding each token as a pair $w = \langle s_w, y_w \rangle$. Additional attributes may be encoded as well: lemmata are useful for abstracting over irregular inflection paradigms, part-of-speech tags and argument frames can help to isolate syntax-dependent phenomena, and document metadata such as author or genre can be used to restrict collocation profiles to a specific corpus subset of particular interest.

Formally, a corpus C is represented as a list of N tokens, each of which is represented as an n_A-tuple of attribute values drawn respectively from the sets $\mathcal{A}_1,\dots,\mathcal{A}_{n_A}$, i.e. $C = t_1 t_2 \dots t_N \in (\mathcal{A}_1 \times \dots \times \mathcal{A}_{n_A})^N$. Additionally, each corpus token t_i

[9] Traditional static tabular display formats suitable for further processing are of course also available.

[10] http://kaskade.dwds.de/diacollo-tutorial

[11] http://kaskade.dwds.de/diacollo/help.perl

must be associated with the date of its occurrence, $\mathrm{Y}(t_i) \in \mathbb{N}$. The set of all types in the corpus modulo occurrence date is denoted by $\mathcal{W} = \bigcup_{i=1}^{N} \{t_i\} \subseteq \mathcal{A}_1 \times \ldots \times \mathcal{A}_{n_A}$, and $\mathcal{Y} = \bigcup_{i=1}^{N} \{\mathrm{Y}(t_i)\} \subset \mathbb{N}$ denotes the set of all occurrence dates in the corpus. I will also use the notation $t[j]$ to represent the projection of the jth attribute from the n-tuple t, $t[j] = a_j$ for $t = \langle a_1,\ldots,a_n \rangle$ and $1 \leq j \leq n$. By extension, $t[J] = \langle t[j_1],\ldots,t[j_{n_J}] \rangle$ denotes the projection of the attribute-list $J = \langle j_1,\ldots,j_{n_J} \rangle$ from t, and for a relation T of arity n (i.e. a set of n-tuples), $T[J] = \bigcup_{t \in T} \{t[J]\}$ denotes the projection of the attribute-list J from T. For an n_J-tuple $u \in T[J]$, $[u]_{T/J} \subseteq T$ denotes the equivalence class modulo J of u in T; $[u]_{T/J} = \{t \in T \mid t[J] = u\}$.

3.3 Co-occurrence frequencies

Traditional collocation discovery methods for homogeneous synchronic corpora based on co-occurrence frequency distributions are well established and well understood. The most important question in the context of the current work is: how can we most effectively bring these methods to bear on heterogeneous diachronic data? Given an operational definition of what exactly it means for two words to "co-occur" in a corpus, conventional synchronic collocation profiling software such as DWDS Wortprofil (Didakowski and Geyken 2013) can compute all supported association scores for each collocation pair represented in the corpus offline, storing these in a static database for efficient runtime retrieval.[12] For DiaCollo, such a static offline database is not feasible: not only should the user be provided with full runtime control of epoch granularity as argued by Gabrielatos et al. (2012), he or she should also be allowed to choose which of the indexed corpus attributes are to be projected onto the result set, and to restrict the profiled corpus subset by means of those attributes. Semantic preferences for example can be summarized well with a lemma-based collocation profile, disregarding differences in part-of-speech tag or surface form. Detection of genre-sensitive phenomena requires the ability to restrict the discovery procedure to one or more target genres, assuming these are appropriately encoded in the corpus data.

In place of a static association score database, DiaCollo offers several different runtime methods for acquiring a "raw" epoch-labeled absolute co-occurrence frequency profile from an arbitrary user request, which must then be split into independent epoch-wise sub-profiles, and the requested association scores computed on-the-fly from the raw frequency data for each collocate item in each subprofile. While the computational load involved is substantially greater than that

[12] In practice, some form of score- or frequency-based filter is usually applied in order to reduce storage requirements and improve access speed.

required for direct static database lookup, use of index structures optimized for sparse natural language data can provide sufficiently speedy access even for large source corpora.[13] The main advantage of the DiaCollo approach is the flexibility offered by user specification of epoch granularity, target collocant selection, and associated collocate grouping. DiaCollo's runtime computation of association scores from raw frequency profiles also makes it very easy to implement support for new association scores without the need for index re-compilation.[14]

Formally, a DiaCollo request is a 6-tuple $Q = \langle q, E, G, H, \varphi, k\rangle$, where:

- q is an expression selecting the target collocant(s) given by the request `query` parameter,
- $E \in \mathbb{N}$ is the size of the epochs into which the collocation profile is to be partitioned given by the request `slice` parameter,
- $G \in \langle g_1, g_2,\ldots,g_{n_G}\rangle$ is an n_G-tuple indicating the attributes to be projected onto the result as specified by the request `groupby` parameter,
- $H : \mathcal{Y} \times \mathcal{W}[G] \to \{0, 1\}$ is a Boolean-valued filter function determined by the request's `date` and `groupby` restriction clause parameters,
- φ: $\mathbb{R}^4 \to \mathbb{R}$ is an association score function given by the request `score` parameter, and
- $k \in \mathbb{N}$ is a non-negative integer specifying the maximum number of collocate items to return per epoch.

A raw frequency profile for a user request Q over epochs from the finite set $\mathcal{E} \subset \mathbb{N}$ is a 4-tuple $R_Q = \langle r_N, r_1, r_2, r_{12}\rangle$, where:[15]

- $r_N : \mathcal{E} \to \mathbb{N}$ maps each epoch to the total number of co-occurrences in that epoch,
- $r_1 : \mathcal{E} \to \mathbb{N}$ maps each epoch to the total independent frequency of the collocant(s) in that epoch,

[13] An 8-epoch query over a 4.4 billion word corpus for example is evaluated using the native co-occurrence relation (Section 3.3.1) in under 5 seconds if the index data are present in the operating system's buffer cache, and in under 20 seconds if the data need to be paged in from disk.

[14] This feature could in principle be extended to support generic user-defined score function scripts in the spirit of *GraphColl* (Brezina et al. 2015).

[15] Note that the values of r_N, r_1, r_2, and r_{12} are *co-occurrence counts* and not necessarily traditional (unigram) frequencies. This is important for association score functions such as pointwise mutual information which properly operate on probabilities, in order to avoid overflow.

- $r_2 : \mathcal{E} \times \mathcal{W}[G] \to \mathbb{N}$ maps each epoch-labeled collocate item to its independent frequency in the respective epoch, and
- $r_{12} : \mathcal{E} \times \mathcal{W}[G] \to \mathbb{N}$ maps epoch-labeled collocate items to the associated co-occurrence frequencies with the requested collocant(s).

The following subsections describe three of DiaCollo's runtime methods for raw frequency profile acquisition. These methods differ not only with respect to the underlying data structures employed, but also with respect to what exactly constitutes a corpus "co-occurrence" to be counted in the profile.

3.3.1 *Native co-occurrence relation*

For efficient profiling of co-occurrences within a fixed-width moving window of $\ell \in \mathbb{N}$ adjacent content tokens, DiaCollo uses a two-level native binary index I_{12} to associate pairs of fully specified attribute n_A-tuples with their absolute co-occurrence frequencies at each attested date unit. To populate I_{12} at index compilation time, the corpus is assumed to be partitioned into n_S contiguous segments[16] $s_1 s_2 \ldots s_{n_S} = C$ with $s_i = s_{i1} s_{i2} \ldots s_{in_{s_i}}$ a list of n_{s_i} corpus tokens, and the native index is populated by counting co-occurrences in the specified window within each corpus segment as in (1):[17]

(1)
$$I_{12} : \mathcal{W} \times \mathcal{W} \times \mathcal{Y} \to \mathbb{N}$$
$$: \langle w,v,y \rangle \mapsto \sum_{i=1}^{n_S} \sum_{j=1}^{n_{s_i}} \sum_{j'=\max\{j-\ell,1\}}^{\min\{j+\ell,\, n_{s_i}\}} \mathbb{1}\,[j \neq j' \;\&\; s_{ij} = w \;\&\; s_{ij'} = v \;\&\; \mathrm{Y}(s_{ij}) = y]$$

Independent occurrence frequencies are stored as true marginals over I_{12}:

(2)
a. $I_1 : \mathcal{W} \times \mathcal{Y} \to \mathbb{N} : \langle w,y \rangle \mapsto \Sigma_{v \in \mathcal{W}}\, I_{12}(w, v, y)$
b. $I_N : \mathcal{Y} \to \mathbb{N} : y \mapsto \Sigma_{w \in \mathcal{W}}\, I_1(w, y)$

[16] If available, sentence boundaries are good candidates for corpus segments. Implicit segment boundaries are inserted before and after each corpus source file. Since DiaCollo never counts co-occurrences crossing segment boundaries, this ensures that whenever a co-occurrence is counted, the co-occurring items do indeed share a common date label.

[17] $\mathbb{1}[\psi] \in \{0,1\}$ is the indicator function for the truth-valued formula ψ; $\mathbb{1}[\psi] = 1$ if and only if ψ holds true for the current variable bindings, otherwise $\mathbb{1}[\psi] = 0$. Equation (1) thus counts exactly one co-occurrence of an ordered pair of terms $\langle w, v \rangle$ for each pair of distinct corresponding tokens ($w = s_{ij}$, $v = s_{ij'}$, $j \neq j'$) within a single corpus segment s which are separated by no more than ℓ intervening items ($-\ell \leq j' \leq \ell$). Note that identity pairs ($w = v$) will still be counted whenever multiple tokens of a single type co-occur within the selected context window.

At runtime, a user collocant query q is first expanded to a set of attribute tuples $[\![q]\!] \subseteq \mathcal{W}$ – for a simple single-value single-attribute query such as "`$lemma=love`" this is simply a matter of selecting the appropriate tuples from the corpus vocabulary, $[\![\texttt{\$lemma=love}]\!] = [\texttt{love}]_{\mathcal{W}/a_{\text{lemma}}}$. Then, a fully specified co-occurrence frequency distribution $\hat{I}_q$ at date-unit granularity can be computed as (3):

(3) $\hat{I}_q : \mathcal{Y} \times \mathcal{W} \to \mathbb{N} : \langle y,v \rangle \mapsto \sum_{w \in [\![q]\!]} I_{12}(w, v, y)$

Next, the co-occurrence distribution must be aggregated by the requested attributes G:

(4) $\hat{I}_{q,G} : \mathcal{Y} \times \mathcal{W}[G] \to \mathbb{N} : \langle y,g \rangle \mapsto \sum_{v \in [g]_{\mathcal{W}/G}} \hat{I}_q(y, v)$

The co-occurrence distribution must then be restricted to that subset of projected tuples satisfying the request filter function H as in (5):

(5) $$\hat{I}_{q,G,H} = \hat{I}_{q,G} \restriction H^{-1}(1) : \mathcal{Y} \times \mathcal{W}[G] \to \mathbb{N} : \langle y,g \rangle \mapsto \begin{cases} \hat{I}_{q,G}(y, g) & \text{if } H(y, g) = 1 \\ \text{undefined} & \text{otherwise} \end{cases}$$

If $E > 0$ is the epoch size requested by the user via the `slice` option, then the function $\tilde{E} : \mathcal{Y} \to \mathbb{N}$ maps each date value y to its associated epoch label $\tilde{E}(y) = E\left\lfloor \frac{y}{E} \right\rfloor$.[18] DiaCollo epochs are thus labeled by their minimum possible element, so for a decade-wise partitioning ($E = 10$) the epoch label "1970" represents the interval 1970–1979. Let $\mathcal{E}_E \subset \mathbb{N}$ be the set of all zero-offset epochs of size E available in the corpus and let $[e]_E \subseteq \mathcal{Y}$ be the set of dates assigned to the epoch $e \in \mathcal{E}_E$; $\mathcal{E}_E = \tilde{E}(Y) = \bigcup_{y \in \mathcal{Y}} \{\tilde{E}(y)\}$ and $[e]_E = \tilde{E}^{-1}(e) = \{y \in \mathcal{Y} \mid \tilde{E}(y) = e\}$. Then, the filtered distribution can be aggregated by epoch as in (6):

(6) $\hat{I}_{q,G,H,E} : \mathcal{E}_E \times \mathcal{W}[G] \to \mathbb{N} : \langle e,g \rangle \mapsto \sum_{y \in [e]_E} \hat{I}_{q,G,H}(y, g)$

The desired raw frequency co-occurrence distribution to be returned is then $\hat{I}_{q,G,H,E}$, and the independent profile frequencies r_1, r_2, and r_{N} can be acquired by consulting the marginal indices for each pair : $\langle e, g \rangle \in \operatorname{dom}(\hat{I}_{q,G,H,E})$:

(7) a. $r_{\mathrm{N}}(e) = \sum_{y \in [e]_E} I_{\mathrm{N}}(y)$

b. $r_1(e) = \sum_{y \in [e]_E} \sum_{w \in [\![q]\!]} I_1(w,y)$

c. $r_2(e,g) = \sum_{y \in [e]_E} \sum_{v \in [g]_{\mathcal{W}/G}} I_1(v,y)$

d. $r_{12}(e,g) = \hat{I}_{q,G,H,E}(e,g)$

Note that the runtime evaluation of Equations (3)–(7) can in most cases be efficiently performed with the help of appropriate auxiliary indices, without the

[18] The special case $E = 0$ is interpreted as a request for a synchronic profile over the entire corpus; $\tilde{0}(y) = 0$ for all $y \in Y$.

need to iterate over all corpus tokens or types. Cached multi-maps for converting attested attribute values to the associated tuples in $\mathcal{W}$ for example enable efficient expansion of $[\![q]\!]$ and $[g]_{\mathcal{W}/G}$. Similarly, since only attested collocations with nonzero frequency[19] are stored in the index, evaluating Equation (3) will require iterating over only those candidate collocates which actually co-occur with the target collocant(s). Since these items are stored sequentially in the native index file, access speed for joint co-occurrence frequencies is improved even for large collocate sets on contemporary operating systems using a read-ahead cache. The filter function H and epoch partitioning $\tilde{E}$ need only be evaluated for attested co-occurrences as well, and no explicit computation of the inverse relations H^{-1} and $\tilde{E}^{-1}$ is required. In practice, the running time of native co-occurrence profiling is typically dominated by the retrieval of attested co-occurrences and their independent frequencies from the indices on disk as described by Equations (3) and (7c).

3.3.2 *Term × document matrix co-occurrence relation*

The fixed-width moving window notion of co-occurrence is not suitable for all applications. Assuming the default sentence segmentation granularity for example, very few potential collocates will be indexed for infrequent terms, leading to hyper-specific but uninterpretable result profiles for these items. In order to ameliorate such sparse data problems, most conventional distributional semantic models (Berry et al. 1995; Blei et al. 2003) represent the underlying corpus as a term × document frequency (TDF) matrix, which is then typically mapped to a low-dimensional approximation in terms of “latent” factors. In a similar spirit, DiaCollo offers a term × document co-occurrence relation drawing on a broader range of candidate collocates than those accessible by a short moving window. Unlike conventional distribution semantic models however, DiaCollo’s TDF co-occurrence relation does not rely on opaque topics or latent factors to estimate term similarities, but provides exact co-occurrence counts for each attested collocation pair.

Formally, the TDF co-occurrence relation is defined *in terms of a corpus partitioning* into a set Doc of documents[20] and a frequency matrix $\mathrm{tdf} : \mathcal{W} \times \mathrm{Doc} \to \mathbb{N}$ such that $\mathrm{tdf}(w, d)$ is the frequency of the term w in document d. At index compilation time, DiaCollo explicitly creates and stores such a matrix together with an auxiliary vector mapping documents to the associated dates $\mathrm{dy} : \mathrm{Doc} \to \mathcal{Y}$,

[19] Arbitrary attribute-wise and co-occurrence frequency threshold values may be specified at index compilation time to further compress the disk index and improve access speed.

[20] By default, DiaCollo uses paragraph boundaries as “documents”.

as well as an independent marginal date-frequency vector $\mathrm{yf} : \mathcal{Y} \to \mathbb{N} : y \longmapsto \Sigma_{w \in \mathcal{W}} \Sigma_{d \in \mathrm{dy}^{-1}(y)} \mathrm{tdf}(w,d)$.

At runtime, a user request q is interpreted independently as a set of term-tuples $[\![q]\!]_{\mathcal{W}} \subseteq \mathcal{W}$ and a set of documents $[\![q]\!]_{\mathrm{Doc}} \subseteq \mathrm{Doc}$. If $[\![q]\!]_{/y} = [\![q]\!]_{\mathrm{Doc}} \cap \mathrm{dy}^{-1}(y)$ represents the subset of queried documents with date $y \in \mathcal{Y}$, then a TDF co-occurrence frequency distribution $\hat{I}_{\mathrm{tdf}:q,G}$ at date-unit granularity over the requested projection attributes G is computed as (8):

$$
\begin{aligned}
(8)\quad & \hat{I}_{\mathrm{tdf}:q,G} : \mathcal{Y} \times \mathcal{W}[G] \to \mathbb{N} \\
& : \langle y, g \rangle \mapsto \Sigma_{d \in [\![q]\!]_{/y}} \min \left\{ \left(\Sigma_{w \in [\![q]\!]_{\mathcal{W}}} \mathrm{tdf}(w, d) \right), \left(\Sigma_{v \in [g]_{\mathcal{W}/G}} \mathrm{tdf}(v, d) \right) \right\}
\end{aligned}
$$

Candidate filtering and epoch aggregation are analogous to the procedures described by Equations (6) and (7) for the native co-occurrence relation, and the final TDF raw frequency profile to be returned is defined for each pair $\langle e, g \rangle \in \mathrm{dom}(\hat{I}_{\mathrm{tdf}:q,G,H,E})$ as:

$$
\begin{aligned}
(9)\quad \text{a.}\quad & r_{\mathrm{N}}(e) = \Sigma_{y \in [e]_E} \mathrm{yf}(y) \\
\text{b.}\quad & r_1(e) = \Sigma_{y \in [e]_E} \Sigma_{w \in [\![q]\!]_{\mathcal{W}}} \Sigma_{d \in [\![q]\!]_{/y}} \mathrm{tdf}(w,d) \\
\text{c.}\quad & r_2(e,g) = \Sigma_{y \in [e]_E} \Sigma_{v \in [g]_{\mathcal{W}/G}} \Sigma_{d \in \mathrm{dy}^{-1}(y)} \mathrm{tdf}(v,d) \\
\text{d.}\quad & r_{12} = \hat{I}_{tdf:q,G,H,E}(e,g)
\end{aligned}
$$

As for native profiling, the runtime evaluation of TDF co-occurrence profiles can be optimized by use of appropriate data storage formats and access methods. DiaCollo employs a pair of Harwell-Boeing offset pointers (Duff et al. 1989) to optimize access to the sparse TDF matrix stored on disk. Index lookup, aggregation, and filtering are performed by optimized routines written in the Perl Data Language "PDL" (Glazebrook and Economou 1997) for manipulation of large numerical data structures, using the operating system's memory-mapping facility for transparent on-demand paging.

3.3.3 *DDC co-occurrence relation*

DiaCollo's DDC co-occurrence relation makes use of the DDC search engine (Sokirko 2003; Jurish et al. 2014) to acquire raw frequency profiles from a running DDC server.[21] Unlike the native co-occurrence index, which implicitly defines a "co-occurrence" of two corpus items to be any occurrence of both items within a fixed-width moving window over the corpus, the DDC co-occurrence relation makes no implicit assumptions regarding which corpus configurations constitute

[21] A companion DDC index for the underlying corpus must be independently configured and created.

"co-occurrences". Rather, DiaCollo's DDC back-end makes use of user-supplied search term subscripts ("match-IDs") to identify which elements of the query-string q represent the collocant(s) and which represent the associated collocates. By convention, collocant terms are identified by the subscript '`=1`' and the associated collocate terms are identified by the subscript '`=2`'. Query terms without an explicit subscript are treated as collocant restrictions and implicitly assigned the subscript '`=1`'.

By using subscripted DDC queries, the definition of what exactly constitutes a "co-occurrence" depends entirely on the query specified by the user, which may be any context query expressible in the DDC query language.[22] In particular, the DDC query language supports arbitrary corpus segmentations (e.g. sentences, paragraphs, files), Boolean expressions on both the token- and the segment-level, phrase- and proximity queries (e.g. immediate predecessor, immediate adjacency), bibliographic metadata filters and grouping, and server-side term expansion pipelines (e.g. thesauri).

This flexibility comes at a price however: since DDC itself was designed as a search engine rather than a collocation database, the entire corpus must be searched and each occurrence explicitly identified in order to acquire a raw co-occurrence frequency distribution. In particular, acquisition of the independent collocate-frequency distribution r_2 may need to process a huge number of individual occurrences, which makes DiaCollo's DDC back-end a comparatively slow and resource-hungry profile acquisition method.[23]

[22] http://odo.dwds.de/~jurish/software/ddc/querydoc.html

[23] Performance of the DDC back-end depends heavily on the user query and the size of the underlying corpus. Simulating the native co-occurrence relation's moving context window for example using the DDC query `NEAR(`$*$=2, q, $\ell - 1$) first requires an inexpensive index lookup of $O(|q| \log(|\mathcal{W}|))$, followed by explicit processing of $f_{12} \approx 2\ell \mathrm{f}(q)$ co-occurrences for evaluation of (10d) and (10b). Heaps' (1978) law indicates that the number of candidate collocates discovered will be approximately $n_2 \approx \sqrt{f_{12}}$, and their total expected independent frequency can be estimated *ceteris paribus* as $f_2 \approx n_2 N/|\mathcal{W}|$, so the processing time for DDC frequency profile acquisition is dominated by (10c) with $O(n_2 \log(|\mathcal{W}|) + f_2)$ Since DDC indices in general store every token in the original corpus – including closed-class items – the candidate token count f_2 can in practice grow very large even for "simple" collocant queries with only a few co-occurrences f_{12}. Evaluating all of (10a)–(10d) using a DDC-based approximation of the native index query from Footnote 13 on the same 4.4 billion word corpus requires more than 5 minutes to complete even for an immediate adjacency query ($\ell = 1$), and causes DiaCollo to return an error message indicating that the underlying search engine query timed out. The epoch-size query of (10a) necessarily covers even more tokens than (10c), but since its `#BY`-clause does not refer to any token-level attributes, it can be evaluated by an optimized subroutine in near-constant time, typically under 100 millisec-

Raw frequency profiles are generated for the DDC co-occurrence relation for a user request Q by means of four separate DDC query requests – one request for each component of the frequency profile:

(10) a. $r_\mathrm{N} = \lambda_q \times \texttt{COUNT(* \#SEP) \#BY[date/}E\texttt{]}$

b. $r_1 = \lambda_q \times \texttt{COUNT(KEYS(}[\![q\&H]\!]\texttt{ \#SEP \#BY[}G\texttt{=1]) \#SEP)}$
$\texttt{\#BY[date/}E, G\texttt{=1]}$

c. $r_2 = \lambda_q \times \texttt{COUNT(KEYS(}[\![q\&H]\!]\texttt{ \#SEP \#BY[}G\texttt{= 2]) \#SEP)}$
$\texttt{\#BY[date/}E, G\texttt{=2]}$

d. $r_{12} = \texttt{COUNT(}[\![q\&H]\!]\texttt{ \#SEP \#BY[date/}E, G\texttt{=2])}$

Here, $[\![q\&H]\!]$ is a DDC query string representing the logical conjunction of the request query q and filter conditions H. λ_q is a scaling coefficient heuristically determined by the query q which ensures that the joint and independent frequencies returned are compatible. A simple immediate-successor query (“`love *=2`”) for example will be assigned $\lambda_q = 1$, since at most one collocate occurrence can be identified for each occurrence of the collocant, while an immediate adjacency query such as `NEAR(love,*=2,0)` will be assigned $\lambda_q = 2$ since up to two co-occurrences (left + right) will be counted for each occurrence of the collocant.

3.4 Scoring and pruning

All association score functions supported by DiaCollo are defined in terms of the frequency values provided by a raw co-occurrence profile $\mathrm{R}_Q = \langle r_\mathrm{N}, r_1, r_2, r_{12}\rangle$ for the user request Q. The default score function used by DiaCollo is the scaled log-Dice ratio as introduced by Rychlý (2008). For details on the relative merits of other supported score functions, the interested reader is referred to the excellent discussion by Evert (2008). Formally, an association score function is a quaternary operation on real numbers φ: $\mathbb{R}^4 \to \mathbb{R}$. Independent score profiles $P_{Q,e}$ are computed for each epoch $e \in \mathcal{E}_E$ by applying φ to each candidate collocate item in turn:

(11) $$p_{Q,e} : \mathcal{W}[G] \to \mathbb{R} : g \mapsto \varphi(r_\mathrm{N}(e), r_1(e), r_2(e,g), r_{12}(e,g))$$

Since the user is typically not interested in an exhaustive profile of all potential collocates, DiaCollo prunes each epoch-local profile $p_{Q,e}$ to its k-best collocates

onds. Since the native and TDF relations do not store or process individual tokens, the performance of (7c) and (9c) is limited for analogous queries only by the number of candidate collocate types to $O(n_2 \log(|\mathcal{W}|))$.

before returning the results.[24] For $k \in \mathbb{N}$ and a real-valued function f, let $\text{best}_k(f) \subseteq \text{dom}(f)$ represent the set of k elements from dom(f) with maximal values under f; $|\text{best}_k(f)| = \min\{k, |\text{dom}(f)|\}$ and $f(x) \geq f(y)$ for all $x \in \text{best}_k(f)$ and all $y \in \text{dom}(f) \setminus \text{best}_k(f)$. The epoch-local profile $\hat{p}_{Q,e} \subseteq p_{Q,e}$ resulting from k-best pruning can then be expressed as (12):

$$(12) \quad \hat{p}_{Q,e} = p_{Q,e} \restriction \text{best}_k(p_{Q,e}) : g \mapsto \begin{cases} p_{Q,e}(g) & \text{if } g \in \text{best}_k(p_{Q,e}) \\ \text{undefined} & \text{otherwise} \end{cases}$$

The final diachronic profile $\hat{P}_Q$ returned by a unary DiaCollo request Q is then simply a function mapping epoch labels to the corresponding pruned sub-profiles:

$$(13) \quad \hat{P}_Q : \mathcal{E}_E \to \mathbb{R}^{\mathcal{W}[G]} : e \mapsto \hat{p}_{Q,e}$$

3.5 Comparisons

In addition to simple requests which return a score profile for the specified collocant(s), DiaCollo also offers a “comparison” or “diff” mode, by means of which the user may request a summary of the most prominent similarities or differences between two independently evaluated queries, e.g. between two different words or between occurrences of the same word in different date intervals, different text genres, or in the works of different authors. The first step in computing a diachronic comparison profile for two independent query requests Q_a and Q_b is to define an epoch-alignment $\mathcal{E}_{a \bowtie b} \subseteq \mathcal{E}_{E_a} \times \mathcal{E}_{E_b}$ supports both trivial alignments with singleton epoch domains as in (14a) and alignments of uniform-sized epoch domains as in (14b), where $X^{\leq}$ is the $|X|$-tuple resulting from sorting the elements of the finite set $X \subset \mathbb{N}$ by the natural order $\leq$.[25] Query pairs whose epoch domains do not satisfy either of these conditions cause DiaCollo to return an error message indicating that the alignment failed.

$$(14) \quad \begin{array}{ll} \text{a.} & \\ \text{b.} & \mathcal{E}_{a \bowtie b} = \\ \text{c.} & \end{array} \begin{cases} \mathcal{E}_{E_a} \times \mathcal{E}_{E_b} & \text{if } \min\{|\mathcal{E}_{E_a}|, |\mathcal{E}_{E_b}|\} \leq 1 \\ \bigcup_{i=1}^{N_{\mathcal{E}}} \left\{\langle \mathcal{E}^{\leq}_{E_a}[i], \mathcal{E}^{\leq}_{E_b}[i] \rangle\right\} & \text{if } |\mathcal{E}_{E_a}| = |\mathcal{E}_{E_b}| = N_{\mathcal{E}} \\ \text{undefined} & \text{otherwise} \end{cases}$$

Given such an epoch alignment $\mathcal{E}_{a \bowtie b}$ and epoch-wise operand profiles $\{p_{Q_a, e_a}\}_{e_a \in \mathcal{E}_{E_a}}$ and $\{p_{Q_b, e_b}\}_{e_b \in \mathcal{E}_{E_b}}$, a diachronic comparison sub-profile $p_{Q_a \ominus Q_b, e_{ab}}$ is computed for

[24] An alternative pruning method allows the user to specify a minimum score threshold for collocate items.

[25] In the simplest case, Q_a and Q_b share the same epoch domain $\mathcal{E} = \mathcal{E}_{E_a} = \mathcal{E}_{E_b}$ and the alignment will be the identity relation over the shared domain, $\mathcal{E}_{a \bowtie b} = \text{Id}(\varepsilon) = \{\langle e, e \rangle\}_{e \in \mathcal{E}}$.

Table 1. Comparison operations supported by DiaCollo in "diff" mode. "Pre-trimmed" comparison operations are defined over the union of pruned operand domains $\mathrm{Dom}_{Q_a \ominus Q_b/e_{ab}} = \mathrm{dom}(\hat{p}_{Q_a,e_a}) \cup \mathrm{dom}(\hat{p}_{Q_b,e_b})$, while "restricted" operations use the intersection of the un-pruned operand domains, $\mathrm{Dom}_{Q_a \ominus Q_b/e_{ab}} = \mathrm{dom}(p_{Q_a,e_a}) \cap \mathrm{dom}(p_{Q_b,e_b})$

Label	**Domain**	$x \ominus y$	**Description**
`diff`	pre-trimmed	$x - y$	raw difference
`adiff`	pre-trimmed	$\lvert x - y \rvert$	absolute difference
`max`	restricted	$\max\{x, y\}$	maximum
`min`	restricted	$\min\{x, y\}$	minimum
`avg`	restricted	$\frac{1}{2}(x + y)$	arithmetic average
`havg`	restricted	$\frac{1}{2}\left(\frac{2xy}{x + y} + \frac{1}{2}(x + y)\right)$	pseudo-harmonic average

each pair of aligned epochs $e_{ab} = \langle e_a, e_b \rangle \in \varepsilon_{a \bowtie b}$ by applying a binary comparison operation $\ominus : \mathbb{R}^2 \to \mathbb{R}$ to each collocate item in turn:[26]

$$(15) \quad p_{Q_a \ominus Q_b, e_{ab}} : \mathrm{Dom}_{Q_a \ominus Q_b/e_{ab}} \to \mathbb{R} : g \mapsto p_{Q_a,e_a}(g) \ominus p_{Q_b,e_b}(g)$$

Here, $\mathrm{Dom}_{Q_a \ominus Q_b/e_{ab}} \subseteq \mathrm{dom}(p_{Q_a,e_a}) \cup \mathrm{dom}(p_{Q_b,e_b}) \subseteq \mathcal{W}[G]$ is the characteristic domain of the comparison profile, which depends on the comparison operation $\ominus$. A list of selected comparison operations supported by DiaCollo and their characteristic domains is given in Table 1. The raw difference operation $\ominus_{\texttt{diff}}$ selects collocate items which associate strongly only with Q_a, while the default comparison operation $\ominus_{\texttt{adiff}}$ selects those items which associate strongly with only one of Q_a or Q_b, regardless of which collocant is preferred. Collocates showing strong associations for both operand profiles can be selected with the $\ominus_{\texttt{havg}}$ operation, which uses the harmonic average of operand scores.[27]

[26] Undefined operand values are treated as zeroes when computing comparison scores via Equation (15). The DiaCollo API also ensures that the projected attributes and score functions of comparison operand requests are compatible, $G_a = G_b$ and $\varphi_a = \varphi_b$.

[27] The actual comparison score value for $\ominus_{\texttt{havg}}$ is computed as shown in Table 1 as the mean of the harmonic and arithmetic averages of the operand scores in order to avoid singularities due to disjoint operand domains and the associated implicit zero score values.

Comparison profiles must then be pruned to their k-best collocates as given in Equation (16), analogous to the unary profile pruning procedure described in Section 3.4. In the case of comparison profiles however, the score values to be returned may be distinct from those used for k-best selection. In particular, the default absolute difference comparison operation $\ominus_{\mathtt{adiff}}$ selects the k collocates with maximally dissimilar association scores between its operand profiles, regardless of which of the operands displays the stronger preference (as expressed by the sign of the raw difference score). For intuitive interpretation of comparison profile results, the returned values should retain the information provided by the sign of the raw difference score. To this end, each comparison operation $\ominus$ is implicitly associated with a companion operation $\boxminus$ for producing final return values. The absolute difference operation is configured to return raw difference scores by setting $\boxminus_{\mathtt{adiff}} = \ominus_{\mathtt{diff}}$. For all other comparison operations δ, $\boxminus_{\delta} = \ominus_{\delta}$ and Equation (16) is completely analogous to the unary profile pruning procedure given by Equation (12).

$$\text{(16)} \quad \hat{p}_{Q_a \ominus Q_b, e_{ab}} = p_{Q_a \boxminus Q_b, e_{ab}} \upharpoonright \mathrm{best}_k(p_{Q_a \ominus Q_b, e_{ab}})$$
$$: g \mapsto \begin{cases} p_{Q_a \boxminus Q_b, e_{ab}}(g) & \text{if } g \in \mathrm{best}_k(p_{Q_a \ominus Q_b, e_{ab}}) \\ \text{undefined} & \text{otherwise} \end{cases}$$

The final diachronic comparison profile $\hat{P}_{Q_a \ominus Q_b}$ is defined by partitioning the set of sub-profiles by aligned epochs, analogous to the procedure for unary profiles given by Equation (13):

$$\text{(17)} \quad \hat{P}_{Q_a \ominus Q_b} : \mathcal{E}_{a \bowtie b} \to \mathbb{R}^{\mathcal{W}[G]} : e_{ab} \longmapsto \hat{p}_{Q_a \ominus Q_b, e_{ab}}$$

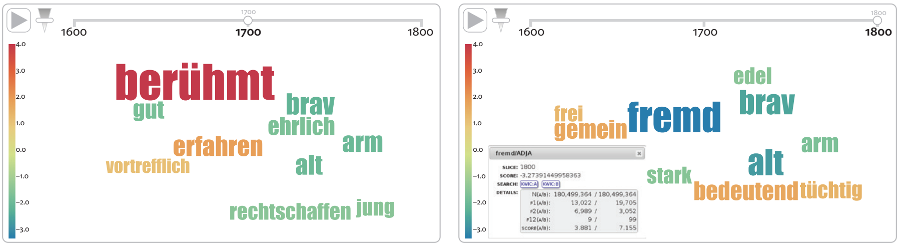
3.6 Output & visualization

In addition to traditional static tabular display formats suitable for further automated processing (JSON) or import into an external spreadsheet program (TAB-separated text, HTML), the DiaCollo web-service plugin also offers several interactive online visualizations of diachronic profile data for exploratory use. Supported visualization formats include two-dimensional time series plots using the Highcharts JavaScript library, flash-based motion charts using the Google Motion Chart library, and interactive tag-cloud and bubble-chart visualizations using the D3.js library. The HTML and D3-based display formats provide an intuitive color-coded representation of the association score associated with each collocate item, as well as hyperlinks to underlying corpus hits for each data point displayed whenever a DDC search engine for the underlying corpus is available.

4. Examples

4.1 Adjectival attribution: What makes a "man"?

Figure 2 contains example tag-cloud visualizations for a simple cross-genre comparison over the *Deutsches Textarchiv* corpus. The DDC back-end was used to acquire raw frequency counts over 100-year epochs for all attributive adjective lemmata immediately preceding an instance of the noun *Mann* ('man') in the genres *Wissenschaft* ('science') and *Belletristik* ('belles lettres'), respectively. The results were collected as a comparison profile using the absolute difference operation $\ominus_{\texttt{adiff}}$ over log-Dice operand profiles to select the most dissimilar association preferences of *Mann* in the respective genres, as described in Section 3.5. In the tag-cloud visualization mode, absolute magnitudes of score differences are mapped to tag font-size, and the raw score differences are mapped to an intuitive color-scale, with warm tones indicating a relative preference for the "science" genre and cool tones indicating a preference for "belles lettres". As Figure 2 shows, men in scientific texts are more likely to be described as *berühmt* ('famous'), *erfahren* ('experienced'), *bedeutend* ('significant'), or *tüchtig* ('capable'); while men in belles lettres are more likely to be designated *brav* ('well-behaved'), *rechtschaffen* ('righteous'), *arm* ('poor'), or *alt* ('old') – presumably reflecting the properties considered most salient in the context of the respective genres.



fremd/ADJA

SLICE: 1800
SCORE: -3.27391449958363
SEARCH: KWIC:A KWIC:B
DETAILS:

N(A/B):	180,499,364 /	180,499,364
F1(A/B):	13,022 /	19,705
F2(A/B):	6,989 /	3,052
F12(A/B):	9 /	99
SCORE(A/B):	3.881 /	7.155

Figure 2. DiaCollo interactive tag-cloud visualization of the 10 most dissimilar adjectives immediately preceding the noun Mann ('man') in the genres "science" (warm colors) and "belles lettres" (cool colors) over the Deutsches Textarchiv corpus for the epochs 1700–1799 (left) and 1800–1899 (right).

4.2 Pronominal adverbs and deictic locality

As closed-class items, pronominal adverbs are good candidates for collocation profiling even in small corpora, since they tend to be highly frequent. Figure 3 shows tag-cloud visualizations for a comparison of pronominal adverb use by text genre over the aggregated *DTA+DWDS* corpus, consisting of the *Deutsches Textarchiv* corpus covering roughly 1600–1900 and the *DWDS Kernkorpus* of 20th century German. Here again, the DDC back-end was used to acquire raw frequency counts for all pronominal adverbs in the genres “science” and “belles lettres”, and the absolute difference operation over log-Dice operand profiles was used to select the most prominent dissimilarities.

Most immediately striking is a strong preference on the part of the adverbs *drüber* (‘there-over’, ‘over which’, ‘about which’) and *drunter* (‘there-under’, ‘under which’, ‘among which’) for the literary genre, especially in younger epochs. Closer inspection of the associated corpus hits via the hyperlinks supplied by DiaCollo shows that this preference is due almost exclusively to the colloquial fixed expression *drunter und drüber* (‘at sixes and sevens’, ‘chaotically disorganized’), rather than any alternative independent lexical senses of these items, for which both academic and literary texts tended in later epochs to employ the uncontracted variants *darunter* rsp. *darüber*. Although a number of laboratory situations might accurately be described as “chaotically disorganized”, such a state is very much at odds with the scientific ideal of sober, systematic investigation. Taken together with the colloquial flavor of the fixed expression, it is unsurprising that it has seen only minimal use in scientific prose.

The tag-cloud animation further reveals divergent trends in the behavior of local vs. non-local deictic adverbials in the respective genres, visible beginning in the second half of the 18th century. Scientific texts show a strong preference for local deictic adverbials such as *hierfür* (‘here-for’, ‘for which’), *hierbei* (‘here-by’, ‘by which’), and *hieraus* (‘here-out’, ‘out of which’) which literary texts lack. The corresponding non-local variants such as *dafür* (‘there-for’), *dabei* (‘there-by’), and *daraus* (‘there-out’) show only minimal differences across the target genres, tending to zero in later epochs.

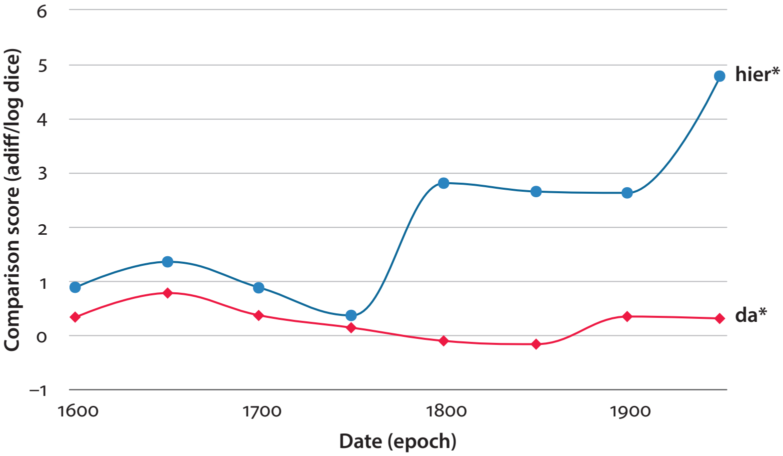


Figure 4. DiaCollo time series plot of selected local (*hier*-) and non-local (*da*-) deictic pronominal adverbs (*-für*, *-bei*, and *-aus*) in the aggregated DTA+DWDS corpus (1600–1999). Plotted axis values are differences in log-Dice association scores for scientific vs. literary texts; higher values indicate a stronger preference for scientific texts.

Focusing our attention on the locality of these three pairs leads to the DiaCollo time series plot in Figure 4, in which the inter-genre divergence between the epochs 1750–1799 and 1800–1849 is immediately apparent, which after a plateau of ca. 150 years increases again in the final epoch (1950–1999). An adequate analysis of this phenomenon is beyond the scope of this contribution, but cursory examination of the associated corpus hits indicates that the *hier-* adverbials are used primarily in their epistemic senses,[28] while the corresponding *da-* variants tend to favor (spatial and/or temporal) locative and telic readings.

If this tendency is representative, the stronger synchronic association of *hier-* adverbials with scientific prose might be explained in terms of the communicative aims underlying the two genres – scientific argumentation relying more heavily on explicit epistemic relations, while narrative prose is more concerned with spatio-temporal and telic exposition. Scientific texts do indeed seem to make more frequent use of epistemic relations, as supported by the genre's stronger association with non-deictic consequentials such as *demnach* ('according to which') and *infolgedessen* ('following from which'). Literary texts on the other hand are more strongly associated with non-deictic spatial and temporal locatives such as *worauf* ('whereupon') and *seitdem* ('since which'), as well as adversative/concessive connectors including *dawider* ('contrary to which') and *trotzdem* ('despite which'). Both Biber et al. (1999) and Herrmann (2013) found that local ("proximate") demonstratives such as *this* and *these* occur more frequently in English-language

[28] I use the term "epistemic" here in its most literal sense 'of or pertaining to knowledge', epistemic readings of (anaphoric) pro-adverbs being those which express a (necessary) logical or epistemic relation between the antecedent proposition and that modified by the pro-adverb (typically a logical consequent), as in (emphasis added):

> *Ursprünglich besagt Wahrheit soviel wie Erschließendsein als Verhaltung des Daseins. Die* **hieraus** *abgeleitete Bedeutung meint die Entdecktheit des Seienden.*
>
> Originally, truth implies as much as opening-up as comportment of the Dasein. The meaning derived **from which** is the discoveredness of the entity. (Heidegger 1927)

If any epistemic modality in the usual linguistic sense is associated with such uses, it would seem to be the necessity of shared knowledge (epistemic state) common to author and reader. Since anaphor resolution in general is often taken to be subject to such constraints (Stalnaker 1974, 2002), no additional lexical characterization of *hier-* adverbials themselves as carriers of epistemic modality need be implied. Such uses may nonetheless be said to be lexicalized (and therefore "senses") to the extent that they impose additional semantic constraints on their antecedents and/or arguments. If on the other hand such ideational usage is taken to be strictly metaphorical (Herrmann 2013), no lexicalization is required and what I have called "epistemic senses" above are simply "epistemic readings."

academic texts than their non-local ("distant") counterparts *that* and *those*: the former study argues that the local variants allow more precise and efficient anaphor resolution by limiting the number of available antecedents, and the latter identifies a strong preference for metaphorical (ideational) uses of local demonstratives in academic text.

I speculate that the shift towards epistemic readings of *hier-* adverbials may have arisen in conjunction with (or perhaps even in response to) their heavier use in academic prose. Use of local propositional deixis in argumentative prose may itself have been further motivated by an active rhetorical strategy on the part of the authors: by using *local* deictic adverbials, the propositional referents – presumably the arguments' premises – were positioned conceptually "closer" to the reader, implicitly encouraging him or her to accept their validity and thus that of the argument as a whole. On the other hand, the use of non-local deixis in literary contexts may in fact serve to support the reader's deictic shift toward the (fictional) narrative index by drawing his or her attention away from the (real but suspended) deictic center (Galbraith 1995).

5. Conclusion

The formal model for diachronic collocation profiling with query-dependent epoch granularity and attribute collation as implemented in the open-source software tool DiaCollo was introduced and some advantages with respect to conventional collocation discovery software were discussed. In its top-level incarnation as a modular web service plugin, DiaCollo provides a simple and intuitive interface for assisting linguists, lexicographers, and humanities researchers to acquire a clearer picture of variation in a word's usage over time and/or corpus subset. The software's capabilities for detecting genre-sensitive phenomena were demonstrated in terms of two example case studies contrasting the behavior of selected items in the genres "science" and "belles lettres" in diachronic corpora of German. Future work will focus on implementation of additional association score functions and a runtime script interpreter, as well as development of a cross-product profile model and associated visualizations suitable for local collocation network analysis. Publicly accessible DiaCollo web-service instances exist for a number of corpora[29] hosted by the DWDS project at institutionthe Berlin-Brandenburg Academy of Sciences, and the DiaCollo source code itself is available via CPAN.[30]

[29] http://kaskade.dwds.de/~jurish/diacollo2017/corpora

[30] http://metacpan.org/release/DiaColloDB

References

Baker, Paul, Gabrielatos, Costas, Khosravinik, Majid, Krzyżanowski , Michał, McEnery, Tony & Wodak, Ruth. 2008. A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society* 19(3): 273–306. https://doi.org/10.1177/0957926508088962

Berry, Michael W., Dumais, Susan T. & O'Brien, Gavin. 1995. Using linear algebra for intelligent information retrieval. *SIAM Review* 37(4): 573–595. <http://www.jstor.org/stable/2132906. https://doi.org/10.1137/1037127

Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan & Finegan, Edward. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Blei, David M., Ng, Andrew Y. & Jordan, Michael I. 2003. Latent Dirichlet allocation. *Journal of machine Learning Research* 3: 993–1022. <http://www.jmlr.org/papers/volume3/blei03a/.pdf>

Brezina, Vaclav, McEnery, Tony & Wattam, Stephen. 2015. Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics* 20(2): 139–173. https://doi.org/10.1075/ijcl.20.2.01bre

Church, Kenneth W. & Hanks, Patrick. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1):22–29.

Davies, Mark. 2012. Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora* 7(2): 121–157. <http://davies-linguistics.byu.edu/ling450/davies_corpora_2011.pdf. https://doi.org/10.3366/cor.2012.0024

Didakowski, Jörg & Geyken, Alexander. 2003. From DWDS corpora to a German word profile – methodological problems and solutions. In *Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information* [OPAL X], Andrea Abel & Lothar Lemnitzer (eds). Mannheim: IDS. <http://www.dwds.de/static/website/publications/pdf/didakowski_geyken_internetlexikografie_2012_final.pdf>

Duff, Iain S., Grimes, Roger G. & Lewis, John G. 1989. Sparse matrix test problems. *ACM Transactions on Mathematical Software (TOMS)*, 15(1): 1–14. https://doi.org/10.1145/62038.62043

Evert, Stefan. 2005. The Statistics of Word Cooccurrences: Word Pairs and Collocations. PhD dissertation, University of Stuttgart. <http://elib.uni-stuttgart.de/opus/volltexte/2005/2371/>

Evert, Stefan. 2008. Corpora and collocations. In *Corpus Linguistics. An International Handbook*, Anke Lüdeling & Merja Kytö (eds), 1212–1248. Berlin: Mouton de Gruyter.

Fielding, Roy T. 2000. Architectural styles and the design of network-based software architectures. PhD dissertation, University of California, Irvine. <https://www.ics.uci.edu/~fielding/pubs/dissertation/fielding_dissertation.pdf>

Firth, John Rupert. 1957. *Papers in Linguistics 1934–1951*. London: OUP.

Gabrielatos, Costas, McEnery, Tony, Diggle, Peter J. & Baker, Paul. 2012. The peaks and troughs of corpus-based contextual analysis. *International Journal of Corpus Linguistics* 17(2): 151–175. https://doi.org/10.1075/ijcl.17.2.01gab

Galbraith, Mary. 1995. Deictic shift theory and the poetics of involvement in narrative. In *Deixis in Narrative: A Cognitive Science Perspective*, Judith F. Duchan, Gail A. Bruder & Lynne E. Hewitt (eds), 19–59. Hillsdale NJ: Lawrence Erlbaum Associates.

Geyken, Alexander. 2013. Wege zu einem historischen Referenzkorpus des Deutschen: Das Projekt Deutsches Textarchiv. In *Perspektiven einer corpusbasierten historischen Linguistik und Philologie* [Thesaurus Linguae Aegyptiae 4], Ingelore Hafemann (eds), 221–234.

Berlin: Berlin-Brandenburgische Akademie der Wissenschaften. <http://nbn-resolving.de/urn:nbn:de:kobv:b4-opus-24424>

Geyken, Alexander, Barbaresi, Adrien, Didakowski, Jörg, Jurish, Bryan, Wiegand, Frank & Lemnitzer, Lothar. 2017. Die Korpusplattform des "Digitalen Wörterbuchs der deutschen Sprache" (dwds). *Zeitschrift für Germanistische Linguistik* 45(2): 327–344. https://doi.org/10.1515/zgl-2017-0017

Glazebrook, Karl & Economou, Frossie. 1997. PDL: The Perl data language. *Dr. Dobb's Journal*, September 1997. <http://www.drdobbs.com/pdl-the-perl-data-language/184410442>

Gries, Stephan Th. & Hilpert, Martin. 2008. The identification of stages in diachronic data: Variability-based neighbor clustering. *Corpora* 3(1): 59–81. <http://members.unine.ch/martin.hilpert/VNCC.pdf. https://doi.org/10.3366/E1749503208000075

Gulordava, Kristina & Baroni, Marco. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, Edinburgh, UK, July 2011, 67–71. Stroudsburg PA: ACL. <http://www.aclweb.org/anthology/W11-2508>

Heaps, H. Stanley. 1978. *Information Retrieval: Computational and Theoretical Aspects*. Orlando FL: Academic Press.

Heidegger, Martin. 1927. Sein und Zeit. In *Jahrbuch für Philosophie und phänomenologische Forschung*, Edmund Husserl (ed.). Tübingen: Neomarius.

Herrmann, J. Bernike. 2013. *Metaphor in Academic Discourse* [LOT Dissertation Series]. Utrecht: Netherlands Graduate School of Linguistics.

Jurish, Bryan. 2015. DiaCollo: On the trail of diachronic collocations. In *CLARIN Annual Conference 2015*, Wrocław, Poland, October 14–16 2015, 28–31. <http://www.clarin.eu/sites/default/files/book%20of%20abstracts%202015.pdf>

Jurish, Bryan, Thomas, Christian & Wiegand, Frank. 2014. Querying the deutsches Textarchiv. In *Proceedings of the Workshop "Beyond Single-Shot Text Queries: Bridging the Gap(s) between Research Communities" (MindTheGap 2014), Berlin, Germany, March 2014*, Udo Kruschwitz, Frank Hopfgartner & Cathal Gurrin (eds), 25–30. <http://ceur-ws.org/Vol-1131/mindthegap14_7.pdf>

Jurish, Bryan, Geyken, Alexander & Werneke, Thomas. 2016. DiaCollo: Diachronen Kollokationen auf der Spur. In *Proceedings DHd 2016: Modellierung – Vernetzung – Visualisierung, University of Leipzig*, March 2016, 172–175. <http://dhd2016.de/boa.pdf#page=172>

Kilgarriff, Adam & Tugwell, David. 2002. Sketching words. In *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*, Marie-Hélène Corréard (ed.), 125–137. <http://www.kilgarriff.co.uk/Publications/2002-KilgTugwell-AtkinsFest.pdf>

Kilgarriff, Adam, Herman, Andrej, Busta, Jan, Rychlý, Pavel & Jakubíček, Milos. 2015. DIACRAN: A framework for diachronic analysis. In *Proceedings of Corpus Linguistics 2015*, Federica Formato & Andrew Hardie (eds), 65–70. Lancaster: UCREL.

Kim, Yoon, Chiu, Yi-K, Hanaki, Kentaro, Hegde, Darshan & Petrov, Slav. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, June 2014, 61–65. Stroudsburg PA: ACL. <http://www.aclweb.org/anthology/W14-2517. https://doi.org/10.3115/v1/W14-2517

Manning, Christopher D. & Schütze, Hinrich. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge MA: The MIT Press.

Mikolov, Tomas, Chen, Kai, Corrado, Greg & Dean, Jeffrey. 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. <https://arxiv.org/abs/1301.3781>

Moretti, Franco. 2013. *Distant Reading*. London: Verso Books.

Rychlý, Pavel. 2008. A lexicographer-friendly association score. In Proceedings of *Recent Advances in Slavonic Natural Language Processing*, RASLAN 2008, 6–9. <http://www.fi.muni.cz/usr/sojka/download/raslan2008/13.pdf>

Sagi, Eyal, Kaufmann, Stefan & Clark, Brady. 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the EACL 2009 Workshop on Geometrical Models of Natural Language Semantics*, March 2009. Stroudsburg PA: ACL. <http://www.aclweb.org/anthology/W09-0214>

Scharloth, Joachim, Eugster, David & Bubenhofer, Noah. 2013. Das Wuchern der Rhizome. Linguistische Diskursanalyse und Data-driven Turn. In *Linguistische Diskursanalyse. Neue Perspektiven*, Dietrich Busse & Wolfgang Teubert (eds), 345–380. Wiesbaden: VS Verlag. https://doi.org/10.1007/978-3-531-18910-9_11

Schiller, Anne, Teufel, Simone & Thielen, Christine. 1995. Guidelines fur das Tagging deutscher Textcorpora mit STTS. Technical report, University of Stuttgart, Institut für maschinelle Sprachverarbeitung and University of Tübingen, Seminar für Sprachwissenschaft.

Sokirko, A. 2003. A technical overview of DWDS/Dialing Concordance. Talk delivered at the meeting *Computational Linguistics and Intellectual Technologies*, Protvino, Russia. <http://www.aot.ru/docs/OverviewOfConcordance.htm>

Stalnaker, Robert C. 1974. Pragmatic presuppositions. In *Semantics and Philosophy*, Milton K. Munitz & Peter K. Unger (eds), 197–213. New York NY: New York University Press.

Stalnaker, Robert C. 2002. Common ground. *Linguistics and Philosophy* 25(5): 701–721. https://doi.org/10.1023/A:1020867916902

Wang, Xuerui & McCallum, Andrew. 2006. Topics over time: A non-Markov continuous-time model of topical trends. In Proceedings of the *12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '06*, *New York*, 424–433. ACM. https://doi.org/10.1145/1150402.1150450

Classical and modern Arabic corpora

Genre and language change

Eric Atwell
The University of Leeds

Our Artificial Intelligence research group at the University of Leeds has collected, analysed and annotated Classical Arabic corpus resources: the *Quranic Arabic Corpus* with several layers of linguistic annotation; the QurAna Quran pronoun anaphoric co-reference corpus; the QurSim Quran verse similarity corpus; the Qurany Quran corpus annotated with English translations and verse topics; the *Boundary-Annotated Quran Corpus*; the *Quran Question and Answer Corpus*; the *Multilingual Hadith Corpus*; the *King Saud University Corpus of Classical Arabic*; and the Corpus for teaching about Islam. We have also developed Modern Arabic corpus resources spanning several genres and language types: *Arabic By Computer*; the *Corpus of Contemporary Arabic*; the *Arabic Internet Corpus*; the *World Wide Arabic Corpus*; the *Arabic Discourse Treebank*; the *Arabic Learner Corpus*; the *Arabic Children's Corpus*; and the *Arabic Dialect Text Corpus*. These corpus resources have informed Arabic corpus linguistics and Artificial Intelligence research, and development of Arabic text analytics tools.

1. Classical Arabic corpora for religious education and understanding

The University of Leeds is unique in bringing together an Artificial Intelligence (AI) research group in its School of Computing with research interests in Arabic text analytics and corpus linguistics, and a department of Arabic Islamic and Middle Eastern Studies (AIMES) with expert Arabic linguists who advise and collaborate with Artificial Intelligence researchers. Corpus linguistics researchers in the Artificial Intelligence research group of the School of Computing at the University of Leeds have collected, analysed and annotated a wide range of Arabic corpus resources, which illustrate genre and language variation in Arabic (Atwell & Alfaifi 2015; Atwell 2018). This chapter reviews the range of Classical and Modern Arabic Corpus resources we have developed, and the range of applications they have been used for.

The Classical Arabic Quran is required reading for all faithful Muslims, who believe it comprises the teachings of God passed on by an angel to the prophet

https://doi.org/10.1075/scl.85.04atw

Mohammad, to be memorised and recited verbatim by all Muslims in unchanged, un-translated original wording. This differentiates the Quran from other religious texts such as the Bible or the Book of Mormon, which are generally read in translated form. Hence the Classical Arabic Quran is probably the single most widely read text ever. Another Classical Arabic text source of major significance to Muslims is the Hadith, the sayings and deeds of the prophet Mohammad, reported by his followers. The Quran and Hadith are the primary corpora of Classical Arabic, although other Classical Arabic texts corpora have also been collated.

To illustrate the range of language and linguistic annotation in Classical Arabic corpora, here are some of the Classical Arabic corpus resources developed by the Artificial Intelligence research group in the School of Computing at the University of Leeds.

1.1 *Quranic Arabic Corpus*

The Quran and other Classical Arabic texts have until recently not attracted much interest among corpus linguists, whose focus is on modern languages and modern language teaching. However, a growing area of Arabic corpus research by computer science and artificial intelligence researchers is in the development of computing tools and resources to aid access to and understanding of key Islamic texts, in particular the Classical Arabic texts of the Quran and Hadith (Atwell et al. 2010). Muslims believe the Quran is the message from God passed on by the angel Gabriel to Mohammad, to teach others to recite. The Quran is divided into chapters and verses, with key themes or tropes running through and linking the text: the roles and attributes of Allah or God; the Day of Judgment when the world ends; stories of previous prophets or religious messengers; and rules and laws for faithful Muslims to obey. Mohammad has a special status in Islam, as the recipient and first reciter of the Quran; and his statements and actions are a secondary source of Islamic knowledge, the Hadith. Hadith are the collection of statements and actions of Mohammad reported by his followers.

One advantage of the Quran as a language data-set is that it has been analyzed, translated, interpreted, annotated and documented by scholars for over a thousand years. Such exegesis or critical explanation and interpretation of the text provides expert knowledge sources for rich corpus linguistic annotation. For example, the Tafsir or Quran exegesis of Ibn Kathir is widely respected and read by Muslims; it provides comments and analysis of each verse, including narrative descriptions of morphological features of words, syntactic dependencies between words, meanings or semantics of key concepts, anaphoric references of pronouns to words and concepts in previous verses, cross-references to other verses with similar meanings, and other linguistic and conceptual insights. Artificial Intelligence researchers have codified this information from the narrative text into formal representations of morphology, grammar, dependency structure, anaphoric co-reference, meanings and ontologies, for use in Natural Language Processing tools for analysis of Arabic text. These formal computational models of Classical Quranic Arabic can then be adapted for analysis of Modern Arabic text.

The Quranic Arabic Corpus (Dukes & Atwell 2012; Dukes et al. 2013), is an online, freely-accessible resource for learning, understanding and research in linguistics, artificial intelligence, and religious studies. There are several other websites offering access to the text of the Quran, in original Arabic and in translations; the *Quranic Arabic Corpus* is unique in also offering several layers of annotation, including part-of-speech tagging and morphological segmentation (Dukes & Habash 2010), syntactic analysis using dependency grammar (Dukes & Buckwalter 2010; Dukes et al. 2010), word-by-word English gloss, several parallel verse-by-verse English translations, audio recordings of recitations, and ontology or index of key Quranic entities and concepts (Dukes & Atwell 2012). There are several Modern Standard Arabic treebanks offering syntactic tags or annotations with each word and sentence, based on modern linguistic theories of grammar (Atwell 2008); the *Quranic Arabic Corpus* treebank is different in that it is built on a deep linguistic model based on the historical traditional grammar known as i'rāb. This traditional description of Quranic Arabic grammatical structure of sentences can engage Quranic Arabic readers as they are more likely to have been exposed to traditional i'rāb than to modern linguistic theories. Furthermore, use of i'rāb allowed us to engage our readers in linguistic annotation of the corpus. The corpus was first annotated with a rule-based tagger program; then online volunteers were invited to collaborate in proofreading the tagging. The *Quranic Arabic Corpus* morphological tagging resulted from "crowdsourcing": about one hundred volunteer annotators proofread sections of the Quran text tags and corrected errors. A small group of expert supervisors reviewed proposed changes made by the crowdsource collaborators; each suggested correction to the computational analysis had to be justified with reference to exegesis. We built a linguistic software platform aimed at collaborative crowdsourcing: LAMP, the Linguistic Analysis Multimodal Platform (Dukes & Atwell 2012).

The *Quranic Arabic Corpus* has been used as a gold standard resource for a range of research on Arabic linguistics, and Arabic Natural Language Processing; for example: Arabic grammatical analysis (Mohammed & Omar 2011; Rabiee 2011), Arabic morphological analysis (Sawalha & Atwell 2009, 2010b, 2011, 2013b; Sawalha et al. 2013; Khaliq & Carroll 2013; Alosaimy & Atwell 2017a, 2017b), Arabic stylometrics (Ali 2012; Alrehaili & Atwell 2013; Alqurneh et al. 2014), coherence analysis in Arabic translation studies (Tabrizi & Mahmud 2013), comparative analysis of the Arabic and English verb systems (Alasmari et al. 2016, 2017), Arabic word stemming (Yusof et al. 2010), Arabic text summarization (El-Haj et al. 2015), Arabic oral-formulaic analysis (Bannister 2014). The *Quranic Arabic Corpus* has achieved greater social impact than typical Corpus Linguistics research projects: the website has attracted millions of visits, including non-Arabic-speakers who want to gain direct insights into the meanings and teachings of the original Classical Arabic text of the Quran through the linguistic annotations.

1.2 QurAna: Quran pronoun anaphoric co-reference corpus

QurAna (Sharaf & Atwell 2012a; Muhammad 2012) is a specialized annotation data-set to accompany the classical *Quranic Arabic Corpus*, showing pronoun co-reference. Each personal pronoun is tagged with its antecedent: the word or phrase it refers to, in the preceding (or occasionally following) text. In addition, ach personal pronoun is also tagged with its "meaning": a link to the the person, entity or concept that pronoun represents, in a separate Quran ontology or knowledge-base of people, entities and concepts. In Classical Arabic, most verbs are morphologically marked with personal pronoun(s) for subject and/or object(s); hence, there is a higher density of personal pronouns in the Classical Arabic text than in English or other translations. The Classical Arabic Quran has nearly 25,000 personal pronouns, and in QurAna each personal pronoun is tagged with antecedent information. QurAna also provides an ontology or term-index of over one thousand persons, entities and concepts, all linked to specific nouns or phrases in the Arabic text which are referred to by the personal pronouns. Deciding on the implicit reference of a personal pronoun in a text is not always straightforward; but for the Quran, we could follow the co-reference analysis in the Tafsir of Ibn Kathir. We could have at least as much confidence in this as in the alternative method widely used in linguistic corpus annotation projects, relying on inter-annotator agreement between two or more casual-worker annotators. The QurAna pronoun anaphoric reference corpus is the first freely downloadable resource of its kind for any type or genre of Arabic. QurAna has been used to guide the analysis and annotation of other Arabic corpora (Zeroual & Lakhouaja 2016; Hammo et al. 2016; Seddik et al. 2015), and in the development of Quran ontologies or knowledge-bases of people, entities and concepts (Alrehaili & Atwell 2014; Hakkoum & Raghay 2015a, 2015b; Alromima et al. 2015; Alqahtani & Atwell 2016; Bentrcia et al. 2017).

1.3 QurSim: Quran verse similarity corpus

QurSim (Sharaf and Atwell 2012b; Muhammad 2012) is another type of corpus research resource based on the Classical Arabic Quran. QurSim shows pairs of Quran verses that are related or similar in meaning, according to the Tafsir or Quranic commentary work of Ibn Kathir. This exegesis examined and commented on each verse of the Quran, and noted links to other verses with related meanings and teachings. We text-mined the Tafsir to extract these cross-references, producing a corpus research resource of over 7,600 pairs of related verses. Users can choose a Quran verse and see verses related to this, which then link on to a network of further related verses. Interestingly, we found that about one third of pairs of related verses shared one or more key words, indicating that "relatedness" of religious text can be partly predicted by lexical matching. However, two thirds of related verse-pairs had no words in common, so predicting semantic relatedness for cases like these is computationally more challenging, requiring artificial intelligence modelling of context and domain knowledge. The QurSim Classical Arabic corpus resource can be used for research on meaning relatedness, similarity and paraphrasing in short texts. Ibn Kathir's commentary was an analysis of the Classical Arabic source text of the Quran, but the verse-relations can also apply to translations: two Quran verses which Ibn Kathir noted as linked in meaning should still be "related" even after translation to another language. Hence, QurSim is a corpus resource for research on textual similarity and relatedness in another language that has a Quran translation (e.g. Basharat et al. 2015). QurSim can also be used for extraction and visualization of topics in the Quran (Panju 2014), and as a component in building further Quran resources such as Quranic Arabic Wordnet linking similar lexical items (AlMaayah et al. 2014) and Quran ontologies or concept-indexes (Alrehaili & Atwell 2014, 2018; Hakkoum & Raghay 2015a, 2015b; Ibrahim et al. 2017; Alqahtani & Atwell 2016, 2018).

1.4 Qurany: Classical Arabic Quran with English translations and verse topics

Qurany (Abbas 2009; Abbas & Atwell 2013) is another corpus research resource based on the Classical Arabic Quran. Qurany was developed to help Quran readers to search for and find verses related to a given concept or concepts. To do this, Qurany encodes the source Arabic text of each verse along with several representations of the meanings or concepts in that verse. Each verse in the Quran is annotated with semantic conceptual category tags, extracted from a respected Quran commentary which includes an index of nearly 1100 concepts or topics with links to the Quran verses, the Mushaf Al Tajweed index. This index shows the main concepts or topics in the Quran, along with the verses each concept appears in. The index was encoded in a Python ontology or knowledge representation formalism. Qurany can be accessed via a web-browser, so that users can navigate the ontology as a tree of main concept-tags, sub-concept-tags, sub-sub-concept-tags etc. Having chosen a specific fine-grained concept-tag, the user can then follow the link to a list of verses tagged with this concept-tag. The concept-tags are available in original Arabic and also in English translation. Each Arabic verse is also annotated with 8 alternative English translations from popular published sources. A verse can be found via Arabic or English keyword-search if any of the original Arabic or English translations contain the keyword(s). Also, the user can opt to see synonyms of keywords, derived from the WordNet synonym-set knowledge-base, to broaden the search-terms. This leads to improved recall: Qurany can show the user more of the verses which are relevant or semantically related to their query. The Qurany dataset is also searchable via standard Google search (or Yahoo, or Bing, or other web-search systems), as it is also online in a single website consisting of a large set of separate web-pages, one per Quran verse. Each verse-webpage displays Arabic source text, 8 English translation texts, and lists of concept-labels or semantic tags relating to the verse, written in both Arabic and English. This website is compatible with standard web search engines. For example, a Google search for "wine" with the site: parameter set to this version of the Qurany website will match all Quran verses containing the word "wine" in at least one of the English translations or concept-labels; and this in turn allows you to see a range of alternative English translations for these verses, along with the Arabic source text, so you can see different translations or interpretation for the word or concept of "wine". In this way, Qurany is useful for showing the range of possible translations and sometimes metaphors and euphemisms for a concept such as "wine" (Gehrels 2016) or "fornication" (Wood 2016). The Qurany resource has been used in research in digital religious studies (Clivaz 2013), Quranic Arabic word meanings or lexical semantics (Al-khalifa et al. 2010), terminology extraction (Mukhtar et al. 2012), formalized knowledge extraction from the Quran (Sharaf & Atwell 2009; Saad et al. 2011, 2013; Muhammad 2012; Atwell 2011; Abed 2015; Ouda 2015; Almaayah et al. 2016; Alrehaili & Atwell 2016, 2017; Bentrcia et al. 2017), Quran recitation methods (Mahmoud & Hassan 2013), combining and merging formal Quran knowledge representations and ontologies (Atwell et al. 2011; Abbas et al. 2013; Alqassem 2013; Dukes et al. 2013; Brierley et al. 2013; Ahmad et al. 2017; ; Alrehaili et al. 2018; Alqahtani & Atwell 2018), and knowledge-based systems for question answering about the Quran (Baqai et al. 2009; Chelli 2012; Abdelhamid et al. 2013; Jilani 2013; Shmeisania et al. 2014; Mohamed et al. 2015; Hakkoum & Raghay 2015a, 2015b; Bakari et al. 2015; Alqahtani & Atwell 2016; Hassan & Atwell 2016a; Kadir & Yauri 2017; Alqahtani & Atwell 2017).

1.5 *Boundary-Annotated Quran Corpus*

Our *Boundary-Annotated Quran Corpus* is another type of corpus research resource based on the Classical Arabic Quran. The number of words and sentences in Arabic text depends on precisely how word-boundaries and sentence-boundaries are defined and counted. For the *Boundary-Annotated Quran Corpus*, we developed a precise computational definition and implementation of sentence and word boundaries, to arrive at a dataset of 77430 words and 8230 sentences of the Classical Arabic text of the Quran. In addition, each word is tagged with phonetic, prosodic and syntactic annotations (Brierley et al. 2012a, 2012b, 2016). The *Boundary-Annotated Quran Corpus* has been used for research in Arabic prosody modeling and visualization (Brierley et al. 2012c, 2014; Sawalha et al. 2012a), Arabic phrase break prediction (Sawalha et al. 2012b, 2012c; Sawalha & Atwell 2012), Arabic speech-to-text transcription (Brierley et al. 2016; Sawalha et al. 2014a, 2014b, 2017).

1.6 *Quran Question and Answer Corpus*

Our *Quran Question and Answer Corpus* is another type of corpus research resource based on the Classical Arabic Quran, but this time extending the Quran text to include questions about the Quran, with answers that include one or more verses from the Quran. In effect, each verse is "annotated" with one or more questions about the verse, and explanatory text linking the question(s) to the verse. Question-Answering systems and chatbots have been developed for a variety of domains (Abu Shawar & Atwell 2007, 2015, 2016). Answering questions about Quran teachings is somewhat different from QA in most other domains, in that the answer is usually expected to be a verse from the Quran, or at least to be based on interpretation of a Quran verse. So, we could model answering a question about the Quran as finding the "best-match" verse for the given input (Abu Shawar & Atwell 2004, 2009). A more general approach is to collect a corpus of attested, reputable answers to questions as a knowledge-base of question-answer pairs, and then the response to a given new question involves finding the "best match" question in the corpus and presenting the corresponding answer to the user (Abu Shawar & Atwell 2005, 2005b). This has led us to collect Quran question-answer corpus collections from Quran scholars (Hamdelsayed & Atwell 2016a, 2016b, 2017) and Islamic websites with Quran Frequently Asked Questions (FAQs) with answers devised by Islamic experts (Hamoud & Atwell 2016a, 2016b, 2016c). These can be used to train a Quran question-answering system (Hamdelsayed & Atwell 2016a, 2016b, 2017).

1.7 *Multilingual Hadith Corpus*

The Quran is believed by Islamic scholars to be the text transcript of messages sent from Allah, and the primary exemplar of Classical Arabic; so the Quran is the focus of both religious and linguistic research. The Hadith are not direct "words of God" but statements about the deeds and saying of Mohammed, and the second most widely used source of Classical Arabic text (Alosaimy & Atwell 2018a). There are many sources of Hadith with varying credentials, depending on credibility of the claimed chain of narrators who passed on the Hadith verbally, at least initially. Because Hadith are not claimed to be the literal words of God, unlike the Quran, it is not so imperative that they are read and understood in original Classical Arabic. We collated a *Multilingual Hadith Corpus*, including parallel texts in Classical Arabic, English, French and Russian (Altoum & Atwell 2016; Hassan & Atwell 2016a, 2016b, 2016c). In information retrieval experiments, we found that search with Arabic keywords in the Arabic original sub-corpus gave slightly higher accuracy results that equivalent searches in the English, French and Russian equivalents. We are adding linguistic annotations to the Arabic Hadith text (Alosaimy & Atwell 2018a, 2018b).

1.8 KSUCCA: *King Saud University Corpus of Classical Arabic*

The Quran is an exemplar of Classical Arabic, and can be used to extract and study examples of Classical Arabic lexis and grammar, using an Arabic-friendly concordance program such as aConCorde or SketchEngine (Roberts et al. 2005, 2006; Wiechmann & Fuhs 2006; Alfaifi & Atwell 2014a, 2016; Kilgarriff et al. 2014a). However, linguists and lexicographers generally see the need for much larger corpora to study less frequent linguistic phenomena such as lexical co-occurrence patterns, collocations and concordance patterns. For research on multi-word units, collocation patterns, rare words, etc., we need a sufficiently large set of examples of each pattern to be studied; but many words and phrases in the Quran occur only once or a handful of times. The Quran contains about seventy thousand words, depending on how word-boundaries are defined and counted; whereas the *British National Corpus*, the first very large corpus developed for British English dictionary research, is much larger, about 100 million words.

So we collaborated with researchers at King Saud University to collect a larger sample of Classical Arabic for lexical pattern research. The 50-million word corpus contains the Quran and some Haddith, and also a range of Classical Arabic texts from around the same period as the Quran was first recited and shortly after. These are predominantly religious texts related to Islam, such as commentaries on the Quran, and biographies of early Islamic scholars. *The King Saud University Corpus of Classical Arabic* (Alrabiah et al. 2013, 2014a, 2014b) allows us to select a word from the Quran and then find many more examples of its use in context in the broader sample of Classical Arabic. The KSUCCA corpus is downloadable from the KSUCCA website, and also searchable online via SketchEngine (Kilgarriff et al. 2014a). KSUCCA has been used for corpus-based study of Arabic historical linguistics (Alrabiah et al. 2014a), and study of distributional lexical patterns around words from the Quran (Alrabiah et al. 2014b).

1.9 Corpus for teaching about Islam

We also used the Web to collect a specialized 'corpus' of texts for university-level teaching about Islam (Atwell et al. 2011), for a project on a Web-as-Corpus approach to populating Wikiversity for teaching about Islam and Muslims in language, linguistics and area studies. The language of this corpus is in fact mainly English, but it contains extracts and references from the core Classical Arabic texts.

2. Modern Arabic corpora for language teaching, lexicography, and text analytics

Modern text corpora have been developed for a wide range of applications (Atwell 1999). The earliest Arabic corpora were developed for modern Arabic language teaching and translation studies to provide representative example texts for teaching and translating modern Arabic. For example, they included magazine articles and similar sources suitable for classroom teaching examples and exercises. A related use of modern Arabic corpora is for dictionary development, for learners and translators; lexicographers exploited modern Arabic corpora to extract concordance examples of lexical items in context to inform the writing of dictionary word and sense definitions and translations. Interesting dictionary items can occur infrequently in a corpus, so lexicographers required much larger corpora.

Artificial Intelligence research, applying Machine Learning to corpus data to build Natural Language Processing models and tools, is a very different use of corpora, and also requires large Arabic corpora. For lexicography and Artificial Intelligence Machine Learning research, size matters more than genre balance, so researchers tended to harvest the most readily available large-scale sources of Arabic text: online news, web-pages, and more recently, internet social media such as Twitter and Facebook.

To illustrate the range of genres and language in Arabic corpora, here are some of the modern Arabic corpus resources we have developed by Artificial Intelligence researchers in the School of Computing at the University of Leeds.

2.1 ABC: *Arabic By Computer*

The first Arabic corpus resource project at Leeds University was ABC (*Arabic By Computer*); we built an Arabic text database and glossary system for Arabic language students (Brockett et al. 1989). In the 1980s and 1990s, UK universities could not afford to provide computers for language teaching, as computers and Knowledge Based Systems (Atwell 1993a,b) were expensive resources which attracted funding only for science and engineering research. We saw a future demand for free open access to Artificial Intelligence and Corpus Linguistics resources for language teaching and research (Atwell 1999, 2005). A major practical hurdle was that computer interfaces at the time only allowed input and output of a very restricted set of characters: Roman alphabet letters A to Z, digits 0 to 9 and a few mathematical symbols. Capture, editing and display of Arabic text required specialist Apple Macintosh hardware and software, including rudimentary Arabic word processing software. Our focus on these technical challenges left less time for linguistic and pedagogical issues such as planning and analysis of text types and genres to be included to match Arabic language syllabuses. The ABC corpus contained a small selection of Arabic magazine articles typed into the Macintosh to provide computer-readable and searchable example texts for Arabic teaching and learning. ABC was a very small example corpus, a taster of things to come.

2.2 CCA: *Corpus of Contemporary Arabic*

Corpus Linguistics research initially concentrated on corpora and tools for English linguistics and language teaching; we wanted to extend corpus linguistics methods and resources to Arabic linguistics and Arabic language teaching. This required an Arabic corpus, so we developed the first freely downloadable million-word *Corpus of Contemporary Arabic* (CCA) (Al-Sulaiti & Atwell 2005, 2006). The million-word LOB and Brown corpora of Modern British English and American English published texts (Atwell 1982; Leech et al. 1983a, 1983b; Johansson et al. 1986) were widely used in English corpus linguistics and English teaching, so we wanted to develop a comparable collection of Arabic published texts. However, the range and balance of text genres in LOB and Brown were decided by Brown University linguistics researchers in the 1970s, and this might not suit the needs of contemporary Arabic researcher and teachers. So before collecting texts, we examined the range of genres covered in other corpora and undertook a survey of prospective Arabic corpus users in Arabic natural language processing and Arabic language teaching to identify user needs and preferences for text-types and genres to be included in our *Corpus of Contemporary Arabic*. This informed our selection of contemporary online sources of text samples: a range of online magazines, websites, newspapers, radio broadcast transcripts, and emails.

The CCA has been widely used by researchers in Arabic language education, Arabic translation, Arabic natural language processing and Arabic corpus linguists; for example in teaching contemporary Arabic (Al-Sulaiti et al. 2005, 2007), Arabic lexical profiling (Attia et al. 2011), the translation of culturally bound metaphors (Merakchi & Rogers 2013; Affeich 2011), learning Arabic spelling and vocabulary (Erradi et al. 2012), lexical differences in world affairs and sports sections in Arabic newspapers (Abdul Razak 2011), corpus-based sociolinguistics (Friginal & Hardy 2014), sentiment analysis of Arabic text (Itani et al. 2017), automated Arabic document classification and clustering (Saad & Ashour 2010; Froud et al. 2010; Alruily 2012; Aly & Kelleny 2014), automated Arabic text summarisation (Froud et al. 2013; Al-Saleh & Menai 2016), improving security and capacity for Arabic text steganography (Al-Haidari et al. 2009), developing and evaluating concordancers for Arabic text (Atwell et al. 2004; Roberts et al. 2006; Alfaifi & Atwell 2014, 2016), Arabic Part-of-Speech tagging (El-Hadj et al. 2009; Sawalha & Atwell 2010a), comparative evaluation of Arabic language morphological analysers and stemmers (Sawalha & Atwell 2008, 2013c), Arabic word sense disambiguation (Zouaghi et al. 2011), and development of Arabic speech recognition systems (Abushariah et al. 2012). The methodology for design and collection of the *Corpus of Contemporary Arabic* was used as a model for corpus development research for other languages and dialects, including Persian (Bijankhan et al. 2011), Kazakh (Makhambetov et al. 2013), Palestinian (Jarrar et al. 2017), Malay (Romli et al. 2016), and Igbo (Onyenwe 2017).

2.3 *Arabic Internet Corpus*

To collect the *Corpus of Contemporary Arabic* text samples, we visited websites and selected sample texts to download "by hand", a time-consuming process; but at least we did not have to type in the text, as was the case for earlier Brown and LOB corpus projects. Collecting a million-word corpus "manually" is labour-intensive and expensive.

The *British National Corpus* (BNC), one hundred million words of British English, was a large-scale research effort by a consortium of several British universities and companies, and became an established gold standard for English corpus linguistics research in the 1990s. In practice, Corpus Linguistics and Artificial Intelligence researchers working on other languages could not get funding to manually collate a large general corpus of the size of BNC. The Web-as-Corpus approach (Baroni & Bernardini 2004; Atwell 2018) offers a more practicable alternative. This approach automates the process of corpus collection from websites: a web-bot visits web-pages and "scrapes" the text. Researchers at Leeds University have used this Web-as-Corpus method to collect Internet corpora for English and many other languages including French, German, Italian, Spanish, Polish, Russian, and Arabic (Sharoff 2006). We can harness these Internet corpora for linguistic research via a web-page with a concordance and collocation search interface. This collection of web-corpora includes the 176-million-word *Arabic Internet Corpus*, which we subsequently lemmatized using the SALMA morphological analysis toolkit (Sawalha & Atwell 2013a). Such a large corpus can be used to examine examples of uses of lexical items, collocations and formulaic sequences or multi-word expressions which occur too infrequently in the smaller *Corpus of Contemporary Arabic* to give sufficient examples (Kilgarriff et al. 2014b; Alghamdi et al. 2016, 2017). Deep Learning researchers use a large corpus to learn categories of words in distributional semantics word-embeddings: capturing the semantic "tag" or category of a word as a word-embedding vector of distributional weights (Atwell 1987a; Atwell & Drakos 1987; Hughes & Atwell 1994). The web-bot collects all texts that match the search terms: a list of about 100 "seed-terms" or common Arabic words, used by the search-engine to find documents containing the seed-terms. Because there is no direct human control over the types of text to be included, a Web-as-Corpus can contain the wide variety of genres found on the World Wide Web. Identification and classification of genre of web-pages is a challenge for Artificial Intelligence research.

2.4 *World Wide Arabic Corpus*

The Web-as-Corpus web-bot can automatically download web-pages that match the "seed-term" list of search terms, and keep collecting matching web-pages until we have reached a target corpus size. We can specify additional constraints on the search, such as limiting matches to a given web address or domain; for example, limiting matches to URLs ending in .SD limits the corpus to web-pages from Sudan. This allows us to run the web-bot repeatedly, changing the national URL constraint each time, to collect a balanced corpus with equal-sized samples from different Arab countries, representing different national dialects of Arabic. We collected a *World Wide Arabic Corpus*, analogous to the *World Wide English Corpus* (Atwell et al. 2007), comprising 200,000-word national sub-corpora, to capture country-by-country linguistic and dialect variation in written Arabic. This has been used to study dialect variation in written Arabic text, for example Arabic dialect variation in connectives (Hassan et al. 2010, 2013) and variation in Arabic and Arab English in the Arab world (Atwell et al. 2008, 2009).

2.5 *Arabic Discourse Treebank*

As outlined above, we added a range of linguistic and semantic annotations to the Quran, to make it a richer corpus resource for research in corpus linguistics and artificial intelligence. Modern Arabic corpus texts can also be annotated with linguistic tagging for research. The *Leeds Arabic Discourse Treebank* (Al-Saif & Markert 2010) is a corpus of 537 news texts where 5651 discourse connectives are tagged with discourse relation type, using a custom discourse annotation program. The argument phrases or clauses they connect are also tagged.

2.6 *Arabic Learner Corpus*

Initial learner corpus research focused on learning English, eg (Herron et al. 1999, Menzel et al. 2000; Atwell et al. 2000). The *Arabic Learner Corpus* (ALC) (Alfaifi & Atwell 2013a, 2013b; Alfaifi et al. 2013) is the first large collection of texts by learners of Arabic in Saudi Arabia. The *Arabic Learner Corpus* includes 282,732 words, in 1585 short student essays or reports, with an average text sample length of 178 words. The texts are student essays or reports on one of two topics: "A vacation trip" (narrative) and "My study interest" (discussion). 942 students from 67 nationalities and 66 different first languages produced the texts. The *Arabic Learner Corpus* also includes rich metadata information, in both English and Arabic, which enables researchers to identify key characteristics of a text and its producer. Each text is stored in a separate file, and key characteristics of the text are also encoded in text sample filenames; e.g. S038_T2_M_Pre_NNAS_W_C shows student identifier S038, text number T2, author gender Male, level of study Pre, nativeness NNAS, text mode Written, and place of text production C. Learner errors are tagged (Atwell 1987b) using an error tag-set adapted to Arabic learners (Alfaifi & Atwell 2012, 2014b, 2015; Alfaifi et al. 2013).

The original hand-written sheets are also downloadable as scanned PDF files. A small portion of the learner texts were spoken by the learner and then transcribed; the original 3.5 hours of MP3 audio recordings are also available to download. The ALC is downloadable and/or searchable from several websites, including SketchEngine (Kilgarriff et al. 2014a; Alfaifi & Atwell 2014a). The corpus has been used for research in Arabic native language identification (Malmasi & Dras 2014, 2015), critical discourse analysis (Haider 2016), and automatic text correction (Mohit et al. 2014; Zaghouani et al. 2015).

2.7 *Arabic Children's Corpus*

The *Arabic Children's Corpus* is a collection of texts written for children (Al-Sulaiti et al. 2016); note this is NOT texts written BY children. The *Arabic Children's Corpus* contains 2950 documents and nearly 2 million words, collected by manually searching the web for suitable texts over a 3-month period. It enabled us to measure variation in Arabic language, and in particular vocabulary, in writing aimed at children compared to an Adult readership, which is represented by most other Arabic corpora.

2.8 *Arabic Dialect Text Corpus*

Most Modern Arabic texts are in Modern Standard Arabic, an international standard written form taught in schools and used in formal writing across the Arab world. Spoken Arabic can vary significantly in language and vocabulary from Modern Standard Arabic; and there is no one standard spoken form, but a wide variety of dialects. Arabic dialect studies have tended to focus on differences in phonetics and phonology, in the variant pronunciations. Speakers of Arabic dialects (including all Arabic native speakers) generally write in Modern Standard Arabic, in effect "translating" from dialect to MSA. There are few Arabic dialect written texts, or standard writing conventions to capture dialect. However, for text analytics research and applications, it is important to be able to document and handle variation in Arabic dialect vocabulary, morphology and grammar. One way to capture written dialect texts is to record dialect speakers and use speech recognition software for an automatic speech-to-text transcription; this was done for the VarDial'2016 contest, to train and test Machine Learning models for Arabic dialect classification (Alshutayri et al. 2016). Another approach is to harvest online sources of spontaneous informal Arabic: online forums and social media where users are likely to write as they speak and not strictly follow Modern Standard Arabic conventions (El-Beltagy & Ali 2013). We harvested Twitter tweets, using location to identify dialect, in an *Arabic Twitter Dialect Text Corpus* (Alshutayri & Atwell 2017); and then extended this by also harvesting online newspaper reader comments and FaceBook comments (Alshutayri & Atwell 2018a, 2018b).

3. Machine learning from the Quran for Modern Arabic text analytics

On the face of it, there is a mismatch in genre and language between the Quran, a collection of religious chapters and verses, and Modern Arabic corpus sources such as online news, web-pages, and Twitter. So how can we equitably compare Classical Arabic to Modern Arabic? And how can we make use of Classical Arabic corpora to aid Natural Language Processing research targeted at Modern Arabic users?

Much current applied NLP research on Modern Arabic is focussed on short text snippets, such as analysis of Twitter tweets, Facebook comments, or Amazon customer reviews, for example (Ahmed et al. 2017). To train supervised Machine Learning models of language processing, we need language data annotated with appropriate target linguistic analyses (Atwell 1996). Despite some vocabulary differences, the Quran verses are also short text snippets; but with the added bonus of semantic and linguistic annotations added to the text snippets derived from centuries of expert study. Linguistically annotated Quran verses provide a rich training set for supervised Machine Learning of language models. For example, the *Quranic Arabic Corpus* morphological annotations were used to train Machine Learning models for Modern Arabic morphological analysis (Khaliq &

Carroll 2013); the QurAna anaphoric coreference annotations were used to train Machine Learning models for Modern Arabic anaphora resolution (Seddik et al. 2015); the *Boundary-Annotated Quran Corpus* phonetic and prosodic annotations have been used to train Machine Learning models for Modern Arabic speech-to-text transcription (Brierley et al. 2016; Sawalha et al. 2017). Despite the differences in genre and language variety, the Classical Arabic text of the Quran can inform Artificial Intelligence and Corpus Linguistics research on Modern Arabic.

References

Abbas, Noorhan Hassan. 2009. Quran Ssearch for a concept tool and website. MRes thesis, University of Leeds, UK.

Abbas, Noorhan Hassan & Atwell, Eric. 2013. Annotating the Arabic Quran with a classical semantic ontology. Proceedings of *WACL2 Second Workshop on Arabic Corpus Linguistics*. Lancaster, UK.

Abbas, Noorhan Hassan, Aldhubayi, Luluh, Al-Khalifa, Hend, Alqassem, Zainab, Atwell, Eric, Dukes, Kais, Sawalha, Majdi & Sharaf, Muhammad. 2013. Unifying linguistic annotations and ontologies for the Arabic Quran. Proceedings of *WACL2 Second Workshop on Arabic Corpus Linguistics*. Lancaster, UK.

Abdelhamid, Yasser, Mahmoud, Mostafa & El-Sakka, Tarek M. 2013. Using ontology for associating Web multimedia resources with the Holy Quran. Proceedings of *Advances in Information Technology for the Holy Quran and Its Sciences*. Medina, Saudi Arabia.

Abdul Razak, Zainur. 2011. Modern media Arabic: A study of word frequency in world affairs and sports sections in Arabic newspapers. PhD thesis, University of Birmingham, UK.

Abed, Qusay Abdullah. 2015. Ontology-based Approach for Retrieving Knowledge in Al-Quran. PhD thesis, Universiti Utara Malaysia.

Abushariah, Mohammad, Ainon, Raja, Zainuddin, Roziati, Elshafei, Moustafa & Khalifa, Othman Omran. 2012. Arabic speaker-independent continuous automatic speech recognition based on a phonetically rich and balanced speech corpus. *International Arab Journal of Information Technology* 9(1): 84–93.

Abu Shawar, Bayan & Atwell, Eric. 2004. An Arabic chatbot giving answers from the Quran. Proceedings of *TALN'2004 Traitement Automatique des Langues Naturelles*. Fez, Morocco.

Abu Shawar, Bayan & Atwell, Eric. 2005a. Using corpora in machine-learning chatbot systems. *International Journal of Corpus Linguistics* 10: 489–516. https://doi.org/10.1075/ijcl.10.4.06sha

Abu Shawar, Bayan & Atwell, Eric. 2005b. A chatbot system as a tool to animate a corpus. *ICAME Journal: International Computer Archive of Modern and Medieval English Journal* 29: 5–24.

Abu Shawar, Bayan & Atwell, Eric. 2007. Chatbots: Sind Sie wirklich nützlich? [Chatbots: Are they really useful?]. *Journal for Computational Linguistics and Language Technology* 22: 31–50.

Abu Shawar, Bayan & Atwell, Eric. 2009. Arabic question-answering via instance based learning from an FAQ corpus. Proceedings of *CL2009 Corpus Linguistics*. Liverpool, UK.

Abu Shawar, Bayan & Atwell, Eric. 2015. ALICE chatbot: Trials and outputs. *Computacion y Sistemas* 19(4): 625–632.

Abu Shawar, Bayan & Atwell, Eric. 2016. Usefulness, localizability, humanness, and language-benefit: Additional evaluation criteria for natural language dialogue systems. *International Journal of Speech Technology* 19(2): 373–383. https://doi.org/10.1007/s10772-015-9330-4

Affeich, Andree. 2011. La métaphore dans le discours technique d'Internet et son passage de l'anglais vers l'arabe. Proceedings of *JéTou'2011 Journées d'études Toulousaines*. Toulouse, France.

Ahmad, Nor Diana, Bennett, Brandon & Atwell, Eric. 2017. Retrieval performance for Malay Quran. *International Journal on Islamic Applications in Computer Science and Technology* 5(2): 13–25.

Ahmed, Saad, Hina, Saman, Atwell, Eric & Ahmed, Farrakh. 2017. Aspect based sentiment analysis framework using data from social media network. *International Journal of Computer Science and Network Security* 17(7): 100–105

Alasmari, Jawharah, Watson, Janet C. E. & Atwell, Eric. 2016. A comparative analysis of the Arabic and English verb systems using a Quranic Arabic Corpus. Proceedings of *IMAN'2016 Islamic Applications in Computer Science and Technologies*. Khartoum, Sudan.

Alasmari, Jawharah, Watson, Janet C. E. & Atwell, Eric. 2017. Using the Quranic Arabic Corpus for comparative analysis of the Arabic and English verb systems. *International Journal on Islamic Applications in Computer Science And Technology* 5(3): 1–8.

Alfaifi, Abdullah & Atwell, Eric. 2012. Arabic Learner Corpora (ALC): A taxonomy of coding errors. Proceedings of *ICCA'2012 International Computing Conference in Arabic*. Cairo, Egypt.

Alfaifi, Abdullah & Atwell, Eric. 2013a. Arabic Learner Corpus v1: A new resource for Arabic language research. Proceedings of *WACL'2 Second Workshop on Arabic Corpus Linguistics*. Lancaster, UK.

Alfaifi, Abdullah & Atwell, Eric. 2013b. Arabic Learner Corpus: Texts transcription and files format. Proceedings of *CORPORA'2013 International Conference on Corpus Linguistics*. St Petersburg, Russia.

Alfaifi, Abdullah, Atwell, Eric & Abuhakema, Ghazi. 2013. Error Annotation of the Arabic Learner Corpus: A new error tagset. Proceedings of *GSCL'2013 German Society for Computational Linguistics: Language Processing and Knowledge in the Web*. Darmstadt, Germany.

Alfaifi, Abdullah & Atwell, Eric. 2014a. Tools for searching and analysing Arabic corpora: An evaluation study. *Proceedings BAAL-CUP'2014 British Association for Applied Linguistics and Cambridge University Press Applied Linguistics Conference*. Leeds, UK.

Alfaifi, Abdullah & Atwell, Eric. 2014b. An evaluation of the Arabic error tagset v2. Proceedings of *AACL'2014 American Association for Corpus Linguistics*. Flagstaff, USA.

Alfaifi, Abdullah & Atwell, Eric. 2015. Computer-aided error annotation: A new tool for annotating Arabic error. Proceedings of *UK Saudi Students Conference*. London, UK.

Alfaifi, Abdullah & Atwell, Eric. 2016. Comparative evaluation of tools for Arabic corpora search and analysis. *International Journal of Speech Technology* 19(2): 347–357. https://doi.org/10.1007/s10772-015-9285-5

Alghamdi, Ayman, Atwell, Eric & Brierley, Claire. 2016. An empirical study of Arabic formulaic sequence extraction methods. Proceedings of *LREC'2016 Language Resources and Evaluation Conference*. Portorož, Slovenia.

Alghamdi, Ayman & Atwell, Eric. 2017. Towards comprehensive computational representations of Arabic multi-word expressions. Proceedings of *EUROPHRAS'2017 European Conference on Computational and Corpus-Based Phraseology*. London, UK.

Al-Haidari, Fahd, Gutub, Adnan, Al-Kahsah, Khalid & Hamodi, Jameel. 2009. Improving security and capacity for Arabic text steganography using Kashida extensions. Proceedings of *CSA'2009 Computer Systems and Applications*. Jeju, Korea.

Ali, Imran. 2012. Application of a mining algorithm to finding frequent patterns in a text corpus: A case study of Arabic. *International Journal of Software Engineering and Its Applications* 6(3): 127–134.

Al-Khalifa, Hend, Al-Yahya, Maha, Bahanshal, Alia, Al-Odah, Iman & Al-Helwah, Nawal. 2010. An approach to compare two ontological models for representing Quranic words. Proceedings of the *12th International Conference on Information Integration and Web-based Applications and Services*. Paris, France.

Almaayah, Manal, Sawalha, Mohammad A. & Abushariah, Majdi. 2014. A proposed model for Quranic Arabic WordNet. Proceedings of *LRE-REL2 2nd workshop on Language Resources and Evaluation for Religious Text*s. Reykjavik, Iceland.

Almaayah, Manal, Sawalha, Mohammad A. & Abushariah, Majdi. 2016. Towards an automatic extraction of synonyms for Quranic Arabic WordNet. *International Journal of Speech Technology* 19(2): 177–189. https://doi.org/10.1007/s10772-015-9301-9

Alosaimy, Abdulrahman & Atwell, Eric. 2017a. Joint alignment of segmentation and labelling for Arabic morphosyntactic taggers. *International Journal of Computational Linguistics* 8(2): 45–58.

Alosaimy, Abdulrahman & Atwell, Eric. 2017b. Tagging classical Arabic text using available morphological analysers and part of speech taggers. *Journal for Language Technology and Computational Linguistics* 32(1): 1–26.

Alosaimy, Abdulrahman & Atwell, Eric. 2018a. Diacritization of a highly cited text: A classical Arabic book as a case. Proceedings of *ASAR'2018 Arabic Script Analysis and Recognition*. London, UK.

Alosaimy, Abdulrahman & Atwell, Eric. 2018b. Web-based annotation tool for inflectional language resources. Proceedings of *LREC 2018 Language Resources and Evaluation Conference*. Miyazaki, Japan.

Alqahtani, Mohammad & Atwell, Eric. 2016. Arabic Quranic search tool based on ontology. Proceedings of *NLDB'2016 Natural Language and Information Systems*. Salford, UK.

Alqahtani, Mohammad & Atwell, Eric. 2017. Evaluation criteria for computational Quran search. *International Journal on Islamic Applications in Computer Science and Technology* 5(1): 12–22.

Alqahtani, Mohammad & Atwell, Eric. 2018. Developing bilingual Arabic-English ontologies of Al-Quran. Proceedings of *ASAR'2018 Arabic Script Analysis and Recognition*. London, UK.

Alqassem, Zainab. 2013. Unifying Quranic analyses into a single database. BSc Research Project Report, School of Computing, University of Leeds, UK.

Alqurneh, Ahmed, Mustapha, Aida, Murad, Masrah & Sharef, Nurfahdlina. 2014. Stylometric model for detecting oath expressions: A case study for Quranic texts. *Literary and Linguistic Computing Journal* 31(1): 1–20.

Alrabiah, Maha, Al-Salman, AbdulMalik & Atwell, Eric. 2013. The design and construction of the 50 million words KSUCCA King Saud University Corpus of Classical Arabic. Proceedings of *WACL'2 Second Workshop on Arabic Corpus Linguistics*. Lancaster, UK.

Alrabiah, Maha, Al-Salman, AbdulMalik, Atwell, Eric & Alhelewh, Nawal. 2014a. KSUCCA: A key to exploring Arabic historical linguistics. *International Journal of Computational Linguistics* 5: 27–36.

Alrabiah, Maha, Alhelewh, Nawal, Al-Salman, AbdulMalik & Atwell, Eric. 2014b. An empirical study on the holy Quran based on a large classical Arabic corpus. *International Journal of Computational Linguistics* 5: 1–13.

Alrehaili, Sameer & Atwell, Eric. 2013. Linguistics features to confirm the chronological order of the Quran. Proceedings of *WACL'2 Second Workshop on Arabic Corpus Linguistics*. Lancaster, UK.

Alrehaili, Sameer & Atwell, Eric. 2014. Computational ontologies for semantic tagging of the Quran. Proceedings of *LRE-Rel2 2nd Workshop on Language Resource and Evaluation for Religious Texts*. Reykjavik, Iceland.

Alrehaili, Sameer & Atwell, Eric. 2016. A hybrid-based term extraction method on the Arabic text of the Quran. Proceedings of *IMAN'2016 Islamic Applications in Computer Science and Technologies*. Khartoum, Sudan.

Alrehaili, Sameer & Atwell, Eric. 2017. Extraction of multi-word terms and complex terms from the classical Arabic text of the Quran. *International Journal on Islamic Applications in Computer Science and Technology* 5(3): 15–27.

Alrehaili, Sameer & Atwell, Eric. 2018. Discovering Qur'anic knowledge through AQD: Arabic Qur'anic Database, a multiple resources annotation-level search. Proceedings of *ASAR'2018 Arabic Script Analysis and Recognition*. London, UK.

Alrehaili, Sameer, Alqahtani, Mohamad & Atwell, Eric. 2018. A hybrid method of aligning Arabic Qur'anic semantic resources. Proceedings of *ASAR'2018 Arabic Script Analysis and Recognition*. London, UK.

Alromima, Waseem, Elgohary, Rania, Moawad, Ibrahim F. & Aref, Mostafa. 2015. Applying ontological engineering approach for Arabic Quran corpus: A comprehensive survey. Proceedings of *ICICIS'2015 International Conference on Intelligent Computing and Information Systems*. Cairo, Egypt.

Alruily, Meshrif. 2012. Using text mining to identify crime patterns from Arabic Crime News Report Corpus. PhD dissertation, De Montford University, UK.

Al-Saif, Amal & Markert, Katja. 2010. The Leeds Arabic Discourse Treebank: Annotating discourse connectives for Arabic. Proceedings of *LREC'2010: Language Resources and Evaluation Conference*. Valletta, Malta.

Al-Saleh, Asma Bader & Menai, Mohammad El Bachir. 2016. Automatic Arabic text summarization: A survey. *Artificial Intelligence Review* 45(2): 203–234. https://doi.org/10.1007/s10462-015-9442-x

Alshutayri, Areej, Atwell, Eric, Alosaimy, Abdulrahman, Dickins, James, Ingleby, Michael & Watson, Janet. 2016. Arabic language WEKA-based dialect classifier for Arabic automatic speech recognition transcripts. Proceedings of *VarDial'2016 Third Workshop on NLP for Similar Languages, Varieties and Dialects*. Osaka, Japan.

Alshutayri, Areej & Atwell, Eric. 2017. Exploring twitter as a source of an Arabic dialect corpus. *International Journal of Computational Linguistics* 8(2): 37–44.

Alshutayri, Areej & Atwell, Eric. 2018a. Creating an Arabic dialect text corpus by exploring Twitter, Facebook, and online newspapers. Proceedings of *OSACT'2018 Open-Source Arabic Corpora and Processing Tools*. Miyazaki, Japan.

Alshutayri, Areej & Atwell, Eric. 2018b. A social media corpus of Arabic dialect text. In *Computer-Mediated Communication and Social Media Corpora*, Ciara R. Wigham & Egon Stemle (eds). Clermont-Ferrand: Presses Universitaires Blaise Pascal.

Al-Sulaiti, Latifa & Atwell, Eric. 2005. Extending the Corpus of Contemporary Arabic. Proceedings of *CL'2005 Corpus Linguistics*. Birmingham, UK.

Al-Sulaiti, Latifa, Roberts, Andrew & Atwell, Eric. 2005. The use of corpora and concordance in the teaching of contemporary Arabic. Proceedings of *EuroCALL'2005 European conference on Computer Assisted Language Learning*. Krakow, Poland.

Al-Sulaiti, Latifa & Atwell, Eric. 2006. The design of a corpus of contemporary Arabic. *International Journal of Corpus Linguistics* 11: 135–171. https://doi.org/10.1075/ijcl.11.2.02als

Al-Sulaiti, Latifa, Roberts, Andrew, Abu Shawar, Bayan & Atwell, Eric. 2007. The use of corpus, concordancer and chatbot in the teaching of contemporary Arabic. Proceedings of *CL'2007 Corpus Linguistics*. Birmingham, UK.

Al-Sulaiti, Latifa, Abbas, Noorhan, Brierley, Claire, Atwell, Eric & Alghamdi, Ayman. 2016. Compilation of an Arabic Children's Corpus. Proceedings of *LREC'2016 Language Resources and Evaluation Conference*. Portorož, Slovenia.

Altoum, S. & Atwell, Eric. 2016. Compilation of an Islamic Hadith Corpus (تجمع مدونة الحديث النبوي الشريف). Proceedings of *ICCA'2016 International Conference on Computing in Arabic*. Khartoum, Sudan.

Aly, Walid Mohamed & Kelleny, Hany Atef. 2014. Adaptation of cuckoo search for documents clustering. *International Journal of Computer Applications* 86(1): 4–10. https://doi.org/10.5120/14947-3041

Attia, Mohammed, Pecina, Pavel, Tounsi, Lamia, Toral, Antonio & Van Genabith, Josef. 2011. Lexical profiling for Arabic. Proceedings of *eLex'2011 Electronic Lexicography in the 21st Century*. Bled, Slovenia.

Atwell, Eric. 1982. *LOB Corpus Tagging Project: Manual Postedit Handbook*. University of Lancaster, UK.

Atwell, Eric. 1987a. A parsing expert system which learns from corpus analysis. In *Corpus. Linguistics and Beyond: Proceedings of the ICAME 7th International Conference on English Language. Research on Computerised Corpora*, Willem Meijs (ed.), 227–235. Amsterdam: Rodopi,

Atwell, Eric. 1987b. How to detect grammatical errors in a text without parsing it. Proceedings of *EACL'1987 Third Conference of the European Chapter of the Association for Computational Linguistics*. Copenhagen, Denmark.

Atwell, Eric & Drakos, Nicos. 1987. Pattern recognition applied to the acquisition of a classification system from unrestricted English text. Proceedings of *EACL'1987 Third Conference of the European Chapter of the Association for Computational Linguistics*. Copenhagen, Denmark.

Atwell, Eric. 1993a. The HEFC's knowledge based systems initiative. *Artificial Intelligence and Simulation of Behaviour Quarterly* 83: 29–34.

Atwell, Eric (ed.). 1993b. *Knowledge at Work in Universities - Proceedings of the second annual conference of the Higher Education Funding Council's Knowledge Based Systems Initiative*. Leeds: Leeds University Press.

Atwell, Eric. 1996. Machine learning from corpus resources for speech and handwriting recognition. In *Using Corpora for Language Research: Studies in Honour of Geoffrey Leech*, Jenny Thomas & Mick Short (eds), 151–166. London: Longman.

Atwell, Eric. 1999. *The Language Machine*. London: The British Council.

Atwell, Eric, Howarth, Peter, Souter, Clive, Baldo, Patrizio, Bisiani, Roberto, Bonaventura, Patrizia, Menzel, Wolfgang, Herron, Daniel, Morton, Rachel & Wick, Juergen. 2000. User-guided system development in Interactive Spoken Language Education. *Natural Language Engineering Journal* 6(3-4): 229–241. https://doi.org/10.1017/S1351324900002473

Atwell, Eric, Al-Sulaiti, Latifa, Al-Osaimi, Saleh & Abu Shawar, Bayan. 2004. Un examen d'outils pour l'analyse de corpus arabes: A review of Arabic corpus analysis tools. Proceedings of *TALN'2004 Traitement Automatique des Langues Naturelles*. Fez, Morocco.

Atwell, Eric. 2005. Sleeping with the enemy: Infiltrating AI into the broader curriculum. Proceedings of *1st UK Workshop on Artificial Intelligence in Education*. Cambridge, UK.

Atwell, Eric, Arshad, Junaid, Lai, Chien-Ming, Nim, Lan, Rezapour Asheghi, Noushin, Wang, Josiah & Washtell, Justin. 2007. Which English dominates the World Wide Web, British or American? Proceedings of *CL'2007 Corpus Linguistics*. Birmingham, UK.

Atwell, Eric. 2008. Development of tag sets for part-of-speech tagging. In *Corpus Linguistics: An International Handbook*, Anke Lüdeling & Merja Kytö (eds), 501–526. Berlin: Mouton de Gruyter.

Atwell, Eric, Abbas, Noorhan, Abu Shawar, Bayan, Alsaif, Amal, Al-Sulaiti, Latifa, Roberts, Andrew & Sawalha, Majdi. 2008. Mapping Middle Eastern and North African diasporas. Proceedings of *BRISMES'2008 British Society for Middle Eastern Studies*. Leeds, UK.

Atwell, Eric, Al-Sulaiti, Latifa & Sharoff, Serge. 2009. Arabic and Arab English in the Arab world. Proceedings of *CL2009 Corpus Linguistics*. Liverpool, UK.

Atwell, Eric, Dukes, Kais, Sharaf, Abdul Baquee, Habash, Nizar, Louw, Bill, Abu Shawar, Bayan, McEnery, Tony, Zaghouani, Wajdi & El-Haj, Mahmoud. 2010. Understanding the Quran: A new Grand Challenge for Computer Science and Artificial Intelligence. Proceedings of *GCCR'2010 Grand Challenges in Computing Research*. Edinburgh, Scotland, UK.

Atwell, Eric. 2011. Exploiting new technology and innovation for detecting terrorist activities. *Counter Terror Expo*. London, UK.

Atwell, Eric, Brierley, Claire, Dukes, Kais, Sawalha, Majdi & Sharaf, Abdul Baquee. 2011. An artificial intelligence approach to Arabic and Islamic content on the Internet. *Proceedings of NITS'2011 National Information Technology Symposium*. Riyadh, Saudi Arabia.

Atwell, Eric & Alfaifi, Abdullah. 2015. أبحاث جامعة ليدز في مجال لسانيات المدونات العربية (Arabic corpus linguistics research at the University of Leeds). In *Arabic Language and Computing*, Ysi Elarian (ed.). Riyadh: King Abdullah bin Abdulaziz International Center for Arabic Language Service.

Atwell, Eric. 2018. Using the Web to model Modern and Quranic Arabic. In *Arabic Corpus Linguistics*, Tony McEnery, Adrew Hardie & Younis Nagwa Ibrahim Abdel-Fattah (eds). Edinburgh: Edinburgh University Press.

Bakari, Wided, Bellot, Patrice & Neji, Mahmoud. 2015. Literature review of Arabic question-answering: Modeling, generation, experimentation and performance analysis. Proceedings of *FQAS'2015 Flexible Query Answering Systems*. Krakow, Poland.

Bannister, Andrew G. 2014. *An Oral-Formulaic Study of the Quran*. Lanham MD: Lexington Books.

Baqai, Sumayya, Basharat, Amna, Khalid, Hira, Hassan, Amna & Zafar, Shehneela. 2009. Leveraging semantic web technologies for standardized knowledge modeling and retrieval from the Holy Qur'an and religious texts. Proceedings of the *7th International Conference on Frontiers of Information Technology*. Abbottabad, Pakistan.

Baroni, Maco & Bernardini, Silvia. 2004. BootCaT: Bootstrapping corpora and terms from the web. Proceedings of *LREC'2004 Language Resources and Evaluation Conference*. Lisbon, Portugal.

Basharat, Asma, Yasdansepas, D. & Rasheed, Khaled. 2015. Comparative study of verse similarity for multi-lingual representations of the Quran. Proceedings of *ICAI'2015 International Conference on Artificial Intelligence*. Las Vegas, USA.

Bentrcia, Rahima, Zidat, Samir & Marir, Farhi. 2017. Extracting semantic relations from the Quranic Arabic based on Arabic conjunctive patterns. *Journal of King Saud University Computer and Information Sciences*.

Bijankhan, Mahmood, Sheykhzadegan, Javad, Bahrani, Mohammad & Ghayoomi, Masood. 2011. Lessons from building a Persian written corpus: Peykare. *Language Resources and Evaluation Journal* 45(2): 143–164. https://doi.org/10.1007/s10579-010-9132-x

Brierley, Claire, Sawalha, Majdi & Atwell, Eric. 2012a. Open-source boundary-annotated corpus for Arabic speech and language processing. Proceedings of *LREC'2012 Language Resources and Evaluation Conference*. Istanbul, Turkey.

Brierley, Claire, Sawalha, Majdi & Atwell, Eric. 2012b. Boundary Annotated Quran Corpus for Arabic phrase break prediction. Proceedings of *IVACS'2012 Inter-Varietal Applied Corpus Studies*. Cambridge, UK.

Brierley, Claire, Sawalha, Majdi & Atwell, Eric. 2012c. Visualisation of prosody in English and Arabic speech corpora. Proceedings of *AVML'2012 Advances in Visual Methods for Linguistics*. York, UK.

Brierley, Claire, Atwell, Eric, Rowland, Chris & Anderson, John. 2013. Semantic pathways: A novel visualization of varieties of English. *ICAME Journal of the International Computer Archive of Modern and medieval English* 37: 5–36.

Brierley, Claire, Sawalha, Majdi & Atwell, Eric. 2014. Tools for Arabic Natural Language Processing: A case study in qalqalah prosody. Proceedings of *LREC'2014 Language Resources and Evaluation Conference*. Reykjavik, Iceland.

Brierley, Claire, Sawalha, Majdi, Heselwood, Barry & Atwell, Eric. 2016. A verified Arabic-IPA mapping for Arabic transcription technology, informed by Quranic recitation, traditional Arabic linguistics, and modern phonetics. *Journal of Semitic Studies* 61(1): 157–186. https://doi.org/10.1093/jss/fgv035

Brockett, Adrian, Atwell, Eric, Taylor, Owen & Page, Matthew. 1989. An Arabic text database and glossary system for students. Proceedings of the *Seminar on Bilingual Computing in Arabic and English*. Cambridge, UK.

Chelli, Assem. 2012. Advanced Search/Indexing in Holy Quran. Magister Thesis, National Higher School of Computer Science, Algeria.

Clivaz, Claire. 2013. Digital religion out of the book: The loss of the illusion of the 'original text' and the notion of a 'religion of a book'. *Scripta Journal* 25: 26–41.

Dukes, Kais & Habash, Nizar. 2010. Morphological annotation of Quranic Arabic. Proceedings of *LREC'2010 Language Resources and Evaluation Conference*. Valletta, Malta.

Dukes, Kais & Buckwalter, Tim. 2010. A dependency treebank of the Quran using traditional Arabic grammar. Proceedings of *INFOS'2010 7th Informatics and Systems conference*. Cairo, Egypt.

Dukes, Kais, Atwell, Eric & Sharaf, Abdul Baquee. 2010. Syntactic annotation guidelines for the Quranic Arabic dependency treebank. Proceedings of *LREC'2010 Language Resources and Evaluation Conference*. Valletta, Malta.

Dukes, Kais & Atwell, Eric. 2012. LAMP: A multimodal web platform for collaborative linguistic analysis. Proceedings of *LREC'2012 Language Resources and Evaluation Conference*. Istanbul, Turkey.

Dukes, K., Atwell, Eric & Habash, Nizar. 2013. Supervised collaboration for syntactic annotation of Quranic Arabic. *Language Resources and Evaluation Journal* 47: 33–62. https://doi.org/10.1007/s10579-011-9167-7

El-Beltagy, Samhaa & Ali, Ahmed. 2013. Open issues in the sentiment analysis of Arabic social media: A case study. Proceedings of *IIT'2013 Innovations in Information Technology Conference*. Abu Dhabi, United Arab Emirates.

El-Haj, Mahmoud, Kruschwitz, Udo & Fox, Chris. 2015. Creating language resources for under-resourced languages: Methodologies, and experiments with Arabic. *Language Resources and Evaluation Journal* 49(3): 549–580. https://doi.org/10.1007/s10579-014-9274-3

El Hadj, Yahja Ould Mohamed, Al-Sughayeir, Imad Abdulrahman & Al-Ansari, Abdullah Mahdi. 2009. Arabic part-of-speech tagging using the sentence structure. Proceedings of the *Second International Conference on Arabic Language Resources and Tools*. Cairo, Egypt.

Erradi, Abdelkarim, Nahia, Sajeda, Almerekhi, Hind & Al-Kailani, Lubna. 2012. ArabicTutor: A multimedia m-learning platform for learning Arabic spelling and vocabulary. Proceedings of *ICMCS'2012 International Conference on Multimedia Computing and Systems*. Tangier, Morocco.

Friginal, Eric & Hardy, Jack A. 2014. *Corpus-based Sociolinguistics: A Guide for Students*. London: Routledge.

Froud, Hahane, Benslimane, R., Lachkar, Abdelmonaime & Ouatik, Said Alaoui. 2010. Stemming and similarity measures for Arabic documents clustering. *Proceedings of ISVC'2010 5th International Symposium on I/V Communications*. Rabat, Morocco.

Froud, Hahane, Lachkar, Abdelmonaime & Ouatik, Said Alaoui. 2013. Arabic text summarization based on latent semantic analysis to enhance Arabic documents clustering. *International Journal of Data Mining and Knowledge Management Process* 3(1): 79–95. https://doi.org/10.5121/ijdkp.2013.3107

Gehrels, Sjoerd. 2016. Liquid hospitality: Wine as the metaphor. In *The Routledge Handbook of Hospitality Studies*, Conrad Lashley (ed.), 247–259. London: Routledge.

Haider, Ahmad S. 2016. A Corpus-assisted Critical Discourse Analysis of the Arab Uprisings: Evidence from the Libyan Case. PhD dissertation, University of Canterbury, New Zealand.

Hakkoum, Aimad & Raghay, Said. 2015a. Ontological approach for semantic modeling and querying the Quran. *International Journal on Islamic Applications in Computer Science And Technology* 4(1): 37–45.

Hakkoum, Aimad & Raghay, Said. 2015b. Advanced search in the Quran using semantic modeling. Proceedings of *AICCSA'2015 Arab International Conference on Computer Systems and Applications*. Marrakech, Morocco.

Hamdelsayed, Mohamed Adany & Atwell, Eric. 2016a. Islamic applications of automatic question-answering. *Journal of Engineering and Computer Science* 17(2): 51–57.

Hamdelsayed, Mohamed Adany & Atwell, Eric. 2016b. Using Arabic numbers (singular, dual, and plurals) patterns to enhance question answering system results. Proceedings of *IMAN'2016 Islamic Applications in Computer Science and Technologies*. Khartoum, Sudan.

Hamdelsayed, Mohamed Adany & Atwell, Eric. 2017. Quran question answering system using Arabic number patterns (singular, dual, plural). *International Journal on Islamic Applications in Computer Science and Technology* 5(2): 1–12.

Hammo, Bassam, Yagi, Sane, Ismail, Omaima & Abushariah, Mohammad. 2016. Exploring and exploiting a historical corpus for Arabic. *Language Resources and Evaluation Journal* 50(4): 839–861. https://doi.org/10.1007/s10579-015-9304-9

Hamoud, Bothaina & Atwell, Eric. 2016a. Using an islamic question and answer knowledge base to answer questions about the Holy Quran. *International Journal on Islamic Applications in Computer Science And Technology* 4 (4): 20–29.

Hamoud, Bothaina & Atwell, Eric. 2016b. Quran question and answer corpus for data mining with WEKA. Proceedings of *IEEE Conference of Basic Sciences and Engineering Studies*. Khartoum, Sudan.

Hamoud, Bothaina & Atwell, Eric. 2016c. Compiling a Quran Question and Answer Corpus. تجميع مدونة اسئلة واجوبة للقرآن الكر. Proceedings of *ICCA'2016 International Conference on Computing in Arabic*. Khartoum, Sudan.

Hassan, Haslina, Daud, Nuraihan Mat & Atwell, Eric. 2010. Connectives in the World Wide Arabic corpus. Proceedings of *IVACS'2010 Inter-Varietal Applied Corpus Studies*. Leeds, UK.

Hassan, Haslina, Daud, Nuraihan Mat & Atwell, Eric. 2013. Connectives in the World Wide Web Arabic corpus. *World Applied Sciences Journal (Special Issue of Studies in Language Teaching and Learning)* 21: 67–72.

Hassan, Samah & Atwell, Eric. 2016a. Concept search tool for multilingual Hadith corpus. *International Journal of Science and Research* 5(4): 1326–1328.

Hassan, Samah & Atwell, Eric. 2016b. Design requirements for multilingual Hadith corpus. *International Journal of Science and Research* 5(4): 494–498.

Hassan, Samah & Atwell, Eric. 2016c. Design and implementing of multilingual Hadith corpus. *International Journal of Recent Research in Social Sciences and Humanities* 3(2): 100–104.

Herron, Daniel, Menzel, Wolfgang, Atwell, Eric, Bisiani, Roberto, Daneluzzi, Fabio, Morton, Rachel & Schmidt, Juergen A. 1999. Automatic localization and diagnosis of pronunciation errors for second-language learners of English. Proceedings of *EUROSPEECH'1999 Sixth European Conference on Speech Communication and Technology*. Budapest, Hungary.

Hughes, John & Atwell, Eric. 1994. The automated evaluation of inferred word classifications. Proceedings of *ECAI-1994 11th European Conference on Artificial Intelligence*. Amsterdam, The Netherlands.

Ibrahim, Eiman, Ataelfadiel, Mohammed & Atwell, Eric. 2017. Provisions of Quran Tajweed ontology. *International Journal of Science and Research* 6(8): 756–761.

Itani, Maher, Roast, Chris & Al-Khayatt, Samir. 2017. Corpora for sentiment analysis of Arabic text in social media. Proceedings of *ICICS'2017 IEEE International Conference on Information and Communication Systems*. Irbid, Jordan.

Jarrar, Mustafa, Habash, Nizar, Alrimawi, Faeq, Akra, Diyam & Zalmout, Nasser. 2017. Curras: an annotated corpus for the Palestinian Arabic dialect. *Language Resources and Evaluation Journal* 51(3): 745–775. https://doi.org/10.1007/s10579-016-9370-7

Jilani, Aisha. 2013. Parallel Corpus Multi Stream Question Answering with Applications to the Quran. PhD dissertation, University of Huddersfield, UK.

Johansson, Stig, Atwell, Eric, Garside, Roger & Leech, Geoffrey. 1986. *The Tagged LOB Corpus - User Manual*. Bergen: Norwegian Computing Centre for the Humanities.

Kadir, Rabiah A. & Yauri, Aliyu Rufai. 2017. Automated semantic query formulation using machine learning approach. *Journal of Theoretical and Applied Information Technology* 95(12): 2761–2775.

Khaliq, Bilal & Carroll, John. 2013. Induction of root and pattern lexicon for unsupervised morphological analysis of Arabic. Proceedings of *IJCNLP'2013 International Joint Conference on Natural Language Processing*. Nagoya, Japan.

Kilgarriff, Adam, Baisa, Vit, Bušta, Jan, Jakubíček, Miloš, Kovář, Vojtěch, Michelfeit, Jan, Rychlý, Pavel & Suchomel, Vit. 2014a. The Sketch Engine: Ten years on. *Lexicography Journal* 1(1): 7–36. https://doi.org/10.1007/s40607-014-0009-9

Kilgarriff, Adam, Charalabopoulou, Frieda, Gavrilidou, Maria, Johannessen, Janne Bondi, Khalil, Saussan, Johansson, Sofie, Lew, Robert, Sharoff, Serge, Vadlapudi, Ravikiran & Volodina, Elena. 2014b. Corpus-based vocabulary lists for language learners for nine languages. *Language Resources and Evaluation Journal* 48(1): 121–163. https://doi.org/10.1007/s10579-013-9251-2

Leech, Geoffrey, Garside, Roger & Atwell, Eric. 1983a. Recent developments in the use of computer corpora in English language research. *Transactions of the Philological Society* 1983: 23–40. https://doi.org/10.1111/j.1467-968X.1983.tb01200.x

Leech, Geoffrey, Garside, Roger & Atwell, Eric. 1983b. The automatic grammatical tagging of the LOB Corpus. *ICAME Journal: International Computer Archive of Modern and medieval English Journal* 7: 13–33.

Mahmoud, Mostafa & Hassan, Iman. 2013. Artificial intelligence techniques for extracting individual recitation of the Holy Quran from its combinations. Proceedings of *Advances in Information Technology for the Holy Quran and Its Sciences*. Medina, Saudi Arabia.

Makhambetov, Olzhas, Makazhanov, Aibek, Yessenbayev, Zhandos, Matkarimov, Bakhyt, Sabyrgaliyev, Islam & Sharafudinov, Anuar. 2013. Assembling the Kazakh Language Corpus. Proceedings of *EMNLP'2013 Empirical Methods in Natural Language Processing*. Seattle, USA.

Malmasi, Shervin & Dras, Mark. 2014. Arabic native language identification. Proceedings of *EMNLP 2014 Empirical Methods in Natural Language Processing Workshop on Arabic Natural Language*. Doha, Qatar.

Malmasi, Shervin & Dras, Mark. 2015. Multilingual native language identification. *Natural Language Engineering Journal* 23(2):163–215. https://doi.org/10.1017/S1351324915000406

Merakchi, Khadidja & Rogers, Margaret. 2013. The translation of culturally bound metaphors in the genre of popular science articles: A corpus-based case study from Scientific American translated into Arabic. *Intercultural Pragmatics Journal* 10(2): 341–372.

Menzel, Wolfgang, Atwell, Eric, Bonaventura, Patrizia, Herron, Daniel, Howarth, Peter, Morton, Rachel & Souter, Clive. 2000. The ISLE corpus of non-native spoken English. Proceedings of *LREC'2000 Language Resources and Evaluation Conference*. Athens, Greece.

Mohammed, Mona Ali Mohammed & Omar, Nazlia. 2011. Rule based shallow parser for Arabic language. *Journal of Computer Science* 7(10): 1505–1514. https://doi.org/10.3844/jcssp.2011.1505.1514

Mohamed, Reham, Ragab, Maha, Abdelnasser, Heba, El-Makky, Nagwa & Torki, Marwan. 2015. Al-Bayan: A knowledge-based system for Arabic answer selection. Proceedings of *SemEval'2015 Workshop on Semantic Evaluation*. Denver, USA.

Mohit, Behrang, Rozovskaya, Alla, Habash, Nizar, Zaghouani, Wajdi & Obeid, Ossama. 2014. The first QALB shared task on automatic text correction for Arabic. Proceedings of the *EMNLP 2014 Empirical Methods in Natural Language Processing Workshop on Arabic Natural Language*. Doha, Qatar.

Muhammad, Abdul Baquee. 2012. Annotation of conceptual co-reference and Text Mining the Quran. PhD dissertation, University of Leeds, UK.

Mukhtar, Tayyeba, Afzal, Hammad & Majeed, Awais. 2012. Vocabulary of Quranic concepts: A semi-automatically created terminology of Holy Quran. Proceedings of *INMIC'2012 International Multitopic Conference*. Islamabad, Pakistan.

Onyenwe, Ikechukwu. 2017. Developing methods and resources for automated processing of the African Language Igbo. PhD dissertation, University of Sheffield, UK.

Ouda, Karim. 2015. QuranAnalysis: A semantic search and intelligence system for the Quran. MSc thesis, University of Leeds, UK.

Panju, Maysum H. 2014. Statistical Extraction and visualization of topics in the Quran Corpus. MMath thesis, University of Waterloo, Canada.

Rabiee, Hajder S. 2011. Adapting standard open-source resources to tagging a morphologically rich language: A case study with Arabic. Proceedings of *RANLP'2011 Recent Advances in Natural Language Processing*. Hissar, Bulgaria.

Roberts, Andrew, Al-Sulaiti, Latifa & Atwell, Eric. 2005. aConCorde: towards a proper concordance of Arabic. Proceedings of *CL'2005 Corpus Linguistics*. Birmingham, UK.

Roberts, Andrew, Al-Sulaiti, Latifa & Atwell, Eric. 2006. aConCorde: Towards an open-source, extendable concordancer for Arabic. *Corpora Journal* 1: 39–57. https://doi.org/10.3366/cor.2006.1.1.39

Romli, Taj, Hassan, Abd Rauf & Mohamad, Hasnah. 2016. Equivalent Malay-Arabic data corpus collection. *European Journal of Language and Literature Studies* 4(1): 65–73. https://doi.org/10.26417/ejls.v4i1.p65-73

Saad, Motaz K. & Ashour, Wesam. 2010. Arabic text classification using decision trees. Proceedings of *CSIT'2010 12th international workshop on Computer Science and Information Technologies*. Moscow and Saint-Petersburg, Russia.

Saad, Saidah, Salim, Naomie & Zainuddin, Suhaila. 2011. An early stage of knowledge acquisition based on Quranic text. Proceedings of *STAIR'2011 Semantic Technology and Information Retrieval*. Putrajaya, Malaysia.

Saad, Saidah, Salim, Naomie & Zainal, Hakim. 2013. Rules and natural language pattern in extracting Quranic knowledge. Proceedings of *Advances in Information Technology for the Holy Quran and Its Sciences*. Medina, Saudi Arabia.

Sawalha, Majdi & Atwell, Eric. 2008. Comparative evaluation of Arabic language morphological analysers and stemmers. Proceedings of *COLING'2008 Computational Linguistics*. Manchester, UK.

Sawalha, Majdi & Atwell, Eric. 2009. Linguistically informed and corpus informed morphological analysis of Arabic. Proceedings of *CL'2009 Corpus Linguistics*. Liverpool, UK.

Sawalha, Majdi & Atwell, Eric. 2010a. Fine-grain morphological analyzer and part-of-speech tagger for Arabic text. Proceedings of *LREC'2010 Language Resources and Evaluation Conference*. Valletta, Malta.

Sawalha, Majdi & Atwell, Eric. 2010b. Constructing and using broad-coverage lexical resource for enhancing morphological analysis of Arabic. Proceedings of *LREC'2010 Language Resources and Evaluation Conference*. Valletta, Malta.

Sawalha, Majdi & Atwell, Eric. 2011. Morphological analysis of classical and modern standard Arabic. Proceedings of *ICCA'2011 International Computing Conference in Arabic*. Riyadh, Saudi Arabia.

Sawalha, Majdi & Atwell, Eric. 2012. Visualization of Arabic morphology. Proceedings of *AVML'2012 Advances in Visual Methods for Linguistics*. York, UK.

Sawalha, Majdi, Brierley, Claire & Atwell, Eric. 2012a. Prosody prediction for Arabic via the open-source boundary-annotated Qur'an corpus. *Journal of Speech Sciences* 2: 175–191.

Sawalha, Majdi, Brierley, Claire & Atwell, Eric. 2012b. Automatic analysis of phrase-break prediction for Arabic التحليل الآلي للوقف والابتداء في نصوص اللغة العربية الحديثة والكلاسيكية. Proceedings of ICCA'2012 International Computing Conference in Arabic. Cairo, Egypt.

Sawalha, Majdi, Brierley, Claire & Atwell, Eric. 2012c. Predicting phrase breaks in classical and modern standard Arabic text. Proceedings of *LREC'2012 Language Resources and Evaluation Conference*. Istanbul, Turkey.

Sawalha, Majdi & Atwell, Eric. 2013a. Accelerating the processing of large corpora: using grid computing for lemmatizing the 176 million words Arabic Internet Corpus. Proceedings of *WACL'2 2nd Workshop of Arabic Corpus Linguistics*. Lancaster, UK.

Sawalha, Majdi & Atwell, Eric. 2013b. A standard tag set expounding traditional morphological features for Arabic language part-of-speech tagging. *Word Structure Journal* 6: 43–99. https://doi.org/10.3366/word.2013.0035

Sawalha, Majdi & Atwell, Eric. 2013c. Comparing morphological tag-sets for Arabic and English. Proceedings of *CL'2013 Corpus Linguistics*. Lancaster, UK.

Sawalha, Majdi, Atwell, Eric & Abushariah Mohammad. 2013. SALMA: Standard Arabic Language Morphological Analysis. Proceedings of *ICCSPA'2013 International Conference on Communications Signal Processing and Applications*. Sharjah, United Arab Emirates.

Sawalha, Majdi, Brierley, Claire & Atwell, Eric. 2014a. Automatically generated phonemic Arabic-IPA pronunciation tiers for the boundary annotated Qur'an dataset for machine learning. Proceedings of *LRE-Rel'2 2nd Workshop on Language Resource and Evaluation for Religious Text*. Reykjavik, Iceland.

Sawalha, Majdi, Brierley, Claire, Atwell, Eric & Dickins, James. 2014b. Text analytics and transcription technology. Proceedings of *IMAN'2014 Islamic Applications in Computer Science And Technology*. Amman, Jordan.

Sawalha, Majdi, Brierley, Claire, Atwell, Eric & Dickins, James. 2017. Text analytics and transcription technology for Quranic Arabic. *International Journal on Islamic Applications in Computer Science and Technology* 5 (2): 45–51.

Seddik, Khadiga M., Farghaly, Ali & Fahmy, Aly Aly. 2015. Arabic anaphora resolution: Corpus of the Holy Quran annotated with anaphoric information. *International Journal of Computer Applications* 124(15): 35–43. https://doi.org/10.5120/ijca2015905709

Sharaf, Abdul Baquee & Atwell, Eric. 2009. A corpus-based computational model for knowledge representation of the Quran. Proceedings of *CL'2009 Corpus Linguistics*. Liverpool, UK.

Sharaf, Abdul Baquee & Atwell, Eric. 2012a. QurAna: Corpus of the Quran annotated with pronominal anaphora. Proceedings of *LREC'2012 Language Resources and Evaluation Conference*. Istanbul, Turkey.

Sharaf, Abdul Baquee & Atwell, Eric. 2012b. QurSim: A corpus for evaluation of relatedness in short texts. Proceedings of *LREC'2012 Language Resources and Evaluation Conference*. Istanbul, Turkey.

Sharoff, Serge. 2006. Open-source corpora: Using the net to fish for linguistic data. *International Journal of Corpus Linguistics* 11(4): 435–62. https://doi.org/10.1075/ijcl.11.4.05sha

Shmeisania, Hashem, Tartir, Samir, Al-Nassaan, Ammar & Najid, Moath. 2014. Semantically answering questions from the Holy Quran. Proceedings of *IMAN'2014 Islamic Applications in Computer Science and Technology*. Amman, Jordan.

Tabrizi, Arash Amini & Mahmud, Rohana. 2013. Issues of coherence analysis on English translations of Quran. Proceedings of *ICCSPA'2013 International Conference on Communications Signal Processing and Applications*. Sharjah, United Arab Emirates.

Wiechmann, Daniel & Fuhs, Stefan. 2006. Concordance software. *Corpus Linguistics and Linguistics Theory Journal* 2: 109–130.

Wood, Paul. 2016. The pen and the sword: Reporting ISIS. Discussion paper, Shorenstein Center on Media Politics and Public Policy.

Yusof, Raja, Zainuddin, Roziati, Baba, Mohd & Yusoff, Zulkifi. 2010. Quranic words stemming. *Arabian Journal for Science and Engineering* 35(2): 37–49.

Zaghouani, Wajdi, Zerrouki, Taha & Balla, Amar. 2015. SAHSOH@ QALB shared task: A rule-based correction method of common Arabic native and non-native speakers' errors. Proceedings of *ANLP'2015 Arabic Natural Language Processing Workshop*. Beijing, China.

Zeroual, Imad & Lakhouaja, Abdelhak. 2016. A new Quranic corpus rich in morphosyntactical information. *International Journal of Speech Technology* 19(2): 339–346. https://doi.org/10.1007/s10772-016-9335-7

Zouaghi, Anis, Merhbene, Laroussi & Zrigui, Mounir. 2011. Word sense disambiguation for Arabic language using the variants of the Lesk algorithm. Proceedings of *WORLDCOMP'2011 World Congress in Computer Science, Computer Engineering, and Applied Computing*. Las Vegas, USA.

PART II

Genre and diachronic corpora

Scholastic genre scripts in English medical writing 1375–1800

Irma Taavitsainen
University of Helsinki

Late medieval scientific and medical writing had several different genres and levels of writing from the beginning. Learned genres, including commentaries, were introduced into English with the vernacularization boom. The Latin "genre script" lists ancient authorities' opinions of a topic, finishing with the commentator's own. Writing conventions were adopted with a time lag, and fully-fletched commentaries emerge when the heyday of Scholasticism was already over. Research became increasingly based on observation and new top genres were based on empirical science. This chapter traces generic features derived from Scholasticism with genre dynamics and meaning-making practices at center stage. The material comes from medical corpora with background metadata.

1. Introduction

A radically new axiom some decades ago, but generally accepted now, states that all language use and every text is framed within genres (Bakhtin 1986). A further comment along the same line adds that the history of English (or of any other language) is composed of the histories of its registers and genres (Diller 2001: 3). Medical writing is one of the prestige registers or domains of writing in Western languages, with its history beginning with two different trends. The rise of universities in the twelfth century had strengthened the position of Latin as the institutional language and vernacularization of learned texts started towards the end of the medieval period. Medicine was, however, somewhat different from other disciplines as, in addition to the academic curriculum, it was a practical field with healers without formal training, and medical texts had been written in vernaculars even earlier (Crossgrove 1998: 81–82).

Top scholastic genres were first translated into Romance languages and then into English and German in the wide pan-European vernacularization boom from the late fourteenth century onwards. The institutional function was, however, lost in vernacular translations as universities were monolingual Latin, and the practical side became enhanced (Taavitsainen 2004: 67, and below). The repertoire of

https://doi.org/10.1075/scl.85.05taa

medical genres at the end of the medieval period included several different levels of writing with vernacular recipes, health guides and prognostications transmitted almost in the same form from Old English on (Voigts 1984). These genres were aimed at heterogeneous lay and professional audiences and remain fairly constant, so that even some much later texts seem to be largely based on their medieval predecessors (Taavitsainen 2011). In contrast, the top genres of Scholasticism for learned readers were on the move. During the period in focus here (1375–1800) the styles of scientific thinking changed several times: the modes and bases of knowledge shifted from logocentric Scholasticism with absolute certainty to Empiricism with observation in a tentative mode of knowing, and further to "enquiry as a thought style", with a shift to more statistical thinking and innovations in laboratory medicine that paved the way to more modern approaches (see Taavitsainen & Hiltunen (eds) fc.).

Consequently, the question of what continued and what changed is central in the historical analysis of scientific thought styles (Crombie 1994: 6).[1] Everything does not change but the processes are long and gradual with several styles of thinking existing side by side at a time; this is the variationist view adapted to the macrolevel of genres.[2] My study is based on English, but it is reasonable to assume that the same processes and lines of development apply to other European languages as well.

I shall begin my chapter by explaining my aim and approach; next I shall elaborate on my theoretical point of departure. My research questions focus on generic features derived from Scholasticism, whether they changed and in what way, and how they continue in later writings.

2. Aim

Recent studies have shown that there was a time lag before scholastic conventions were established in English and that the scholastic style of writing continued in use beyond the medieval period (Taavitsainen 2009, 2017). This chapter examines

[1] At an early phase of our project, we devised a chart of changing scientific thought styles with their modes of knowing, source of knowing and reliability. In pseudo-science the mode is belief and reliance high; in Scholasticism the mode is hearsay, the source is language and the level of reliance high; in Empiricism the mode of knowing is induction by observation, with low reliance (Taavitsainen & Pahta 1998: 162).

[2] Several variants are found at one time, competing with one another. Some gain and some lose, and eventually, some variants go out of use, while others continue and win over (see e.g. Milroy 1992).

the developments of the top genres of Scholasticism in the vernacular in a long diachronic perspective, stretching the timeline over four hundred years. By focusing on the core question of continuity and change, I hope to cast some new light on the afterlife of medieval Scholasticism as expressed on the macrolevel of genres in English medical writing over the centuries. I shall use the notion of *genre script* defined as an underlying structural pattern as my tool for tracing genre continuity in a long diachronic perspective (see below). My hypothesis is that genre scripts serve to identify texts, point out affinities and show modifications through time. It should be possible to reveal some aspects of cultural genre dynamics, how the old and the new become intertwined and how they develop. An important insight into the nature of genres is that they need to be considered in relation to their neighbours within their own domains of writing. I shall deal with both spearhead genres of medieval learning, commentaries and compilations, and relate them to their uses.

3. Approach

I shall approach the task from the angle of cognitive historical pragmatics in a sociohistorical frame taking e.g. the sociolinguistic background parameters of authors and audiences into account.[3] The multilayered context includes sociohistorical conditions with their situational constraints, textual practices and societal, political, ideological and material aspects (see Pahta & Taavitsainen 2010a: 551, Taavitsainen & Jucker 2015). Genres can be defined as groupings of texts with a shared function, created to meet the needs of discourse communities whose participants have more or less precise knowledge of what the genre labels imply and what their typical linguistic features are (Swales 1990: 45–58). Genres are best viewed in company with their neighbouring genres (see Fowler 1982 and Taavitsainen 2016). From a more cognitive point of view, they are inherently dynamic cultural schemata, used by discourse communities to organise knowledge and experience through language, and they display variation and undergo change in response to their users' sociocultural needs (Taavitsainen 2001, see also Kuna 2016). And further, the cognitive discourse approach of genres sees them as conceptual and semantic reconstructions of the world where texts function as semantic units that map their phenomena with more or less conventionalized linguistic expressions viable to change over time or across cultures. Alterations can

[3] Similar approaches have been developed for other languages as well, e.g. German *Fachprosa* is discussed with a historical pragmatic and philological slant by Habermann (2014).

be triggered by various developments, such as new communicative goals or novel media that may cause or at least contribute to the formation of even radically new genres like experimental reports in *The Philosophical Transactions*, the first scientific journal in English (1665–).

Cultural scripts have been defined as sets of assumptions that relate to particular situations or activity types in a culture (Spencer-Oatey 2000: 61–62). Genre scripts are related and refer to typical patterns of language use and sequences in macrolevel structures; this approach brings concrete evidence and accuracy to genre change as it enables empirical studies that can point out resemblances between various texts in a new way. Thus the notion of genre script is related to the prototype approach to genres: it can indicate affinities between texts and genres, but there may also be distinguishing features that tell adjoining genres apart (cf. Taavitsainen 1997). The prototype approach to genres builds on family resemblance with blurred edges, overlaps and fuzzy boundaries (see Rosch & Mervis 1975). Texts exhibit genre features to different extents, multiple membership in more than one genre is also possible; and in genre change, everything does not change, but there are stable features that remain, thus providing continuity.

For historical genre analysis we need a broad repertoire of genres so that they can be assessed in relation to one another within the whole domain. We know that in the course of time sociocultural needs change and genres change accordingly: old genres become adapted to new functions, new genres are created, and genres that have lost their function cease to exist (Fowler 1982; Swales 1990). Changes take place at different rates in different genres and different layers of writing, and we need more empirical evidence of various aspects for a more comprehensive picture of generic variation and change. In the field of medical writing the picture is much more diversified than depicted in traditional language histories, and the most important discovery so far is that instead of one line of development, there are several (Taavitsainen 2016).

Individual scripts and their elements are activated during interaction between text participants, writers and readers, and the role of a genre script in the creation and reception of texts and genres is of interest. Scripts enable the creation of meaning as even a few words in combination with the situational and functional factors as well as socio-pragmatic and cultural background items can trigger genre expectations. Traditional openings like "once upon a time" signal the genre of a fairy tale to the audience, and a real master like Chaucer can play with genre reversals from the very beginning of a storyline. In all their complexity, scripts provide a convenient framework for the examination of historical changes.

4. Data

The data for the present study come from the electronic corpora compiled by the Scientific thought styles team at the University of Helsinki. The corpus *Middle English Medical Texts 1375–1500* (MEMT; Taavitsainen, Pahta & Mäkinen 2005) is in this chapter complemented by a Hippocratic text that became available only after the corpus was finished. MEMT comprises about half a million words primarily of editions of medical treatises. The scope is wide, from texts of highest learning to practical health guides and remedy books with recipes written for heterogeneous lay audiences, but only the learned end of the scale was employed for the present study; the corpus contains a wide scale of specialized treatises such as compendia, uroscopies and surgical treatises showing the variety of medieval medical texts in the vernacular. Middle English spellings were standardized with the Vard program (also used for EMEMT, see Lehto et al. 2010) and checked by hand.

Early Modern English Medical Texts 1500–1700 (EMEMT; Taavitsainen et al. 2010) begins where MEMT ends and the finishing point was set according to changes in medical thought, institutional developments and the final breakthrough of the vernacular (see note 11). It provides continuation to both learned and more popular traditions, and introduces the new categories of scientific writing at the end of the period. These corpora are designed to contain representative samplings of medical writing, informed by medical history as the texts were selected in collaboration with medical historians (Pahta & Taavitsainen 2010b: 4–6). EMEMT provides a comprehensive database with background metadata for tracing stylistic developments across the register. The corpus contains about two million words in some 230 texts printed in these two centuries ranging from theoretical treatises rooted in academic traditions to popularized and utilitarian texts. They are grouped into six categories. The following were assessed in detail as they were most likely to contain discourse loci with the commentary form: Category 1 (General treatises and textbooks), and Category 2 (Specialized treatises).[4] Normalized EMEMT text files (included with the corpus files) were used for the searches, but all quotations display original spellings.

[4] Its subcategories are: (2a) specific diseases; (2b) methods of diagnosis and treatment; (2c) therapeutic substances; (2d) midwifery and children's diseases; and (2e) texts on plague. In addition, the corpus contains Cat. 3 recipe collections and *materia medica*; Cat. 4 regimens and health guides; Cat. 5 surgical treatises; and Cat. 6, the *Philosophical Transactions* (PT hereafter, 1665-). Texts labeled as Medicine in society are given in the Appendix.

A part of *Late Modern English Medical Texts* (LMEMT; Taavitsainen et al. forthcoming) was also searched to detect whether scholastic genre scripts continued in the eighteenth-century. The three corpora are designed to provide continuity to one another, and this is the first time that the new LMEMT corpus is tested in a larger scale.[5]

5. Methodology

The characteristics of learned writing in MEMT were established in our earlier corpus linguistics studies with top-down methods taking features established by previous studies as their point of departure (Taavitsainen & Pahta 1998).[6] The method proved successful, and as a preliminary measure similar corpus-linguistic searches were made with WordSmith 6 tools (Scott 2012) in EMEMT to locate texts with high frequencies and similar features for qualitative analysis. First, searches focused on authorities, obtained from MEMT and complemented by EMEMT wordlists. They were searched for both directly via their names (in various spelling forms gleaned from the wordlists) and in more indirect ways, through prepositions (*after, according to*) and through logocentric verbs (*say, tell, command,* etc.). Second, prescriptive impersonal phrases were searched for (*it is to be noted*, *it is to be said*, *it is to wit, it bihovith*, etc.), and deontic verbs such as *must* and *shall*, especially in collocations like *thow shalt* + a mental verb (*hope, doubte, deme* or *dreede*), often in the negated form, too, were assessed. Pronouns of the second-person singular *thow* and imperative forms with cognitive verbs like *vndistonde thow*, *knowe thow* are also typical. In addition, boosters like *forsooth* contribute to the absolute certainty of the scholastic style.[7]

In a recent corpus linguistic study our earlier research questions of the distribution of scholastic language features were revisited from a novel perspective. We used methods of Digital Humanities combining Computer Science with Philology (see Taavitsainen & Schneider forthcoming) and applied a new data-driven method of Document Classification with a program called

[5] We employed the same subcategories of Category 2 (Specialized treatises) of LMEMT, except (2d) midwifery and children's diseases, as it was not ready for use.

[6] Top-down methods are deductive, whereas bottom-up methods rely on what the material yields and are inductive (see Jucker & Taavitsainen 2013: 43).

[7] This line of investigation has inspired other scholars to further studies on evidentiality with detailed results (see Whitt 2016a; Landert 2018 fc).

LightSide.[8] This method was chosen as it has the potential of revealing pertinent features that have gone unnoticed before, and it also gives a ranking order for all features typical of the thought style (see Section 10 below).

As a preliminary step for the study, the data were divided into two binary groups, scholastic and non-scholastic. The selection of texts for the positive group was made according to the earlier corpus-linguistic studies and qualitative philological assessments. The results show that the frequencies of scholastic features vary a great deal in MEMT and EMEMT texts.[9] The timeline for the positive group begins c. 1375 and extends slightly beyond 1600 and contains 272,000 words in all.[10] The non-scholastic group was easier to decide upon: it comes from the end of the seventeenth century and consists of texts that, by definition, belong to the new thought style of Empiricism. Texts from *The Philosophical Transactions* (EMEMT Category 6) were written for the new type of discourse community, the Royal Society, within which Empiricism became the dominant trend. Additionally, two longer texts by spearhead empirical scientists from the same community were included. They are Hooke's *Micrographia* (1664) and Boyle's *Electricity & Magnetism* (1675–6).

6. Commentary scripts in the vernacular

The aim of Scholasticism was to reconstruct original knowledge of ancient authorities, conceptualized as texts rather than persons (Parkes 1976: 116). Furthermore according to the logocentric thought style, language had an objective status of a

[8] <http://ankara.lti.cs.cmu.edu/side/> The program is easy to use and offers a wide range of machine learning algorithms, including logistic regression from the LIBLINEAR library. It also performs cross-validation automatically.

[9] My preliminary studies on pragmatic research questions of meaning change in scholastic style dealt with MEMT and EMEMT data (Taavitsainen 2009, 2017).

[10] I used this selection first for my oral presentations at conferences in Kalamazoo in 2015 and in Nottingham in 2016. The following texts were included: (MEMT) Galen, *De ingenio sanitatis*; Trevisa, *Of the properties of things*; Anon., *Phlebotymy*; Anon., *Book of Surgery*; Benvenutus Grassus (ophthalmology); Henry Daniel (urinoscopy); John of Burgundy (pestilence); Canutus (pestilence); (EMEMT) Anon., *Seyng of uryns* (1526); Brauschweig, *Surgery* (1525); Vigo, *Surgery* (1540); Geminus, *Anatomy* (1553); *BATMAN vppon Bartholome* (1582); Moulton, *Myrrour or glasse of helth, plague (1539)*; Galen's *Book of elements* (1574); Laurentius, *Preservation of sigh*t (1599); Lodge, *Treatise of the plague* (1603); Bullein's *Bullwarke (*1562); Thomas Gale, *Enchiridion of chirurgie (*1563); Thomas Gale, *Institution of a chirurgien (*1563); Vicary, *Anatomy* (1577); Harward, *Phlebotomy* (1601); Guillemeau, *Child-birth* (1612).

natural sign (McLean 1980: 4), which contributed to the absolute reliance on the written word. Commentaries were research-based teaching aids that explicated religious, philosophical or scientific base texts with the aim of establishing their original, uncorrupt meanings. The contents of commentary texts became standardized in Latin in the twelfth century (Parkes 1976: 116),[11] but textual processes in vernacular versions have not been studied as commentary text are scarce and not well known. By definition, commentaries build upon passages quoted from various authorities, comparing them, and finishing with a summary and the author's own opinion that reconciliated conflicting views at the end (Minnis 1979).

6.1 Middle English

The manuscript index of medieval English scientific texts (Voigts and Kurz, eVK 2014) is a useful tool for tracing English translations of commentary texts. Alchemy is prominently present and Aristotle's *Parva naturalia* "or Commentary" (as the ms title), but a closer inspection revealed that they contained regimen texts or recipes, and one was a compilation of alchemical aphorisms in "notabilia philosophorum" (see Taavitsainen 2012: 189). The updated version of eVK lists two medical texts, a urinoscopy, which is a practical text, and the Hippocratic prognostication (Tavorimina 2006). This fifteenth-century translation gives an English version of one of the canonical Articella texts, Hippocrates' *Prognostica* (in MS TCC R.14.52, ff. 62–104), presumably the earliest medical representative of the genre in English. The style of writing is distinctive and very different from e.g. regimen texts. Metatext is frequently employed to guide the reader with detailed instructions, in accordance with the statement of the utilitarian nature of vernacular commentaries (Taavitsainen 2004: 37). All the most important authorities are quoted, as e.g.

(1) And in like wise of ache prikyng in sides in pleuresy nat only **I denye,** whiche bifore better knowith the pacient; <***Aristotilus***> **but forwhi after Aristotil**, the busynes of **the leche is to fynde** in nat biholdyng of tyme and bi jugementis of the pacient. Thus saiþ he. <***Galienus***> Also note after Galien in the prolog of *Pronosticaciouns,*<***Ipocras***> that Ipocras vsed nat this name prouisioun or forsight but to the place of pronosticacioun.
(f. 64v, emphases in all examples mine)

[11] Refined presentations are found in Latin manuscripts from that time, and with a time lag in English, see the commentary of John Walton's verse translation of Boethius *De consolatione philosophie* in Copenhagen, Royal Library, MS Thott 304, see <http://www.kb.dk/permalink/2006/manus/627/eng/>

Guidance is often given in a distanced way in the third person (*leche* – *he*). The passage quoted above continues in the same mode with an enumertative text strategy, which is also typical of Scholasticism:

(2) And bi pronosticacioun helpith **the leche** whan he vsith pronosticacioun in so moche that he purchasiþ of **the knowlache therof iij thynges**. Wherof **i is that the leche**, whan he vsith pronosticacioun, receiven of the sike of so moche that thei comaunden and commytten hem gretly in their handis. And **the secunde is forwhi** whan he knowith bifore what shal fall of the sike in tyme to come bifore that it fall, bifore that tyme long, and arraieth hymsilf to withstande it. And **the iij is** that he ne doute ne esteme of so moche that he be cause of his deth whiche dieth of infirmite.[i] (f. 64v)

As in the above passage, the noun *leche* occurs frequently identifying medical practitioners as the target readership. The author's own opinion is also stated, e.g. soon after the above passage:

(3) <***Galienus***> Also, after Galien in his prolog, <***Ipocras***> that hevenly thyng of the whiche Ipocras spekith in his bookis – and in like wise **I dar sey, nat doutyng nor dredyng** that that **it is nat possible** that this thyng that Ipocras hath named … (f. 64v)

Thus the text is in accordance with the prototypical commentary technique, but not consistently (see 7.1). Yet this is the only known medical text from the late medieval period that follows the commentary script in English (see also note 10).

6.2 Sixteenth-century texts

Walter Bailey (1529–1593) was a highly learned medical doctor with a successful medical and academic career as professor at Oxford, Fellow of the College of Physicians and one of Queen Elizabeth's medical doctors. The broadening of the world was documented by eyewitness accounts of plants and animals that were different from their traditional descriptions by ancient authorities of logocentric scholastic science. The observations created scepticism and made attitudes change (cf. Gloning 2011; Whitt 2016b), or as Bailey put it:

(4) The nauigations in these latter yeeres…, hath made manifest to vs, **how greatly the old authors**, I meane **Dioscorides, Galen, Plinie, Auicenna, Serapio, and other writers of the former time were deceiued** … But by the nauigations of the Portingals, and of the Spaniards into those countries, in which these pepper trees do growe, it is euident and well knowen,… . For **the trauellers** into those countries which **haue seene** the trees, and **gathered the fruits, do witnes**, that not one and the same tree, but diuers and different do beare these spices.

(EMEMT, Bailey (1588), *Three peppers*, ff.A5r-v)

Learned authors were familiar with Latin commentary texts,[12] and with both Latin and vernacular models the genre was at the disposal of the learned discourse community, ready to be used as a text-guiding principle. Proof for this statement can be found in Bailey's texts. *Mithridatium* (1585) deals with a poison that had gained medical uses and become famous already in the Antiquity. The text begins with a narrative of its history, followed by a sequence of sections with the commentary script. It summarises earlier texts and gives the author's own opinion at the end, in accordance with the commentary script:

> (5) IT shall not be amisse in this place, to declare the composition of this medicine. And it is to be noted, that where **all writers doe greately co~mend** it, yet certayne it is, that they do not in one sorte describe the makinge of the same. But almost **euery authore hath a seuerall description, differinge** in the number of the simples, and also in y=e= proportions and quantities. …
>
> **In time past** y=e= Apothecaries in making of Mithridatium, folowed most the description of Nicolaus Prepositus, of Auicenna, and of Nicolaus Mirepsicus, some of Ætius, some of Paulus. …Wherefore **in these our later daies**, in which learned men haue examined euery thing perfitly, the most part haue commended one of y=e= three compositions expressed by Galen in 2. de Antid. of the which, two were taken (as Galen wryteth) out of the bookes of Andromachus. … **Wherefore we may reasonably conclude, that the first recept transcribed by Galen in 2. de Antid. out of Andromachus workes, is in truth the selfe same that Mithridates vsed, and in mine opinion, in that respect the better to be liked, and the rather to be followed**. For what better assurance can we haue of the true and perfect confection of this medicine, then that which was deliuered by Mithridates? …
>
> (EMEMT, Bailey (1585) *Mithridatium,* f.B1r-v)

Bailey's later text on *Three peppers* (1588) consists of three different parts with three different genre scripts. A recipe at the beginning adheres to the conventions of its own genre, opening with the imperative *Take* that serves a signal triggering genre expectations. The author adds a commentary with a completely different genre script at the end of the recipe; this is new and perhaps unique. He comments on different versions of the ingredients and their measures and finishes with his own opinion: "And where **our authors do differ in opinion** … **I adiuge it better** …" (Bailey 1588: B7v-8r). The third genre script at the end of the text section forebodes medical advertisements: "So you haue both … sold in the Apothecaries shops, vnder the name of Diatrion pipereon" (*ibid.*). Three different genre

[12] Latin prevailed as the language of science until c. 1700 in England, after which the majority of printed medical books were in English (Webster 1975: 267).

scripts in a short text is exceptional, and shows rare mastery of writing conventions (see also Taavitsainen 2009: 43).

7. Compilations and combinations of genre scripts

The adjoining genre of compilations is more didactic and purely practical even in Medieval Latin. Texts of this genre list and discuss opinions of ancient authorities, often considering their differences and similarities, but the author's own opinion is not expressed (Minnis 1979). Instead, excerpts from various authorities are placed in thematic wholes as a survey of the previous literature, pre-digested for the readers' benefit. The main function of the genre was to disseminate knowledge and make it easily accessible to those who did not have access to the originals (Minnis 1979: 402–3). The two top genres started to overlap in the fifteenth century and the converging development was evident by the end of the medieval period (Minnis 1979).

7.1 Middle English

The usefulness of the genre was praised by Guy de Chauliac at the end of the fourteenth century:

> The resoun of þis exposicioun or gadryng togedre was noght defaute of bookes, but raþer **onhede and profit**. Euery man may not haue alle bookes, and if he hadde, it were irksome or noye to rede hem … (Chauliac, ed. Ogden 1971: 2)

Compilations are common in Middle English medical literature, and passages with the script are found in the Hippocratic text in some sections. Encyclopeadic texts display medieval learning at its height and are basically compilations of learned texts. There is, however, a sliding scale to the more popular type, such as *The Wise Book of Philosophy and Astronomy* (anon.) that circulated widely in various forms in late Middle English. It follows a set pattern at the beginning, but astrological prognostications and other popular materials are added at the end so that the versions are very different in scope (Taavitsainen 2012: 188; Griffin 2013).

7.2 Sixteenth-century texts

A text called *Dial for agves* (2a) was written in 1566 by John Jones, a physician of Welsh origin, believed to have studied at Oxford and Cambridge and practised medicine near Nottingham and Derby. He was well versed with classical authorities and translated a Galenic text (*Galens bookes of elementes* 1574, see Pahta et al. 2011).

In addition, he wrote medical books, e.g. on child-care and bathing. The *Dial for agves* is a trilingual learned text in English, Greek and Latin on acute fevers. It contains opinions of ancient authors given one after the other and is primarily a compilation. Yet there are passages with the author's own opinion, in agreement with commentaries:

> (6) The principall humor is blud, the~ flegme, after choler, last melancholie. The distemperance of blud hapneth by one of thother humors, through the inordinate or superfluous mixture of them, and not of him self, for blud is temperate of his proper quality **as saith Hippocrates** [/28./] hauing no contrariety exceding either in heate or cold, or in cold or heate, or in moist or drith, or in drith or moisture, **as confyrmeth Galen** in his own language, this [/29./] left writen [^GREEK OMITTED^] The which in our tung may be vndersta~ded this, that by aucthoritye and good reason bloud to be neyther hotte nor moyste, but temperate. **Yet here I must saye agreing to** [/30./] **Galen**, although blud be neither hot, cold, moist or dry, yet y=e= blud of y=e= veines to be lesse luke warm then the blud of the artiers, and therfore it is likened to the temperament of the spring which is neither so hot as sommer, nor so cold as winter, **in my Methode** [/31./] **to the Phisition I haue shewed this more large, so that I wyshe the Englyshe readers to resorte thither for my opinion** farther. Fleume is of two sortes, natural and vnnaturall, cold and moiste, whyte and swete.
>
> (Jones 1566: f.E1r-v)

8. Seventeenth-century afterlives of scholastic treatises

The scholastic style of writing continues until later in the early modern period, but it seems advisable to treat commentaries and compilations as one category (see above). Nevertheless, the developments seem to diverge but in different ways, as new meaning-making practices emerge in popular texts.

8.1 Professional audiences

There are several books from the beginning of the seventeenth century with scholastic features. They are found e.g. in a text called *Phlebotomy* by Simon Harward, a clergyman, educated at Cambridge and Oxford, who specifies the audience as "well minded Chirurgians" and adds the commonplace "generally to all men". Mediated information from ancient sources is also strongly present in Petrus Pomarius' *Enchiridion medicum* from 1609. It is a dialogue between a doctor and a student who is questioned and displays his learning. The text is targeted at novices in the field. Thomas Lodge was a physician educated at Oxford and the University of Avignon, and practiced medicine in both countries. His fame rests mainly on his literary works. His plague text, called *Treatise of the plague* (1603), discusses the nature, prevention and curing of the disease from the point of view of both individuals and society. It shows a critical attitude which must have been common among educated people at the time. The form of his text is, however, in accordance with the old style (10,613 words with 45 references to authorities, ratio 0,42), but the attitude is very critical:

(7) An other cause of the Plague **saith Auicen**, proceedeth from the celestiall formes, that is to say, the starres and their configurations and malignant aspects, which by their influences cause such sicknesses full of contagion and Pestilence, as **in generall all other Astrologians testifie: But in truth as touching mine owne opinion** which is grounded vpon the diuine determination of Plato in his Epinomides, and his Timæus, of Plotinus his chiefe follower, of Iamblichus, Proclus, Mercurius, Trismegistus, Aristotle, and Auerrhois, **I finde that this opinion, is both false and erronious**; as namely, to thinke that any contagion or misfortune, incommoditie or sicknesse whatsoeuer may by reason of the starres befall man. Because as Plato witnesseth in his Dialogue intituled Epinomis, … (Lodge 1603: B4v)

8.2 The "debased" trend of scholastic argumentation

The line between science and pseudo-science was not clear in the early modern period, as e.g. alchemy still had a place among sciences even in Newton's circles (see Mandelbrote 2001), but the anonymous pseudo-Aristotelian *Masterpiece* first printed in 1684 falls clearly into the category of pseudo-science. This book has been labeled as a "debased" text to cater for "the almost prurient interest in the physiological and psychological types" (Wear 2000: 192), and was a best seller for centuries. Its topics were highly appealing to wide audiences and the work circulated in several versions in the early modern period and far beyond in multiple editions (see Nichols 2015: 421–422)). References to ancient authorities like Galen, Hippocrates, Pliny, Plato and others, serve to lend an aura of learning to the passages where they are employed, and Biblical figures seasoned with pious wishes and prayers are also present. Some of the passages summarise inherited wisdom to the wide public and follow the genre script of compilations. References such as the following are sprinkled all over the text:

(8) **The learned Hippocrates** … And of the same opinion in this matter was **Bartholinus** … **Columbus** is of opinion … **Hipocrates, that Famous and Learned Physician,** is of Opinion… (p. 8) … IT is the opinion of **learned Physicians**, grounded upon reason … **Lactantius** is of opinion …. **saith famous Galen** …. (p. 24)

9. Eighteenth-century texts

The practical purpose of providing a shortcut to useful knowledge did not lose its appeal, but carries on. Chauliac's praise is echoed several centuries later by John Wesley (1703–1791), a Church of England clergyman and a founder of Methodism :

> (9) … the following is little more than an Extract from others: I intend it so to be. I designed only to **collect together** the Substance of **the most celebrated Writings on the Subject**; and to place them in one connected View, **for the Use of those who have little Time or Money to spare**. (1760: vii)

9.1 Texts for professional audiences

Several LMEMT texts for professional audiences use the genre script deriving from previous periods, but with a difference. The authors referred to are contemporary and often discussed in separate sections that survey the earlier literature on the topic. For example, John Smith M. D. (1723) argues for common water as a universal remedy and surveys contemporary authors with very precise indications of the sources:

> (10) The first Commendation of Common Water that I shall mention … by **Dr. Manwaring** in his *Method and Means of enjoying Health*…(p. 4), … **Dr. Keill** … in his *Abridgement of the Anatomy of Human Bodies,* saith …(p. 5) …
>
> with which **Dr. Baynard** does agree… **Dr. Prat**, in his *Treatise of Mineral Water,* shews it to be his Judgement…(p. 5) … Tis said also by **Dr. Duncan**, in his *Treatise of Hot Liquors*…(p. 6)… **Sir John Flyer** also, in the Preface to his *Treatise of Cold Baths*, does affirm, pag. 109. Edit.5.…
>
> … but since I found out … *which is above Forty Years*, I never have been sick for Two days together… I can persuade to try the Experiment; which is such, *that no Physician whatever can advice a better to the King himself, should he fall sick*…. (p. 18)

After the precise references he draws his own conclusions, enhancing his own experience. Thus the script is in accordance with the main line of argumentation, but has been modernized to meet the requirements of the time.

A very similar structure and treatment is found in the latter half of the century by Peter Clare, a London surgeon with international fame, as his treatises were translated into French. His text on gonorrhea from 1783 is highly argumentative and follows the above pattern of quoting contemporary authors on the topic. He sets out to confute "this ill-founded notion … by several arguments supported by quotations from Authors of distinguished credit…" (p. 2). He lists opinions by contemporary physicians, and finally expresses his own with a firm conviction: "**I have asserted** … (p. 16)… that of being a more cleanly, safe, and agreeable method of cure, than any other whatsoever" (p. 23).

A professional text by Rowland Jackson M. D. introduces a new genre context in the employment of the script that has direct continuation until the present day. His work is entitled *A Physical Dissertation* (1746) with the genre label "Dissertation"; the term occurs only few times in text titles in LMEMT.[13] He begins his overview of the earlier literature starting from Antiquity, but recent works are prominently present and works of contemporary doctors are discussed with exact references, e.g. "*Dr. Brubier*, in 1745, publish'd a small Pamphlet…" (p. 6) or "The celebrated *Kunckel*, in *Ephmerid, Nat, Curios.* informs us …" (p. 7). The literature survey precedes the body of the dissertation. In this form, compilations have preserved their usefulness. Even today, academic dissertations conventionally contain a section on the earlier literature much in accordance with the old compilation genre.

9.2 Pseudo-science

A different line of development is represented by a text about "Magnetical Cures" from 1743 spuriously ascribed to Boerhaave. It crosses the border over to pseudoscience and magical thinking which according to the tradition were areas of secret knowledge. The agenda contains the "Examination and Knowledge of these Qualities and hidden Virtues in Nature, is call'd natural magick … able to produce many marvellous Effects, appearing to the Vulgar to be contriv'd by the Help of Daemons…"(p. 19). The last statement is in accordance with popular beliefs and links with some passages in *Aristotle's Masterpiece* on monsters and monstrous births (Chapter V, see Taavitsainen 2017). The text mentions authorities like Galen, Pliny, Bapt. Porta, Tully, Hippocrates, Solomon, Moses, Empedocles, Pythagoras, Democritus, and Plato whose knowledge of magic allegedly came from Egypt.

10. A new ranking order of scholastic features

Our earlier lists of scholastic features (Section 3) were based on qualitative reading and corpus linguistic searches. The new method of Document Classification and its application by the Lightfoot program, contrasting scholastic and non-scholastic texts (see Section 5), gives us a new ranking order of linguistic features used in scholastic argumentation and sorts these features by decreasing feature weight.[14] In isolation they are usually neither good descriptive features nor good

[13] E.g. *Dissertation upon the nervous system* (Anon. 1780), but ECCO contains many more examples in its vast collection of texts; the adjoining genre is essay (see Taavitsainen 2017 fc).

[14] Every word is considered a feature by the program if its frequency is above a threshold value (5 tokens in our study), which led to 6,182 features.

discriminators between the classes, but in combination they can achieve very high classification accuracy. When the list of keywords is organized according to the semantic fields of the items, they reveal interesting features of the scholastic way of writing science:

1. Argumentation with *because, therefore, wherefore* rises to be the most important characteristic of scholastic style. Evidence comes from individual texts, e.g. the frequency of *for whi* 'because' is very high in the Hippocratic commentary.
2. Deontic modality comes next in importance with linguistic items, such as the verbs *must, shall,* and the phrase *(take) heed.* (cf. phrases in our preliminary list).
3. Moral concerns are prominent in scholastic texts, with binary adjectives like *good* vs *evil.* Interestingly, this feature becomes enhanced in later texts.
4. Certainty with adverbs like *truly, forsoth, sothely*, etc. (cf. our preliminary lists).
5. Authorities *Galen, Avicenna* and *Hippocrates* were among the most frequently cited figures. This is in accordance with our pilot assessment of MEMT (Taavitsainen & Pahta 1998).
6. Topic words: *humour, sickness, medicine* reflect the contents of medieval medicine that relied on humoral theory.
7. Logocentric verbs like *say, call* are connected with authorities (cf. our preliminary list)

11. The diachronic line in a new perspective

At the end of the seventeenth century, the Royal Society coined a new style of writing science, and genres at the top changed. Experimental reports were a new medium of empirical science and follow very different genre scripts, as can be verified in reports and essays published in *The Philosophical Transaction*s (1665–), but also more broadly in medical and scientific writing. Research became increasingly based on observation and experimental science and built on totally different principles, developed for the use of the new discourse community of the Royal Society (1662–). A major guideline was the matter-of-fact principle, according to which everybody could agree upon what happens in nature, even if they might disagree about the causal explanations (Dear 1991: 161). Empirical evidence became the prerequisite for verifying the laws of nature, and it was agreed that experiments should be replicable and objective. Instead of mediated knowledge from ancient authors, the new style of writing science favoured a personal narrative of empirical study with low modality.

The attitudes to knowledge and language are also relevant in this context. The common belief was still valid that the perfect knowledge of Paradise had become corrupt by time but by studying books it could be revived (Shapin 1996: 74). The attitudes to language also changed in this period. The *Masterpiece* (1684–) and the pseudo-Boerhaave (1743) give empirical evidence of changes in the appropriation of authoritative sayings and show how inherited wisdom was watered down to spurious knowledge. The old genre script is partly maintained, but partly it blends with new patterns of argument and gains new functions. The goal of this study was to achieve more precise knowledge of diachronic developments and plural manifestations of thought styles.[15] This survey shows that the old patterns linger on and several variant thought styles exist at any one time in different layers of writing. The developments diverge according to the purpose and the intended audiences. But changes are gradual, and the genre map becomes more complicated in the course of time. The learned top genres of scholasticism merged in Latin but the English commentaries reach their full form without contamination in the sixteenth century and show a time lag. For professional audiences, the compilation principle proved its usefulness and became adopted to academic dissertations for displaying knowledge of the previous literature. In contrast, the style of argumentation with references to ancient and religious authorities was used for new applications in pseudo-science for moral ends in order to influence common opinion in matters of sexual conduct.

12. Conclusions

Genres constitute dynamic systems and in a long diachronic perspective, the diversity of styles during the transition periods becomes evident. Old mental modes and habits of thought break down and gradually give room to new forms of comprehension. In the case of scientific and medical writing, early modern authors reveal first-hand evidence of the changes that were taking place in people's attitudes to ancient authorities and inherited wisdom.

A larger historical and cultural context is needed for interpretations. Genres are created in response to the needs of their users. The sociocultural context is important, and the afterlife of the commentary genre can yield us new knowledge for a fuller picture of genre dynamics. Changes in the meaning-making

[15] The scientific thought styles project at the University of Helsinki aims at describing stylistic changes in medical English in a long diachronic perspective in a sociocultural framework, see <http://www.helsinki.fi/varieng/domains/scientific%20thought.html>

practices can be detected with discourse analytic methods. Texts need to be consulted and analysed in their sociocultural context, in relation to their users, but corpus linguistic assessments show the larger picture and give us a reliable basis for further assessments.[16] In this respect the early modern period provides plenty of fruitful material as different genres undergo changes at different rates at different times. The old prestigious genre script became applied to different types of writing, but the script itself changes, too. In early and late modern professional writing, the references are to contemporary and recent authors, whereas in popular texts ancient authorities hold their place and their function becomes modified to moralistic guards in popular sex guides with spurious information. The notion of genre script proved helpful in tracing these changes, and it can well be added to the analyst's toolkit for future use.

Corpora

Taavitsainen, Irma, Pahta, Päivi & Mäkinen, Martti (compilers). 2005. *Middle English Medical Texts*. CD-ROM with MEMT Presenter software by Raymond Hickey. Amsterdam: John Benjamins. https://doi.org/10.1075/z.131

Taavitsainen, Irma, Pahta, Päivi, Hiltunen, Turo, Mäkinen, Martti, Marttila, Ville, Ratia, Maura, Suhr, Suhr & Tyrkkö, Jukko (compilers). 2010. CD-ROM with EMEMT Presenter software by Raymond Hickey. Irma Published together with Irma Taavitsainen and Päivi Pahta (eds), *Early Modern English Medical Texts*. *Early Modern English Medical Texts: Corpus description and studies*. Amsterdam: John Benjamins. https://doi.org/10.1075/z.160

Taavitsainen, Irma, Hiltunen, Turo, Lehto, Anu, Mäkinen, Martti, Marttila, Ville, Oinonen, Raisa, Pahta, Päivi, Ratia, Maura, Suhr, Carla & Tyrkkö, Jukka (compilers). Forthcoming. *Late Modern English Medical Texts 1700–1800*. Amsterdam: John Benjamins.

References

Bakhtin, Mikhail M. 1986[1953]. *Speech Genres and Other Late Essays*. Austin TX: University of Texas Press. https://doi.org/10.1075/z.131

Crombie, Alistair C. 1994. *Styles of Scientific Thinking in the European Tradition*, 3 Vols. London: Duckworth.

Crossgrove, William. 1998. Introduction. *Early Science and Medicine* 3(2): 81–87. Special issue ed. by William Crossgrove, Margaret Schleissner & Linda E. Voigts. https://doi.org/10.1163/157338298X00220

[16] It was interesting to notice that our earlier qualitative assessments were partly confirmed and partly revised by the novel bottom-up Digital Humanities method of computer science.

Dear, Peter. 1991. Narratives, anecdotes, and experiments: Turning experience into science in the seventeenth century. In *The Literary Structure of Scientific Argument: Historical Studies*, Peter Dear (ed.). 135–163. Philadelphia PA: University of Pennsylvania Press. https://doi.org/10.9783/9781512801590

Diller, Hans-Jürgen. 2001. Genre in linguistic and related discourses. In *Towards a History of English as a History of Genres*, Hans-Jürgen Diller & Manfred Görlach (eds), 3–43. Heidelberg: C. Winter.

Fowler, Alastair. 1982. *Kinds of Literature: An Introduction to the Theory of Genres and Modes*. Oxford: Clarendon Press.

Gloning, Thomas. 2011. Spielarten der Quellenkennzeichnung in Fachtexten des Mittelalters und der Frühen Neuzeit. In *Textsortentypologien und Textallianzen des 13.und 14. Jahrhunderts*, Mechthild Habermann (ed.), 303–332. Berlin: Weidler.

Griffin, Carrie. 2013. *The Middle English Wise Book of Philosophy and Astronomy: A Parallel-Text Edition*. Heidelberg: Winter Universitätsverlag.

Habermann, Mechthild. 2014. Mittelalterlich-frühneuzeitliche Fachprosa als Gegenstand historischer Pragmatik. In *Fachtexte des Spätmittelalters und der frühen Neuzeit. Tradition und Perspektiven der Fachprosa- und Fachsprachenforschung*, Lenka Vanková (ed.), 11–30. Berlin: De Gruyter

Jucker, Andreas H. & Taavitsainen, Irma. 2013. *English Historical Pragmatics*. Edinburgh: EUP.

Kuna, Ágnes. 2016. Genre in a functional cognitive framework: Medical recipe as a genre in 16th and 17th century Hungarian. In *Genre in Language, Discourse and Cognition*, Ninke Stukker, Wilbert Spooren & Gerard Steen (eds), 193–224. Berlin: De Gruyter.

Landert, Daniela. Forthcoming. Function-to-form mapping in corpora. Historical corpus pragmatics and the study of stance expressions. In *From data to evidence in English language research*, Carla Suhr, Terttu Nevalainen & Irma Taavitsainen (eds). Leiden: Brill.

Lehto, Anu, Baron, Alistair, Ratia, Maura & Rayson, Paul. 2010. Improving the precision of corpus methods: The standardized version of Early Modern English Medical Texts. In *Early Modern English Medical Texts: Corpus description and studies*, Irma Taavitsainen & Päivi Pahta (eds), 279–289. Amsterdam: John Benjamins.

Mandelbrote, Scott. 2001. *Footprints of the Lion: Isaac Newton at Work*. Cambridge: Cambridge University Library.

McLean, Ian. 1980. *The Renaissance Notion of Woman. A Study in the Fortunes of Scholasticism and Medical Science in European Intellectual Life*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511562471

Milroy, James. 1992. *Linguistic Variation and Change: On the Historical Sociolinguistics of English*. Oxford: Blackwell.

Minnis, Alastair J. 1979. Late-medieval discussions of *compilatio* and the rôle of the *compilator*. *Beiträge zur Geschichte der deutschen Sprache und Literatur*, 101(3): 385–421. https://doi.org/10.1515/bgsl.1979.101.3.385

Minnis, Alastair J., Scott, A. B. & Wallace, David (eds). 1988. *Medieval Literary Theory and Criticism c. 1100–c.1375: The Commentary Tradition*. Oxford: Clarendon Press.

Nichols, Marcia D. 2015. Attitudes, tropes, satire: The Aristotle texts, sex, and the American woman. In *The Secrets of Generation: Reproduction in the Long Eighteenth Century*, Raymond Stephanson & Darren N. Wagner (eds), 417–437. Toronto: University of Toronto Press.

Pahta, Päivi, Hiltunen, Turo, Marttila, Ville, Ratia, Maura, Suhr, Carla & Tyrkkö, Jukka. 2011. Communicating Galen's *Methodus medendi* in Middle and Early Modern English.

In *Communicating Early English Manuscripts*, Päivi Pahta & Andreas H. Jucker (eds), 178–196. Cambridge: CUP.

Pahta, Päivi & Taavitsainen, Irma. 2010a. Scientific discourse. In *Historical Pragmatics*, Andreas H. Jucker & Irma Taavitsainen (eds), 549–586. Vol. VIII in Bubliz, Wolfram, Jucker, Andreas H. & Schneider, Klaus P. (eds), *Handbook of Pragmatics*, Volumes I–X. Berlin: Mouton de Gruyter.

Pahta, Päivi & Taavitsainen, Irma. 2010b. Scientific discourse. In *Historical Pragmatics*, Andreas H. Jucker & Irma Taavitsainen (eds), 549–586. Berlin: Mouton de Gruyter.

Parkes, Malcolm Beckwith. 1976. The influence of the concepts of *ordinatio* and *compilatio* on the development of the book. In *Medieval Learning and Literature: Essays Presented to Richard William Hunt*, ed. Jonathan James Graham Alexander & Margaret T. Gibson (eds), 115–141. Oxford: Clarendon Press.

Rosch, Elinor & Mervis, Carolyn B. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology* 7: 303–322. https://doi.org/10.1016/0010-0285(75)90024-9

Scott, Mike. 2012. *WordSmith Tools*, Version 6. Liverpool: Lexical Analysis Software.

Shapin, Steven. 1996. *The Scientific Revolution*. Chicago IL: Chicago University Press. https://doi.org/10.7208/chicago/9780226750224.001.0001

Spencer-Oatey, Helen. 2000. *Culturally Speaking: Managing Rapport through Talk across Cultures*. London: Continuum.

Swales, John M. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge: CUP.

Taavitsainen, Irma. 1997. Genre conventions: Personal affect in fiction and non-fiction in Early Modern English. In *English in Transition: Corpus-based Studies in Linguistic Variation and Genre Styles*, Matti Rissanen, Merja Kytö & Kirsi Heikkonen (eds), 185–266. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110811148.185

Taavitsainen, Irma. 2001. Changing conventions of writing: The dynamics of genres, text types, and text traditions. *European Journal of English Studies* 5(2): 139–150. https://doi.org/10.1076/ejes.5.2.139.7309

Taavitsainen, Irma. 2004. Transferring classical discourse conventions into the vernacular. In *Medical and Scientific Writing in Late Medieval English*, Irma Taavitsainen & Päivi Pahta (eds), 37–72. Cambridge: CUP.

Taavitsainen, Irma. 2009. The pragmatics of knowledge and meaning: Corpus linguistic approaches to changing thought-styles in early modern medical discourse. In *Corpora: Pragmatics and Discourse*, Andreas H. Jucker, Daniel Schreier & Marianne Hundt (eds), 37–62. Amsterdam: Rodopi. https://doi.org/10.1163/9789042029101_004

Taavitsainen, Irma. 2011. Discourse and genre dynamics in Early Modern English medical writing. In *Medical Writing in Early Modern English*, Irma Taavitsainen & Päivi Pahta (eds), 13–27. Cambridge: CUP. https://doi.org/10.1017/CBO9780511921193

Taavitsainen, Irma. 2012. Disseminating learning: Linguistic features of the commentary tradition and other learned texts in Middle English. In *Congrés internacional Icrea. Ciència i societat a la Corona d'Aragó a l'època de Llull i Eiximenis (Barcelona, 20–22 d'octubre de 2009)*, Anna Alberni, Lola Badia, Lluís Cifuentes & Alexander Fidora (eds), 183–200. Barcelona: Publicacions de l'Abadia de Montserrat.

Taavitsainen, Irma. 2016. Genre dynamics in the history of English. In *Cambridge Handbook of Historical Linguistics*, Merja Kytö & Päivi Pahta (eds), 271–285. Cambridge: CUP. https://doi.org/10.1017/CBO9781139600231.017

Taavitsainen, Irma. 2017. Meaning-making practices in the history of medical English: A socio-pragmatic approach. *Journal of Historical Pragmatics* 18(2): 252–270. https://doi.org/10.1075/jhp.00005.taa

Taavitsainen, Irma. 2017. The essay in Early Modern and Late Modern English medical writing. In Jean-Jacques Chardin, Albert Hamm and Anne Bandry-Scubbi (eds.), *Discourse, Boundaries and Genres in English Studies.* Special Issue of *Ranam* 50 (Strassbourg), 15–30.

Taavitsainen, Irma & Hiltunen, Turo. Forthcoming. *Late Modern English Medical Texts 1700–1800: Corpus description and studies.* Amsterdam: John Benjamins.

Taavitsainen, Irma & Jucker, Andreas H. 2015. Twenty years of historical pragmatics: Origins, developments and changing thought styles. *Journal of Historical Pragmatics* 16(1): 1–24. https://doi.org/10.1075/jhp.16.1.01taa

Taavitsainen, Irma & Pahta, Päivi. 1998. Vernacularisation of medical writing in English: A corpus-based study of scholasticism. *Early Science and Medicine* 3(2): 157–185. Special issue ed. by William Crossgrove, Margaret Schleissner & Linda E. Voigts. https://doi.org/10.1163/157338298X00266

Taavitsainen, Irma & Schneider, Gerold. Forthcoming. Scholastic argumentation in early English medical writing and its afterlife: New corpus evidence. In *From data to evidence in English language research*, Carla Suhr, Terttu Nevalainen & Irma Taavitsainen (eds). Leiden: Brill.

Tavormina, M. Teresa (ed.). 2006. *Sex, Aging & Death in a Medieval Medical Compendium: Trinity College Cambridge MS R.14.52, its Texts, Language, and Scribe*, Vol. 1 [Medieval and Renaissance Texts and Studies 292]. Tempe AZ: Arizona Center for Medieval and Renaissance Studies.

Voigts, Linda Ehrsam. 1984. Medical prose. In *Middle English Prose: A Critical Guide to Major Authors and Genres*, Anthony S. G. Edwards (ed.), 315–335. New Brunswick NJ: Rutgers University Press.

Voigts, Linda Ehrsam & Deery Kurz, Patricia. 2014. *An expanded and revised version of Scientific and Medical Writings in Old and Middle English: An Electronic Reference, CD*. Ann Arbor MI: University of Michigan Press.

Webster, Charles. 1975. *The Great Instauration: Science, Medicine and Reform, 1626–1660*. London: Duckworth.

Wear, Andrew. 2000. *Knowledge and Practice in English Medicine 1550–1680*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511612763

Whitt, Richard J. 2016a. Evidentiality in Early Modern English medical treatises (1500–1700). *Journal of Historical Sociolinguistics* 2(2): 235–263. https://doi.org/10.1515/jhsl-2016-0014

Whitt, Richard J. 2016b. Using corpora to track changing thought styles: Evidentiality, epistemology and Early Modern English and German scientific discourse. *Kalbotyra* 69: 265–291. <http://www.zurnalai.vu.lt/kalbotyra/article/view/10376>

Academic writing as a locus of grammatical change

The development of phrasal complexity features

Bethany Gray & Douglas Biber
Iowa State University / Northern Arizona University

Based on large-scale corpus analysis, this study challenges the notion that academic writing is conservative and resistant to change by documenting linguistic innovations that have emerged in academic writing over the past 200 years. The study explores the dramatic patterns of change that have culminated in the present-day phrasal discourse style of academic writing. The study demonstrates that academic writing today employs a dense use of phrasal complexity features which were minimally used in earlier historical periods. Cross-register comparisons show that these features have largely not been adopted in other spoken and written registers, and none to the extent as in academic writing. The results, which illustrate that these changes have been both quantitative and functional in nature, thus challenge not only the view that academic writing is resistant to change, but also the claim that grammatical innovation originates primarily in speech.

1. Introduction

The use of large corpora in the study of linguistic change has increased in recent years as historical texts have become more readily available in electronic form. It is thus unsurprising that, as written records of scientific and medical knowledge have been well-preserved, many corpora of historical academic texts have been compiled and analyzed by linguists interested in how language has developed over the past centuries. This research has often compared academic writing to other registers[1] for which there are historical records available, such as fiction, news, legal writing, and drama (as a written representation of speech).

[1] We use the term *register* here in the sense of Biber and Conrad (2009), in which *register* refers to a variety of language that can be defined based on its non-linguistic or 'situational' characteristics, such as mode, communicative purpose, topic, degree of interactivity,

https://doi.org/10.1075/scl.85.06gra

In research on the historical development of English, academic writing has often been characterized as a language variety that is particularly resistant to linguistic change, failing to adopt linguistic innovations witnessed in other registers, both written and spoken (e.g., Hundt & Mair 1999). These characterizations of academic writing as resistant to change have stemmed from both theoretical arguments about the nature of language change, as well as empirical documentation of the use of certain linguistic features across different registers over time.

From a theoretical perspective, spoken language has often been positioned as the primary form of language while writing is an artificial, written derivative of speech (e.g., Aronoff 1985; Fillmore 1981; McWhorter 2001). This focus on the primacy of speech has extended to the area of language change, with a common view that language change is driven by conversational interaction (e.g., Hopper & Traugott 2003; Bybee & Hopper 2001; Croft 2000; see the discussion in Biber & Gray 2011: 223–225 and Biber & Gray 2016) and then adopted into written language. Both of these views (the primacy of speech and the lack of change in writing) are exemplified by McWhorter:

> Writing, however, is an artificial, conscious activity, and thus it is easy to resist language change in writing. We are taught to do just this, and therefore most written language is an artificial representation, omitting the signs of change which the real language, the spoken one, is full of. Indeed, writing slows language change down somewhat even on the spoken level, as writing reinforces our sense of "language" as a disembodied blueprint to be followed or flouted….No matter what the authority of the written form, or how tenaciously it holds on to the past, or how absurd the gulf between the written and the spoken form becomes, the spoken form always, always keeps on changing – and ultimately drags the written form reluctantly with it. (McWhorter 2001: 17)

The goal of this chapter is to challenge these two perceptions of academic writing: (1) that academic writing is resistant to change, and (2) that changes adopted in writing originate in speech. To counter these perceptions, we employ large-scale corpus analysis to demonstrate how academic writing has changed, often dramatically, over the past two centuries. Analyzing features related to the distinctive phrasal or nominal discourse style of present day academic writing, we demonstrate that academic writing has in fact been the leader of specific innovations in the use of such grammatical features. Along the way, we demonstrate the importance of register and sub-register as an important consideration in corpus-based

relationships between participants, ability to edit, and so on. In a register perspective on language variation, differences in linguistic structure can be attributed to these non-linguistic features of a language or text variety.

studies of language change, showing that the exact trajectories of the observed changes vary by academic writing sub-registers and disciplines.

1.1 Colloquialization in writing

Given the view that language change originates in speech discussed above, it is not surprising that many empirical investigations of language change and grammaticalization have focused on features which have arisen in and are characteristic of spoken language (e.g., Krug 2000 and Tagliamonte 2004 on semi-modals). The increase in use of spoken features (such as modals and semi-modals, progressive aspect verbs, contractions, personal pronouns, the *get*-passive, and phrasal verbs) has remained a focus in studies of change in writing, despite the fact that these features are relatively rare in writing generally, and in academic writing particularly. Such adoption of spoken features into writing has been referred to as *colloquialization* (Hundt & Mair 1999; Mair 2006; Leech et al. 2009), the *drift* toward a more oral style (Biber & Finegan 1989), and *informalization* (Fairclough 1992).

Research that has empirically supported the claim that academic writing is resistant to change has often focused on colloquialization. This has been a productive area of research, and has shown that while processes of colloquialization or drift *have* occurred in written registers in English, this shift has not occurred consistently across all written registers (Biber & Finegan 1997/2001; Hundt & Mair 1999). In particular, academic writing has participated the least in these changes, leading Hundt and Mair (1999: 236) to label academic writing as an ‘uptight’ (rather than ‘agile’) register, in that it seems to be relatively closed to innovation and more prone to retain traditional or conservative forms.

Indeed, it is easy to see why academic writing is characterized as a minimally-changing register based on the results of such studies. For example, using corpora of fiction, newspapers, and science writing from 1750–1990, Biber and Gray (2016) examine Pearson correlations between year of publication and rates of occurrence for features such as contractions, semi-modals, core modals, progressive aspect verbs, the *get*-passive, phrasal verbs, and personal pronouns and year. Table 1 summarizes these correlations, with strong positive correlations indicating an increase in frequency over time (for full correlational results, see Biber & Gray 2016: 135).

While fiction exhibited strong positive correlations for eight of these features (indicating an increase in frequency over time), news reportage had fewer meaningful positive correlations, and academic writing had only one – for progressive passive verbs (and actually exhibited negative correlations for four features). These findings, which are consistent with other studies, show

Table 1. Historical trends in the adoption of colloquial features in three written registers from 1750–1900 based on Pearson correlation coefficients** between rates of occurrence and date*

Colloquial Features	**Fiction** ***N* = 215 texts**	**News** ***N* = 1140 texts**	**Science prose** ***N* = 524 texts**
Contractions	++	+	
Semi-modals (all)	+++		
Core modals (all)	-		
Simple progressive	+++	++	
Perfect progressive	++	+	
Progressive passive	++		++
GET passive	++		
Phrasal verbs	++		-
Help + bare infinitive	++	+	+
1st person pronouns			--
2nd person pronouns			--
3rd person pronouns			--

*Adapted from Biber and Gray (2016: 135).
** where ______ indicates a Pearson correlation from __________

+++	.60 to .99
++	.30 to .59
+	.20 to .29
-	.20 to −.29
--	.30 to −.59

that colloquialization has indeed been a major trend in historical patterns of language use in writing – but not in the register of academic writing. The fact that the use of colloquial features has increased in some written registers but not academic writing provides important insights into the processes of language change. In particular, this finding contributes to our understanding of how such changes are mediated by register.

However, these findings only give a partial picture of academic writing and its development over the recent past; they only illustrate that academic writing has not exhibited substantial changes in the use of colloquial features more typically associated with spoken registers – features which are rare in present-day academic writing (see, e.g., Biber et al. 1999). We argue that studies of historical development should also consider register features that characterize modern instantiations of that register in order to gain a fuller understanding of how the present-day discourse style has arisen over time.

Register features are lexical or grammatical constructions that are particularly frequent and pervasive in a register, occurring (a) with a markedly higher frequency in the register when compared to other registers and (b) throughout texts representing the register (Biber & Conrad 2009: 53). That is, in order to fully account for linguistic developments in academic writing, we must also consider the core grammatical characteristics that the register exhibits today and investigate to what extent the use of those features has evolved over time to arrive at the current-day discourse style. So what grammatical features *are* particularly characteristic of academic writing today?

1.2 Register features of present-day academic writing

A great deal of research has focused on the grammatical characteristics of present-day academic writing. This research has often focused on individual grammatical features that carry out specialized functions in academic writing, such as shell nouns (e.g., Aktas & Cortes 2008, Charles 2003; Flowerdew & Forest 2014), passive voice (e.g., Baratta 2009), extraposed constructions with anticipatory *it* (e.g., Groom 2005; Hewings & Hewings 2002), and stance markers (e.g., Biber 2006; Hyland 1998, Hyland & Tse 2005), to name a few.

On the other hand, other research focuses on establishing the core grammatical characteristics of academic writing, identifying linguistic features which are especially common in academic writing (or are particularly more common in academic writing than in other registers) – what we can consider to be register features of academic writing (Biber & Conrad 2009). *The Longman Grammar of Spoken and Written English* (Biber et al. 1999) documents that many of these characteristic features are associated with nouns and noun phrases, including nouns, nominalizations, demonstrative determiners before nouns, noun phrases with pre- and post-modifiers, attributive adjectives, noun complement clauses, and participle clauses as noun modifiers (Biber et al. 1999; for a fuller survey, see Biber & Gray 2016: 79–82).

Indeed, the nominal style of academic writing has long been recognized and is well-documented (Wells 1960; Biber 1988; Biber et al. 1999; Halliday 1988/2004). Example 1 illustrates this nominal style, in which there is a high density of nouns and many complex noun phrases. In Example 1, nouns that are the head of a noun phrase are **bolded**; adjectives, nouns and participles as pre-modifiers are in *italics*; prepositional phrases and appositive noun phrases as post-modifiers are underlined:

(1) The *lysosome-like* **vacuole** of *budding* **yeast** Saccharomyces cervisiae is a *robust* **model** for studying the *cell biological* **aspects** of *regulated membrane* **flux**. Several **principles** of *vesicle* **targeting** and *membrane* **fusion** have been established through *genetic* and *cell biological* **studies** of *vacuole* **biogenesis** and *biochemical* **analysis** of *isolated* **vacuoles**.

-Biology research article

The dense use of nouns and noun phrase modifiers is in contrast to the relatively infrequent use of finite (2 occurrences) and non-finite (1 occurrence) verbs, which are indicated in SMALL CAPS. The result is discourse which is densely packed with information. Most of the information in this excerpt is conveyed through nouns and the many words and structures that modify those nouns. This nominal packaging condenses complex information into a few words.

We have argued elsewhere (e.g., Biber & Gray 2010, 2016) that this condensing of information into nominal structures results in implicit meaning relationships that require specialist knowledge to unpack and comprehend the meaning of. For example, *regulated membrane flux* might be paraphrased as *flux in membranes that is regulated*, which is slightly less condensed but still requires specialist knowledge to understand the exact meaning. In fact, we had to consult with a biologist to understand the precise meaning of this noun phrase, in part because of the use of technical vocabulary (i.e., *flux*), but also because the relationships between the head noun *flux* and its modifiers is not explicit. The meaning of this phrase as understood by a biologist would be "the flow of molecules across a membrane that is controlled" in a fine-tuned and very specific manner (Nitya Jacob, personal communication, 6/15/2017). Further, biologists would have the background knowledge to understand that this control involves "a protein in the membrane that opens or closes in response to the conditions that should permit or inhibit the flow of those specific molecules" (Nitya Jacob, personal communication, 6/15/2017). Of course, the full meaning of this three-word phrase can be explained explicitly with language (since the biologist we consulted could explain it readily!), but those explanations require much more elaborated language expressed in full clauses.

The condensed, nominal style in Example 1 is in direct contrast to conversation (Example 2), where pronouns (rather than nouns and complex noun phrases), verbs, and embedded clauses (indicated with square brackets) are much more frequent and pervasive:

(2) Well the **reason** [$_1$ I SAID [$_2$ it was too good [$_3$ to BE true]]] IS [$_4$ when **Pete** and I WERE TALKING the other **day**,] I ASKED [$_5$ what HAPPENED [$_6$ when it HITS **rust**]] and he SAID [$_7$ it, it CHANGES to *molecular* **structure**.]
-Conversation (*Longman Corpus of Spoken and Written English*)

This distinction between the characteristic nominal style of academic writing and the clausal style of conversation has been validated by empirical, descriptive grammars such as the *Longman Grammar of Spoken and Written English* (Biber et al. 1999), through qualitative and theoretical investigations under the Systemic Functional Linguistics umbrella (e.g., Halliday 1988/2004), and through multi-dimensional analysis studies that employ statistical measures to identify co-occurrence patterns of linguistic features across registers (e.g., Biber 1988 on patterns of register variation in English; Xiao 2009 on spoken and written registers in World English varieties; Biber 2001 on historical registers; Biber 2006 on spoken and written academic registers).

By comparing academic writing to other written and speech-based registers, such studies have consistently documented that the register of academic writing is distinctive in its reliance on nouns and associated noun phrase modifiers. For example, Biber and Gray (2016: 93–94) show that complex noun phrase structures such as attributive adjectives, nouns as nominal pre-modifiers, non-finite relative clauses, and prepositional phrases as noun post-modifiers are significantly more common in academic writing than in conversation (with high effect sizes, as indicated by r^2 values ranging from .52 to .94), while clausal features such as finite adverbial clauses, finite complement clauses, and adverbs as adverbials are significantly more common in conversation.

These contrasting phrasal and clausal discourse styles have been associated with two complementary types of grammatical complexity. The first type, which is largely aligned with traditional views of complexity, is clausal in nature, and is characteristic of spoken, interactional language (Biber & Gray 2010, 2011; Biber 1988, 1992). The second type, on the other hand, is associated with nouns and noun phrase modifiers and is characteristic of written, informational language like academic writing (Biber & Gray 2010, 2011, 2016; Biber 1988, 1992). The highly frequent use of phrasal complexity features distinguishes academic writing not only from spoken registers, but also from other written registers. Figure 1 illustrates of few of these phrasal features based on present-day corpora of novels, newspaper writing, and academic prose with conversation as a final comparison point (see Biber & Gray 2016). Figure 1 shows that these features are much more common in academic writing than any of these spoken and written registers.

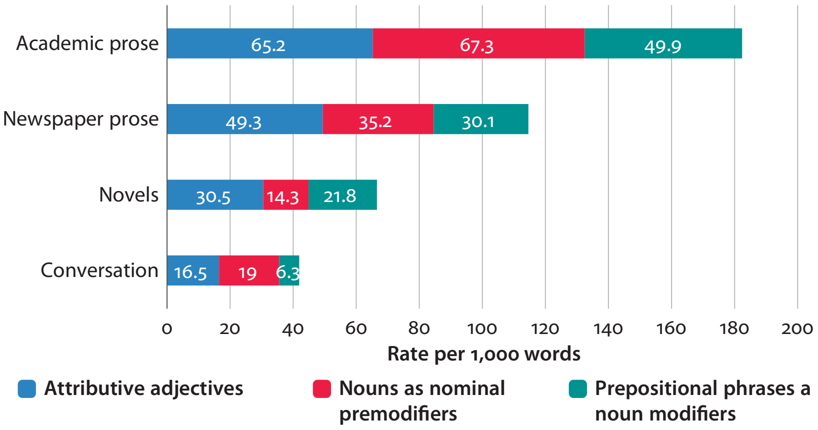


Figure 1. Selected phrasal complexity features in written and spoken registers

The predominance of phrasal complexity features in academic writing can be explained by taking a register theoretical approach to language variation. In a register perspective, patterns of use of lexical and grammatical features can be associated with or explained by a consideration of the non-linguistic characteristics of the situations in which the language occurs (i.e., the register). Registers can be described according to factors external to the language itself, including mode (spoken vs. written), communicative purpose, degree of interactivity and relationships among participants, the extent to which production is on-line or edited, topic, setting, and so on (see Biber & Conrad 2009).

Thus, we can explain the typical phrasal style of academic writing by considering the situational characteristics of the register. Academic writing is highly informational in purpose, is written by an expert with specialist knowledge for a specific audience who shares that specialist knowledge, is heavily edited, and is produced over a long period of time. These factors, combined with publishing constraints requiring conciseness, can explain why academic writing has adopted the informationally-dense noun phrase structures illustrated in Example 1. Complex information can be encoded in complex noun phrases when writers have the time to produce and edit complex structures, when writers and readers share specialist knowledge that does not have to be explicitly stated using multiple clauses, and when there is a need for conciseness.

Figure 1 above showed that newspaper writing is the most similar to academic writing in terms of its use of three of the phrasal complexity features, yet the magnitude of the use of these features is still quite higher in academic writing. Again, we can explain this trend relative to the situational characteristics of the newspaper register. Like academic writing, newspaper prose has an informational purpose, is edited (albeit on a much shorter timescale), and requires conciseness – thus explaining the higher use of these noun phrase complexity features compared to conversation and fiction. However, unlike academic writing, newspaper prose is written for a popular audience, one who should not require any specialized knowledge to understand the article contents. Hence, the use of the condensing, inexplicit phrasal complexity features is not nearly as frequent in newspaper writing as in academic prose. We claimed above that to fully account for linguistic change in a register, we must account for how frequent and pervasive register features came to be characteristic of that register. For academic writing, this means tracking the use of phrasal complexity features over time (with clausal complexity features offering an important comparison point), as these features distinguish academic writing from both written and spoken registers.

Importantly, such an investigation must consider multiple types of changes, including changes in the frequency of particular features, shifts in the functional uses grammatical resources, and how those linguistic developments relate to the evolution of situational characteristics of registers.

1.3 Two types of historical development: The need for quantitative corpus-based research

Recent historical changes in the grammatical system of English have generally fallen into two types of developments. The first type of developments, often labeled 'grammaticalization', represent new grammatical features that have emerged over time (e.g., the development of the *get* passive, *well* as a discourse marker) or new grammatical uses of a specific lexical item (e.g., the use of *because* as a preposition). This type of change is typically highly salient, as language users become aware of innovative uses of language that were not previously in use.

The second type of change is less salient, as it represents shifts in the typical use and functions of core grammatical features. That is, the linguistic resources are already in use in the language, but undergo shifts in the exact patterns of use. These shifts may occur in terms of (a) increased or decreased frequency of use, such as the increasing use of progressive aspect, (b) changes in the set of lexical items that a grammatical structure co-occurs with, such as an increasing number of verbs being associated with progressive aspect, and (c) increases in the number and types of meanings or functions that a structure carries out in discourse.

As Fennell (2001) and Denison (1998) point out, many of the changes in the last 300 years have been of this second type, often representing changing frequencies of use rather than the emergence of new grammatical forms. Consequently, the use of corpus-based methods for quantifying such patterns of use is particularly needed in order to describe changing patterns of use that are less salient or intuitively recognized. Corpus methods also enable the researcher to identify quantitative findings and then "subject quantitative results to qualitative scrutiny" (Leech et al. 2009: 32), leading to the discovery of functional shifts in the uses of core grammatical features. These quantitative and functional shifts in language use can then be interpreted relative to "factors external to corpus data – cognitive, social or historical factors, for example" (Leech et al. 2009: 34).

In the present study, we focus on this second type of change, exploring both quantitative and functional shifts in the use of core linguistic features, as these types of changes are most representative of the linguistic change that has occurred in academic writing.

1.4 Goals of the study

The goal of this study is to demonstrate that the register of academic writing has in fact undergone substantial linguistic change over the past 200 years. We take as our starting point the core characteristics of present-day academic writing and trace the developmental path that has resulted in today's phrasal discourse style. By comparing academic writing to other registers, we show that academic writing has actually been the locus of these changes toward increased phrasal complexity, exhibiting both quantitative and functional shifts in the use of linguistic features. We intend to show that the context in which academic writing takes place, including situational factors like purpose, production circumstances, and audience, has the potential to lead to grammatical change,[2] and that these changes are mediated by sub-registers within academic writing.

To accomplish these goals, we present a series of case studies drawn from a larger project reported in Biber and Gray (2016). In Section 3.1, we provide an overview of historical change in the use of phrasal complexity features, with selected clausal complexity features presented as a comparison point. We compare diachronic corpora of academic writing, fiction, and newspaper writing to show that academic writing has adopted the phrasal complexity features to a much greater extent than other written registers.

In Sections 3.2 and 3.3, we then select two phrasal features which have exhibited particularly strong increases in frequency over time for further analysis (nouns as nominal pre-modifiers and prepositional phrases as post-modifiers), exploring how these quantitative changes also encompass important functional shifts in the use of these features. Along the way, we consider how these linguistic developments have been mediated by different types (i.e., sub-registers) of academic writing.

2. Corpora and analytical methods

The analyses reported here are based on a series of corpora to enable the investigation of both recent and longer-term historical change. These analyses rely on several established corpora of academic texts: ARCHER (*A Representative Corpus of Historical English Registers*; Biber, Finegan & Atkinson 1994; Yáñez-Bouza 2011), *The 20th Century Research Article Corpus* (Biber & Gray 2010), and CETA

[2] Halliday and Martin (1993: 7–12) offer similar claims, arguing that innovations in science writing began as functional shifts that persisted as the social context also changed. However, Halliday and Martin also argue for a more cyclical relationship, in that at the same time, "the grammar has been reconstruing the nature of experience" (1993: 12). That is, they argue that the rise of technical vocabulary and specialized grammatical structures in science writing, while beginning as a way to fill a functional need, now function to construct and maintain "complex ideological edifices" (1993: 11).

(the *Corpus of English Texts on Astronomy*, see Crespo García & Moskowich-Spiegel Fandiño 2010).

We supplemented these with corpora compiled for the purposes of the larger project reported on in Biber and Gray (2016), including samples from *The Philosophical Transactions of the Royal Society* (published since 1665) and *Science* (published since 1880), as well as a collection of history monographs from 1850–1900 (to represent humanities writing).

Because we wanted to account for possible variations in the historical development of academic writing due to sub-register, we organize these corpora into four categories of academic writing based on the audience (specialist or non-specialist/multi-disciplinary) and discipline group (science, social science, and humanities). Table 2 summarizes how the corpora were divided into these four categories:

Table 2. Historical corpora of academic writing

Type of academic writing	Corpora	Time periods represented	Number of texts	Number of words
Specialist science	20th Century Research Article Corpus *(Biology, medicine, ecology, and physiology)*	1965–2005	167	c. 1.1 million
Specialist social science	20th Century Research Articles *(Education, psychology)*	1965–2005	173	c. 1.1 million
Specialist humanities	20th Century Research Article Corpus (*History*), History monographs	1850, 1965–2005	157	c. 2.5 million
Non-specialist (multi-disciplinary) science	ARCHER, CETA, *Philosophical Transactions* and *Science*	1700–2005	422	c. 2.4 million

Note: for full descriptions of these corpora, see Biber and Gray (2016: Chapter 2).

In order to demonstrate that academic writing has initiated linguistic innovations in English that have not been adopted to the same extent in other written registers, we also compare the grammatical characteristics of academic writing to historical corpora of fiction and newspapers, as summarized in Table 3:

Table 3. Historical corpora of other written registers

Register	Corpora	Time period	Number of texts	Number of words
Fiction	ARCHER, samples from *Gutenberg* novels	1700–1990	215	c. 1.2 million
Newspapers	ARCHER, samples from *New York Times*	1700–1990	340	c. 250,000

All corpora were tagged with the Biber tagger, which annotates each word in a text with grammatical information using both rule-based and probabilistic information. Words are tagged for specific part of speech, along with additional grammatical information (e.g., tense, aspect and voice; subordination, etc.). The Biber tagger, originally developed in the 1990s and then expanded during the work on the *Longman Grammar of Spoken and Written English* (Biber et al. 1999), is quite reliable on academic texts, with precision and recall rates above .90 for most features (see Gray 2015, Appendix B). To address less reliable tags, automatic scripts[3] were developed to correct systematic tagging errors and increase the reliability of the tagging system.

Further programs were developed to process the tagged texts, identifying a full complement of general grammatical features, along with features specifically related to phrasal and clausal complexity. Table 4 summarizes the features that we focus on in this chapter, and includes core grammatical features, phrasal structures that contribute to noun phrase complexity (both pre- and post-modifiers), and clausal structures (see Biber & Gray 2016: Chapter 2 for a complete listing of features included in the larger study).

While most of these features could be identified with a high degree of reliability automatically, prepositional phrases required manual coding to distinguish prepositional phrases as noun modifiers versus other syntactic functions (e.g., adverbial). These hand analyses were conducted on a subsample of texts (see Biber & Gray 2016: 142 for details on the subsampling process).

Rates of occurrence were normalized to 1,000 words to enable comparisons across corpora of varying sizes (see Biber, Conrad & Reppen 1998: 263–264). Rates of occurrence were calculated for each text in the corpora (rather than a total for each corpus), thus enabling the use of statistical analyses such as Pearson correlations,[4] which measure the extent to which two variables (i.e., year and rate of occurrence) have a linear relationship. Quantitative analyses were followed up with qualitative analyses of functional and semantic shifts.

[3] These scripts were developed in Perl based on an analysis of tagging errors in academic texts from a range of disciplines, as reported in Gray (2015).

[4] Pearson correlations are used here despite the fact that historical change is often non-linear in nature because non-parametric alternatives (such as Spearman) may inflate the estimate of the relationship. Pearson correlations are thus a more conservative statistical measure.

Table 4. Linguistic features utilized in the study

Linguistic features	Examples or description
Core grammatical characteristics	
Nouns*	all words tagged as a noun
Nominalizations	*development, examination, happiness*
Average word length	length in number of characters
Lexical verbs	all main verbs (not auxiliary verbs)
Prepositions	all prepositions
Phrasal noun phrase pre-modifiers	
Attributive adjectives	*emotional* injury, *conventional* practices
Nouns as pre-modifiers	*trial transfer* sessions, *computer* desk
N-xxing as pre-modifiers	*law-making* powers, *ribosome-binding* activities
N-xxed as pre-modifiers	*age-related* change, *membrane-embedded* spindle pole body
Phrasal noun phrase post-modifiers	
Of-phrases as post-modifiers	the effect *of quality differences*
In-phrases as post-modifiers	changes *in the diversity of adjective types*
On-phrases as post-modifiers	a poem *on the virtue of a laurel leaf*
For-phrases as post-modifiers	the scores *for male and female target students*
With-phrases as post-modifiers	the cohort *with the pertinent diagnosis*
Clausal noun phrase post-modifiers	
Relative clauses (all)	the amount of experimental error *that could be expected to result from using cloze tests of various lengths*a ring *which limits a central electrotransparent space*transfer tests *following over-training*the results Tables IV and V
That noun complement clauses	The fact *that no tracer particles were found in or below the tight injunction (zonula occludens)* indicates that…
To noun complement clauses	The project is part of a massive plan *to complete the section of road…*
Of + -ing clause	The result *of omitting the variable from the study…*
Clausal clause constituents	
Finite adverbial clauses	*If I don't put my name,* she doesn't know who wrote it, *although she might guess.*
Finite complement clauses (all)	I would hope *that we can have more control over them.*I don't know *how they do it.* It is evident *that the virus formation is related to the cytoplasmic inclusions.…*despite the fact *that they are monologic and produced in writing.*

*The category of nouns includes any word tagged as a noun regardless of form or syntactic position. The frequency for nouns thus subsumes the frequencies of more specific noun-related features (e.g., nominalizations, nouns as nominal modifiers). However, the more specific categories represent only a small portion of the overall frequency of nouns.

3. The historical evolution of academic writing: Quantitative increases and functional extensions of phrasal complexity features

3.1 General patterns of historical change: Phrasal and clausal complexity features

Table 5 presents Pearson correlations measuring the relationship between the use of linguistic features and the progression of time. Overall, Table 5 reflects a pattern of overall decline of the use of clausal subordination and overall increase in the use of phrasal features. While these trends hold to some extent in all three written registers examined here, these changes have been most pronounced in academic research writing. That is, academic writing is distinctive in the extent to which phrasal complexity features have been adopted over the past 200 years.

Table 5 shows strong positive correlations for many of the phrasal features related to noun phrase complexity, indicating increases in frequency for these features over time. These increases include the use core grammatical features, pre-modifiers, and post-modifiers. Nouns and nominalizations have increased substantially in academic research writing, increased to a lesser extent in newspaper writing, and stayed the same or even decreased (in the case of nominalizations) in fiction.

In terms of phrasal pre-modifiers, nouns as nominal pre-modifiers have increased substantially in all three registers, with the strongest increases in academic research writing, developing from a feature that was quite rare in the 1700s to being on par with attributive adjectives in terms of frequency (see Section 3.2 below). The combination of nouns with a participle as pre-modifiers (e.g., *law-making powers, age-related change)*, has also increased over time (although they remain a lower frequency feature; see Biber & Gray 2016: 113). The only pre-modifier examined here which does not exhibit strong increases over time is attributive adjectives; this may be because attributive adjectives were already frequent and have remained a common feature in written registers.

Noun phrase post-modifiers that are phrasal in nature have also increased over time, with strong increases in science writing and some increases in newspaper writing and fiction. These increases in the post-modifier function of prepositional phrases occur despite a negative correlation between time and the use of prepositional phrases in all three registers. This finding indicates that while the overall use of prepositional phrases has declined, the use of prepositional phrases as noun phrase modifiers is in fact increasing – that is, prepositional phrases are increasingly being used to embed complex information into noun phrases while other uses are declining.[5]

[5] Like attributive adjectives, *of* prepositional phrases modifying nouns were already quite frequent in the 1700s and do not show increases over time. In fact, *of*-phrases decrease in use in all three registers, but decrease the least in science writing.

Table 5. Historical trends in the use of complexity features in three written registers from 1750–1990 based on Pearson correlation coefficients** between rates of occurrence and date*

	Science writing (N = 524 texts)		News writing (N = 1,140 text)		Fiction (N = 215 texts)	
	trend	**r**	**trend**	**r**	**trend**	**r**
Core grammatical characteristics						
Nouns	+++	.77	++	.34		.17
Nominalizations	++	.48		.04	--	−.55
Average word length	+++	.60	+	.29	--	−.32
Lexical verbs	--	−.55	+	.28	++	.33
Prepositional phrases	--	−.42	--	−.55	--	−.40
Phrasal noun phrase pre-modifiers						
Attributive adjectives	+	.29		.09	-	−.25
Nouns as pre-modifiers	+++	.75	++	.53	++	.51
N-xxing as pre-modifiers	+	.28		.12	+	.22
N-xxed as pre-modifiers	++	.33	+	.21	+	.26
Phrasal noun phrase post-modifiers						
Of-phrases as post-modifiers	-	−.20	--	−.57	--	−.47
In-phrases as post-modifiers	+++	.60		.17	+	.26
On-phrases as post-modifiers	++	.41	+	.20	++	.43
For-phrases as post-modifiers	+++	.60		.11		−.10
With-phrases as post-modifiers		.13	+	.29		−.02
Clausal noun phrase post-modifiers						
Relative clauses (all)	--	−.61		−.05	--	−.47
That noun complement clauses	+	.24		.04	+	.25
To noun complement clauses		−.11		.14		−.13
Of + -ing clause		−.12		−.10		−.12
Clausal clause constituents						
Finite adverbial clauses	-	−.22	+	.24		.13
Finite complement clauses (all)		−.15		.12		.16

*Adapted from Biber & Gray (2016: Chapter 4).

** where _____ indicates a Pearson correlation from _________

+++	.60 to .99
++	.30 to .59
+	.20 to .29
-	−.20 to −.29
--	−.30 to −.59

The strong correlations between the use of many noun and noun phrase features and time is in contrast to verb and clause features, even clauses that modify nouns. These features have generally shown little change over the past 200 years or have actually decreased in some written registers. Lexical verbs overall decreased in academic research writing, while increasing modestly in news and fiction. Embedded clauses such as finite adverbial clauses and finite complement clauses show little change across the three written registers investigated here (with small correlations ranging from −.22 and +.24), although they have decreased only in science writing.

Perhaps more surprising, however, is the overall decrease in relative clauses, with the strongest decreases in use occurring in academic writing. Since relative clauses post-modify nouns, we might expect this feature to increase along with other features associated with noun phrase complexity (e.g., phrasal post-modifiers, which exhibited marked increases in frequency of use). Upon further examination, it turns out that this trend is largely accounted for by a decrease in use of *wh*-relatives, the most frequent type of relative clause. Finite relatives with *that* have stayed consistent in fiction and science writing (although they have increased modestly in news), and non-finite relatives have remained consistent or decreased slightly in all three registers. Other clauses associated with noun phrases (*that*- and *to*-noun complement clauses, noun + *of* + *-ing* clauses) have also remained relatively stable in terms of frequency over time.

Overall, Table 5 reflects a pattern of overall decline or stability in the use of clausal subordination and overall increase in the use of phrasal features, with phrasal features which were not already common in the 1700s showing the largest increases. Examples 3 and 4 illustrate these changes in science prose. Head nouns are **bolded** (excluding proper nouns), pre-modifiers are *italicized*, post-modifiers are underlined, and verb phrases are in small caps. (Note: both examples contain 57 words).

Example 3 (from 1825) contains 10 nouns (excluding proper nouns), often with non-finite relative clauses (e.g., *results thus obtained, instruments described in this paper, a clock made by….*), but few of these nouns occur with phrasal pre- or post-modifiers. There are 8 verb phrases in this example, corresponding to independent and dependent clauses.

(3) According to the **results** thus obtained the **piers** of the **transit** were placed; and when, in June 1824, the **instrument** was put upon them, were found to be placed with *considerable* **exactness**. From the *above* **time observations** have been constantly made with the **instruments** described in this **paper**, and with a **clock** made by Molyneux and Cope. [Science RA, 1825]

In contrast, Example 4 (from 2005) contains fewer verbs and clauses, but a higher density of pre- and post-modified nouns, with almost every noun phrase occurring with at least one modifier. While Example 3 contained only 2 phrasal pre-modifiers (*above, considerable*), Example 4 contains fourteen! There are relatively few clausal post-modifiers, but an abundance of prepositional phrases as post-modifiers.

(4) An **example** of such an *enigmatic* **pattern** of *egg size* **variation** is the *strong positive* **correlation** between *body* **size** and *egg* **size**: a **pattern** that occurs in many *animal* **taxa** (reviewed in REFS). Given a *tight* **correlation** between *egg* **size** and *offspring* **success**, suboptimally reduced *egg* **size** in *small* **adults** imposes a *reproductive* **cost** to those **adults**. [Science RA, 2005]

Importantly, while these trends toward increased use of phrasal features hold to some extent in all three written registers examined here, these changes have been most pronounced in academic research writing. That is, academic writing is distinctive in the extent to which phrasal complexity features have been adopted over the past 200 years. The variation across written registers is particularly informative when it comes to interpreting possible functional and situational motivations for such changes. Thus, consider Table 6, which summarizes key situational characteristics of the three written registers examined here:

Table 6. Key situational characteristics of academic writing, news, and fiction

Register	Purpose	Audience
Academic writing	Informational	Expert/specialized
News writing	Informational	Popular
Fiction	Entertainment	Popular

Of the two comparison registers, news reportage shares an important situational characteristic with academic prose: an informational purpose. It is thus not surprising that while the increases in use of phrasal features was most marked in academic writing, newspaper prose also underwent large increases in the use of many of these features (yet still smaller increases than seen in academic writing). The primary situational difference between these two registers is audience: while academic writing has specialized audience with expert knowledge, newspapers are written for a popular audience who is not expected to have specialized background knowledge. Fiction, with a popular audience and a purpose to entertain rather than inform, has adopted some of these feature but to a much lesser extent either of the informational registers.

The results presented in this section indicate general trends of increasing, decreasing, or stable use of these features. However, this analysis does not provide information about the overall frequency of these features, which is important for understanding the magnitude of the changes that registers have undergone. In addition, understanding developments in the typical meanings and functions of features which have increased in use are important for understanding why quantitative shifts have occurred. Thus, in the following sections, we explore descriptive frequencies and functional analyses of two features that have shown to be particularly important in the development of academic writing: nouns as nominal pre-modifiers and prepositional phrases as post-modifiers. Both of these features were relatively rare in the 18th century but have become highly frequent, characteristic features of academic writing.

3.2 Nouns as noun pre-modifiers across written registers

Nouns as nominal pre-modifiers (e.g., *<u>influenza</u> vaccination, <u>daughter</u> cell, <u>life</u> goals, <u>frequency</u> shift, <u>patient</u> history*) exhibited one of the strongest positive correlations with time (.75 in academic writing). Looking at the descriptive frequencies of nouns as pre-modifiers over time also shows that the magnitude of this change is quite high. Figure 2 plots the normalized rate of occurrence for this feature. The rates for attributive adjectives, as well as the rates of occurrence for these two features in news writing, are plotted for the sake of comparison. Attributive adjectives were already frequent in both registers in 1750, and have remained frequent despite some fluctuations over time. In contrast, nouns as nominal pre-modifiers were quite rare in 1750 (occurring less than 10 times per 1,000 words). A steady increase of use was witnessed into the early 1900s, when the frequency began increasing at a much faster rate and continued to the present. In academic writing at the beginning of the 21st century, nouns as pre-modifiers are on par with attributive adjectives, occurring at a rate 8 times more frequent than in 1750. While the use of this feature also steadily increased in newspaper writing, this increase has been much less dramatic and has levelled off in recent decades.

The marked increase in the use of nouns as nominal pre-modifiers has been restricted to academic writing, and to a much lesser extent other informational registers like news. For example, fiction, a non-informational written register, has not witnessed this increase in frequency (with a frequency of 6.2 per 1,000 words in 1750 increasing only to 14.3 in 2000). However, even within academic writing, there are varied patterns in terms of the extent to which this feature has been adopted. Figure 3 presents short term change for the last half of the 20th century and early 21st century (the time period of fastest growth in the use of nouns

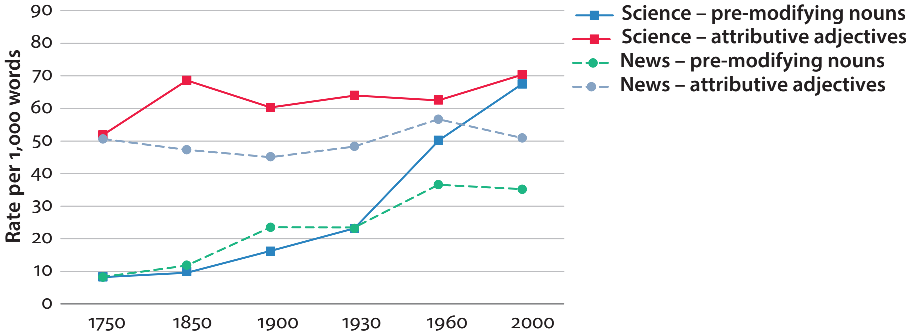


Figure 2. Frequency of pre-modifying nouns and attributive adjectives in academic prose and news from 1750–2000

as pre-modifiers) in four sub-registers of academic writing: specialist science, specialist social science, specialist humanities (operationalized as history texts in this study), and multi-disciplinary (non-specialist) science.

Specialist science has adopted this feature to the greatest extent, following by social sciences. This feature has also increased in multi-disciplinary science; however, this register has lagged a bit behind the specialist registers in terms of overall frequency. Humanities texts written for specialist audiences have also shown small increases in this feature, but the increases have been much more modest: rates for nouns as pre-modifiers increased from 15.8 instances per 1,000 words in 1850 to 24.3 in 2005 for specialist humanities texts (see Biber & Gray 2016: 165). Thus, we can say that specialist science writing appears to be the most 'agile' (in Hundt & Mair's 1999 terminology) of the academic sub-registers in terms of its adoption of this feature, while humanities texts are less so.

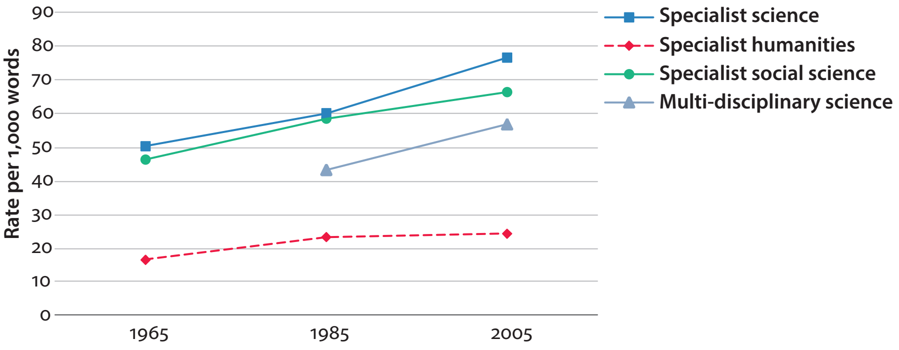


Figure 3. Recent change (1965–2005) in the use of nouns as nominal pre-modifiers across sub-registers of academic writing

Of all of the linguistic developments that we have investigated, the massive influx of nouns as nominal pre-modifiers in academic writing is perhaps the most dramatic change from a quantitative perspective. It turns out that this change is noteworthy not only quantitatively, but also functionally. As the frequency of nouns as nominal pre-modifiers expanded, so did the meaning relationships and discourse functions, moving from straightforward and literal meaning relationships to more abstract and less transparent ones. Over time, noun-noun sequences developed to include a wider range of structural and semantic types of nouns as pre-modifiers and an expanded set of possible meaning relationships. We provide a brief overview of several of these functional extensions here.[6]

One major shift occurred in the semantic classes of nouns that are used to pre-modify other nouns. In the seventeenth and eighteenth centuries, noun-noun sequences were relatively rare, but when they were used, the typically fell into three general semantic categories: titles (Example 5), places and locations (Examples 6, 7), and concrete/tangible nouns (Example 8).

(5) Titles: *Doctor Coulter, Lord Archbishop of Dublin*

(6) Places: *Dumbarton Castle, Greenwich Park*

(7) Locations: *frontier garrisons*, *town wall*

(8) Concrete nouns: *canon ball, goose eggs, tea commissioners*

In the late 1800s, as frequencies of noun-noun sequences began to increase, the typical semantic types of nouns occurring as noun modifiers expanded to include institutions (Example 9), conditions or states of being (Example 10, especially in medical writing), and other abstract/intangible nouns (Example 11). The range of intangible nouns particularly increased in the twentieth century, to the point that almost any noun can now be used to modify another noun.

(9) Institutions: *union member, school proposal*

(10) Conditions: *maternity hospitals, smallpox eruption*

(11) Intangibles: *currency troubles, temperature chart, news agency, sex differences, wage increases*

A second way that noun-noun sequences functionally expanded was to allow nominalizations in either noun slot (i.e., as the modifier or as the modified head noun). Nominalizations in these structures include both morphological derivations

[6] For additional examples of these functions, along with a detailed discussion of noun-noun sequences used in place of ‘s-genitives and of-genitive phrases, see Biber and Gray (2016: Chapter 5).

(Example 12) as well as conversions (Example 13), typically representing actions or processes nominalized from verb forms. Like the expansion in semantic types of nouns, shifts toward the use of nominalizations in these structures began in the 1800s, and continued to expand throughout the 1900s, to include nominalizations from adjectives to express abstract qualities (Example 14).

(12) Derived nominalizations (from verbs):
extradition treaty, inoculation experiments, regression analysis, correlation coefficients

(13) Conversions (from verbs):
murder trials, transport unions, study period

(14) Nominalized adjectives to express abstract qualities:
intelligence agencies, security interests, freedom movement, majority group

As an increasing array of structural and semantic types of nouns came to be used as noun phrase pre-modifiers, the meaning relationships between nouns in noun-noun sequences was also extended, often becoming less transparent. For example, it is fairly straightforward to understand the meaning relationship of a noun phrase like *ground floor* – a floor that is located at ground level. But what about *casualty department* or *safety officials* or *membrane flux?* The relationship between the two nouns vary, and often require background knowledge to explicitly state. These meaning relationships could vary even in the 1800s (e.g., *hen eggs* are eggs that come from a hen, while *sea captains* are captains who specialize in navigating the sea). However, the range of types of relationships greatly expanded during the 20th century. Examples 15–21 illustrate a few of the typical meaning relationships in our 20th century data (for additional meaning relationships, see Biber & Gray 2016: Chapter 5):

(15) A person (N2) who belongs to an institution (N1)
government official, union member

(16) A text (N2) about a topic (N1)
family history, psychology lecture, biology textbook

(17) A person or institution (N2) that regulates N1
awards bureau, safety officials, price commission

(18) An institution (N2) that maintains information about N1
casualty department, intelligence agency

(19) N1 is patient/theme affected by the process indicated in N2, where N1 is the logical subject of N2
eye movement (cf. the eye moved)
justice department intervention (cf. the justice department intervened)

(20) N1 is a patient/theme affected by the process indicated in N2, where N1 is the logical direct object of N2
weight loss (cf. someone lost weight)
trade legislation (cf. someone legislated trade)

(21) N1 is the purpose or topic of N2
trade agreement (cf. an agreement about trade)
freedom movement (cf. a movement about freedom)

A final functional extension of nouns as nominal pre-modifiers occurs with respect to the extent of modification. To this point, we have illustrated this structure with sequences of two nouns. However, during the late 1800s, sequences with three nouns begin to appear (albeit infrequently). In the latter half of the 1900s, however, three-word (Example 22) and even four-word (Example 23) sequences begin to increase in use.

(22) Noun-noun-noun sequences
justice department spokesman, trade boycott campaign, blood glucose level, sinus node dysfunction, chromosome gene product

(23) Noun-noun-noun-noun sequences
life table survival curves, mean plasma glucose value, emergency cabinet committee meetings

The quantitative and functional data presented in this section demonstrate that the use of nouns as nominal pre-modifiers represents a major development that has arisen primarily in written informational language, specifically academic writing. The changing frequencies of use represent much more than simple stylistic shifts, as demonstrated by the magnitude of these changes and the major extensions in the semantic and functional characteristics of these structures.

3.3 Prepositional phrases as noun post-modifiers across written registers

In Section 3.1, we showed that noun phrase post-modifiers exhibited strong patterns of change over time, with relative clauses and *of*-phrases as noun modifiers decreasing in academic writing. In contrast, the use of prepositional phrases with prepositions other than *of* have shown strong *increases* in use as noun phrase post-modifiers. Figure 4 illustrates the magnitude of this shift in science writing, showing that the overall frequency of prepositional modifiers with prepositions other than *of* are as frequent as *of*-phrases in science writing today – occurring 31.5 times per 1,000 words in science writing (cf. a frequency of 30.2 per 1,000 words for noun + *of*-phrases). Figure 4 also includes data points for the rate of prepositional phrases as noun modifiers in present-day news, showing that these structures are substantially more frequent in academic writing than in news. The rates of occurrence for prepositional phrases as noun modifiers in fiction (not displayed) are even lower (14. 4 and 7.4 occurrences per 1,000 words for *of*-phrases and other prepositional phrases respectively).

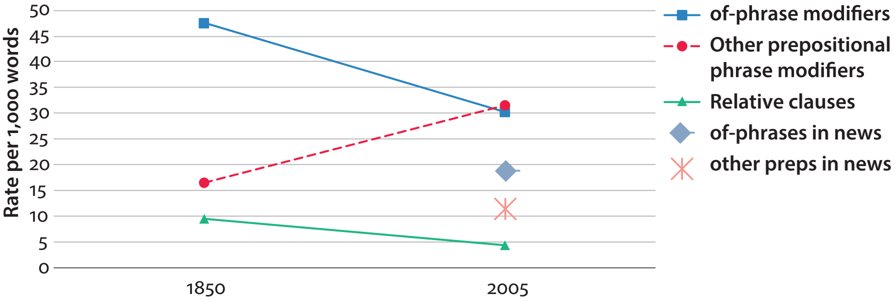


Figure 4. Frequency of prepositional phrases and relative clauses as post-nominal modifiers in science writing (1850–2005), with comparison data for present-day news

In fact, prepositional phrases as noun modifiers are so frequent in academic writing today (i.e., more than 60 per 1,000 words when *of*-phrases and other prepositions are combined) that it is common to see multiple prepositional phrases as noun modifiers in a single clause in academic writing, often in a series. In a series of prepositional phrases, one head noun may be modified by multiple prepositional phrases (as is the case with *difference* in Example 24), or prepositional phrases may be embedded within other prepositional phrases (as in both instances in Example 25). In these examples, nouns modified by prepositional phrases are **bolded**; post-modifying prepositional phrases are bracketed:

(24) Specifically, we were interested in the qualitative ecological **difference** [in **emphasis**] [between **changes** [in composition] vs. **changes** [in relative abundance]].

(25) This **belief** [in a near magical **efficacy** [of effort]] is also manifest at the lower **levels** [of **differentiation** [of luck and skill]].

Figure 4 showed the overall frequency of prepositional phrases as noun post-modifiers at two time periods: 1850 and 2005. Figure 5 displays the frequency of four of the most common prepositions[7] used for this function in science and medical writing from 1750–2005, showing that (like nouns as nominal pre-modifiers)

[7] Figure 5 is based on a hand-coding of twenty texts from science/medical writing in each time period. Common prepositions (*in, on*) were sampled while all occurrences of *for* and *with* were coded for the grammatical function of the prepositional phrase (noun modifier or adverbial/other), and then rates of occurrence were extrapolated.

prepositional phrases as noun modifiers (excluding *of*) were quite rare in 1750. These noun modifiers began increasing in the 1800s, and then increased at a faster rate during the 1900s – with these four prepositions more than doubling by 2005.

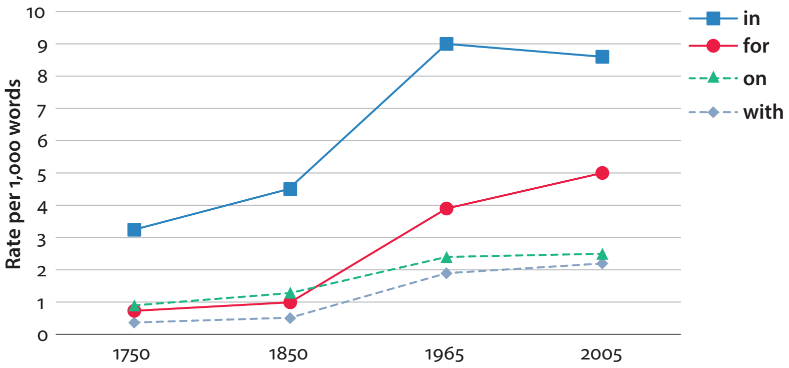


Figure 5. Frequency of selected prepositional phrases as noun modifiers in science/medical writing from 1750–2005

In and *for* have been particularly important for this historical trend, as indicated by strong correlations (see Table 5) and the highest individual frequencies (see Figure 5). It turns out that when we consider sub-registers of academic writing, post-modifying phrases with *in* are particularly common in science research writing (Example 26), while *for* phrases are more common in social science writing (Example 27):

(26) Specifically, we were interested in the qualitative ecological **difference** [IN emphasis] between **changes** [IN composition] vs. **changes** [IN relative abundance]. - Science (Ecology)

(27) A special quantitative attribute is the value of the information **function** [FOR the **items** [FOR a given examinee]] under the item-response theory IRT model - Social science (Education)

Like nouns as nominal pre-modifiers, the quantitative increases in the frequency of prepositional phrases as noun modifiers have been accompanied by extensions in the meaning relationships that the structures convey. One of the major developments that has taken place is the extension of prepositional phrases to convey abstract meanings rather than concrete, often locative, meanings. Biber and Gray (2011, 2016) document that during the 1500–1700s, most instances of *in* and *on* as noun post-modifiers in medical prose were locative[8] (e.g., *a wynde in the head,*

[8] Most prepositional phrases in present-day conversation also have a concrete, locative meaning.

the oil in the thermometer, a postume on the longes), including textual location (e.g., *his judgement and candor in his writings*). Only abstract meanings related to 'topic' were identified in this data (e.g., *a learned author on this subject).*

This is in direct contrast to present-day academic writing, in which more than half of all instances of *in* and *on* as post-nominal modifiers are abstract in meaning (for quantitative data and extended discussion, see Biber & Gray 2016: 190–202). While abstract meanings related to 'topic' remain common, several additional abstract meanings have become prevalent:

(28) Nominalization + *on*
a focus on measures of student outcomes
its reliance on observational instruments

(29) Process noun + *on/in*
its influence on equilibrium density
a causal impact on their behavior at home
increase in the rate of incorporation of radioactivity
a significant change in the metabolism of glucose

(30) Noun + *on/in* + *-ing* clause
an effect on determining choice
emphasis on providing support
assistance in recording electrocardiographs
problems in determining observer reliability

We have illustrated the quantitative and functional shifts of prepositional phrases as noun modifiers with a few common prepositions. However, it should be noted that similar increases can be observed for other prepositions as noun modifiers (e.g., *about, between, from,* and *to*), thus resulting in the overall rate of occurrence of prepositional phrases as noun phrase modifiers seen in Figure 4. Other prepositions also occur in this function with lower frequencies, such as *across, after, against, along, among(st), around, as, at, before, behind, by, into, like, near, over, toward(s), under, upon, within* and *without.* Many of these prepositions also have abstract meanings:

(31) *We examine the relationships between class and country…*

(32) *Immediately after exposure to wasps, pupae were weighed.*

(33) *We used corrected-item-mean imputation to handle missing data at the item level*

(34) *…then the great fight against intuition would be won.*

(35) *…the monopoly over knowledge*

Due to the wide range of prepositions that now occur in the noun-post-modifier function, we argue that one of the major developments in academic writing has been the generalization of this function to the full set of prepositions, rather than being restricted to a small set of more specialized prepositions (e.g., *in* and *on*; see extended discussion in Biber & Gray 2016: 190–192).

The increased use of prepositional phrases to post-modify noun phrases may be complementary to the decreasing use of clausal post-modifiers like relative clauses. Prepositional phrases (like nouns as nominal modifiers) are often more concise and condensed structures, and can usually be paraphrased with a full relative clause. Examples 36–40 illustrate the full, explicitly stated meaning between a head noun and its prepositional post-modifier:

(36) *experiments in India*
c.f. experiments that were conducted in India

(37) *experiments in biological science*
c.f. experiments that are focused on the study of biological science

(38) *restrictions on the underground injection of chemicals*
c.f. rules that restrict the underground injection of chemical

(39) *the returns on these investments*
c.f. the returns that investments produced

(40) *the literature on lactose metabolism*
c.f. literature that contains information about lactose metabolism

In addition to being more concise expressions of information, these prepositional modifiers are less explicit. It is up to the reader to understand the precise meaning relationship between the head noun and the modifier (which a relative clause structure would explicitly state). This task is further complicated by the fact that the same preposition has the potential to express a variety of meanings. In Examples 36–40, although only 2 prepositions were used in the examples (*in* and *on*), each example represents a different type of meaning relationship. Thus, prepositional phrases in this function represent compressed and inexplicit expressions of information.

In sum, prepositional phrases as post-nominal modifiers represent a substantial shift in discourse style in academic writing, both in terms of quantitative increases of use (which have not occurred to the same extent in other written or spoken registers) as well as massive functional extensions. These functional extensions include both the adoption of the structure to convey abstract meaning relationships, as well as an expansion in the set of prepositions which can be utilized in this structure. In addition, the use of multiple embedded prepositional phrases (which represents both a quantitative and functional shift) characterizes academic writing.

4. Summing up: Academic writing as a locus of historical change

In this chapter, we have surveyed general patterns of change in academic writing over the past 200 years, showing that previous portrayals of academic writing as resistant to change were based on a restricted set of features that are not particularly characteristic of academic writing today. We have shown that when register features of present-day academic writing are analyzed from a diachronic perspective, the findings are quite different. Academic writing has changed – and many of those changes have been dramatic quantitative and functional shifts. These changes have been concentrated in the use of phrasal structures embedded in noun phrases, which has resulted in a nominally and informationally dense discourse style.

We further argue that academic writing has in fact been the locus of such historical change. That is, these grammatical changes represent major historical innovations which have been primarily adopted in academic writing, and which encompass both quantitative and functional shifts. When they have occurred in other registers, they have largely been restricted to other written, informational registers like news, yet even news has not adopted these features to the same extent. In addition, there are variations in the exact trajectories of these developments across sub-registers of academic writing. The highest magnitude changes occur in specialist science, followed by specialist social science and then multidisciplinary science. Specialist humanities texts have participated the least of the academic sub-registers that we investigated.

In taking a register theoretical approach, we can explain these findings by considering the unique situational characteristics of academic writing. Above, we illustrated how highly compressed phrasal structures such as nouns and prepositional phrases as nominal modifiers condense information into fewer words. The precise meanings of these compressed phrases are often implicit, requiring the reader to infer the relationships between the head noun and its modifiers. As a register, academic writing has an audience of highly specialized readers, who have a high degree of background knowledge that enables them to quickly decode those meaning relationships. Over time, that specialized audience has become more and more specialized. In the 17th and 18th centuries, scholars publishing research in science and medicine produced articles (often narrative reports) that were read by an entire community of scientists, and sometimes the wider literate public.

During the 19th century, a diversification of academic disciplines and increasingly specialized sub-disciplines began to emerge (e.g., biology, geology, astronomy, physics, chemistry), and this diversification increased dramatically in the 20th century. As the publication of academic research proliferated during the 20th century, research articles came to be read by a narrower audience who shared a specific discipline (and often sub-discipline) with the author. This context, paired with advancements in technology that further facilitated writing and editing processes, provided an environment in which the dense use of phrasal features to encode complex information thrived. News writing, which did not share the specialized audience and increasing specialization, thus did not adopt these compressed and inexplicit features to the same extent.

Our goal in this paper has been to argue that the unique situational characteristics of any register have the potential to lead linguistic change, countering the stereotype that most linguistic change originates in speech and that academic writing has largely resisted linguistic developments that have occurred in other spoken and written registers. The evidence presented here supports this argument, and emphasizes the crucial role of register in research on historical change in language. Specifically, we argue that register and the selection of linguistic features for diachronic studies must be aligned to enable comprehensive accounts of linguistic change. For academic writing, this means considering not only colloquial and clausal features which are characteristic of spoken language, but also grammatical features which are phrasal in nature and represent the major discourse style of present-day academic writing.

References

Aktas, Rahime Nur & Cortes, Viviana. 2008. Shell nouns as cohesive devices in published and ESL student writing. *Journal of English for Academic Purposes* 7: 3–14. https://doi.org/10.1016/j.jeap.2008.02.002

Aronoff, Mark. 1985. Orthography and linguistic theory: The syntactic basis of Masoretic Hebrew punctuation. *Language* 61(1): 28–72. https://doi.org/10.2307/413420

Baratta, Alexander. 2009. Revealing stance through passive voice. *Journal of Pragmatics* 41: 1406–1421. https://doi.org/10.1016/j.pragma.2008.09.010

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511621024

Biber, Douglas. 1992. On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes* 15: 133–163. https://doi.org/10.1080/01638539209544806

Biber, Douglas. 2001. Dimensions of variation among 18th century registers. In *Towards a history of English as a History of Genres*, Hans-Jürgen Diller & Manfred Görlach (eds), 89–110. Heidelberg: C. Winter. (Reprinted in Susan Conrad & Douglas Biber (eds) 2001, 200-214).

Biber, Douglas. 2006. Stance in spoken and written university registers. *Journal of English for Academic Purposes* 5: 97–116. https://doi.org/10.1016/j.jeap.2006.05.001

Biber, Douglas & Conrad, Susan. 2009. *Register, Genre, and Style*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511814358

Biber, Douglas & Finegan, Edward. 1989a. Drift and the evolution of English style: A history of three genres. *Language* 65: 487–517. https://doi.org/10.2307/415220

Biber, Douglas & Finegan, Edward. 1997. Diachronic relations among speech-based and written registers in English. In *To Explain the Present: Studies in Changing English in Honor* *of Matti Rissanen*, Terttu Nevalainen & Leena Kahlas-Tarkka (eds), 253–276. Helsinki: Société Néophilologique. [reprinted in Susan Conrad & Douglas Biber (eds) 2001, 66–83]

Biber, Douglas, Finegan, Edward & Atkinson, Dwight. 1994. ARCHER and its challenges: Compiling and exploring A Representative Corpus of Historical English Registers. In *Creating and Using English Language Corpora*, Udo Fries, Gunnel Tottie & Peter Schneider (eds), 1–14. Amsterdam: Rodopi.

Biber, Douglas & Gray, Bethany. 2010. Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes* 9: 2–20. https://doi.org/10.1016/j.jeap.2010.01.001

Biber, Douglas & Gray, Bethany. 2011. Grammatical change in the noun phrase: The influence of written language use. *English Language & Linguistics* 15(2): 223–250. https://doi.org/10.1017/S1360674311000025

Biber, Douglas & Gray, Bethany. 2016. *Grammatical Complexity in Academic English: Linguistic Change in Writing*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511920776

Biber, Douglas, Conrad, Susan & Reppen, Randi. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511804489

Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan & Finegan, Edward. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Bybee, Joan & Hopper, Paul. 2001. *Frequency and the Emergence of Linguistic Structure* [Typological Studies in Language 45]. Amsterdam: John Benjamins. https://doi.org/10.1075/tsl.45

Charles, Maggie. 2003. 'This mystery…': A corpus-based study of the use of nouns to construct stance in theses from two contrasting disciplines. *Journal of English for Academic Purposes* 2: 313–326. https://doi.org/10.1016/S1475-1585(03)00048-1

Crespo García, Begoña & Moskowich-Spiegel Fandiño, Isabel. 2010. CETA in the context of the *Coruña Corpus*. *Literary and Linguistic Computing* 25(2): 153–164. https://doi.org/10.1093/llc/fqp038

Croft, William. 2000. *Explaining Language Change: An Evolutionary Approach*. London: Longman.

Denison, David. 1998. Syntax. In *The Cambridge History of the English Language*, Vol. IV, 1776–1997, Suzanne Romaine (ed.), 92–329. Cambridge: CUP.

Fairclough, Norman. 1992. *Discourse and Social Change*. Cambridge: Polity.

Fennell, Barbara. 2001. *A History of English*. Malden MA: Blackwell.

Fillmore, Charles. 1981. Pragmatics and the description of discourse. In *Radical Pragmatics*, Peter Cole (ed.), 143–166. New York NY: Academic Press.

Flowerdew, John & Forest, Richard W. 2014. *Signalling Nouns in Academic English: A Corpus-Based Discourse Approach*. Cambridge: CUP.

Gray, Bethany. 2015. *Linguistic Variation in Research articles: When Discipline Tells only Part of the Story* [Studies in Corpus Linguistics 71]. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.71

Groom, Nicholas. 2005. Pattern and meaning across genres and disciplines: An exploratory study. *Journal of English for Academic Purposes* 4: 257–277. https://doi.org/10.1016/j.jeap.2005.03.002

Halliday, Michael A. K. 1988. On the language of physical science. In *Registers of Written English*, Mohsen Ghadessy (ed.), 162–178. London: Pinter. (reprinted in Halliday 2004).

Halliday, Michael A. K. & Martin, James R. 1993. *Writing Science: Literacy and Discursive Power*. London: Falmer.

Hewings, Martin & Hewings, Ann. 2002. "It is interesting to note that...": A comparative study of anticipatory 'it' in student and published writing. *English for Specific Purposes* 21: 367–383. https://doi.org/10.1016/S0889-4906(01)00016-3

Hopper, Paul J. & Traugott, Elizabeth C. 2003. *Grammaticalization*, 2nd edn. Cambridge: CUP. https://doi.org/10.1017/CBO9781139165525

Hundt, Marianne & Mair, Christian. 1999. "Agile" and "uptight" genres: The corpus-based approach to language change in progress. *International Journal of Corpus Linguistics* 4: 221–242. https://doi.org/10.1075/ijcl.4.2.02hun

Hyland, Ken & Tse, Polly. 2005. Hooking the reader: A corpus study of evaluative *that* in abstracts. *English for Specific Purposes* 24: 123–139. https://doi.org/10.1016/j.esp.2004.02.002

Hyland, Ken. 1998. Boosting, hedging and the negotiation of academic knowledge. *Text* 18(3): 349–382. https://doi.org/10.1515/text.1.1998.18.3.349

Krug, Manfred. 2000. *Emerging English Modals: A Corpus-based Study of Grammaticalization*. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110820980

Leech, Geoffrey, Hundt, Marianne, Mair, Christian & Smith, Nicholas. 2009. *Change in Contemporary English: A Grammatical Study*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511642210

Mair, Christian. 2006. *Twentieth-century English: History, Variation and Standardization*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511486951

McWhorter, John. 2001. *The Word on the Street: Debunking the Myth of "Pure" Standard English*. New York NY: Basic Books.

Tagliamonte, Sali. 2004. Have to, gotta, must: Grammaticalisation, variation and specialization in English deontic modality. In *Corpus Approaches to Grammaticalization in English* [Studies in Corpus Linguistics 13], Hans Lindquist & Christian Mair (eds), 33–55. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.13.04tag

Wells, Rulon. 1960. Nominal and verbal style. In *Style in Language*, Thomas A. Sebeok (ed.), 213–220. Cambridge: CUP.

Xiao, Richard. 2009. Multidimensional analysis and the study of world Englishes. *World Englishes* 28(4): 421–450. https://doi.org/10.1111/j.1467-971X.2009.01606.x

Yáñez-Bouza, Nuria. 2011. ARCHER: past and present (1990–2010). *ICAME Journal* 35: 205–36.

PART III

Genre-based analyses of linguistic phenomena

The importance of genre in the Greek diglossia of the 20th century

A diachronic corpus study of recent language change

Georgia Fragaki & Dionysis Goutsos
National and Kapodistrian University of Athens

Greek diglossia has been mainly studied so far by focusing on linguistic attitudes rather than investigating actual use. This paper aims at studying evidence from a number of genres in the *Diachronic Corpus of Greek of the 20th Century*, including academic texts, public speeches, film scripts, newsreels, literature, song lyrics and private letters. The frequencies of the high vs. low variants of two pairs of basic grammatical items are compared across the nine decades of the corpus. It is suggested that recent language change in Greek largely depends on genre, which can account, among other factors, for the direction and timing of developments. Dimensions of text production like formality, conventionality, planning and (perceived) spontaneity contribute to the specificities of each genre.

1. Introduction

Similar to many other periods in the language's history, twentieth century Greek has been characterized by the split between a low and a high variety, that is *dhimotikí* or *demotic* (L, the language of the 'people') and *katharévousa* (H, the 'purified' language). This was the outcome of the so-called "language question", the issue of standardization of the Modern Greek state's language, which was to be created after the heroic and devastating War of Independence (1821–1832). Whereas before 1821 there was no standard, but there seemed to be shared norms,[1] the language question becomes especially prominent at the end of the 19th century owing to, as Mackridge puts it, "the desire to develop a written language that would reflect an ideal national image that would in turn embody and express the relationship of

[1] Petrounias believes that a "common dialect evolved in the 19th c. which was used both in oral and written communication" (1978: 201) and calls this Pre-Common Modern Greek (PCMG).

https://doi.org/10.1075/scl.85.07fra

the modern Greeks to the ancients" (2009: 2). Thus, the beginning of the 20th century coincides with the bloody clashes of the University of Athens students with the police over the translation of the Gospels into the demotic, while more than a century of debate comes to an end in 1976 with the declaration of the demotic as the official language of education in Greece.

The split between the demotic and *katharévousa* came to be understood in terms of Ferguson's (1959) definition of diglossia, which takes Greek to be one of its "defining languages", along with Standard vs. Swiss German, Standard Arabic vs. Arabic vernaculars and Standard French vs. Creole in Haiti. In this classic description, diglossia refers to "a relatively stable language situation", in which there is a superposed variety "which is learned largely by formal education and is used for most written and formal spoken purposes but is not used by any section of the community for ordinary conversation" (Ferguson 1959: 336). However, the vast literature and the debates that followed from Ferguson's conception of diglossia and its application to the Greek case (see e.g. the entries in Fernández 1993) has anything but clarified our understanding of what Greek has been like in the twentieth century. Instead, a number of important questions still remain without a satisfying answer.

First of all, provided that different language varieties can be distinguished for Greek, how many of these have there been in the twentieth century? Whereas Fergusonian diglossia recognizes only two varieties, i.e. the high vs. the low variety, in a very influential paper Alexiou has pointed out that Greek demotic and *katharévousa* "cannot be divided into two absolutely watertight compartments, each with a separate set of syntactic, morphological and lexical structures" (1982: 162–163; cf. Holton 2002, Mackridge 2009: 29). This clearly implies the possibility that no specific answer to the above question can be provided in terms of number of varieties. The need to clearly separate varieties has led Mirambel (1937) to make a five-part distinction between "états de langue" for the time before World War II, while in the 1960s Pappageotes and Macris (1964) distinguish four types of spoken and seven types of written Greek, and Papatzikou-Cochran believes that "Greece was essentially a triglossic environment from the turn of the twentieth century through its third quarter" (1997: 42). On the other hand, Frangoudaki (2002: 106) claims that the Greek case since the beginning of the twentieth century cannot be analyzed in terms of diglossia at all, but as a function of symbolic control and social power, something which again implies that twentieth century Greek cannot be analyzed in terms of diglossia at all.

In relation to the above, Mackridge suggests that "actual language used in Greece covered a continuum of linguistic registers ranging from 'pure' demotic to 'extreme' *katharévousa*, with hybrid varieties in between" (2009: 29). Assuming that it is possible to isolate clearly identifiable registers at distinct in-between linguistic levels, what exactly has been their range, as well as what were their domains of application and their defining characteristics? Could it be that the crucial distinction between H and L is that of mode, according to Halliday's parameters of register (Halliday & Hasan 1990), that is, do the Greek demotic and *katharévousa* correspond to the spoken and the written mode? Mackridge (2004: 112), for instance, thinks that Greek diglossia is the coexistence of two varieties of Greek for written purposes, whereas there has been "a fundamental misunderstanding of Greek diglossia, which has usually been presented as a conflict between spoken and written usage".

A third set of questions concerns the timing of changes (cf. Nevalainen & Raumolin-Brunberg 2003: 56). Was *katharévousa* superseded by the demotic in the course of the twentieth century in a gradual fashion or at one fell swoop? When did specific changes take place? Most pronouncements on when changes happened in descriptions of Greek diglossia seem to be based on individual speculation.

Finally and most importantly, the vast majority of studies of Greek diglossia refer to speaker attitudes rather than their linguistic practices or, in Hudson's (2002) terms, to sociological or contextual rather than grammatical or linguistic aspects of diglossia. However, as Daltas correctly points out, "what is societally recognized as diglossia is not directly related to how actual people implement the relevant contrasts" (1994: 348). We thus lack specific details on what actual people did in specific contexts throughout the previous century.[2] As a result, we also do not know how what Daltas calls the public aspect of diglossia, that is the "legal" system, influenced the private aspect or "the way people respond to it in actual life" (1994: 341–342). How did language attitudes, policies and prescriptions really affect the way people spoke or wrote in specific circumstances? Did the public perception of diglossia correspond to the private use of Greek by individual speakers and writers?

Furthermore, these confounding parameters must be placed within the historical context of their era. Greek in the 20th century appears as the ending point of a long and complicated history, dominated, as Horrocks notes (2010: 3), by the issue of diglossia, which first appeared as a split between the spoken and the written varieties of the language because of the "fossilizing effect" that the early emergence of the Attic classical canon had on Greek (ibid.). The subsequent Atticist movements from the Hellenistic Koine to the Byzantine and post-Byzantine

[2] It is characteristic that both Alexiou (1982: 156) and Mackridge (2009: 29) refer to "the actual spoken and written usage of Greek" and the "actual language used in Greece", without, however, referring to any language data.

periods are related to the origins of modern diglossia (ibid: 100), which appeared when the issue of the language's standardization came up along with the formation of the Modern Greek state, as noted above. As a result, issues of language choice became inextricably linked with issues of national and political identity. As by the end of the nineteenth century the varieties of *katharévousa* and the demotic became polarized (Mackridge 2009: 242ff.), Greek language standardization seems to have been left to the vagaries of the political situation. For instance, *katharévousa* becomes Greece's official language and language of education in 1911, 1920 and later by right-wing governments, whereas the demotic is introduced in public education in 1917, 1923 and 1964 and only becomes the state's official language in 1976, following the demise of what was to be the longest and last dictatorship of the 20th century (1967–1974).

It would be presumptuous to suggest that complex questions like those concerning Greek diglossia can have a simple answer, especially on the basis of data concerning only few linguistic items. Our purpose in this paper is rather to investigate how evidence from a number of genres in a diachronic corpus of authentic texts can shed light on the study of Greek in the 20th century. This is the first analysis of its kind since, as mentioned above, we lack systematic evidence from the actual use of Modern Greek in various contexts in the recent past. As Goutsos et al. (1994) have already noticed, the tradition of empirical linguistic research on Modern Greek is rather limited; thus, in relation to diglossia, only Papatzikou-Cochran (1997) and Iordanidou (e.g. 1999, 2009) make use of spoken and written (journalistic) corpora respectively. However, these are not electronic corpora and as such are not easily accessible. In addition, their data are restricted to a period from the 1980s to the 2000s, precluding any long-scale investigation. Iordanidou (2002: 242) herself notes that the confusion around issues of standardization arises from the lack of evidence from corpora. This lack of evidence seriously impinges both on the question of standardization and the broader issues of the diachronic development of Greek.

Our research is based on the *Diachronic Corpus of Greek of the 20th Century*, developed in order to gain a more informed understanding of the language's history in the 20th century. Our approach is thus a historical corpus linguistic one and adopts a variationist view of language change (Milroy 1992: 123) by paying special attention "to the role played by textual and discourse factors across the centuries" (Rissanen et al. 1997: v). In the following sections we present our data and methodology and then proceed to the analysis of our findings from the corpus with regard to two pairs of grammatical items across seven genres and along nine decades of the 20th century (1900–1989). In the final section of this paper we discuss the implications of our findings for the role of genres in recent language change.

2. Data and methodology

The *Diachronic Corpus of Greek of the 20th Century* (Greek Corpus 20) has been developed at the National and Kapodistrian University of Athens for the study of recent language change in Greek. The research project for its development has had the following aims:

a. to examine the issues involved in the compilation of a diachronic corpus of Greek of the 20th Century, including the availability of data across decades, the availability and continuity of text types, and the issue of representativeness;
b. on the basis of exploration of data sources, to collect data from a variety of genres of 20th century Greek from the 1900s to the 1980s, designed to be integrated with the existing synchronic 30 million word *Corpus of Greek Texts* (CGT, Goutsos 2010), which includes data from 1990 to 2010; and
c. to analyze the corpus with a view to drawing basic conclusions on linguistic change across the decades of the 20th century.

In this paper we analyze approximately 3 million words from seven different genres of Greek Corpus 20, including spoken news (newsreels), public speeches, film scripts, literature, song lyrics, academic texts and private letters. We compare the frequencies of the high vs. low variants of two pairs of grammatical items, namely the purposive preposition *διά* /ðiˈa/ vs. *γιά/για* /ja/ 'for' and the locative/allative preposition *εἰς* /is/ vs. *σέ/σε* /se/ 'in/at' respectively. We searched for all possible variations of these items, including their occurrence in the polytonic or the monotonic orthographic system, that is with or without diacritics (e.g. *γιά* vs. *για*, *εἰς* vs. *εις*), as well as their contracted versions (e.g. *δι', γι', σ'*) and their composite forms (e.g. *στον, στη, στις* etc, by searching for *στ**).

We have decided to concentrate on recent language change in the grammatical words of Greek, since, as is well-known, these are extremely frequent; for instance, the variants studied here are all found, at least in their basic form, among the first fifty items of the corpus' frequency word list. In addition, grammatical items cover a wide range of linguistic functions, including purpose, reference etc. for *ði'a* and *ja* and location, movement, recipient etc. for *is* and *se*, and are thus difficult to avoid in extended texts. Although we have studied several grammatical items, we have decided to focus here only on the two most frequent pairs.[3]

[3] The items studied include *ἵνα* /ˈina/ 'in order to', *ποῖος* /pˈios/ vs. *ποιος* /pços/ 'who', *εἷς* /ˈis/ vs. *ένας* /ˈenas/ 'a, one', *διατί* /ðiaˈti/ vs. *γιατί* /jaˈti/ 'because, why', which all appear among the top 200 most frequent items of the corpus. The pairs studied here belong to the first 50 most frequent items of the corpus.

Table 1 gives details on the genres and their specific text types which have been selected from Greek Corpus 20 to be studied in this paper, including the number of words in each genre.

Table 1. Data from *Greek Corpus 20* analyzed in this paper

Genres	Text types	Number of words
Spoken news	Newsreels	78,441
Public speeches	Parliament, Academic, Other	339,194
Conversation	Film scripts	208,206
Literary	Novels, Short stories, Poetry, Drama	1,086,687
Literary	Song lyrics	202,863
Academic	Humanities, Social/Finance, Science	1,044,200
Private	Letters	178,720
TOTAL		**3,138,311**

As can be seen, the number of words for each genre is not equal and this is why we have studied both raw and normalised (per 10,000 words) frequencies. Our findings in the following section are based on normalised frequencies.

While it is true that the text types studied here mainly belong to non-spontaneous genres, it is important to note that most of these belong to speech-related genres, in terms of Culpeper and Kytö (2010). In particular, private letters and song lyrics belong to *speech-like* genres, public speeches are *speech-based*, while films and newsreels belong to *speech-purposed* genres. The latter applies to dramatic texts that are included in the literature genre, whereas prose fiction with speech presentation from literary texts belongs to speech-like genres, according to Culpeper and Kytö (2010: 18). These genres have been related to spoken interaction and are thus as close as possible to the "actual" language used throughout the previous century, for which extensive archives of spoken data are missing.

3. Grammatical words in diachrony

Our findings from the investigation of the two pairs of H and L variants in Greek Corpus 20 are presented in the form of curves, based on normalised frequencies (per 10,000 words) as they progress in the intervals of nine decades which are covered in the corpus. Figures 1 and 2 present the overall distribution of the frequencies of the two pairs respectively in all genres over real time. Red lines are used for the H variants and blue lines for the L ones, which eventually came to replace the former.

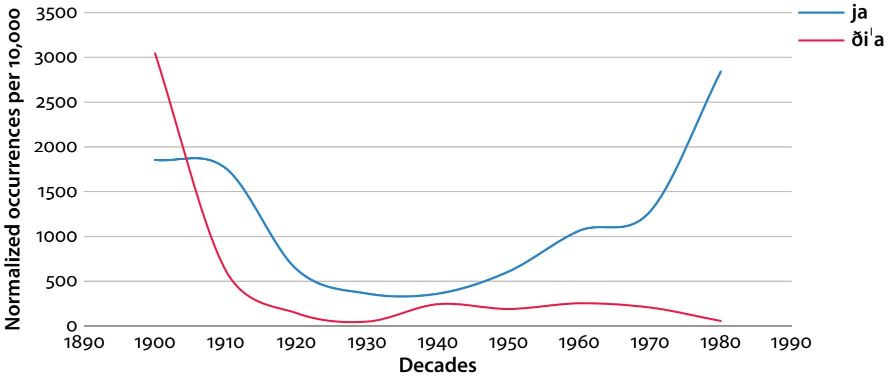


Figure 1. Frequency of variants *ði'a* (H) and *ja* (L) in all genres of *Greek Corpus 20*

Starting from the pair of variants *ði'a* (H) and *ja* (L), an interesting picture emerges as regards language change in Greek over the course of the 20th century. As illustrated in Figure 1, the two variants do not start off at opposing positions at the beginning of the century, but, although the H form is clearly more frequent, both are found in mid-range as items in competition. A sharp decline of both variants then follows that almost equalises their frequencies in the next three decades (1910s to 1930s) with a slight preference for the L variant, while the 1940s mark the beginning of a sharp increase for the L variant with the H variant slowly fading out to disappearance, especially after the 1970s.

This is a surprising picture of language change, since the item to be linguistically entrenched does not follow the expected pattern of incipient to completed change, but is found in vigorous competition at the beginning of the twentieth century and declines before its frequency starts picking up towards the end of the century. Thus, instead of the familiar S-curve in sociolinguistics (Labov 1994: 65; Nevalainen & Raumolin-Brunberg 2003: 53 ff.; cf. Aitchison 1981: 100), the variant to prevail shows a U-curve with its high points at the beginning and end of the century and its low point in the middle.

That the U-curve is not restricted to this pair is suggested by Figure 2, which illustrates the change in frequency for the pair of variants *is* (H) and *se* (L) in all genres studied here.

A U-curve appears in the prevailing L variant of this pair, which, although it starts slightly higher than the H variant at the beginning of the century, slumps towards the middle and then picks up in frequency so as to end at an extreme high

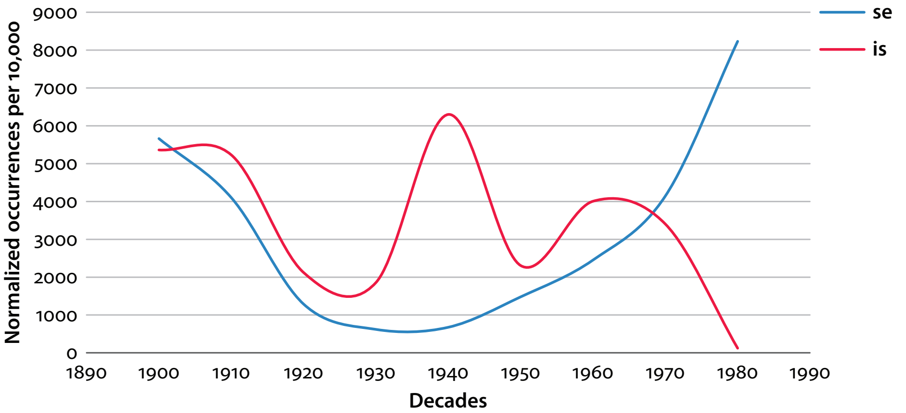


Figure 2. Frequency of variants *is* (H) and *se* (L) in all genres of *Greek Corpus 20*

by the end of the century. Instead, the H variant of the pair shows a "roller-coaster" pattern, starting from a high in the 1900s and then remains in prevalence but gradually declines, although with two unexpected peaks, a sharp one in the 1940s and a smaller one in the 1960s. The 1970s is a crucial turning point after which the H variant shows a sharp decline and the L one a sharp increase.

It is perhaps not unusual that language change in diglossic situations would diverge from expected patterns. In addition, it must be constantly kept in mind that our investigation only looks at the most recent stage in what is a long history of change; thus, the variants *is* and *se* co-occur in the same texts since at least late medieval Greek,[4] while here we are focussing on the ultimate stages of diglossia before its official resolution. However, the precise ways in which language change happens, as well as the specific points in time in which it occurs, can only be found with recourse to corpus data, as is the case with our study. In addition, one might expect the H and L variants of a pair to be in competition with each other throughout a time span, but their common decline at certain periods and the particular direction of change are not observable without recourse to corpus data.

How can we account for these unexpected patterns of change with regard to grammatical items in Greek of the 20th century? It is important to clarify at the outset that patterns like the ones presented in Figures 1 and 2 above emerge from the contribution of several genres, whose individual development may be widely different. As the lines representing language change constitute aggregates for what happens in individual genres, we should refrain from generalizing about what

[4] For instance, in the first few lines of *Erotokritos*, the early 17th century romance in verse, we find *στο Καλό κ' εις το Κακό* /sto ka'lo c is to ka'ko/ 'to the Good and the Evil'.

happened in language as a whole and turn our attention to individual genres. For instance, Figure 3 presents the change in frequencies of both pairs of H and L variants *ði'a* vs. *ja* and *is* vs. *se* in only one genre, that of academic texts.

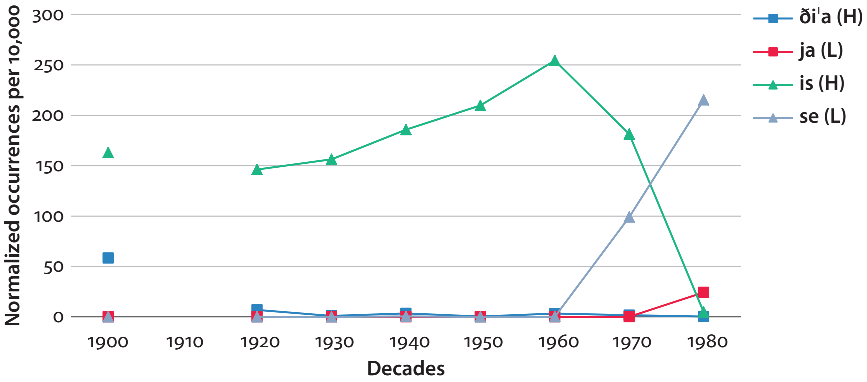


Figure 3. Frequency of H and L variants in academic texts

The picture here is quite different from the overall pattern seen earlier.[5] The H variants seem to be the only options available to writers of academic texts up to the 1960s (for *is* and *se*) or the 1970s (for *ði'a* and *ja*). At those points there is a simultaneous sharp increase of the L variants and sharp decline of the H variants. This seems to suggest relatively stable conventions of this text type over time, which place it at the H end of a possible continuum, with sudden change taking place over two decades that ends up with the complete reversal of the situation in preference of the L variants.

The observed sudden change for the grammatical items studied reflects the movement from an entirely H variety, as in Example (1) from the 1940s, to an entirely L variety, as in Example (2) from the 1980s.

(1) Δυστυχῶς ὅμως αἱ προσπάθειαι τοῦ Βαλκανικοῦ Συμφώνου καὶ τῆς Βαλκανικῆς συνεννοήσεως, εἰς οὐδὲν θετικὸν ἀποτέλεσμα κατέληξαν, ἕνεκα λόγων οὓς δὲν δυνάμεθα σήμερον καὶ εἰς τὴν παροῦσαν μελέτην νὰ ἐξετάσωμεν. (WAB14-1940-0001)
Unfortunately, however, the endeavours of the Balkan Treaty and the Balkan entente ended up to (H) no positive outcome, because of reasons which we cannot examine at present and in (H) this current study.

[5] As in the following figures, the lines are interrupted at those decades when there is not enough data in the genre concerned to allow for statistical comparisons.

(2) Προς αυτήν την κατεύθυνση βοήθησαν τόσο η δημοσίευση του Κώδικα της Ομοσπονδιακής Φορολογίας (1913), όσο και η δημιουργία του Ομοσπονδιακού Συστήματος Αποθεμάτων (Federal Reserve System, 1914) με συνέπεια την αύξηση των πληροφοριακών αναγκών για δημοσίευση χρηματοοικονομικών καταστάσεων και μάλιστα βελτιωμένου περιεχομένου. (WAB14-1980-0001)
To this direction contributed both the publication of the Code of Federal Taxation (1910) and the creation of a Federal Reserve System (1914), with the result of an increase in information needs for (L) the publication of finance states, and more so of improved content.

For instance, in Example (1) various H variants, such as *αί* 'the' instead of *οι* (L), *προσπάθειαι* 'endeavours' instead of *προσπάθειες* (L), *συνεννοήσεως* 'entente' instead of *συνεννόησης* (L), οὐδὲν 'no' instead of *κανένα* (L), *ἕνεκα* 'because' rather than *εξαιτίας* (L), *οὓς* 'which' instead of *τους οποίους* or *που* (L), *δέν δυνάμεθα* 'we cannot' instead of *δεν μπορούμε* (L), co-occur with the H variant *is*. The reverse can be seen in Example (2) from the 1980s, in which no H variant is found. The L variant *ja* co-occurs with items such as *προς* 'to' rather than *εις*, *αυτήν την* 'this' instead of *τοιαύτην*, *κατεύθυνση* 'direction' instead of *κατεύθυνσιν*, *δημοσίευση* 'publication' instead of *δημοσίευσις* etc.

Similar trends appear in the genre of public speeches, as illustrated in Figure 4.

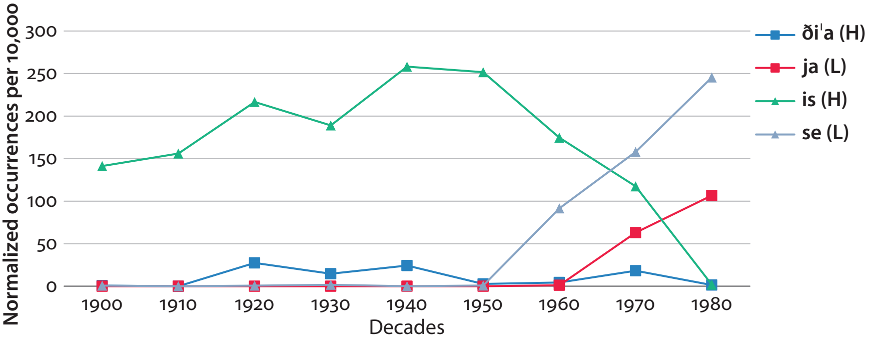


Figure 4. Frequency of H and L variants in public speeches

In public speeches the L variants are practically non-existent up to the 1950s in the case of *is* and *se* and the 1960s for *ði'a* and *ja*; throughout this whole period the H variants seem to be the only options available in the genre. At those points in time there is a sharp increase of the former and a sharp decline of the latter. Again, what seems to be suggested here is a complete reversal of conventions for this genre, which replaces H with L variants. This is a publicly performed genre and, as in the case of academic texts, it is expected to follow certain rigid assumptions of formality. As a result, the relation between the H and L varieties is closer to an absolute dichotomy rather than a gradual displacement of one by the other.

The opposite picture emerges when one looks at the genre of literary texts. Figure 5 presents the change in frequency of both pairs of H and L variants in the genre of literature, including novels, short stories, poetry and drama, but excluding song lyrics.

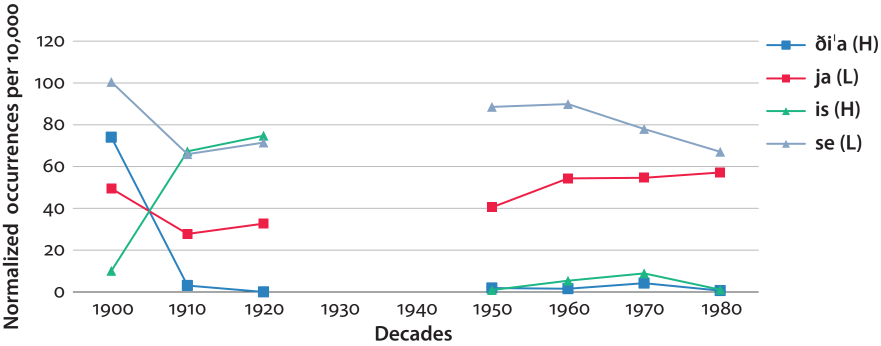


Figure 5. Frequency of H and L variants in literary texts

As can be seen, at the beginning of the century H and L variants seem to be in competition with each other and then, starting from the 1910s/1920s, H variants seem to vanish and L variants remain as the only option available to literary writers. (Since data is missing for the decades of the 1930s and 1940s we have to extrapolate on the basis of the general shape of the lines involved.) This implies an early standardization of the language used in literature in preference of the demotic variety, which clearly prevails throughout the 20th century. However, a few occurrences of H variants are found in the 1960s and the 1970s, presumably for stylistic effects.

This picture is even more extreme for the text type of song lyrics, the data for which is presented in Figure 6. This is clearly the other end of the continuum, with almost no occurrences of the H variants throughout the whole of the century. It appears that, irrespective of singular fluctuations in the frequency of L variants, song lyrics follow clearly established conventions that know of no alternative variants.

A similar patterning appears in film dialogues, which represents simulated conversations. As seen in Figure 7, there is a steady higher frequency of L variants in lines that are almost flat, something that implies that special conventions have been developed in this genre aiming at imitating authentic speech. At the same time, some occurrences of H variants are found at the beginning of the genre

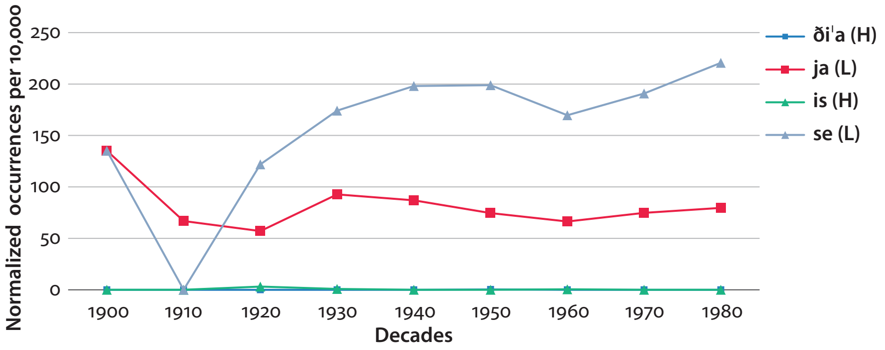


Figure 6. Frequency of H and L variants in song lyrics

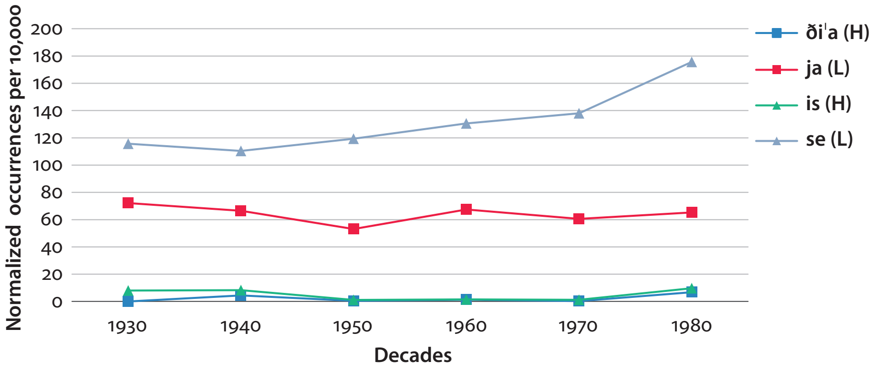


Figure 7. Frequency of H and L variants in film dialogues

(sound films made their appearance in Greece in the 1930s) up to 1950s, as is expected, whereas the slight rise of H variants in the 1980s are due to the inclusion of a specific film in the data which caricatures *katharévousa*.[6]

Although Examples (3) and (4) come from two different decades of the 20th century, namely the 1930s and the 1970s respectively, they do not exhibit differences regarding their preference towards L variants such as *se* and *ja*.

(3) <Μηνάς> ΓΕΙΑ ΣΟΥ ΑΦΕΝΤΙΚΟ <Λουκάς> ακούστε (.) αύριο επειδή είναι γιορτή δουλειά δεν έχει (.) σας περιμένω λοιπόν όλους το βράδυ στο τραπέζι που δίνω για τη γιορτή της γυναίκας μου <Όλοι> ΝΑ ΣΑΣ ΖΗΣΕΙ (.) ΚΑΙ ΒΕΒΑΙΑ ΑΦΕΝΤΙΚΟ ΚΑΙ ΒΕΒΑΙΑ @@ <Λουκάς> ευχαριστώ (.)

[6] As such, this is a parallel to the slight rise of H variants in literary texts at the end of the century for stylistic effects, indicating perhaps a relaxing attitude towards the H variety.

άκουσε:: Θανάση (.) πάρε αυτούς τους δύο για απόψε το βράδυ στο σπίτι
να φάνε και να κοιμηθούνε <Θανάσης> πολύ καλά (SFF19-1930-0001)
<Minas> HI BOSS <Lukas> listen (.) tomorrow is a holiday there's no work
(.) so I'm inviting you all to (L) the dinner I throw for my wife's name day
<All> MANY HAPPY RETURNS (.) OF COURSE, BOSS, OF COURSE
<Lukas> thank you (.) listen Thanassis (.) take these two to (L) your place to
have something to eat and sleep for (L) tonight <Thanassis> very well

(4) <Φάνης> πάψε μωρέ μη γρουσουζεύεις και μας πάθουνε κανένα κακό (.)
πήγαιν' τα πράματα μέσα (.) και ρίξε και μια ματιά στο φούρνο που έχω
τις πατάτες να μη μου καούνε <Στέφανος> ΔΕΝ ΞΕΡΩ ΑΠΟ ΟΙΚΙΑΚΑ
(SFF19-1970-0003)
<Fanis> shut up, you, don't put a jinx on them and something bad happens
(.) take these things inside (.) and have a look at the potatoes in (L) the oven
so they don't burn <Stefanos> I'M NOT FAMILIAR WITH HOUSEHOLD
MATTERS

It is worth noting that verbal inflection, e.g. in *κοιμηθούνε* 'sleep' instead of *κοιμηθούν* in (3) and *καούνε* 'burn' instead of *καούν* in (4), phonological variation such as *πράματα* 'things' instead of *πράγματα*, as well as the collocations found in both texts, e.g. *δίνω τραπέζι* 'throw dinner' in (3), *ρίξε μια ματιά* 'have a look' in (4), are steadily related to L variety regardless of the decade the text belongs to.

Turning now to newsreels, a genre that is written with the purpose of being publicly performed, the situation is rather different as can be seen in Figure 8.

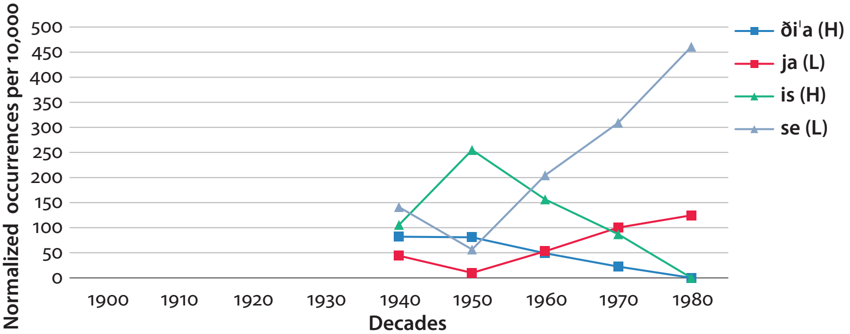


Figure 8. Frequency of H and L variants in newsreels

Newsreels were short documentary films prevalent between the 1940s to the 1980s, regularly released in Greek (winter or summer/open) cinemas and later on television. They last around 10 minutes, cover a variety of topics, with an emphasis on the promotion of government propaganda and were produced by military or government offices dedicated to this purpose.[7] As such, they are expected to exhibit strong prescriptive bias in the language used, according to the prevailing political situation.

The overall picture appearing in Figure 8 largely confirms this, as intercrossing lines seem to correspond to heavy prescription, especially in the 1950s, a decade marked by rigid conservative governments. However, L variants co-occur with H variants in the 1940s and, despite their decline in the 1950s, they gradually rise in the 1960s with a parallel decline of their competitors.

As this was a gradual process we should expect both types of variants to have been used not just in the same period, but also in the same text. This is confirmed by examples such as (5) and (6) below, which are not at all infrequent in their respective decade:

(5) Μετά την άφιξιν εις την πόλιν, τα παιδιά δέχονται την ευλογίαν της Εκκλησίας. Προσφέρονται εις τα παιδιά διάφορα δώρα. Τα παιδάκια φθάνουν εις παιδόπολιν της Θεσσαλονίκης. Με τη χαρά ζωγραφισμένη στα πρόσωπά των παίζουν με τα δώρα που τους προσεφέρθησαν […]. (SRF01-1950-0019)

After the arrival at (H) the city, the children receive the blessing of the Church. Various gifts are offered to (H) the children. The small children arrive at (H) a child city of the Thessaloniki area. With joy radiant on (L) their faces they play with the gifts they were given […]

(6) Στον αερολιμένα του Ελληνικού συνεχίζεται με εντατικότατο ρυθμό στα πλαίσια του Τεχνικού Προγράμματος του 1969 η ανέγερσις νέων κτηριακών εγκαταστάσεων [….] Στο Τεχνικό Πρόγραμμα του 1969 προβλέπεται […] διά της χρησιμοποιήσεως ραντάρ εις τον έλεγχον της εναερίου κυκλοφορίας. […] να ανταποκρίνεται ανέτως εις τας υποχρεώσεις του ως διεθνούς αεροπορικού κόμβου και εις τας ανάγκας διακινήσεως εμπορευμάτων και επιβατών (SRF01-1960-0050)

At (L) the Hellenikon airport the construction of new buildings is continued at an intense pace in (L) the frame of the 1969 Technical Plan. […] In (L) the 1969 Technical Plan it is provided […] through the use of radar in (H) the control of air traffic. […] in order to easily respond to (H) its obligations as an international air traffic hub and to (H) the needs of moving merchandise and people.

Although in Example (5), which comes from a 1950s newsreel text, the L variant seems to be employed for stylistic purposes (e.g. to show affection for the

[7] *Greek Corpus 20* data was obtained from the archives of the Hellenic Broadcasting Corporation and the Hellenic National Audio Visual Archive <http://mam.avarchive.gr/portal/>. Since no scripts are available, special care was taken in the transcription, especially with regard to points that could easily be confused e.g. *is to* vs. *sto* etc.

small children), it would be difficult to suggest a similar functional differentiation for Example (6), which comes from a 1960s text. The whole issue merits further investigation, especially since the text was written during the colonels' ultra-conservative and oppressive dictatorship, which was apparently in favour of *katharévousa* (see also the discussion of the example in the following section). Although further discussion is beyond the scope of this paper, examples like the ones above are indicative of why there has been talk of mixed varieties of Greek throughout the 20th century.

Finally, the genre of private letters merits separate investigation, as it both conforms to strict conventions and involves interpersonal communication that is reminiscent of spontaneous speech. Figure 9 presents the findings for both pairs of variants in the course of the nine decades under study.

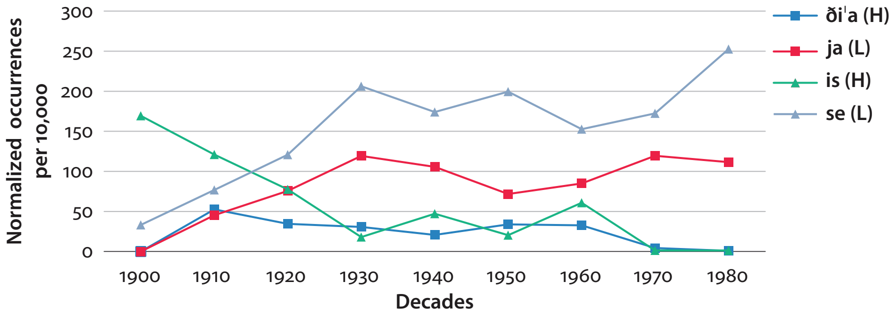


Figure 9. Frequency of H and L variants in private letters

As can be seen, this genre presents a more "typical" or expected development, with steady decline of H variants, accompanied by steady preference for L variants. As a result, the latter exhibit the gradual S-curve, evidenced in language change, whereas the former show a reverse S-curve, as expected. It seems then that data from private letters suggest a slow process of change in favour of the L variants which were to prevail and parallel co-existence of variant forms, which in certain cases as e.g. in the 1910s and the 1920s may be thought of as free variants.

Overall, frequency data from the examination of the variants for the two central pairs of grammatical items corresponding to the prepositions 'at/in' and 'for' in Greek suggest a complicated view of Greek in the 20th century. Taking the corpus examined as a whole it appears that H variants show a steady decline throughout the period studied, which is predicted by their gradual replacement by L variants. The latter, however, are prominent both at the beginning and the end of the period and significantly decline in the middle of the 20th century, when H variants also are in decline. At the same time, this is a somewhat simplified picture to the extent that it arises from the contribution of different genres, in which the frequency of variants may widely differ. Thus, there are such genres as academic texts and public speeches in which H variants are predominant up to a certain point of time, following which they are rapidly substituted by L variants; genres such as literature, song lyrics and film dialogues in which L variants are predominant throughout; and genres like newsreels and private letters which show a gradual replacement of H by L variants. Figure 10 illustrates the placement of genres with respect to the H and L varieties.

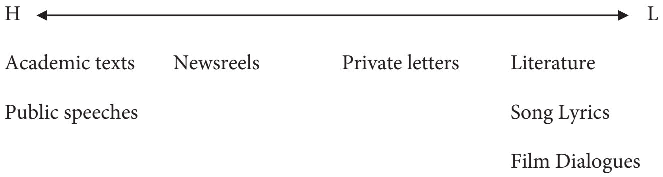


Figure 10. Continuum of genres from the H to L varieties

The following section discusses these findings and draws their implications for a data-based account of Greek in the 20th century.

4. Discussion and conclusions

The approach we have followed in this paper belongs to historical corpus linguistics, since we have looked into changes in the frequencies of grammatical items on the basis of evidence from a diachronic corpus. We have also adopted a variationist view of language change, as we investigated variants and possible sociolinguistic parameters that influence the occurrence of one over another. This is the first study of corpus-based historical sociolinguistics on issues related to Greek; for this reason, our findings cannot be easily discussed against other empirical work on Greek, as this is rather scarce, but provides evidence for judging claims made in descriptions based on linguistic intuition.

Corpus-based evidence has allowed us to gain considerable insight as regards the direction of language change in Greek of the 20th century. In particular, it was observed that, if all genres are taken together, L variants seem to follow a U-curve pattern overall, whereas H variants show a pattern of typical decline (e.g. a reversed S-curve) or may be found in "roller-coaster" patterns, showing sharp rises followed by sharp falls (see Figure 2, as well as Figures 5 and 9 to a lesser extent). These two observations, taken together, seem to suggest, first of all, that the L variants that finally prevailed were present, and even predominant in certain cases, at the beginning of the 20th century. Furthermore, they seem to compete with H variants throughout this period, although the latter manifest sudden peaks at specific periods. This may indicate that, during these periods speakers are aware of H variants and use them, following specific prescriptive injunctions for emblematic purposes. In this way, they seem to function as stereotypes, rather than markers or indicators in Labov's (1972) terms. However, it is difficult to draw conclusions on the basis of quantitative data alone and further analysis of the socio-historical context is necessary. For instance, the drop in frequency of both H and L variants for a long period of time in mid-20th century (see the low point of the U-curve in Figures 1 and 2) may suggest that speakers felt insecure about choosing one of the variants during this period and thus tended to avoid or underuse both, a hypothesis that obviously requires further investigation. Similarly, the unexpected peaks in H variant frequency in the 1940s and 1960s may indicate a conservative backlash at the face of prevailing L variants, but again we should be careful to separate the effects of the corpus composition and aggregation of individual genres from what may have happened in the language as a whole.

As regards the timing of changes, corpus data allow us to pinpoint the important periods of shift. According to the evidence presented here, at the beginning of 20th century there is strong competition between H and L variants, while during the inter-war period up to the 1950s there is a backlash of H items, followed by radical changes in favour of L items in the 1960s and 1970s. Near the end of the century (1980s) L variants have been established in all genres as the exclusive or the vastly dominant option. This description ties in nicely with major socio-political events in Greece at the respective periods such as the series of ultra-conservative or repressive Greek regimes in the middle of the 20th century or the shift to a left-wing government in the 1980s, but may also challenge a facile one-to-one correspondence, by suggesting e.g. that language change continued regardless of the dictatorship's public policy in the 1960s in favour of *katharévousa*.[8]

One of the most important findings of our work concerns the central role played by genre in our understanding of language change in Greek of the 20th century. Specifically, we have noted the large divergence between genres and text types as regards patterns of change. There seem to be several dimensions involved in genre variation that influence language variation, including the public vs. private audience of texts, as well as the degree of planning involved in their composition and their formal or conventional type. Thus, genres including academic texts and public speeches are designed to follow strict formal guidelines. This would

[8] For a comprehensive list of socio-political events in conjunction with developments in language planning and attitudes in 20th century Greek, along with a discussion of how these may have influenced linguistic choices, see Goutsos & Fragaki (2014).

seem to account for their conservative linguistic choices and the sudden replacement of one set of injunctions with another in the 1960s.

Literary genres, including the traditional text types of novels, short stories, poetry etc., as well as song lyrics and film dialogues seem to be aligned in Greek with innovative linguistic choices, remain relatively stable over time, and manifest minimal changes; this implies their strictly conventional character in the sense that they rely on a well-developed idea of what "authentic" or "spontaneous" language is like.

Text types such as newsreels, on the other hand, seem to depend on *ad hoc* planning and less well-established conventions and thus seem to follow the prescriptions of the day; this accounts for the haphazard patterns found in them, as well as the late rise of L variants, which suggests that they also succumb to language change.

In contrast to the above text types, the genre of private letter writing is the only one in which the expected gradual rise of L variants throughout the 20th century is found. It is also the only genre studied here that has a private audience and thus is assumed to reflect actual language interaction to a greater extent than others. Evidence from this genre seems to support the gradual prevalence of language change rather than an abrupt shift.

Table 2. Text types and language change in Greek Corpus 20

Dimensions of text production		Text types	Language change
public	formal	Academic Public speeches	conservative with sudden changes
	conventional	Literature Song lyrics Film dialogues	innovative with minimal changes
	planned	Newsreels	following language planning
personal	spontaneous	Letters	gradual language change

The discussion above is summarized in Table 2. It is significant that our findings from a Greek diachronic corpus concur with the understanding of the role of genre in recent language change in other languages e.g. with regard to the contribution of speech-like genres (Culpeper & Kytö 2010) and especially private letters (cf. Dossena & Del Lungo Camiciotti 2012). It also confirms the importance of using diachronic corpora (see e.g. Taavitsainen et al. 2015).

Our findings, moreover, can provide some preliminary evidence for a better understanding of recent language change in Greek. We can thus attempt to sketch what it would take to reply to the unanswered questions of Greek diglossia, pointed out in the Introduction. For instance, as regards the question of how many varieties appear to have existed in the Greek of the 20th century, our analysis concurs with the existence of two poles (demotic and *katharévousa* or L and H), although the possibility of co-existing variants within text types or even individual texts cannot be excluded. However, the main purport of our study has been to underline the central role of genre in trying to formulate an answer to this question. Thus, from the perspective of academic texts or literature there seems to be no choice of varieties available, since for most of the century one or the other appears exclusively, and this would seem to confirm Ferguson's strict separation of domains, although this is not the case with other genres, as we have seen.

As regards the range of registers available, we have pointed out the importance of dimensions such as conventionality, planning and (perceived) spontaneity, in addition to the well-known parameter of formality. Greek diglossia seems to have been an effect of factors belonging to multiple dimensions and, for this reason, cannot be collapsed to a simple spoken vs. written dichotomy. Although the distinction between spoken and written language may seem clear, e.g. in the description of Fergusonian diglossia, recent developments like mediated communication and the observed conversationalization of public discourse (Fairclough 1995) have put into question the value of such a strict distinction and have led to an understanding of spoken and written as comprised by a cluster of characteristics which are not shared by all text types (cf. Culpeper & Kytö 2010). As we have seen in the genres studied, there have also been, among other things, different conceptions of what 'speech' looks like. Our investigation of Greek genres indicates e.g. that public speeches followed different norms than literary or film dialogue and, as a result, were designed to widely varying representations of real interaction.

We have already discussed the timing of changes that is of central importance to any account of Greek diglossia and have suggested that time periods such as the beginning of the 20th century, the 1960s to 1970s, and the 1980s have been crucial points in terms of language change.

Finally, our research points to the need for further empirical studies on what actual speakers of Greek did in the course of the 20th century before we start exploring the question of how attitudes affected language use. If we could attempt to generalise in terms of a tentative answer, we would have to point out that, although language planning and prescriptions have had, to a certain extent, important influence on several genres, they could not have curbed the eventual language change towards the demotic. This is true even in text types heavily influenced by prescription, such as newsreels. It is also indicated by genres with gradual shift towards the L variety such as private letters.

It is obvious that much further work is needed in order to gain a better view of such complicated issues as Greek diglossia. First of all, our findings in this paper are based on the study of few linguistic items, namely the high vs. low variants of only two pairs of grammatical items. The study of more items, both grammatical and lexical, can both extend our findings and provide a firmer basis of evidence for discussing the theoretical issues surrounding Greek diglossia.

In addition, it must be pointed out that the grammatical words studied cover a variety of functions. For example, *is* and *se* can be used to specify the location or movement towards an entity, to denote the recipient of an action or relation etc. It remains to be studied whether the various functions of the grammatical items concur with particular choices of H or L variants, i.e. whether, for example, there is a preference for the L variant as regards the function of location and for the H variant as regards the function of relation, as happens in Example (6) above. It would also be interesting to investigate in more detail phrases which function as prefabricated chunks of language and thus are less infected by temporary changes. For instance, is the use of the L variant *se* in the set phrase στα πλαίσια του 'in the frame of' (see Example 6) a general pattern in the corpus or in the particular text type?

Finally, a fully-fledged exploration of these issues would require looking at more genres, such as journalistic texts or spontaneous conversation, which are especially difficult to gather in the compilation of a diachronic corpus of Greek (see Goutsos et al. 2017), as well as comparing with data from the last two decades of the century, available from the *Corpus of Greek Texts*.

Our current research is concerned with a full-scale investigation of variants, including both grammatical and lexical items, as well as the exploration of key lexical changes which reflect cultural change in the manner of Baker (2010; 2011) or Marchi (2010). This research is expected to contribute to an understanding of the range of registers and language varieties in the Greek of the 20th century and a better informed view of standardisation in Modern Greek. As must have become clear from our paper, the discussion of recent language change in Greek cannot move forward without specific evidence from diachronic corpora that allow us to appreciate the role of genre and text variation.

Acknowledgments

We would like to thank Richard Jason Whitt for his generosity and his help during the editorial process, as well as two anonymous reviewers for their comments.

References

Aitchison, Jean. 1981. *Language Change: Progress or Decay?* London: Fontana Paperbacks.

Alexiou, Margaret. 1982. Diglossia in Greece. In *Standard Language, Spoken and Written*, William Haas (ed.), 156–192. Manchester: Manchester University Press.

Baker, Paul. 2010. Will Ms ever be as frequent as Mr? A corpus-based comparison of gendered terms across four diachronic corpora of British English. *Gender and Language* 4(1): 125–149. https://doi.org/10.1558/genl.v4i1.125

Baker, Paul. 2011. Times may change, but we will always have money: Diachronic variation in recent British English. *Journal of English Linguistics* 39(1): 65–88. https://doi.org/10.1177/0075424210368368

Culpeper, Jonathan & Kytö, Merja. 2010. *Early Modern English Dialogues: Spoken Interaction as Writing*. Cambridge: CUP.

Daltas, Periklis. 1994. The concept of diglossia from Ferguson to Fishman to Fasold. In *Themes in Greek Linguistics* [Current Issues in Linguistic Theory 117], Irene Philippaki-Warburton, Katerina Nicolaidis & Maria Sifianou (eds), 341–348. Amsterdam: John Benjamins. https://doi.org/10.1075/cilt.117.50dal

Dossena, Marina & Del Lungo Camiciotti, Gabriella. 2012. *Letter Writing in Late Modern Europe* [Pragmatics & Beyond New Series 218]. Amsterdam: John Benjamins. https://doi.org/10.1075/pbns.218

Fairclough, Norman. 1995. *Media Discourse*. London: Edward Arnold.

Ferguson, Charles A. 1959. Diglossia. *Word* 15: 325–340. https://doi.org/10.1080/00437956.1959.11659702

Fernández, Mauro. 1993. *Diglossia: A Comprehensive Bibliography, 1960-1990 and Supplements* [Library and Information Sources in Linguistics 23]. Amsterdam: John Benjamins. https://doi.org/10.1075/lisl.23

Frangoudaki, Anna. 2002. Comment. Greek societal bilingualism of more than a century. *International Journal of the Sociology of Language* 157: 101–107.

Goutsos, Dionysis. 2010. The Corpus of Greek Texts: A reference corpus for Modern Greek. *Corpora* 5(1): 29–44. https://doi.org/10.3366/cor.2010.0002

Goutsos, Dionysis & Fragaki, Georgia. 2014. Πρόσφατη γλωσσική αλλαγή στα ελληνικά: Σχεδιασμός του Διαχρονικού Σώματος Ελληνικών Κειμένων του 20ού αιώνα (Recent language change in Greek: Planning the Diachronic Corpus of Greek of the twentieth Century). In *Selected Papers of the 11th International Conference on Greek Linguistics*, Giorgos Kotzoglou, Kalomoira Nikolou, Eleni Karantzola, Katerina Frantzi, Ioannis Galantomos, Marianthi Georgalidou, Vasilia Kourti-Kazoullis, Chrysoula Papadopoulou & Evangelia Vlachou (eds), 318–329. Rhodes: University of the Aegean.

Goutsos, Dionysis, Fragaki, Georgia, Florou, Irene, Kakousi, Vasiliki & Savvidou, Paraskevi. 2017. The Diachronic Corpus of Greek of the 20th Century: Design and compilation. In *Proceedings of the 12th International Conference on Greek Linguistics (ICGL 12)*, Thanasis Georgakopoulos, Theodossia-Soula Pavlidou, Miltos Pechlivanos, Artemis Alexiadou, Jannis Androutsopoulos, Alexis Kalokairinos, Stavros Skopeteas, Katerina Stathi (eds), 369–381. Berlin: Edition Romiosini/CeMoG, Freie Universität Berlin.

Goutsos, Dionysis, King, Philip & Hatzidaki, Rania. 1994. Towards a corpus of spoken Modern Greek. *Literary and Linguistic Computing* 9(3): 215–223. https://doi.org/10.1093/llc/9.3.215

Halliday, Michael A. K. & Hasan, Ruqaiya. 1990. *Language, Context, and Text: Aspects of Language in a Social-semiotic Perspective*. Oxford: OUP.

Holton, David. 2002. Modern Greek: Towards a standard language or a new diglossia? In *The Interplay of Internal, External and Extra-linguistic Factors*, Mari C. Jones & Edith Esch (eds), 169–179. Berlin: De Gruyter. https://doi.org/10.1515/9783110892598.169

Horrocks, Jeffrey. 2010. *Greek. A History of the Language and its Speakers*, 2nd edn. Chichester: Wiley & Blackwell. https://doi.org/10.1002/9781444318913

Hudson, Alan. 2002. Outline of the theory of diglossia. *International Journal of the Sociology of Language* 157: 1–48. https://doi.org/10.1515/ijsl.2002.039

Iordanidou, Anna. 1999. Ζητήματα τυποποίησης (standardisation) της σύγχρονης νεοελληνικής (Issues of standardization in Modern Greek). In *"Strong" and "Weak" Languages in the European Union*, Anastasios-Phoevos Christidis (ed.), 835–854. Thessaloniki: Centre for Greek Language.

Iordanidou, Anna. 2002. Επιτεύχθηκε ή επετεύχθη η τυποποίηση της νέας ελληνικής; (Has the standardization of Modern Greek been accomplished?). In *Recherches en Linguistique Grecque*, Christos Clairis (ed.), 239–242. Paris: L'Harmattan.

Iordanidou, Anna. 2009. Γλωσσική ποικιλία και δημοσιογραφικός λόγος (Language variation and journalistic discourse). *Ζητήματα Επικοινωνίας* 9: 102–114.

Labov, William. 1972. *Sociolinguistic Patterns*. Oxford: Blackwell.

Labov, William. 1994. *Principles of Linguistic Change, Vol. 1: Internal Factors*. Oxford: Blackwell.

Mackridge, Peter. 2004. Diglossia and the separation of discourses in Greek culture. *Teoreticheskie problemy yazykoznaniya. Sbornik Statey k 140-letiyu Kafedry obshchego yazykoznainya Filologicheskogo fakulteta Sankt-Peterburgskogo gosudarstvennogo universiteta*, 112–130. St Petersburg: Filologicheskiy fakul'tet Sankt-Peterburgskogo gosudarstvennogo universiteta.

Mackridge, Peter. 2009. *Language and National Identity in Greece, 1766–1976*. Oxford: OUP. https://doi.org/10.1093/acprof:oso/9780199214426.001.0001

Marchi, Anna. 2010. 'The moral in the story': A diachronic investigation of lexicalised morality in the UK press. *Corpora* 5(2): 161–189. https://doi.org/10.3366/cor.2010.0104

Milroy, James. 1992. *Linguistic Variation and Change: On the Historical Sociolinguistics of English*. Oxford: Blackwell.

Mirambel, André. 1937. Les états de langue dans la Grèce actuelle. *Conférences de l'Institut de Linguistique de l'Université de Paris 5*. Paris.

Nevalainen, Terttu & Raumolin-Brunberg, Helena. 2003. *Historical Sociolinguistics: Language Change in Tudor and Stuart England*. London: Routledge.

Papatzikou-Cochran, Effie. 1997. An instance of triglossia? Codeswitching as evidence for the present state of Greece's 'language question'. *International Journal of the Sociology of Language* 126: 33–62. https://doi.org/10.1515/ijsl.1997.126.33

Pappageotes, George C. & Macris, James. 1964. The language question in Modern Greece. *Word* 20(2), 53–59. https://doi.org/10.1080/00437956.1964.11659850

Petrounias, Evangelos. 1978. The Modern Greek language and diglossia. In *The "Past" in Medieval and Modern Greek Culture*, Speros Vryonis (ed.), 193–220. Malibu, CA: Undena.

Rissanen, Matti, Kytö, Merja & Heikkonene, Kirsi (eds). 1997. *English in Transition: Corpus-Based Studies in Linguistic Variation and Genre Styles*. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110811148

Taavitsainen, Irma, Kytö, Merja, Claridge, Claudia & Smith, Jeremy (eds). 2015. *Developments in English: Expanding Electronic Evidence*. Cambridge: CUP.

"You can't control a thing like that"

Genres and changes in Modern English human impersonal pronouns

Florian Haas
University of Jena

While there is ample evidence showing that the impersonal use of second-person singular pronouns has increased in several languages, the recent history of impersonal *you* in English has not yet received much attention in the literature. The present investigation presents corpus evidence from Modern English indicating that this strategy has indeed gained in frequency, independently of changes in the general frequency of second-person pronouns and the evolution of genres. Tracing specific functions of impersonal *you* diachronically reveals that *you* simulating the hearer's membership in the set generalized over and encoding hidden self-reference are relatively new uses, supporting the view that this impersonal strategy has undergone semantic extensions comparable to developments found in other languages.

1. Introduction

English employs an array of strategies to encode 'impersonalization', as illustrated in (1)–(3).[1]

(1) **They** didn't raise the taxes on the middle class. [COCA,[2] Spoken]

(2) And if **one** looks throughout the history of our country and other great countries, it is the latter category that tend to be the difficult ones. [COCA, Spoken]

[1] The impersonalizing strategies under discussion here should not be confused with other argument structure constructions that have been called 'impersonal', such as clauses with experiencer subjects or ones featuring 'dummy' pronouns for the description of weather conditions. I will exclusively deal with structures involving a human impersonal participant.

[2] *Corpus of Contemporary American English* (Davies 2008–).

https://doi.org/10.1075/scl.85.08haa

(3) There's an essay in the Times that – it talked to a lot of couples. And they say **you** can ask thirty-six questions to find out if you're in love.
[COCA, Spoken]

In contrast to many other languages, English does not have a formal marker or construction that is exclusively dedicated to impersonalization. Instead, the passive construction and a number of pronouns have impersonalization as one of their functions. Whereas for the passive the impersonalizing function can be considered a central one (Seoane Posse 2000: 108), impersonal uses of the personal pronouns *they*, *one* and *you* are clearly secondary both in terms of their frequency and the communicative function at issue (cf. Gast et al. 2015: 149–150 on *you*, for instance).

When empirically investigating the history of impersonalization in English, the above-mentioned coexistence of what we may call the 'canonical' functions of pronouns and their impersonal uses becomes a practical problem – impersonal instances of *they*, *one* and *you* cannot be extracted from corpora automatically. This is why for the current study on impersonal strategies in Modern English, all impersonal uses of the expressions at issue have been extracted manually from ARCHER, a diachronic corpus of British and American English.[3]

Another challenge to be dealt with when tracing the impersonal use of these pronouns is to do with the nature of historical corpora such as ARCHER. As has been shown by Biber and Finegan (1989, 2001), the corpus does not provide us with a stylistically homogenous picture of how Modern English has developed. Biber and Finegan alert us to the fact the genres[4] represented in the corpus for the different periods have themselves undergone changes, independently of changes in the linguistic system. Any linguist interested in how the language system has developed will therefore have to take changes in genres into account. This also applies to the study of human impersonal pronouns (HIPs), and impersonalization more generally, not least because HIPs are and have been sensitive to genre

[3] ARCHER 3.1 (Bamberg) (1990/1993/2002/2007/2010/2013). *A Representative Corpus of Historical English Registers*. Originally compiled under the supervision of Douglas Biber and Edward Finegan at Northern Arizona University and University of Southern California; modified and expanded by subsequent members of a consortium of universities. Current member universities are Bamberg, Freiburg, Heidelberg, Helsinki, Lancaster, Leicester, Manchester, Michigan, Northern Arizona, Santiago de Compostela, Southern California, Trier, Uppsala, Zurich. Examples of usage taken from ARCHER were obtained under the terms of the ARCHER User Agreement (available on the Documentation page of the ARCHER website, <http://www.manchester.ac.uk/archer/>.

[4] I will not differentiate between 'genres' and 'registers' (Biber & Conrad 2007: 15–23) in this paper.

differences. In this chapter, I will focus on changes in the frequency and usage conditions of the HIP *you* and discuss whether the observed changes are due to the above-mentioned shifts of English genres in the period covered by ARCHER, or if there are arguments for assuming a change in how the expression has been used. In this context, I will consider to what extent *you* has been used to achieve 'simulation', i.e. simulating the hearer's membership in the category generalized over, and 'self-reference', where the speaker expresses a hidden kind of reference to him- or herself.

The structure of the chapter is as follows. Section 2 introduces the grammatical concept of impersonalization, the system of impersonal strategies in present-day English (PDE) and some central facts regarding its earlier history. Section 3 provides the results of my corpus study, concentrating on the development of impersonal *you*. In Section 4, Biber and Finegan's findings concerning changes in English genres throughout the ARCHER periods are summarized and related to the HIP *you*. Section 5 brings into play a number of arguments that I take to support the idea that there has been a change in the use of impersonal *you* after all. Section 6 sums up the discussion, offers conclusions and outlines questions for further research.

2. Human impersonal pronouns

2.1 Introduction

Defining impersonal constructions in a cross-linguistically sensible way is far from trivial (see Siewierska 2008: 116–126). For the HIPs under discussion in this chapter, I will subscribe to Siewierska's (2011: 57–60) notion of 'reference-impersonals' (or 'R-impersonals'), i.e. constructions in which a human core participant – typically, but not always, the agent – is non-referential.[5]

What the three central HIPs of PDE, as illustrated in (1)–(3), have in common is that the exact identity of the agents in the events described is left open. It is in this sense that they are "non-referential". To be sure, *they*, *one* and *you* each provide different hints as to who is meant. *They* in (1) excludes speaker and hearer from the set of people responsible for not raising the taxes on the middle class, and in fact the verb phrase narrows the set down to a collective (the 'corporate'

[5] A similar definition is proposed by Gast and van der Auwera (2013: 124): "IMPERSONALIZATION is the process of filling an argument position of a predicate with a variable ranging over sets of human participants without establishing a referential link to any entity from the universe of discourse."

use in Cabredo-Hofherr 2006: 244; Siewierska & Papastathi 2011: 580–584). The impersonal use of *one* in (2) is truly generic in making a general claim to the effect that *anyone* who happens to be in the situation described – including speaker and hearer – would come to similar conclusions. Impersonal *you* as in (3) is used in generalizing sentences, i.e. sentences not denoting specific, temporally bound situations. It also differs from *one* in being less formal and in that the deictic semantics of the second-person pronoun seem to persist in its impersonal use: the hearer is invited to identify with the group at issue more than with comparable uses of *one*. Still, in none of these cases does the speaker make reference to an individual that the hearer can unambiguously identify. It is a correlate of the latter that there is no discourse referent that HIPs make available for the ensuing discourse (Gast & van der Auwera 2013: 122–124). Consider sentence (2) and the modified version in (4), for instance. *One* in (2), in contrast to the existential quantifier *someone*, does not introduce a referent that could function as an antecedent of an expression like *this person*.[6]

(4) And if **one** looks throughout the history of our country and other great countries, ***this person** will find that it is the latter category that tend to be the difficult ones.

2.2 Human impersonal pronouns in earlier English

The most striking difference between the PDE system of HIPs and earlier English is the central role that the pronoun *man* had in Old English (cf. Example [5]) and most of the Middle English period, where the reduced variant *me* had become the predominant form (Rissanen 1997: 518–519).

(5)	Her	onginneþ	seo	boc	þ	**man**	Orosius	nemneþ
	here	begins	that	book	that	one	Orosius	calls

‘ Here begins the book that is called Orosius.’ [Orosius, 1]

In contrast to all the PDE HIPs, this was a pronoun functionally dedicated to impersonalization (on the pronominal status of *man* in Old English see Fröhlich 1951: 30–45 and van Bergen 2000). It was used very similarly to the impersonal pronoun *man* in German (Zifonun 2000), Swedish (Egerland 2003) and other Germanic languages. Like these modern *man*-pronouns, the early English expression was functionally less constrained than its PDE counterparts. Whereas impersonal *you*, as mentioned above, involves special effects concerning the role of the hearer

[6] An anonymous reviewer and Richard J. Whitt have pointed out that in American English *he* would be possible here (see also Trudgill & Hannah 2008: 76–77).

(see Section 5.2), and the impersonal use of *they* excludes speaker and hearer from the group about which a generalization is made, *man* was free of such constraints. Even though this general applicability in the impersonal sphere made the expression particularly useful and common in early English (Fröhlich 1951: 113; Los 2002), it fell into disuse, became obsolete by the end of the fourteenth century (Rissanen 1997: 520) and eventually disappeared in the course of the fifteenth century.[7] The reasons for its demise are not yet entirely clear (see e.g. Meyer 1953: 235–240; Jud-Schmid 1956: 110–123; Los 2002: 182; Light & Wallenberg 2015). It is likely, however, that it correlated with changes in the distribution of the remaining and new pronominal impersonals. The relevant use of *one* (see [6] for an early example), which arose in the fifteenth century can be counted as a new strategy, although it probably arose too late to be considered a direct replacement of *man* (Los 2002: 182).

(6) Doo thus from be to be; thus wol thai lede **oon** to thair dwelling place.
[ca. 1420; Meyer 1953: 38]

Another development which has been related to *man* not being available for impersonalization anymore is the frequency of the passive (e.g. Visser 1973: 2102–2103; Denison 1985: 195). Again, it has to be noted that linguists who have worked on the respective historical periods do not agree as to what exactly triggered the increase of passive use. While Los (2002, 2009) and Seoane (2006) claim that changes in information-packaging options following the loss of the verb-second rule in English were eventually responsible, Light and Wallenberg "find no evidence that this change can be directly tied to the loss of V2-like word orders in English" (2015: 242). Instead, based on a comparison of a Northern and a Southern Middle English text, they claim that the passive construction indeed compensates for the loss of *man* as an impersonal (2015: 241). In what ways the loss of *man* may also have been responsible for speakers using the second-person singular pronoun impersonally more often is difficult to determine without a comprehensive corpus investigation of late Middle English impersonalization.

For which periods are second-person impersonals first attested? Fröhlich (1951: 66) finds no evidence for their use in Old English. Meier (1953: 227) considers the strategy to have been very rare in Middle English and makes similar remarks on Early Modern English (1953: 40). Meyer's observations on Early Modern English are only based on small text samples and therefore have to

[7] Interestingly, Cheshire (2013) sees first signs in some English varieties that the noun *man* is undergoing grammaticalization to a HIP again.

be treated with caution.[8] The OED's earliest attestation of impersonal *you* is from c1555 (cf. Wales 1985: 8–9 for references to further examples from the same time). Jud-Schmid (1956: 84), in her work on Early Modern English, sees signs of second-person impersonals becoming more and more popular in that period, however.

3. A corpus study on the Modern English HIP *you*

3.1 The corpus and data extraction

In order to document Modern English changes in the expression of impersonalization, all impersonal uses of the expressions *one*, *you* and *they* have been extracted from the ARCHER corpus (version 3.1). The corpus contains 674 texts from 1650 to present-day British and American English, coming from the genres drama, fiction, sermons, journals, medicine, news, science and letters. The data are partitioned into seven periods[9] covering 50 years each: 1650–99 (2), 1700–49 (3), 1750–99 (4), 1800–49 (5), 1850–99 (6), 1900–49 (7), 1950–99 (8). Since American data are only available for periods 4, 6 and 8, I will consider a sample of only British data for all periods and a sample of combined British and American texts coming from periods 4, 6 and 8 separately.

Having available a wide variety of genres and a relatively fine-grained diachronic differentiation suggests that a detailed picture of how impersonal strategies were used could be obtained. Unfortunately, however, the overall frequency of impersonals in the corpus is too low to arrive at statistically significant generalizations concerning their distribution across genres through time (see Table 2). Note also that the different genres are not equally well represented in ARCHER; while sermons and letters each only account for roughly half of the words that the genres drama, journal or diaries, medicine, science and news provide for a given period, fictional texts double the word counts for these latter genres. This holds for both the British and the American texts included in the corpus.

[8] The same holds for Seoane Posse's (2000) detailed investigation of impersonalising strategies in a selection of Early Modern English texts amounting to 153,000 words from the Helsinki Corpus of English Texts. Her data suggest that *you* was already the most frequent impersonalizing device early on. However, the examples that one finds in the corpus are few in total and come from a small number of texts, so we may not see a real change here.

[9] The first period envisaged for the corpus (1600–1649) was not yet available for ARCHER 3.1. Period 2 (1650–1699) is therefore the earliest period under discussion in this paper.

As mentioned earlier, there has turned out not to be any reliable procedure of automatically extracting impersonal uses of *one*, *you* and *they* from ARCHER. The version of the corpus used is not tagged for parts of speech. Yet even if it were tagged for parts of speech, one would still need to inspect all instances of the relevant expressions and their context to tell apart impersonal from non-impersonal uses. Since this chapter concentrates on changes in impersonalization involving the pronoun *you*, I will illustrate the problem using examples of that pronoun.

You is canonically employed as a deictic singular and plural second-person pronoun. For the 16,652 hits of *you* in the corpus, the relevant sentences and – if necessary – their wider context were checked with respect to the question whether the speaker made deictic reference to the hearer(s) or rather included the hearer in a generalization over a set of referents (see Gast et al. 2015 and Deringer et al. 2015 for a more detailed description of the semantics and pragmatics of second-person singular impersonals).[10]

Given that impersonal *you* is semantically rather close to deictic *you*, it is not surprising that there are examples involving *you* that allow both a deictic and an impersonal reading. Consider (7):

(7) Remsen, like a family physician, does very well for ordinary cases. When **you** want a legal surgeon, **you** call in Jitt; he's an amputator. [Drama, 1871]

The example illustrates the difficulty of making clear-cut coding decisions. In, as in a number of others in my sample, neither a deictic, nor an impersonal, interpretation can be ruled out with absolute certainty. I have coded it as 'impersonal' because such a reading would be well compatible with the context in which the sentence occurs, and indeed seems more likely than a deictic interpretation of *you*. All controversial instances were discussed with a second coder and only included if we both took the impersonal reading to be more likely in the given context. For (7), the utterance containing the instances of *you* at issue was thus interpreted as answering the implicit question 'Whom should one generally call when one needs a legal surgeon?' (the speaker is metaphorically referring to a lawyer). Yet, no one could rule out a deictic interpretation with full certainty. After all, the speaker could give the hearer a purely personal piece of advice. We are here dealing with

[10] Plural *you* is not used impersonally in the data. This is in line with what other authors have found for earlier and PDE (Meyer 1953: 95; Malamud 2012: 3) and with the facts of languages where singular and plural terms of address are formally distinguished. The older form *thou* only occurs twice in impersonal use – in the same text, a sermon from period 6 – and was not included in the statistics.

two interpretations that are not only hard to tell apart in terms of coding data, but also conceptually. It is argued in Gast et al. (2015), for example, that an impersonal interpretation of second person pronouns crucially builds on their canonical (i.e. deictic) semantics. One could thus assume that using *you* either deictically or impersonally does not always imply a consciously categorical decision on the part of the speaker.

3.2 Quantitative observations

Tracing the frequency of impersonal second-person pronouns throughout the ARCHER periods suggests an increase. As we saw earlier, the second-person impersonal was relatively rare before Modern English. My data show rising frequencies both for the British English data, in periods 2 (1650–99) to 8 (1950–99) (see Figure 1), and for the periods that contain both British and American English texts: periods 4 (1750–99), 6 (1850–99) and 8 (1950–99) (see Figure 2). Table 1 contains the absolute frequencies for British English and the combined frequencies for periods 4, 6 and 8. Since the subcorpora for the different periods do not all contain an equal amount of words, the frequencies have been normalised to rate per 100,000 words for the counts in this section (see Table 2 for the absolute frequencies). The ratio of HIPs to non-impersonal instances of *you* is discussed in Section 5.1.

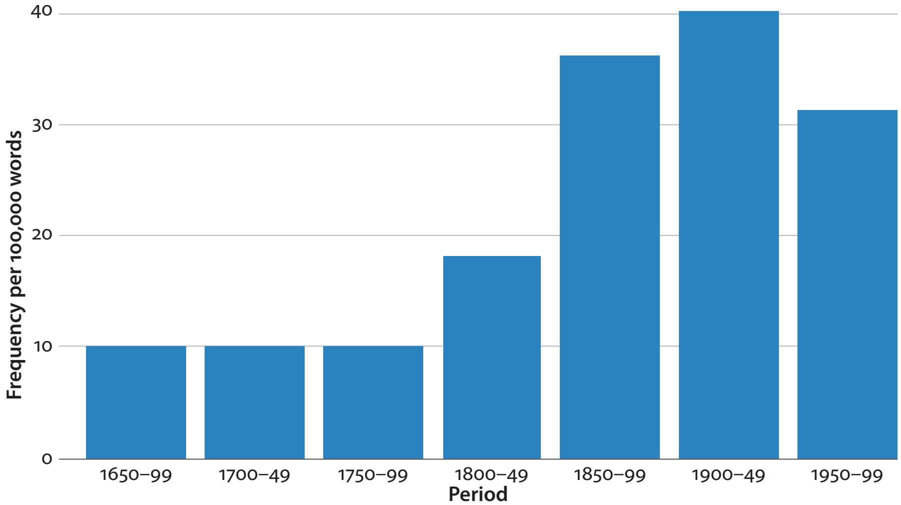


Figure 1. Normalised frequency of impersonal *you* in ARCHER (British English)

There are not enough hits overall to investigate the distribution of the expression across periods *and* genres. Table 2 provides the numbers of hits per genre.

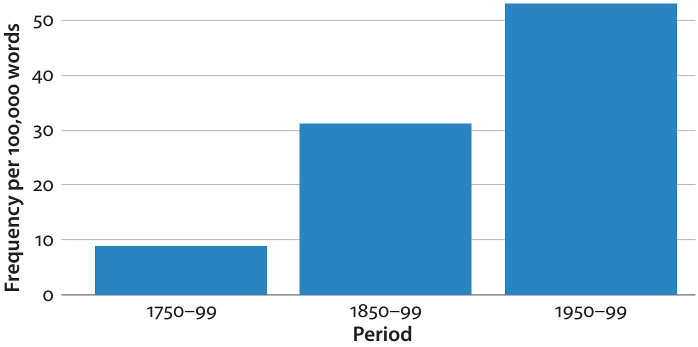


Figure 2. Normalised frequency per 100,000 words of impersonal *you* in ARCHER (British and American English, Periods 4, 6 and 8)

Table 1. Impersonal *you* in ARCHER (normalized frequency per 100,000 words)

Period	British English	British & American English
1650–99	10	
1700–49	10	
1750–99	10	9
1800–49	18	
1850–99	36	31
1900–49	40	
1950–99	31	53

Unsurprisingly, fiction and drama are the genres that represent the large majority of impersonal *you* uses, the other genres containing only a few relevant examples. For this reason, I will not be able to make statistically useful statements on changes for individual genres, even though it would be interesting to explore shifts in the use of second-person impersonals for specific text types, not least because the genres themselves have changed throughout the periods covered by the corpus.[11]

[11] An anonymous reviewer suggests to group genres according to Biber and Finegan's (2001) distinction between 'popular' and 'specialist' texts (see below) and consider frequency changes for the two groups separately. For the popular genres, the figures displayed in Table 1 would change only slightly and leave the general picture of an increase intact. For the specialist genres, the numbers are simply too small to trace historical changes.

Table 2. Frequency of impersonal *you* across genres (the numbers for British and American English in periods 4, 6 and 8 are given separately)

	2	**3**	**4b/4a**	**5**	**6b/6a**	**7**	**8b/8a**
Drama	3	10	2/0	0	8/7	23	9/50
Fiction	3	4	2/1	5	20/8	14	33/67
Journal or Diaries	0	2	5/1	6	17/0	19	10/3
Letters	0	1	3/1	3	14/2	4	1/9
Medicine	0	1	0/1	0	0/0	0	0/0
News	1	0	1/0	1	0/0	1	1/0
Science	2	0	0/8	1	2/0	0	0/0
Sermons	9	0	6/0	17	6/23	13	3/0
Total	**18**	**18**	**19/12**	**33**	**67/40**	**74**	**57/129**

4. Changes in English genres

4.1 Genres throughout Modern English

Biber and Finegan (1989, 2001) demonstrate that from Early Modern English to PDE literary and non-literary genres have undergone changes along several of the dimensions that Biber (1989) determines as crucial for the characterization of genres in general:

> The specialist registers have followed an essentially steady development towards even more ‘literate’ styles […]. In contrast, the popular registers developed towards more ‘literate’ characterizations in the earlier periods, but then reversed this trend in the more recent periods as they shifted towards more ‘oral’ characteristics. (Biber & Finegan 2001: 76)

These observations provide interesting insights into changing textual conventions, readerships and purposes of a given genre. It turns out, for instance, that medical, scientific and legal texts have become more and more ‘literate’, presumably in response to an increasing expertise on the part of the respective readers (and writers). ‘Popular’ genres, such as fictional prose, drama and personal letters, by contrast, seem to have moved closer towards the ‘oral’ pole: they have taken on more characteristics of spoken language, having become “more involved and situated, but less abstract” (Culpeper & Kytö 2010: 101).

As insightful as these generalizations are for a comprehensive diachronic description of written English, they also raise the question of whether, and to what degree, we can trust historical corpus data when we aim at reconstructing the characteristics of spoken English in a particular period. We can only trace grammatical change throughout modern English by relying on written texts, after all. The ARCHER corpus makes it possible to track lexical and grammatical developments through more than three centuries. Biber and Finegan's findings suggest, however, that a number of linguistic features are not distributed equally across historical periods because their use in a given genre has undergone changes independent of changes in the linguistic system. Reasons for such genre shifts are rather language-external (cf. Biber & Finegan 1989: 512–515) and change the degree to which a given genre approximates the 'oral' or the 'literate' pole of an assumed stylistic continuum (see also Koch & Oesterreicher 2007).

4.2 The role of second-person pronouns

Biber (1989: 8) identifies second-person pronouns as a positive feature in his dimension 1 "Involved vs. Informational Production", one of the three dimensions along which genres are shown to change. In other words, the more frequent second-person pronouns are in a given text, the more involved the text will be. Since Biber and Finegan (1989: 499–501) show that 'popular genres' like letters become more involved from the eighteenth century on, one might be tempted to conclude that the changes in the use of impersonal *you* described in Section 3.2 are simply concomitant of this genre shift. Biber and Finegan (1989, 2001) themselves do not distinguish between deictic and impersonal uses of second-person pronouns and thus leave it open whether, firstly, impersonal *you* indicates involved production in the same way as deictic *you* and, secondly, whether the use of impersonal relative to deictic second-person pronouns has changed.

In what follows, I will firstly investigate how far the frequency of impersonal *you* in ARCHER has increased relative to that of deictic *you*. Secondly, more specific use types of the impersonal expression will be analysed diachronically. I will then argue that these diachronic developments suggest genuine changes in the meaning and use of the English second-person HIP, even though a correlation with an increasing readiness to mark involvement in certain text types cannot be excluded conclusively.

5. Has impersonal *you* changed, after all?

In order to find out whether the increase of impersonal second-person uses merely supports Biber and Finegan's (1989, 2001) analysis of 'popular' genres becoming more involved, one procedure would be to test if impersonal *you* clusters with the other predictors that characterize stereotypically oral genres. This is beyond the scope of the present chapter. I have instead tried to tease apart the diachrony of the impersonal type of *you* and that of the "canonical" deictic type targeted in the historical genre studies cited above. Furthermore, I have investigated the development of two special uses of the HIP *you* which have been argued to be semantically further removed from their deictic source: simulation and self-reference (Deringer et al. 2015: 325–328; Gast et al. 2015: 158; Tarenskeen 2010: 54).

5.1 Impersonal vs. deictic *you*

Table 3 contains the absolute frequencies of impersonal and deictic *you* throughout periods 2–8, as well as the proportion of the former for each period. Figure 3 provides a bar plot for these data.

Table 3. Proportion of impersonal uses in all second-person pronouns for British English

Period	HIP *you*	Deictic *you*
1650–99	18 (1.28%)	1,392
1700–49	18 (1.18%)	1,508
1750–99	19 (1.12%)	1,684
1800–49	33 (2.15%)	1,505
1850–99	67 (3.20%)	2,027
1900–49	74 (3.94%)	1,804
1950–99	57 (3.15%)	1,750

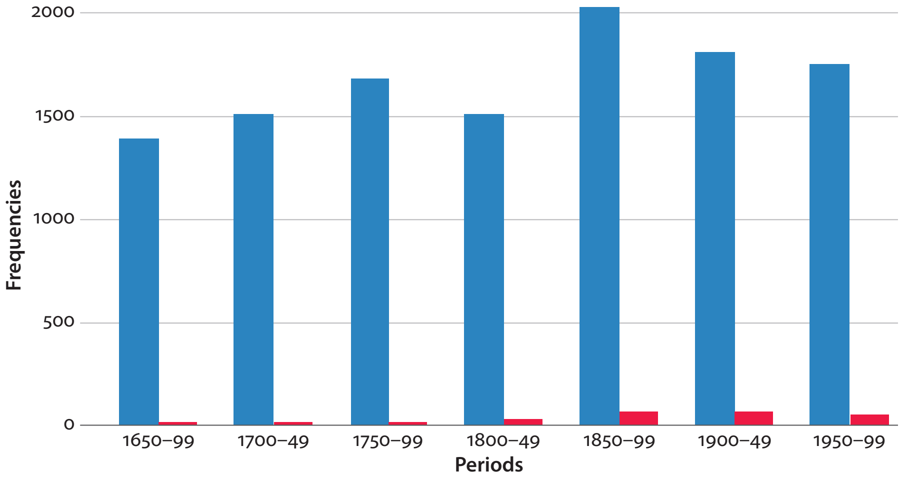


Figure 3. Absolute frequencies of deictic (blue) and impersonal (red) second-person pronouns

Given the much higher frequency of deictic *you* overall, Figure 3 is not particularly revealing, even though the general rise in the frequency of the impersonal is visible. Let us therefore consider an association plot (Figure 4) indicating in how far the share of impersonal uses for a given period is significantly below or above its statistically expected frequency.[12]

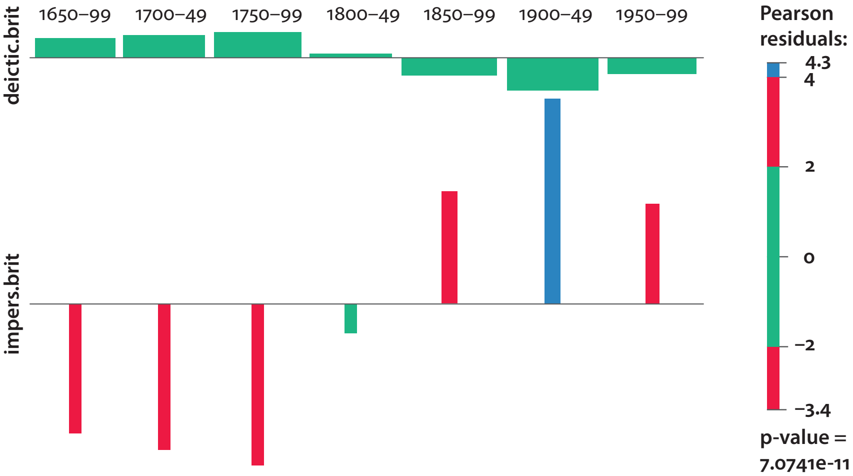


Figure 4. Association plot of deictic and impersonal *you* for British English through periods 2–8. (Pearson's chi-squared test $\chi^2 = 59.033$, d.f. = 6, $p < 0.001$)

The bars for deictic *you* are wider for all periods since their absolute frequency is higher. Crucially, however, the height of the bars and their shading indicate deviations from the respective expected frequencies. In this way, we can see that the proportion of impersonal *you* for periods 2–4 is lower and that of periods 6–8 is higher than expected.[13]

It has to be borne in mind that these figures concern all genres taken together, which is why the lack of a frequency change in deictic *you* should not be surprising. After all, Biber and Finegan (2001) show how popular and specialized genres have developed in opposite directions with respect to dimension 1 (involved vs informational production), so that changes in the frequencies of second-person pronouns in

[12] All statistical analyses presented in this paper were performed with the open-source software *R*, version 3.2.1 (R Development Core Team 2015).

[13] Standardized residuals are crucial here. They indicate the degree to which a given frequency deviates from the expected frequency in one or the other direction. Values above 2.0 and below −2.0 are generally taken to be significant.

different genres can be expected to cancel each other out to a certain degree. Note also that the relevant variable in Biber and Finegan's work includes second- as well as first-person pronouns, whereas the latter have not been taken into account here.

As I mentioned earlier, ARCHER provides American English data for periods 4, 6 and 8 only. Nevertheless, comparing just these three to each other (see Table 4, which combines British and American data for these periods) again shows a significant increase in the proportion of impersonal *you* uses.

Table 4. Proportion of impersonal uses in all second-person pronouns for British and American English

Period	HIP *you*	Deictic *you*
1750–99	31 (1.10%)	2,793
1850–99	107 (2.82%)	3,682
1950–99	186 (5.06%)	3,492

An association plot for these counts (Figure 5) shows the three periods to involve statistically significant deviations from the expected frequencies in the anticipated directions.

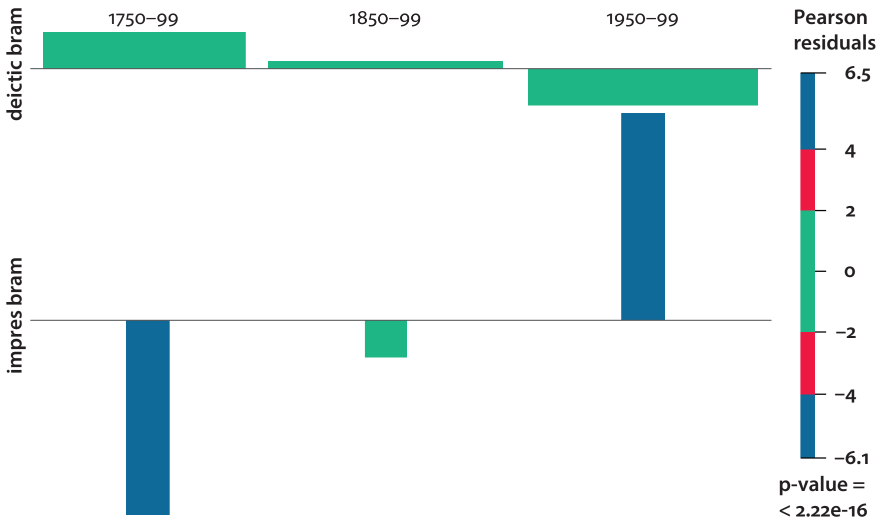


Figure 5. Association plot of deictic and impersonal *you* for British and American English in periods 4, 6 and 8. (Pearson's chi-squared test $\chi^2 = 84.197$, d.f. = 2, $p < 0.001$)

5.2 Simulation

Assuming the semantic analysis of second-person singular impersonals put forward in Gast et al. (2015), the kind of generalization brought about in the relevant contexts is based on the pronoun's deictic meaning: the hearer is addressed as a member of a group over which the sentence generalizes. This analysis is in line with the observation that second-person impersonals including English *you* more directly concern the hearer than impersonal constructions not based on the second person (Kitagawa & Lehrer 1990: 744; Malamud 2012: 13; Deringer et al. 2015: 323–331; de Hoop & Tarenskeen 2015: 166).

Cases of 'simulation' are special in that the hearer is *not* a member of the group that is generalized over. It is argued in Gast et al. (2015: 158) and Deringer et al. (2015: 325–328), therefore, that simulated uses of impersonal *you* are one step further removed from their underlying deictic meaning than 'valid' generalizing impersonals, where the hearer is also included in a truth-conditional sense. From a diachronic point of view, the question arises whether the secondary, derived nature of simulated second-person impersonals is also reflected in the frequency of their use through time. In other words, one may expect a development in which simulated impersonals are only rarely used in earlier periods and gain ground later, as their innovative semantics gets conventionalized as a standard subtype of impersonal meaning.

Sentence (8) is one of the simulated uses of impersonal *you* we can find in the data.

(8) What do they call boiling the skin off a cat? I call it murder, that's what I call it. They say it was hit by a car and Janice just scooped it up and before **you** could say bingo it was screaming in a pot of boiling water. [Drama, 1970]

(8) is a clear simulation case because the speaker talks about a past event. Since the addressee did not witness the event and is now being told about it, she was in no position to say bingo at that past point in time. (9) is similar in that the addressee, a doctor investigating the speaker and having him report on his past, is truth-conditionally not a member of the set established by *you*.

(9) The workhouse where they put me. They beat **you** there like a drum. [Drama, 1979]

Let us now consider in how far the incidence of simulated uses has changed diachronically. Table 5 shows a clear surge in period 6 (1850–99), where more than half of all impersonal *you*-instances turn out to be simulated. The percentage of simulated uses goes up to 77 in period 8 (1950–99).

It is also noteworthy that not a single example of the simulation type was found for the earliest period. Instances of impersonal *you* from that period are exclusively of the valid type, as illustrated in (10)–(11).

(10) **You** might as well expect to have a fish live out of the water as to expect her to be without acting some of these falsities, and in all these things she was as false as Hell. [Fiction, 1673]

Table 5. Proportion of simulated uses in all impersonal second-person singular pronouns for British English

Period	Simulated	Valid
1650–99	0 (0.0%)	18
1700–49	3 (16.67%)	15
1750–99	6 (31.58%)	13
1800–49	8 (24.24%)	25
1850–99	37 (55.22%)	30
1900–49	39 (52.70%)	35
1950–99	44 (77.19%)	13

(11) ’tis a horrid age that we live in, so that an honest man can keep nothing to himself. If **you** have a good estate, every covetous rogue is longing for it […]. [Drama, 1680]

Examples like (10) and (11) are categorized as ‘valid’ second-person HIPs because in these cases the hearer/reader is in fact part of the set over which the speaker/writer generalizes. In (10), the relevant predicate (*expect to have a fish live out of the water*) does not necessarily apply to the hearer, but by adding the possibility modal *might* the hearer becomes one of the many individuals who are now included. That impersonal *you* occurs in the protasis of a conditional construction in (11) makes the denotation of the pronoun universal in a similar way. It applies to *everyone* who has a good estate – including the hearer – that every covetous rogue is longing for.

Given our diachronic findings, we would be much less likely to find a sentence like (9) in period 2. Figure 6 sounds a note of caution as far as the assumption of a steady development is concerned, however; only the observed frequencies of periods 2 and 8 are significantly lower and higher than the expected frequencies of the respective periods. Thus, although the anticipated change in the use of simulated *you* appears to be in line with what the figure indicates, only standardized residuals below –2.0 or above 2.0 are statistically reliable indicators. Therefore, the assumption of a neatly linear increase of simulated *you* remains speculative at this point.

5.3 Self-reference

Another subtype of impersonal *you* which goes beyond the ‘valid’ generalizing type, and which should be worth considering from a diachronic point of view, is the well-known use we could describe as “hidden” self-reference by the speaker

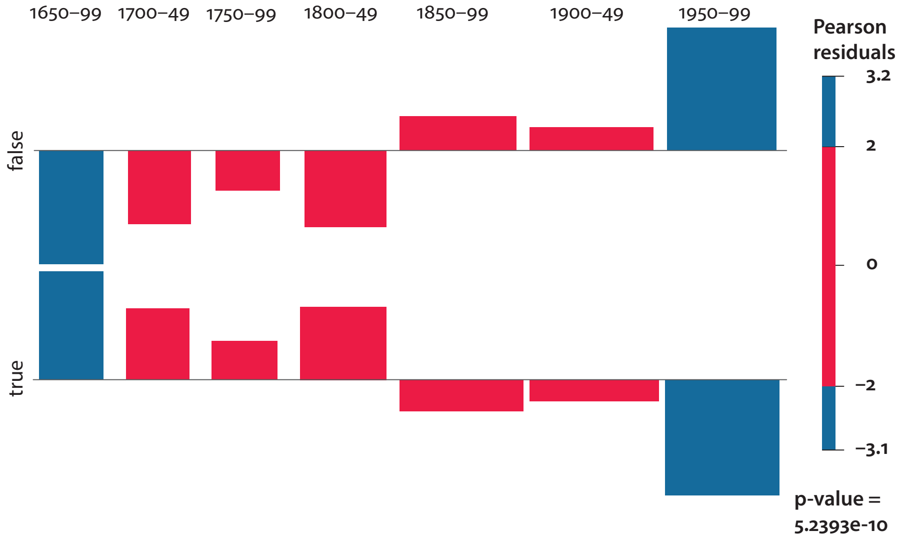


Figure 6. Association plot of simulated ('hearer included' = false) and non-simulated ('hearer included' = true) uses of impersonal *you* in British English for periods 2–8 (Pearson's chi-squared test $\chi^2 = 54.7366$, d.f. = 6, $p < 0.001$)

(see Tarenskeen 2010 for detailed discussion of self-reference). Instead of employing a first-person pronoun, the speaker makes impersonal reference to a set of individuals, including speaker and hearer, which only the speaker him- or herself is actually a member of. Example (12) is a case in point:

(12) Hard to find any place to work now where people don't bother **you** and will let **you** work. [Letters, 1961]

Even though the assertion in (12) is framed as a generalization over anyone who happens to look for a quiet place to work, the larger context of the sentence makes it clear that what the writer of the letter is really reporting on is his personal experience. I checked all sentences in the sample for this feature of 'self-reference', testing whether replacing the impersonal by an explicit reference to self (i.e. a first-person singular pronoun) would be a possible paraphrase. Table 6 indicates that such uses of the impersonal are not at all found in periods 2 and 3, and only once in period 4. They seem to be more common in periods 5 to 8.

Again, we are of course interested in the relative frequency of self-reference impersonals in a given period compared to the total number of second-person HIPs in that period. Figure 7 plots the two options against each other and reveals that self-reference is underrepresented in periods 2–4 and over-represented in period 7. No statistically significant claims can be made about the remaining periods.

Table 6. Proportion of self-referential uses in all impersonal second-person singular pronouns for British English

Period	Self-reference	Non-self-ref.
1650–99	0 (0.0%)	18
1700–49	0 (0.0%)	18
1750–99	1 (5.26%)	18
1800–49	7 (21.21%)	26
1850–99	27 (40.30%)	40
1900–49	40 (54.05%)	34
1950–99	25 (43.86%)	32

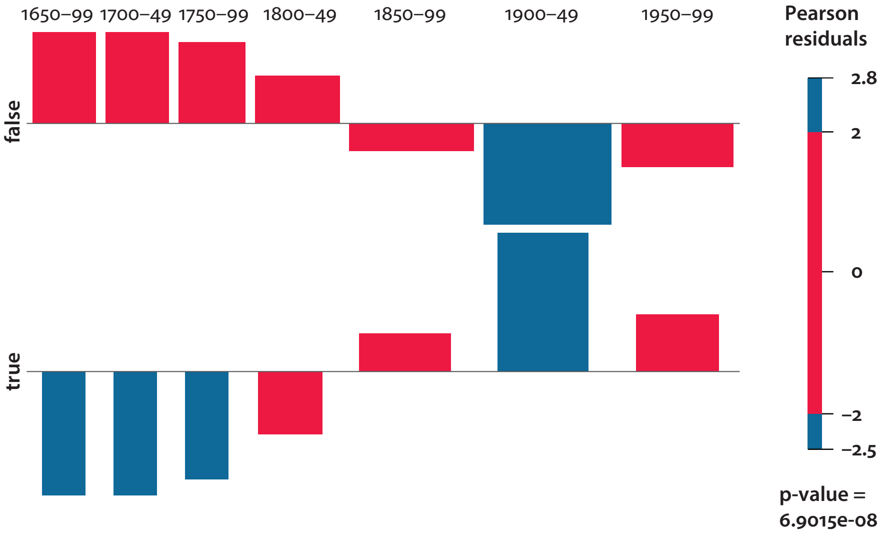


Figure 7. Association plot comparing the presence and absence of self-referential uses of impersonal *you* in British English for periods 2–8 (Pearson's chi-squared test $\chi^2 = 44.15038$, d.f. = 6, $p < 0.001$)

To sum up, we have observed changes that directly concern the semantics and pragmatics of impersonal *you* and not merely the frequency of HIPs in a given period.

5.4 A comparative view

In view of the question of whether the increase in the frequency of English impersonal *you* that can be observed in ARCHER is a genuine development of the relevant impersonal construction or rather a side-effect of historical shifts in genres, it could be revealing to consider diachronic findings on second-person HIPs that have been reported for other languages.

Jensen (2009: 95–99, 103–104), as well as Nielsen et al. (2009), have worked out the recent diachrony of the Danish second-person singular impersonal *du* 'you.SG' on the basis of a large sample of interview data. This text type would seem to be independent of how 'oral' written genres were conceived to be at a particular point in time. One of the authors' main findings is that *du* has gained in frequency in recent decades, crucially also before the middle of the twentieth century, hence before influence from English could have played a major role.

Nielsen et al. (2009) have also investigated the relevant utterances in context and conclude that the Danish second-person HIP is used for 'enactment', i.e. making some state of affairs come to life at the moment of speech (2009: 123), and 'involvement' more than the expression *man* (see also Jensen and Gregersen 2016). The latter turns out to be a more neutral generic pronoun. At this point, we could wonder if the fact that an expression creating 'enactment' and 'involvement' gains in popularity should not be viewed as a change corresponding to what Biber and Finegan describe for English genres. Note again, however, that Jensen (2009) as well as Nielsen et al. (2009) investigated spoken interview data, which should not be subject to the above-mentioned historical shifts, which have affected written registers, in the same way.

An increase in second-person impersonal usage has also been observed for French (Laberge 1980: 85; Coveney 2003), Finnish (Leino & Östman 2008: 39–43) and Spanish (Posio 2016). Posio (2016) analyses interview data involving Spanish speakers of different age groups in an apparent-time study and demonstrates "a significant negative correlation between the speakers' age and the normalized frequency of 2SG-IMP in their speech […] in line with the claims of an increase in the use of 2SG-IMP" (7).

These cross-linguistic observations indicate two things. Firstly, there are clearly changes in how generally second-person impersonals are used in a given language. The relevant Finnish construction, for instance, seems to have been much constrained regionally and stylistically until very recently (cf. Kaiser 2015: 18, fn. 6), becoming more and more popular only in present-day Finnish (Leino & Östmann 2008: 42). This implies that, although the impersonal interpretation of second-person pronouns is transparently related to and derived from their deictic meaning via pragmatic interpretive processes, the question of whether and in which contexts this construction is actually used for the expression of impersonality depends on the degree of its conventionalization, which in turn seems to depend on the availability and distribution of alternative impersonal constructions, and possibly also on more general developments in society. For the latter, concepts such as 'intimization' and 'individualization' have been proposed (Nielsen et al. 2009; Posio 2016: 12).

Secondly, these developments, which do not appear to be correlated with changes in written genres, lend support to the idea that the English development under discussion has also taken place independently of how written genres as such have changed, albeit earlier than in the languages reported on above.

5.5 How 'involved' are second-person impersonals?

In Biber and Finegan's (1989, 2001) work, first- and second-person pronouns are part of the set of features making an English text more 'involved'. I have tried to show that the changes in frequency that we can observe for impersonal *you* in the recent history of English are not just a side effect of 'popular' genres like drama, letters and novels becoming more involved.

One might now object that a rise in the frequency of impersonal *you* and the development of new use types such as simulation and self-reference are in fact even *better* indicators of involvedness than the use of first- and second-person pronouns in general. In other words, the function of involving the hearer in a generalization by using a second-person pronouns impersonally could be expected to be another feature defining Biber and Finegan's (1989: 493) dimension 'Informational vs. involved production'.

As far as the functions of HIPs, including second-person impersonals, are concerned, however, it should be noted that not all of them should be associated with involvement. In the main, HIPs generalize over individuals or states of affairs (Deringer et al. 2015: 314–318), something which in itself is rather a characteristic of abstract style in text types aiming at objectivity, such as academic prose (Biber & Finegan 1989: 492; Biber & Conrad 2009: 116–117). As for dialogues, Scheibman (2007: 131–133), Stirling and Manderson (2011: 1598–1599) and Sorlin (2015: 138–139) show how impersonal *you* can serve to distance the speaker from the events described and involve the hearer at the same time (see also de Hoop and Tarenskeen 2015: 166 on Dutch).

In the case of impersonal *you*, the generalizing function is hence combined with the presupposition of empathy and thus the creation of solidarity between interlocutors. Even though the latter function may well be compatible with an increasing oral character of 'popular' genres, one must not forget the above-mentioned role of generalization; to generalize over individuals or states of affairs by inviting the hearer to identify with the group generalized over is not the same as deictically referring to the hearer. Further research will have to go beyond these rather speculative remarks and determine how exactly this double role of second-person HIPs is conceptually and statistically correlated with other predictors of involved production.

6. Conclusion

On the basis of the findings reported in this chapter, I think that there are reasons to believe that the increasing popularity of impersonal *you* in English is a development independent of how second-person pronouns have been used in different genres through time. First, the share of impersonal *you* in the total number of second-person pronouns has not only risen, but also developed functional subtypes (simulation and self-reference) that appear to have been much more restricted, if not non-existent, in earlier periods. Second, a rise in frequency of second-person impersonals has been observed for several languages on the basis of spoken data. It is not unlikely that impersonal *you*, including its more recent semantic developments, indeed contributes to linguistic interactions becoming more involved. Yet, given the changes in impersonal *you* described in this chapter, this changing role of involvement is likely not to merely be a consequence of shifts in popular written genres. It may as well form part of language-external developments (see Section 5.4) and ongoing language-internal rearrangements in the system of impersonal constructions.

Acknowledgments

The research reported in this paper was carried out in the context of the project *Towards a typology of impersonal pronouns* (2012–2015), funded by the *Deutsche Forschungsgemeinschaft* and *L'Agence nationale de la recherche*. I would like to thank Eliane Lorenz for helping me extract and code the data and Manfred Krug for making the ARCHER corpus available to me. I am also very grateful to Volker Gast, Lisa Deringer and Olga Rudolf for fruitful discussions during our joint work on impersonalization.

References

Biber, Douglas. 1989. A typology of English texts. *Linguistics* 27: 3–43. https://doi.org/10.1515/ling.1989.27.1.3

Biber, Douglas & Conrad, Susan. 2009. *Register, Genre and Style*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511814358

Biber, Douglas & Finegan, Edward. 1989. Drift and the evolution of English style: A history of three genres. *Language* 65: 487–517. https://doi.org/10.2307/415220

Biber, Douglas & Finegan, Edward. 2001. Diachronic relations among speech-based and written registers in English. In *Variation in English: Multidimensional Studies*, Douglas Biber & Susan Conrad (eds), 66–83. Harlow: Pearson Education.

Cabredo-Hofherr, Patricia. 2006. "Arbitrary" pro and the theory of pro-drop. In *Agreement and Arguments*, Peter Ackema, Patrick Brandt, Maaike Schoorlemmer & Fred Weermann (eds), 230–257. Oxford: OUP.

Cheshire, Jenny. 2013. Grammaticalization in social context: The emergence of a new English pronoun. *Journal of Sociolinguistics* 17: 608–633. https://doi.org/10.1111/josl.12053

Coveney, Aidan. 2003. 'Anything *you* can do, *tu* can do better': *tu* and *vous* as substitutes for indefinite *on* in French. *Journal of Sociolinguistics* 7: 164–191. https://doi.org/10.1111/1467-9481.00218

Culpeper, Jonathan & Kytö, Merja. 2010. *Early Modern English Dialogues: Spoken Interaction as Writing*. Cambridge: CUP.

Davies, Mark. 2008–. The Corpus of Contemporary American English (COCA): 520 million words, 1990–present. <http://corpus.byu.edu/coca/>

de Hoop, Helen & Tarenskeen, Sammie. 2015. It's all about *you* in Dutch. *Journal of Pragmatics* 88: 163–175. https://doi.org/10.1016/j.pragma.2014.07.001

Denison, David. 1985. Why Old English had no prepositional passive. *English Studies* 3: 189–204. https://doi.org/10.1080/00138388508598384

Deringer, Lisa, Gast, Volker, Haas, Florian & Rudolf, Olga. 2015. Impersonal uses of the second person singular and generalized empathy: An exploratory corpus study of English, German and Russian. In *The Pragmatics of Personal Pronouns* [Studies in Language Companion Series 171], Laure Gardelle & Sandrine Sorlin (eds), 311–334. Amsterdam: John Benjamins. https://doi.org/10.1075/slcs.171.15der

Egerland, Verner. 2003. Impersonal pronouns in Scandinavian and Romance. *Working Papers in Scandinavian Syntax* 71: 75–102.

Fröhlich, Jürg. 1951. *Der indefinite Agens im Altenglischen, unter besonderer Berücksichtigung des Wortes man*. Winterthur-Töß: Paul Gehring.

Gast, Volker & van der Auwera, Johan. 2013. Towards a distributional typology of human impersonal pronouns, based on data from European languages. In *Languages across Boundaries: Studies in the Memory of Anna Siewierska*, Dik Bakker & Martin Haspelmath (eds), 119–158. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110331127.119

Gast, Volker, Deringer, Lisa, Haas, Florian & Rudolf, Olga. 2015. Impersonal uses of the second person singular: A pragmatic analysis of generalization and empathy effects. *Journal of Pragmatics* 88: 148–162. https://doi.org/10.1016/j.pragma.2014.12.009

Jensen, Torben Juel. 2009. Generic variation? Developments in use of generic pronouns in late 20th century spoken Danish. *Acta Linguistica Hafniensia* 41: 83–115. https://doi.org/10.1080/03740460903364128

Jensen, Torben Juel & Gregersen, Frans. 2016. What do(es) you mean? The pragmatics of generic second person pronouns in modern spoken Danish. *Pragmatics* 26: 417–446. https://doi.org/10.1075/prag.26.3.04jen

Jud-Schmid, Elisabeth. 1956. *Der indefinite Agens von Chaucer bis Shakespeare. Die Wörter und Wendungen für "man"*. Meisenheim am Glan: Anton Hain.

Kaiser, Elsi. 2015. Impersonal and generic reference: A cross-linguistic look at Finnish and English narratives. *Eesti ja Soome-Ugri Keeleteaduse Ajakiri* 6: 9–42. https://doi.org/10.12697/jeful.2015.6.2.01

Kitagawa, Chisato & Lehrer, Adrienne. 1990. Impersonal uses of personal pronouns. *Journal of Pragmatics* 14: 739–759. https://doi.org/10.1016/0378-2166(90)90004-W

Koch, Peter & Oesterreicher, Wulf. 2007. Schriftlichkeit und kommunikative Distanz. *Zeitschrift für Germanistische Linguistik* 35: 346–375. https://doi.org/10.1515/zgl.2007.024

Laberge, Suzanne. 1980. The changing distribution of indefinite pronouns in discourse. In *Language Use and the Uses of Language*, Roger W. Shuy & Anna Shnukal (eds), 76–87. Washington DC: Georgetown University Press.

Leino, Pentti & Östmann, Jan-Ola. 2008. Language change, variability and functional load: Finnish genericity from a constructional point of view. In *Constructional Reorganization* [Constructional Approaches to Language 5], Pentti Leino (ed.), 37–54. Amsterdam: John Benjamins. https://doi.org/10.1075/cal.5.03lei

Light, Caitlin & Wallenberg, Joel. 2015. The expression of impersonals in Middle English. *English Language and Linguistics* 19: 227–245. https://doi.org/10.1017/S1360674315000076

Los, Bettelou. 2002. The loss of the indefinite pronoun *man*. In *English Historical Syntax and Morphology. Selected Papers from 11 ICEHL, Santiago de Compostela, 7-11 11 September 2000* [Current Issue in Lingustic Theory 223], Teresa Fanego, Maria Jose Lopez-Couso & Javier Perez-Guerra (eds), 181–202. Amsterdam: John Benjamins.

Los, Bettelou. 2009. The consequences of the loss of verb-second in English: Information structure and syntax in interaction. *English Language and Linguistics* 13: 97–125. https://doi.org/10.1017/S1360674308002876

Malamud, Sophia. 2012. Impersonal indexicals: *One*, *you*, *man* and *du*. *Journal of Comparative Germanic Linguistics* 15: 1–48. https://doi.org/10.1007/s10828-012-9047-6

Meyer, Hans-Heinrich. 1953. *Der indefinite Agens im Mittelenglischen (1050–1350). Wörter und Wendungen für "man"*. Bern: Francke.

Nielsen, Søren, Fogtmann, Christina & Jensen, Torben Juel. 2009. From community to conversation — and back: Exploring the interpersonal potentials of two generic pronouns in Danish. *Acta Linguistica Hafniensia* 41: 116–142. https://doi.org/10.1080/03740460903364151

Posio, Pekka. 2016. You and we: Impersonal second person singular and other referential devices in Spanish sociolinguistic interviews. *Journal of Pragmatics* 99: 1–16. https://doi.org/10.1016/j.pragma.2016.04.014

Rissanen, Matti. 1997. Whatever happened to Middle English indefinite pronouns. In *Studies in Middle English Linguistics*, Jacek Fisiak (ed.), 513–529. Berlin: De Gruyter. https://doi.org/10.1515/9783110814194.513

Scheibman, Joanne. 2007. Subjective and intersubjective uses of generalizations in English conversations. In *Stancetaking in Discourse* [Pragmatics & Beyond New Series 164], Robert Englebretson (ed.), 111–138. Amsterdam: John Benjamins. https://doi.org/10.1075/pbns.164.06sch

Seoane, Elena. 2006. Information structure and word order change: The passive as an information-rearranging strategy in the history of English. In *Handbook of the History of English*, Ans van Kemenade & Bettelou Los (eds), 360–391. Oxford: Blackwell. https://doi.org/10.1002/9780470757048.ch15

Seoane Posse, Elena. 2000. Impersonalising strategies in Early Modern English. *English Studies* 18: 102–116. https://doi.org/10.1076/0013-838X(200003)81:2;1-T;FT102

Siewierska, Anna. 2008. Introduction: Impersonalization from a subject-centered vs. agent-centered perspective. *Transactions of the Philological Society* 106: 115–137. https://doi.org/10.1111/j.1467-968X.2008.00211.x

Siewierska, Anna. 2011. Overlap and complementarity in reference impersonals: *man*-constructions vs. third-person plural impersonals in the languages of Europe. In *Impersonal Constructions: A Cross-linguistic Perspective* [Studies in Language Companion Series 124], Andrej L. Malchukov & Anna Siewierska (eds), 57–90. Amsterdam: John Benjamins. https://doi.org/10.1075/slcs.124.03sie

Siewierska, Anna & Papastathi, Maria. 2011. Towards a typology of third-person impersonals. *Linguistics* 49: 575–610. https://doi.org/10.1515/ling.2011.018

Sorlin, Sandrine. 2015. Breaking the fourth wall: The pragmatic functions of the second person pronoun in *House of Cards*. In *The Pragmatics of Personal Pronouns* [Studies in Language Companion Series 171], Laure Gardelle & Sandrine Sorlin (eds), 125–145. Amsterdam: John Benjamins. https://doi.org/10.1075/slcs.171.07sor

Stirling, Lesley & Manderson, Lenore. 2011. About *you*: Empathy, objectivity and authority. *Journal of Pragmatics* 43: 1581–1602. https://doi.org/10.1016/j.pragma.2010.12.002

Tarenskeen, Sammie. 2010. From You to Me (and Back): The Flexible Meaning of the Second-person Pronoun in Dutch. MA thesis, Radboud University Nijmegen.

Trudgill, Peter & Hannah, Jean. 2008. *International English: A Guide to the Varieties of Standard English*, 5th edn. London York: Arnold.

van Bergen, Linda D. 2000. The indefinite pronoun *man*: 'nominal' or 'pronominal'? In *Generative Theory and Corpus Studies: A Dialogue from 10 ICEHL*, Ricardo Bermùdez-Otero, David Denison, Richard M. Hogg & Chris B. McCully (eds), 103–122. Berlin: De Gruyter. https://doi.org/10.1515/9783110814699.103

Visser, Fredericus Theodorus. 1973. *An Historical Syntax of the English Language, Part Three, Second Half*: *Syntactical Units with Two and with more Verbs*. Leiden: Brill.

Wales, Kathleen. 1985. Generic *your* and Jacobean drama: The rise and fall of a pronominal usage. *English Studies* 66: 7–24. https://doi.org/10.1080/00138388508598364

Zifonun, Gisela. 2000. "Man lebt nur einmal." Morphosyntax und Semantik des Pronomens *man*. *Deutsche Sprache* 28: 232–253.

Concessive conjunctions in written American English

Diachronic and genre-related changes in frequency and semantics

Ole Schützler
University of Bamberg

Based on the *Corpus of Historical American English* (COHA; Davies 2010–), this chapter inspects the frequencies and semantics of the concessive conjunctions *although*, *though* and *even though* from the 1860s to the present day. In the data, *although* and *though* predominantly express what Sweetser (1990) has called *speech-act concessives*, while *even though* mainly expresses *content concessives*. However, there is a general development towards a higher proportion of speech-act concessives. Further, *although* and *even though* increase in frequency, while *though* decreases over time. Semantic properties and the double function of *though* (conjunction and conjunct) are proposed as explanations. Frequency changes progress equally through all genres, but the semantic change seems to have pervaded all genres only as far as *although* is concerned.

1. Introduction

This chapter focuses on understudied aspects of concessive constructions in English, more specifically the concessive conjunctions *although*, *though* and *even though* in written American English. While a number of studies address the structure and semantics of concessive clauses in general (e.g. König 1985; Hermodsson 1994; Azar 1997; di Meola 1998), corpus-based studies of a quantitative nature are lacking (exceptions being Aarts 1988 and Hilpert 2013), in particular studies that take an interest in the semantic variability of concessives. In this chapter, the different semantic types of concession that are typically encoded by *although*, *though* and *even though* as well as the frequencies of the markers themselves are inspected in diachrony and across genres. The central question is how far the three syntactically equivalent conjunctions express – or have come to express – different semantic relations between propositions, and whether or not those differences are variable across different genre-related contexts.

https://doi.org/10.1075/scl.85.09sch

Syntactically, *although*, *though* and *even though* are subordinating conjunctions, whose clausal complements may be finite, non-finite or verbless, as in (1)–(3).

(1) **Although** Paul Taylor once studied painting, he is not a painter on canvas.
(COHA, 1985, newspapers)

(2) Thus, **although** retreating, we were always ready to fight.
(COHA, 1892, fiction)

(3) **Although** overlong, the picture has a fair measure of jolts and surprises.
(COHA, 2003, newspapers)

The connective *though* can also occur as what is called a 'connective adjunct' by Huddleston and Pullum (2002: 736), a 'linking adverbial' by Biber et al. (1999: 850–851), and a 'conjunct' by Quirk et al. (1985: 632), namely an element that has a linking function between sentences, not intra-sententially between clauses. This use of the form *though* is shown in Example (4).

(4) Fred Shepley said he hadn't ever paid any attention to the lions. He wished he had, **though**. (COHA, 1940, fiction)

The conjunct is included in some of the analyses as it can help explain certain patterns of change. However, it is excluded from all analyses at the semantic and stylistic levels.

The conjunctions *although*, *though* and *even though* are grammatically interchangeable and etymologically (and morphologically) related via the Old English form *þēah*, from which diverse variants were derived (OED online, *s.v.* 'though'; cf. Burnham 1911: 12–14; Chen 2000: 104–105). In dialect contact scenarios, this kind of co-existence of several seemingly equivalent variants in a language or dialect may point to reallocation (Trudgill 1986; Britain & Trudgill 2000: 73–74). That is, different variants are retained because they have taken on special socio-stylistic or linguistic (including semantic) functions. Although not concerned with dialect contact, the present chapter nevertheless explores similar questions: Why should forms continue to coexist if, on a surface level, they appear to have similar functions and are moreover formally and etymologically related as in the present case?

The chapter is structured as follows: First, the relevant research background is outlined (Section 2); this includes a discussion of the three relevant semantic types of constructions that are investigated, a summary of what has been claimed concerning existing stylistic differences between markers, and a formulation of the research questions. The remaining sections of the chapter discuss corpus-methodological issues (Section 3), present and interpret the results (Section 4) and, finally, summarise the main findings and point to possible future research (Section 5).

2. Research background

2.1 Three semantic types of concessives

In this section, the three relevant semantic types of concessives will be illustrated. Examples are based on constructed examples, since they will be more immediately comparable this way and differences between them will become clearer than would be the case if corpus examples were used. Examples from actual American English usage will be discussed when presenting results.

Three different categories of concessives are described in Sweetser (1990: 76–77): *content*, *epistemic* and *speech-act concessives*.[1] Content concessives are constructed and decoded based on so-called *topoi* (Azar 1997: 306; Anscombre 1989). A topos is a set of presuppositions based on world knowledge or contextual knowledge that is shared by language producer (i.e. speaker/writer) and recipient (addressee/reader). Example (5), for example, operates on a topos according to which hard work will generally result in success, a special case of success being promotion. The connection between the two propositions is construed as a real-world causal or conditional relation.

(5) **Although** <u>Jonathan worked very hard last year</u>, he did not get promoted.

Example (6) is an example of Sweetser's second type, *epistemic concessives*. In this case, there is no real-world relation of cause and effect, since not getting promoted can hardly be regarded as resulting in somebody's having failed to work hard in the past. Rather, the fact that somebody is overlooked when it comes to promotion may lead to the conclusion that he or she probably did not make a sufficient effort – a conclusion which in this case turns out to be false.

(6) Jonathan worked very hard last year, **although** <u>he did not get promoted</u>.

Content and epistemic concessives are both based on topoi, but in epistemic concessives the two propositions are held together not by an 'if-then' relation, but by an inference or conclusion based on the proposition that is headed by the conjunction and thus marked as the subordinate clause.

[1] Crevels (2000: 315–317) identifies *textual concessives* as a fourth type, in which the clause (or sentence) marked for concession does not obviously build upon another clearly delineated syntactic unit but 'stretches over a whole series of preceding utterances […]'. While this type differs in syntactic or quantitative terms, it will still belong to one of the three categories sketched above. In the data analysed for this chapter, textual concessives were quite rare and therefore excluded from the analyses.

Concessives of the third type described by Sweetser (1990) – so-called *speech-act concessives* – function rather differently, as shown in Example (7).

(7) Jonathan is a very hard worker, **although** <u>he is not very talented</u>.

The proposition in the matrix clause ('Jonathan is a very hard worker') suggests a positive evaluation of the subject, Jonathan, and a number of possible positive consequences (e.g. success in general, high earnings, or promotion). The proposition in the subordinate clause ('Jonathan is not very talented') makes contrasting negative scenarios more likely (e.g. a general lack of success, lower earnings, not getting promoted). There is no topos according to which a lack of talent correlates with a lack of effort. Rather, the contrast is between two different pragmatic stances. It could be argued that in speech-act concessives, the two propositions do not belong to the same topos but trigger two independent (and contrasting) topoi.

It has been claimed that the types of concessives discussed by Sweetser (1990; cf. Crevels 2000) correspond to different degrees of subjectivity (Hilpert 2013: 165; Traugott & Dasher 2002: 89–99). In content concessives, the contrast between propositions could be said to be objective in the sense that it is based on real-world likelihoods and expectations. While they are usually still based on topoi, epistemic concessives hinge upon conclusions or inferences, not on real-world causes and effects. In concessives of this type, the stance of the speaker (or writer) becomes more transparent – or, in other words, the speaker/writer becomes more visible as a reasoning subject. Accordingly this semantic type can be regarded as more subjective. Finally, speech-act concessives neither operate on a topos-based direct relation between propositions, nor are reasoning processes of the speaker/writer crucial in meaning-making. Instead, the contrast between propositions corresponds to two different pragmatic stances. The contrast is thus of a purely pragmatic nature, which is why this type is regarded as the most subjective of the three.[2]

The only quantitative approach to the semantic structure of concessive constructions in English is Hilpert's (2013: 181ff.) study of *although* and *though* (together with other connectives) in written twentieth-century American English. Hilpert shows that, at least as far as full clauses are concerned, *although* is predominantly found in content concessives. This is also (but less clearly) true for *though*, which is associated with a considerably higher proportion of epistemic concessives. The focus of Hilpert's study is not on the correlation of subordinators

[2] Rather than placing the three types of concessives on a subjectivity cline, it might be more appropriate to speak of a subjectivity-intersubjectivity cline, since the sense of a speech-act concessive will in many (if not all) cases have to be established by the addressee/reader. This theoretical and terminological aspect will be pursued further in future research.

and semantic types more generally, but on the testing of a specific hypothesis concerning the emergence of a particular construction, parenthetical concessives. Only concessives with co-referential subjects in both clauses are relevant in this context, and results can therefore not necessarily be assumed to hold for concessives more generally.

2.2 The stylistics of concessive conjunctions

Some differences in formality between *although*, *though* and *even though* are described in the literature, but they are rarely based on quantitative analyses. For example, Quirk et al. (1985: 1098) comment that *though* is less formal than *although*, a view that is essentially echoed by Huddleston and Pullum (2002: 736). *Even though*, on the other hand, is regarded as an emphatic variant that also expresses unexpectedness (Quirk et al. 1985: 1099). Biber et al. (1999: 842) find that *although* is much more frequent than *though* and *even though* in academic writing, while in conversation and fictional writing *though* (and *even though*) are more frequent; this pattern essentially confirms what Quirk et al. say about the difference in formality between the conjunctions. Aarts (1988) investigates the stylistic functions of *although*, *though* and *even though* in written British English and shows that *although* is more frequent in formal styles and that, in consequence, it is less evenly distributed across different text categories than the other two conjunctions. According to Aarts (1988), *although* responds most strongly to different styles or genres, followed by *though*, while *even though* is least sensitive. In the present chapter, style or genre will be viewed in combination with semantics. That is, the question is not only whether the markers differ in frequency between types of text, but also whether their preferred semantic structures vary along the stylistic dimension.

2.3 Research questions

Beyond certain stylistic differences, very little is known about functional differences between *although*, *though* and *even though*. Particularly the semantic differences between markers and their diachronic development (or stability), both in frequency and semantics, have not been quantitatively investigated. To fill this gap, this chapter addresses the following research questions, based on written American English data from the middle of the 19th century to the present day:

1. What are the frequencies of the three concessive conjunctions in the early 21st century, and how have they changed over the preceding 150 years?
2. Which semantic types of concessives do the three conjunctions preferably encode – in particular: Are there significant differences between them in this respect, and has this changed diachronically?

3. How does genre affect both the frequencies and the semantics of the conjunctions under investigation, i.e. do certain genres serve as catalysts in changes in frequency or semantics?

Underlying processes that may play a role are the ongoing grammaticalization of one or several of the markers (cf. Hopper 1991; Hopper & Traugott 2003), the reallocation of different markers to different semantic functions (cf. research in social dialectology, e.g. Trudgill 1986; see above), or – if semantic change happens in a certain direction – a process of (inter-)subjectification. The latter point is of particular importance: On the one hand it has been argued that there is a semantic hierarchy within classes of adverbials, e.g. an ordered set of semantic subtypes of concessives (cf. Crevels 2000; Hilpert 2013: 164–165), which is often taken to imply that those types emerge in a certain order. On the other hand, this hypothesis has not been tested quantitatively.

3. Methodology

All analyses in this chapter are based on data from the *Corpus of Historical American English* (COHA; Davies 2010–). COHA contains approximately 406 million words in four written genres (fiction, non-fiction, magazines, and newspapers), arranged in successive decades from the 1810s to the 2000s (see appendix for details). For the present study, only material dating from the 1860s to the 2000s is used, since there are no newspaper texts in the earlier decades and genre-based analyses would therefore be limited for that period. In the selected part of the corpus (1860s–2000s) there are 352 million words (i.e. between 17.1 and 29.5 million per decade). The dominant genre is fiction, which accounts for approximately 50% of the data throughout the corpus. This imbalance makes it all the more important to inspect the different genres separately.

General inspections of frequency developments in the corpus or in specific genres are based on all occurrences of the three conjunctions, regardless of semantic type. When the focus is on diachronic changes in the frequencies of semantic types, or on semantic types in interaction with genre, two subcorpora are used, one for the late 19th century (1860s–1900s) and one for the late 20th century (1960s–2000s). Each subcorpus consists of 480 randomly selected tokens balanced by conjunction and genre. That is, a total of 960 tokens were semantically disambiguated, as detailed in Table 1.

Frequencies are reported in words per million (pmw), normalised within each decade of the corpus. For some plots, frequency was scaled logarithmically in order to give a more accurate impression of relative diachronic changes over time. For

Table 1. Numbers of randomly sampled and manually disambiguated tokens by time period, conjunction and genre

Period	Conjunction	Genre			
		Fiction	Popular magazines	Newspapers	Non-fiction books
1860s–1900s	*although*	40	40	40	40
	though	40	40	40	40
	even though	40	40	40	40
1960s–2000s	*although*	40	40	40	40
	though	40	40	40	40
	even though	40	40	40	40

example, an increase in frequency from 200 to 250 occurrences (pmw) would at first seem to be equivalent to an increase from 50 to 100 occurrences (pmw), when in fact the latter is much more substantial, being an increase by 100%, while the former is a mere 25% increase. In the respective plots in this chapter, normalised, easy-to-read frequencies are shown, but the spacing of values on the y-axis is logarithmic, showing differences based on ratios, not raw differences.

In some cases, diachronic trends were tested using Kendall's τ (cf. Hilpert & Gries 2009). In this approach, the fifteen consecutively ordered decades in COHA are correlated with the normalised frequencies of conjunctions, the correlation indicating whether or not there is a significant trend. For example, if we assign the ascending numbers 1–15 to consecutive decades (1860s–2000s), and if those numbers correspond to increasing frequencies, this will show up as a positive correlation, signifying a positive diachronic trend. Whether or not the correlation is statistically significant depends on how regular this correspondence between temporal categories and frequencies is. In like fashion, there will be a negative correlation if ascending numbers of decades correspond to descending frequencies. When comparing distributional patterns of the three semantic types between groups (cf. Figures 3 & 5), the chi-squared (χ^2) test was applied (cf. Agresti & Finlay 2009: 224–229). It was performed on the raw scores (counts) of variants in the respective groups, while the plots only show proportions.

As explained in the discussion of semantic types above, higher rates of epistemic and speech-act concessives may be regarded as more subjective. To gauge this meta-property of constructions more precisely, a scale was used in some of the analyses, assigning a subjectivity value of zero to content concessives, a value of 0.5 to epistemic concessives, and a value of 1 to speech-act concessives. The average degree of subjectivity was then calculated for different time periods and conjunctions. This subjectivity scale is used to complement the more fine-grained inspection of the relative frequencies of specific semantic types.[3] It constitutes a case of aggregative analysis, in which multi-faceted results are expressed more simply on a meta-level (cf. Szmrecsanyi 2013) – in this case, proportions of three different semantic types of concessives are aggregated into a single subjectivity value.

The data were extracted from the corpus via the online interface of COHA (http://corpus.byu.edu/coha/). Statistical analyses were conducted in R (R Development Core Team 2016) and RStudio (RStudio Team 2009–2015); for graphical displays the R-package *lattice* (Sarkar 2014) was used.

4. Results

4.1 Corpus examples

In contrast to the constructed examples used earlier, this section shows cases from actual usage as found in COHA. In principle, all three semantic types (content, epistemic and speech-act concessives) can occur in connection with all three conjunctions (*although*, *though* and *even though*), as shown in examples (8)–(16). Not all combinations are equally likely, of course, and some must even be considered rare, as the quantitative analyses below will demonstrate. Nevertheless, the evidence suggests that the semantic functions of the three conjunctions overlap. Examples (8)–(10) illustrate constructions that belong to the category of content concessives.

(8) George tells me that **although** he is long ago of age, he has as yet received no portion of his father's estates. (COHA, 1897, fiction)

(9) When through with his day's work, **though** his bones ached and his eyes were drowsy, he seldom went to sleep without first studying awhile [*sic*] […]. (COHA, 1866, fiction)

(10) **Even though** I had cash in hand, nobody would rent an apartment to me […]. (COHA, 1999, magazines)

In (8), the underlying topos is that, for a young man in the given context, coming of age will normally result in his being given a share of his inheritance. This example from the late nineteenth century is interesting in using a topos that is no

[3] The relative degree of subjectivity of the three semantic types is of course rather difficult to assess; for example, it is not clear whether epistemic concessives are closer to content or to speech-act concessives in subjectivity. The scale is therefore used only sparingly and for certain purposes in this chapter and in other research on this topic.

longer valid today, thus illustrating that topoi need not be universal but may well be restricted to a certain time, culture or context. In Example (9), being exhausted from a hard day's work is likely to result in one's going to bed straight away rather than making the effort of extra study – a topos that will be rather widely shared. Finally, in (10), it can generally be presupposed that someone who demonstrably has the money to pay the rent will be an attractive tenant; thus, the two propositions are not in harmony.

Examples (11)–(13) illustrate epistemic concessives. As will be shown in the quantitative analyses below, this semantic type is rather rare overall, but it nevertheless occurs in combination with all three conjunctions.

(11) […**A**]**lthough** five feet seven inches by seven feet in size, it is a prayer rug […]. (COHA, 1904, magazines)

(12) [… W]hen I heard the child was dead I did not care so very much, **though** I wrote to her kindly enough […]. (COHA, 1871, fiction)

(13) The man had only fainted, **even though** his eyes stared upward, open and unseeing. (COHA, 1961, fiction)

The construction shown in (11) is a particularly clear-cut case of epistemic concession: If one observes that a rug is 5′7″ by 7′ in size (ca. 3.6 m^2), one might draw certain conclusions as to its functions, but that of a prayer rug would probably not be among them, since such rugs will normally be considerably smaller.[4] Thus, in the example the proposition marked by *although* triggers certain conclusions and inferences which turn out to be false. In Example (12), writing a kindly letter in the given scenario will normally be expected to result from kindly feelings of pity or compassion. However, the writer in the example states that he or she did not have such feelings, and thus the inference based on the observed behaviour turns out to be false. Finally, in Example (13), upward-staring, open but unseeing eyes are likely to lead to the conclusion that a person is dead. In this example, however, the person referred to by the subject of the matrix clause has merely fainted. So, once again, conclusions drawn on the basis of observed evidence turn out not to be in harmony with reality, which is why the entire construction is marked as a concessive. Constructing the sentences the other way round (e.g. 'Although I did not care, I wrote to her kindly') results in relatively straightforward content concessives. Indeed, it often seems possible to construct pairs of content and epistemic concessives that are equivalent in propositional content but differ in semantic polarity, which is effected by changing the status of clauses (matrix ↔ subordinate).

[4] This topos basically works everywhere except in those societies where large prayer rugs of this type are in fact used, and of course in communities that know nothing about prayer rugs.

The final three examples present cases of speech-act concessives found in COHA.

(14) [… C]ellulose has no nutritional value **although** it may serve a useful purpose as roughage. (COHA, 1982, non-fiction)

(15) […] Sears […] is well on its way to catching up with Arizona, analysts say, **though** the company would not disclose sales figures […]. (COHA, 1997, newspapers)

(16) [… H]e is not blind to their little weaknesses, but these he can forgive **even though** he refuses to forget […]. (COHA, 1878, fiction)

In (14), the two propositions ('cellulose has no nutritional value' and 'cellulose is useful as roughage') are neither linked at the content nor at the epistemic level. What this construction does is present two pragmatic perspectives on the same situation or, in this case, object: cellulose. The proposition in the matrix clause suggests a negative evaluation of cellulose, low nutritional value meaning that it is rather worthless as fodder, while the proposition in the subordinate clause indicates that it does have a certain value (even if this is unrelated to its nutritional value). Thus, the construction as a whole supplies negative and positive evidence concerning the evaluation of cellulose, without recourse to any kind of 'if-then' relation. In (15), two competing brands of jeans (Sears and Arizona) are under discussion, and the proposition in the matrix clause suggests that Sears is gaining ground in this competition. However, the subordinate clause qualifies this by stating that no sales figures are available to substantiate this impression, which effectively makes the first statement somewhat less convincing. One proposition is qualified by another, but once again there is no underlying topos. Finally, in (16), the matrix clause (beginning with *but*) states that weaknesses are forgiven, while the subordinate clause makes the qualifying addition that forgiving does not automatically mean forgetting. The latter qualification takes away from the positive first statement, but it does not depend on an *a priori* assumption to the effect that someone who forgives should also automatically forget.[5]

As stated above, the corpus examples collected in those paragraphs illustrate that it is possible for all three conjunctions to encode all three semantic types of concessives. A more detailed quantitative picture of the semantic preferences that characterise the different conjunctions will be provided below.

4.2 Frequencies

Based on COHA, Figure 1 traces diachronic changes in the frequency of the three conjunctions *although*, *though* and *even though*, complemented with data for the

[5] Example (16) also clearly plays with the fixed phrase/collocation *to forgive and forget*.

conjunct *though*, as shown in Example (4) above. Correlation tests (cf. Hilpert & Gries 2009; also see above) reveal that all four connectives undergo significant changes between the 1860s and the 2000s: *although* increases, *though* decreases, and both *even though* and the conjunct *though* increase considerably.[6]

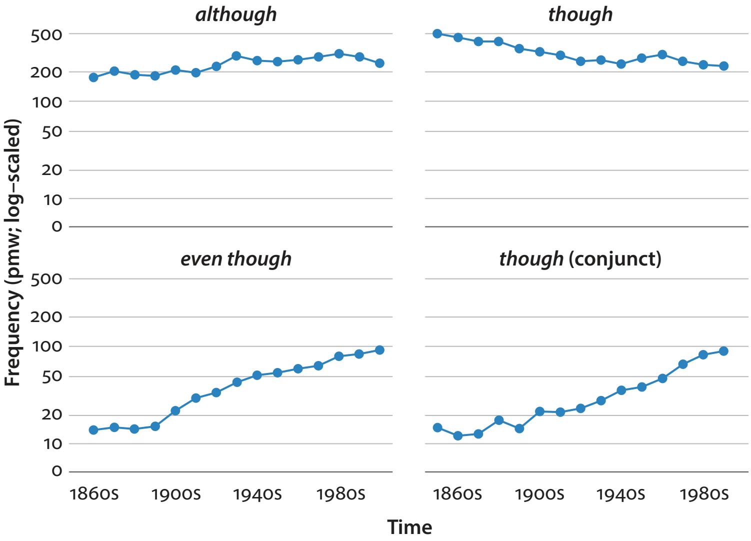


Figure 1. Frequency changes of *although*, *though*, *even though* and *though* (conjunct) in COHA

At the beginning of the period, *though* is quite clearly the most popular of the three conjunctions, while *even though* is very much a minority form. The conjunct *though* is often regarded as a feature of speech rather than writing (cf. Biber et al. 1999: 850–851), and, to some extent at least, its dramatic increase in frequency in the written COHA data may therefore be due to a process of colloquialisation, whereby structures typically found in speech become more readily available in written genres as well (cf. Hundt & Mair 1999; Mair 2006: 187; Smitterberg 2014; Smitterberg & Kytö 2015). Apart from a general process of colloquialisation, the genre patterns shown below will also help to account for this phenomenon.

That there should be some changes in the system of concessive conjunctions does not come as a surprise. After all, as stated initially, the three have the same

[6] *although*: $\tau = .619$, $p = .001$; *though*: $\tau = -.810$, $p = .000$; *even though*: $\tau = .981$, $p = .000$; *though* (conjunct): $\tau = .905$, $p = .000$.

syntactic function, and it would only be expected that some simplification (in the sense of a reduction of available different forms) should take place (cf. Hopper 1991: 22). However, the nature of the changes is not easily explicable. For example, if we assume that Quirk et al. (1985: 1098) are correct in claiming that *though* is less formal than *although*, and if we equally accept that present-day English is characterised by processes of colloquialisation, then the development of *though* – a decrease in frequency – would mean that the less formal (i.e. the more colloquial) form decreases in frequency. In consequence, we either have to assume that Quirk et al.'s (1985) statement does not in fact hold true, or we must assume that it is perhaps valid for British English only; as a third possibility, we could look for alternative explanations. The decrease of the supposedly less formal conjunction (*though*) in combination with the increase of the supposedly more formal one (*although*) may simply suggest that colloquialisation is not a regular and exceptionless process; with regard to the conjunctions investigated in this chapter it seems to have little explanatory power.

To account for the observed changes in frequency, one can also invoke the process of specialisation (Hopper 1991: 22), a sub-process of grammaticalization, whereby the number of formal options within a grammatical domain (here: concessive conjunctions) tends to be reduced, and the domain becomes more regular. This explanation is of limited use for two reasons: First, the conjunction *though* is the most frequent one to start with, at a frequency of 498 occurrences (pmw) in the 1860s as compared to 177 and 14 occurrences (pmw) for *although* and *even though*, respectively (cf. Figure 1). It is hard to see why specialisation should not continue in this direction, i.e. by making the most frequent marker even more frequent. Secondly, *though* would be the better candidate for economic reasons, being shorter than *although* and *even though*.

One possible explanation supported by the corpus evidence shown in Figure 1 would be that in the period under investigation, the form *though* is in the early stages of developing from a subordinating conjunction into a conjunct. In the 2000s, the conjunction *though* is still much more frequent than the conjunct, but the difference is very much reduced, with 498 vs 15 occurrences pmw in the 1860s as compared to 232 vs 92 occurrences pmw in the 2000s. In sum, one factor that can help account for the decline in frequency of the conjunction *though* is the concomitant increase in frequency of the conjunct of the same form, the underlying mechanism perhaps being a tendency for the language system to allocate one meaning to one form.

Next, changes in frequency will be inspected in the individual genres of COHA. Figure 2 is arranged along the same lines as Figure 1 but highlights differences between the four genres in the corpus. The lines have been smoothed to reduce noise in the data and thus make results easier to interpret. All trends shown in Figure 2 are significantly positive or negative, with the exception of *although* in newspaper texts.[7]

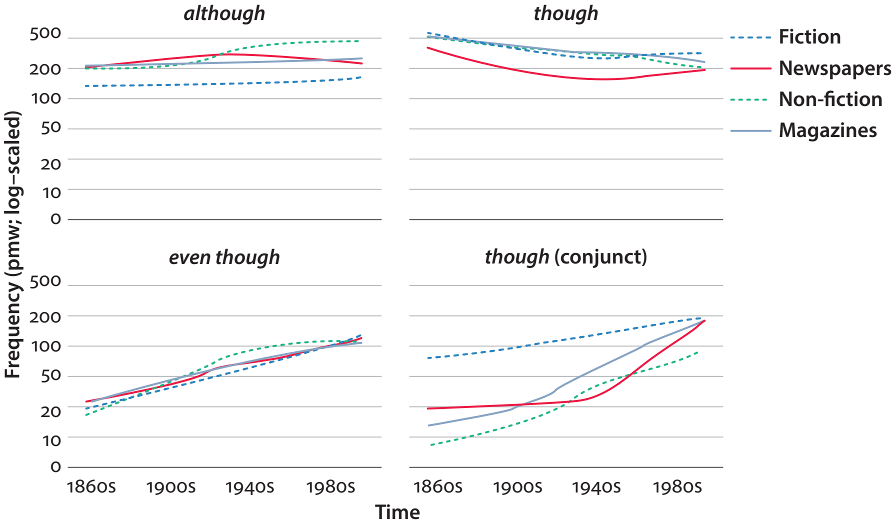


Figure 2. Frequency changes of *although*, *though*, *even though* and *though* (conjunct) in four genres of COHA

Genre patterns for the conjunction *though* and particularly for *even though* are remarkably regular, i.e. there are no marked differences between genres both as regards frequencies and rates of change, except that *though* is considerably less frequent in newspapers throughout the period under investigation. The conjunction *although* appears to increase most dynamically in non-fiction books and to be less frequent overall in fictional texts (cf. top-left panel of Figure 2).

The diachronic pattern found for the conjunct *though* (in the bottom-right panel of Figure 2) is more striking. From the 1860s well into the twentieth century, this marker is much more frequent in fictional texts than in other genres. However, the other three genres can be observed to catch up, and by the 2000s the frequency of the conjunct is remarkably similar in fictional texts, newspapers and magazines. It is possible that this change is a symptom of colloquialisation, i.e. the spreading of a feature predominantly used in (informal) conversation – and thus occurring in the fictional dialogues that can make up relatively large parts of fictional texts – to other types of text. Of course, this would need to be tested by checking how

[7] Once again, the correlation method suggested by Hilpert and Gries (2009) was used to test for significant diachronic trends.

many instances of *though* (conjunct) in fictional texts in COHA actually occur in dialogue. The use of the conjunct may also simply be a feature of narrative texts in general, irrespective of the amount of dialogue they contain.

In sum, as far as frequencies are concerned, genre patterns in COHA are fairly parallel and homogeneous for the most part (with the exception of the very interesting patterns displayed by the conjunct) – a finding that is perhaps unsurprising, considering that COHA contains written language from a single standard variety of English.

4.3 Semantics

The second aspect of the diachronic and genre-related changes of the three conjunctions *although*, *though* and *even though* concerns the semantic types of concessives that they preferably encode. As discussed in the second subchapter ('Research background'), three different types are assumed to exist (content, epistemic and speech-act), which, moreover, can be ranked on a continuum of smaller or greater subjectivity.

Figure 3 shows the proportions of the three semantic types. The three panels of the plot correspond to the three conjunctions, and each panel shows the diachronic difference between the subsamples from the late nineteenth and the late twentieth centuries, respectively (cf. the discussion of the subsampling procedure in the section on methodologies).

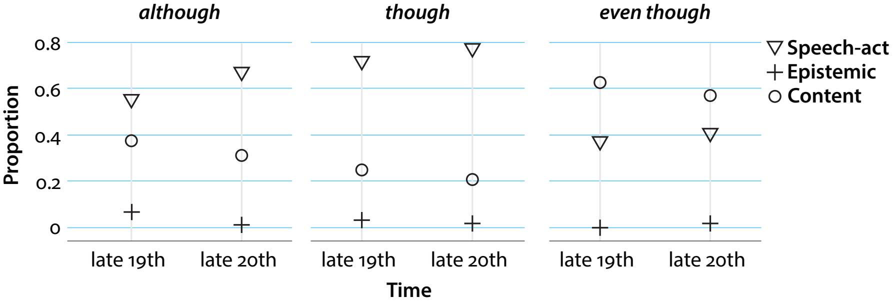


Figure 3. Proportions of semantic types in diachrony

What is readily apparent is a fundamental difference between *although* and *though* on the one hand and *even though* on the other. While *although* and *though* predominantly encode speech-act concessives (as well as a sizeable proportion of content concessives), *even though* predominantly encodes content concessives (as well as a fair number of speech-act concessives). Compared to *although*, *though* is somewhat more specialised towards the encoding of speech-act concessives. This basic pattern appears to hold for the entire COHA, i.e. it does not change between the earlier and the later sample. Whatever the stylistic or emphatic value of *even though* (cf. Quirk et al. 1985: 1099), there is also a clear difference in semantic function that sets this conjunction apart from the other two.

The second noteworthy aspect is that epistemic concessives appear to be quite low in frequency overall. Regarding this semantic type, the diachronic evidence is conflicting: The proportions shown in Figure 3 correspond to an absolute decrease from eleven to two epistemic cases in combination with *although* (a change from 6.9% to 1.2% of the respective total subsamples) and an absolute decrease from five to three cases in combination with *though* (a change in proportion from 3.1% to 1.9%). By contrast, there is an increase from zero to three cases in combination with *even though* (i.e. this type rises to a proportion of 1.9% in the subsample from the late 20th century). However, it has to be said that the rarity of this type does not allow strong generalisations, which is why the epistemic type plays a relatively marginal role in the discussion of results.

Although the general semantic preferences of the three conjunctions do not change radically over time, there are certain diachronic developments, which, moreover, are quite regular. Leaving aside the rather marginal epistemic concessives, for *although* and *though* the gap between the proportions of speech-act concessives and content concessives widens over time. This is due to a relative increase in the number of speech-act concessives found in combination with those two conjunctions, as well as a concomitant decrease in the number of content concessives. The change is statistically significant in the case of *although* (χ^2 (2) = 8.97; p = .011; φ = .167), while for *though* it is not (χ^2 (2) = 1.51; p = .470; φ = .069). As can be seen, the effect size is rather modest even for *although*, so we certainly cannot speak of a particularly dramatic change.[8] The same general tendency – i.e. a diachronic increase in the proportion of speech-act concessives and a decrease in the proportion of content concessives – is evident in the third conjunction, *even though*. Here, however, the result is a reduction in the gap between those two semantic types, i.e. the preference of content-type constructions in combination with *even though* becomes less marked over time. This development, too, is not statistically significant, and the effect size is small (χ^2 (2) = 3.71; p = .157; φ = .108).

Using the subjectivity scale that was proposed, Figure 4 translates proportions of different semantic types into degrees of subjectivity, again showing diachronic developments by conjunction and time period. Firmly based on the assumption that the three semantic types correspond to different degrees of

[8] The effect size is measured using Cramers Phi (φ), which is a standardised assessment of differences, in this case between two distributions (late 19th century and late 20th century), with φ taking values between zero and one.

subjectivity (speech-act concessives being the most subjective and content concessives being the least subjective type), this representation abstracts away from the specific semantic types and aggregates their proportions into a single score. Subjectivity thus functions as a meta-variable which reduces complexity at the cost of ignoring detail.

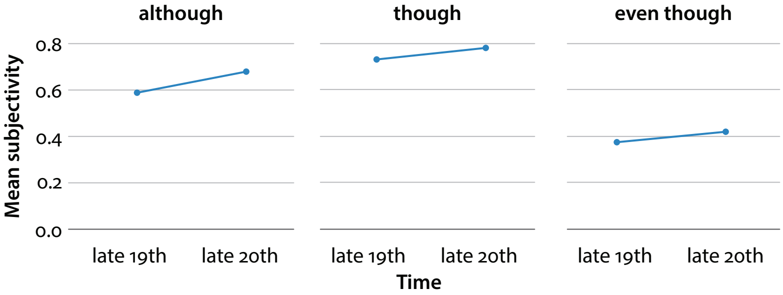


Figure 4. Subjectivity in diachrony

Finally, proportions of semantic types in the four different genres in COHA were investigated, again including the diachronic dimension. Results are shown in Figure 5, where the three tiers correspond to the three conjunctions, the four panels on each tier correspond to the four genres in COHA (fiction, non-fiction, newspapers, magazines), and within each individual panel the difference between the earlier and the later data is indicated. None of the individual differences within each of the twelve panels is statistically significant, partly of course due to the relatively modest size of the subsample (n = 80 within each panel). Nevertheless, a few noteworthy patterns will be discussed below.

If we accept the approach of mapping semantic types onto much more abstract quantitative degrees of subjectivity, the general trend for concessives constructed with all three conjunctions is to become more subjective in the period under investigation. Directly comparing Figures 3 and 4, it is evident that the detected increases in subjectivity correspond mainly to increases in the proportion of speech-act concessives, combined with decreases in the proportion of content concessives, as discussed above. In other words, the relationship between concrete semantic patterns and abstract degrees of subjectivity is relatively straightforward, since there are only relatively minor changes in the proportion of the (relatively rare) epistemic concessives. This is particularly true for *though* and *even though*; as far as *although* is concerned, the relative number of epistemic concessives decreases somewhat more substantially, which affects but does not fundamentally alter the general trend.

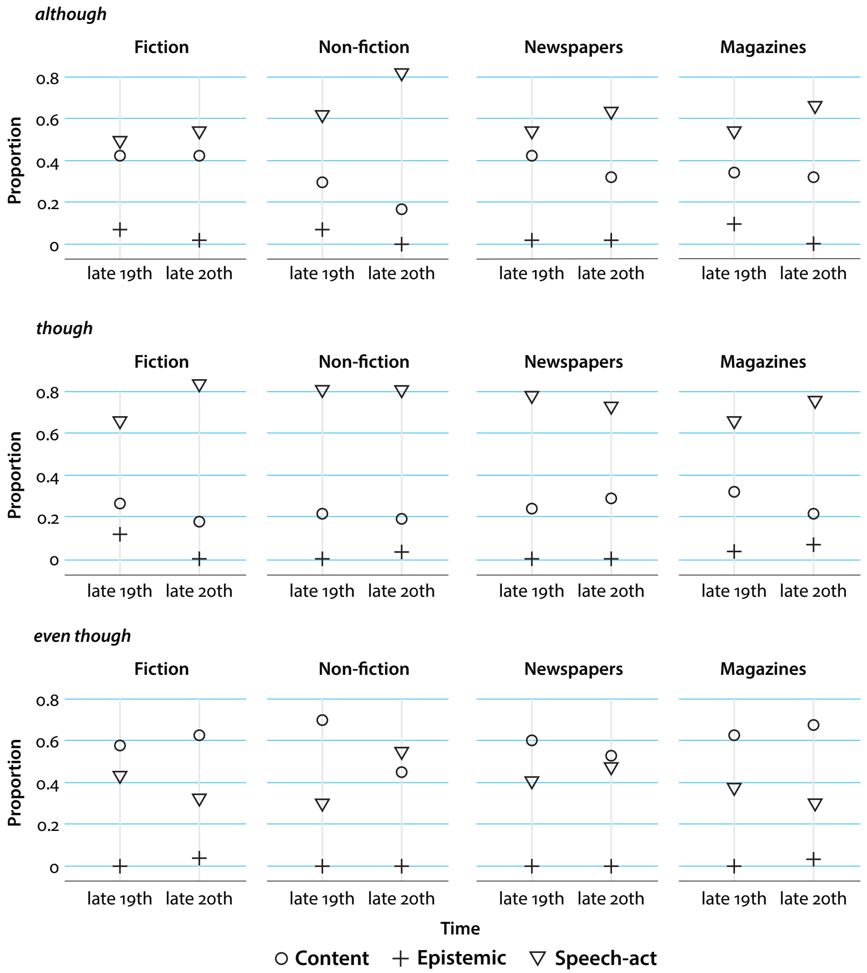


Figure 5. Proportions of semantic types in four genres in diachrony

The patterns generated by the simultaneous inspection of semantics, genre and diachrony are most homogeneous for the conjunction *although*. The general trend shown in Figure 3 above, i.e. an increase in the proportion of speech-act concessives and a decrease in the proportion of content concessives, is mirrored in each of the four genres. The genre of non-fiction books stands out, and this particular pattern is only marginally non-significant, with a medium effect size (χ^2 (2) = 5.42; p = .067; φ = .260). The genre of magazines also produces a medium effect (φ = .238), which is more clearly non-significant (χ^2 (2) = 4.55, p = .103), however. The effects for fiction and newspapers are only small (φ = .117 and .104, respectively).

Concerning *though*, overall patterns are much less orderly. The general trend (an increase in speech-act concessives and a decrease in content concessives) is even reversed in texts published in newspapers, and hardly any change is detected in non-fiction books. On the other hand, the overall trend is confirmed in magazines and in fiction, in the latter constituting a medium effect that is marginally non-significant (χ^2 (2) = 5.36; p = .069; φ = .259).

Finally, the conjunction *even though* patterns against the trend in fiction and magazines, while it appears to pattern strongly with the trend in non-fiction and newspaper texts. However, those two genres were not tested due to the non-occurrence of the epistemic type.

Certain genres can be argued to be mainly responsible for the general trends described above and shown in Figure 3, but they differ between the different conjunctions. *Although* is exceptional in showing the same general trend in all four genres; *though* only shows the main trend in fiction and magazines; and *even though* only follows the main trend in newspapers and non-fiction books. It appears that in the case of *although*, a slow-moving semantic change towards more subjective concessives mainly based on the speech-act type has already pervaded all genres. This can also be interpreted as indicating that *although* is leading the change – if we assume that there is a change that affects the entire system of concessive conjunctions. Concerning *though* and *even though*, it is as yet unclear whether the development is truly going in that same direction; the very least we can say is that, although the average tendency is towards the marking of speech-act concession, the underlying genre-specific tendencies are far from uniform, or may even be conflicting for *though* and *even though*. This is in line with the finding that, at a surface level, only *although* could be shown to have undergone a statistically significant semantic change.

5. Summary and outlook

This chapter investigated the semantics of concessive constructions involving the subordinating conjunctions *although*, *though* and *even though* in written American English from the 1860s to the present day. It focused firstly on the frequencies of the three markers and how they changed diachronically; secondly, it investigated the preferred semantic relations within constructions, again inspecting the diachronic dimension. And, thirdly, possible effects of genre on developments in frequency and semantics were inspected.

In the period under investigation, there is an increase in the frequency of *although*, a decrease in the frequency of *though*, and a marked increase in the frequency of *even though*. It was argued that neither colloquialisation nor ongoing grammaticalization seem particularly suitable theories to account for those changes. *Though*, rather than *although*, would have been the better candidate for an increase in frequency in models based on specialisation (in the sense of Hopper 1991) or colloquialisation: It is morphologically more compact, regarded as less formal (cf. Quirk et al. 1985), and it is the most frequent of the three conjunctions to start with. One alternative explanation of the observed changes in frequency is that the marked increase in frequency of the conjunct *though* has an effect on the development, with *though* beginning to be used less as a conjunction and more as a conjunct. The increase of *even though*, on the other hand, is perhaps best explained in terms of semantics: Since *although* and particularly *though* predominantly encode concessives of the speech-act type, *even though* is needed for the coding of content concessives. Thus, there is probably much more to the story of *even though* than its simply being an emphatic (or otherwise stylistically marked) variant (cf. Quirk et al. 1985: 1099). Finally, it is also possible that *even though* is gradually losing its emphatic value, as part of an ongoing process of grammaticalization, which can account for its general increase in frequency. In fact, semantic motivations and grammaticalization may interact and exert a joint influence on the conjunction *even though*, which might, in the long-term perspective, even become a serious competitor of the declining conjunction *though*. Speculating on developments in the distant future, a possible trend might therefore be for *although* to be mainly used to express concession at the speech-act level, for *though* to be used predominantly as a conjunct, and for *even though* to express content concession (see below) and to keep increasing in frequency. As briefly stated above, the epistemic sense of concessives is (and appears to remain) relatively infrequent, and does not show any inclination to enter into a closer association with any of the three conjunctions under discussion.

Genre has relatively undramatic effects on frequency developments, with the exception of the conjunct *though*, which played a marginal role in the chapter as a whole. The conjunct appears to spread from fictional texts to other genres. It was tentatively proposed that fiction leads the change because it contains a high proportion of fictional dialogue (cf. Biber et al. 1999: 850–851), but of course the use of conjuncts (as one way of generating coordinated syntactic structures) may be a characteristic of fiction more generally.

Semantic differences and developments are the most interesting aspect of the analyses presented in this chapter. Concessive constructions in which the subordinate clause is headed by *although* and *though* are mostly of the speech-act type, while constructions involving *even though* are predominantly of the content type. Using the more abstract semantic concept of subjectivity, those findings mean that *although* and *though* are found in more subjective constructions dominated by speech-act concessives. For all three conjunctions, there is a general tendency for the proportion of speech-act concessives to increase over time, and thus for the constructions to become somewhat more subjective, again given that we accept the proposed subjectivity scale.

The analysis of the four genres in COHA in combination with semantic changes revealed patterns that are partly ambiguous. Most importantly, the conjunction *although* is characterised by the same diachronic semantic developments in all genres – an increase in speech-act concessives and a decrease in content concessives – while the other two conjunctions behave much less uniformly; in some genres, the trend even seems to be going in the opposite direction. While we can only speculate as to why certain genres should behave in a certain way – particularly if, as is the case in COHA, the genres are relatively closely related – the general difference between *although* and the other two conjunctions suggests that the slow-moving semantic change towards a greater use of speech-act concessives has progressed further in *although*, since it truly pervades all genres.

Overall, the three conjunctions appear to be subject to slow-moving, long-term changes in written American English, many of which can be attributed to the development towards a somewhat tidier mapping of form onto function – e.g. the competition between *though* as a conjunction and *though* as a conjunct, or the increase in frequency of *even though*, which, it was argued, is needed as a conjunction marking concession at the content level. Genre patterns are for the most part not particularly prominent, but their uniformity or non-uniformity may be one additional piece of evidence that can help assess the status of an ongoing diachronic change.

On a methodological note, it might be worth considering if (and how) semantic classes can be simplified, seeing that the rather rare epistemic type complicates the analysis considerably; on the other hand, it seemed sensible for theoretical reasons to consider all three types in this chapter. It would of course be rather difficult to decide on an alternative category for those epistemic concessives that were found. Further, the subjectivity scale used in some analyses might also be in need of reconsideration or refinement, since it contains a strong arbitrary element concerning the relative degree of subjectivity of the three semantic types of concessives.

Future research on concessive markers in English could (and should) be taken in several directions. Firstly, British English needs to be investigated in addition to American English, both in synchrony and diachrony. To this end, British corpora based on the original design of the Brown corpus (cf. Francis & Kucera 1979) could be used, accessible via the *Corpus Query Processor* at Lancaster University (CQPweb; <https://cqpweb.lancs.ac.uk>; cf. Hardie 2012). Corpora based on Brown, however, at present only cover the twentieth century and are relatively small. An alternative corpus is the HANSARD corpus (Alexander & Davies 2015–). However, containing only speeches from the British Parliament, it is difficult to make generalisations based on results from HANSARD, and the corpus may not be entirely unproblematic for other reasons as well (cf. Mollin 2007). Varieties of English beyond British and American English should of course also be inspected to relate research on concessives to other research within the World Englishes paradigm (cf. Schützler 2017). Finally, more markers will need to be investigated, first and foremost the concessive prepositions *in spite of* and *despite*, to gain more insights into semantic and stylistic patterns within the domain of English concessive markers more generally.

References

Aarts, Bas. 1988. Clauses of concession in written present-day British English. *Journal of English Linguistics* 21(1): 39–58. https://doi.org/10.1177/007542428802100104

Agresti, Alan & Finlay, Barbara. 2009. *Statistical Methods for the Social Sciences*. Upper Saddle River NJ: Pearson Education.

Alexander, Marc & Davies, Mark. 2015–. *Hansard Corpus 1803–2005*. <http://www.hansard-corpus.org>

Anscombre, Jean-Claude. 1989. Théorie de l'argumentation, topoï, et structuration discursive. *Revue Québécoise de Linguistique* 18(1): 13–55. https://doi.org/10.7202/602639ar

Azar, Moshe. 1997. Concessive relations as argumentations. *Text* 17(3): 301–316. https://doi.org/10.1515/text.1.1997.17.3.301

Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan & Finegan, Edward. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Britain, David & Trudgill, Peter. 2000. Migration, dialect contact, new-dialect formation and reallocation. In *Dialect and Migration in a Changing Europe*, Klaus J. Mattheier (ed.), 73–78. Frankfurt: Peter Lang.

Burnham, Josephine M. 1911. *Concessive Constructions in Old English Prose*. New York NY: Henry Holt.

Chen, Guohua. 2000. The grammaticalization of concessive markers in Early Modern English. In *Pathways of Change: Grammaticalization in English* [Studies in Language Companion Series 53], Olga Fischer, Annette Rosenbach & Dieter Stein (eds), 87–110. Amsterdam: John Benjamins. https://doi.org/10.1075/slcs.53.06che

Crevels, Mily. 2000. Concessives on different semantic levels: A typologocal perspective. In *Cause – Condition – Concession – Contrast. Cognitive and Discourse Perspectives*, Elisabeth Couper-Kuhlen & Bernd Kortmann (eds), 313–339. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110219043.4.313

Davies, Mark. 2010–. *The Corpus of Historical American English: 400 million words, 1810–2009*. <http://corpus.byu.edu/coha/>

Di Meola, Claudio. 1998. Zur Definition einer logisch-semantischen Kategorie: Konzessivität als „versteckte Kausalität". *Linguistische Berichte* 175: 329–352.

Francis, W. Nelson & Kucera, Henry. 1979. *Manual of Information to Accompany A Standard Corpus of Present-Day Edited American English, for Use with Digital Computers*. Providence RI: Brown University Department of Linguistics. <http://clu.uni.no/icame/manuals/BROWN/INDEX.HTM> (27 February 2015).

Hardie, Andrew. 2012. CQPweb - combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17(3): 380–409. https://doi.org/10.1075/ijcl.17.3.04har

Hermodsson, Lars. 1994. Der Begriff „konzessiv". Terminologie und Analysen. *Studia Neophilologia* 66: 59–75. https://doi.org/10.1080/00393279408588131

Hilpert, Martin. 2013. *Constructional Change in English: Developments in Allomorphy, Word formation, and Syntax*. Cambridge: CUP. https://doi.org/10.1017/CBO9781139004206

Hilpert, Martin & Gries, Stefan Th. 2009. Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing* 24(4): 385–401. https://doi.org/10.1093/llc/fqn012

Hopper, Paul J. 1991. On some principles of grammaticization. In *Approaches to Grammaticalization, Vol I: Focus on Theoretical and Methodological Issues* [Typological Studies in Language 19], Elizabeth Closs Traugott & Bernd Heine (eds), 17–35. Amsterdam: John Benjamins. https://doi.org/10.1075/tsl.19.1.04hop

Hopper, Paul J. & Elizabeth Closs Traugott. 2003. *Grammaticalization*. Cambridge: CUP. https://doi.org/10.1017/CBO9781139165525

Huddleston, Rodney D., & Pullum, Geoffrey K. 2002. *The Cambridge Grammar of the English Language*. Cambridge: CUP. https://doi.org/10.1017/9781316423530

Hundt, Marianne & Mair, Christian. 1999. 'Agile' and 'uptight' genres: The corpus-based approach to language change in progess. *International Journal of Corpus Linguistics* 4(2): 221–242. https://doi.org/10.1075/ijcl.4.2.02hun

König, Ekkehard. 1985. On the history of concessive connectives in English: Diachronic and synchronic evidence. *Lingua* 66(1): 1–19. https://doi.org/10.1016/S0024-3841(85)90240-2

Mair, Christian. 2006. *Twentieth-Century English. History, Variation and Standardization*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511486951

Mollin, Sandra. 2007. The Hansard hazard: Gauging the accuracy of British parliamentary transcripts. *Corpora* 2(2): 187–210. https://doi.org/10.3366/cor.2007.2.2.187

Oxford English Dictionary Online. Oxford: OUP. <http://www.oed.com/> (22 February 2017).

Quirk, Randolph, Greenbaum, Sidney, Leech, Geoffrey & Svartvik, Jan. 1985. *A Comprehensive Grammar of the English Language*. London: Arnold.

R Development Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Version 3.2.4. [computer program]. <http://www.R-project.org>

RStudio Team. 2009–2015. *RStudio: Integrated Development for R*. Version 0.99.486. Boston MA: RStudio. <http://www.rstudio.com/>

Sarkar, Deepayan. 2014. *lattice: Lattice Graphics*. R-package version 0.20-29. <http://cran.r-project.org/web/packages/lattice/lattice.pdf>

Schützler, Ole. 2017. A corpus-based study of concessive conjunctions in three L1-varieties of English. In *Language Variation – European Perspectives VI: Selected papers from the Eighth International Conference on Language Variation in Europe (ICLaVE 8), Leipzig, 2015* [Studies in Language Varietion 19], Isabelle Buchstaller & Beat Siebenhaar (eds), 173–184. Amsterdam: John Benjamins. https://doi.org/10.1075/silv.19.11sch

Smitterberg, Erik. 2014. Syntactic stability and change in nineteenth-century newspaper language. In *Late Modern English Syntax*, Marianne Hundt (ed.), 311–29. Cambridge: CUP. https://doi.org/10.1017/CBO9781139507226.023

Smitterberg, Erik & Kytö, Merja. 2015. English genres in diachronic corpus linguistics. In *From Clerks to Corpora: Essays on the English Language Yesterday and Today*, Philip Shaw, Britt Erman, Gunnel Melchers & Peter Sundkvist (eds), 117–33. Stockholm: Stockholm University Press. https://doi.org/10.16993/bab.g

Sweetser, Eve E. 1990. *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511620904

Szmrecsanyi, Benedikt. 2013. Analysing aggregated linguistics data. In *Research Methods in Language Variation and Change*, Manfred Krug & Julia Schlüter (eds), 433–455. Cambridge: CUP. https://doi.org/10.1017/CBO9780511792519.028

Traugott, Elizabeth C. & Dasher, Richard B. 2002. *Regularity in Semantic Change*. Cambridge: CUP.

Trudgill, Peter. 1986. *Dialects in Contact*. Oxford: Blackwell.

Appendix

Wordcounts in COHA by decade and genre

Decade	Number of words				
	Fiction	Popular magazines	Newspapers	Non-fiction books	Σ
1860s	9,450,562 (55.2%)	4,437,941 (25.9%)	262,198 (1.5%)	2,974,401 (17.4%)	17,125,102
1870s	10,291,968 (55.3%)	4,452,192 (23.9%)	1,030,560 (5.5%)	2,835,440 (15.2%)	18,610,160
1880s	11,215,065 (53.7%)	4,481,568 (21.5%)	1,355,456 (6.5%)	3,820,766 (18.3%)	20,872,855
1890s	11,212,219 (52.9%)	4,679,486 (22.1%)	1,383,948 (6.5%)	3,907,730 (18.4%)	21,183,383
1900s	12,029,439 (53.4%)	5,062,650 (22.5%)	1,433,576 (6.4%)	4,015,567 (17.8%)	22,541,232
1910s	11,935,701 (52.7%)	5,694,710 (25.1%)	1,489,942 (6.6%)	3,534,899 (15.6%)	22,655,252
1920s	12,539,681 (48.9%)	5,841,678 (22.8%)	3,552,699 (13.9%)	3,698,353 (14.4%)	25,632,411
1930s	11,876,996 (48.6%)	5,910,095 (24.2%)	3,545,527 (14.5%)	3,080,629 (12.6%)	24,413,247
1940s	11,946,743 (49.5%)	5,644,216 (23.4%)	3,497,509 (14.5%)	3,056,010 (12.7%)	24,144,478
1950s	11,986,437 (49.1%)	5,796,823 (23.8%)	3,522,545 (14.4%)	3,092,375 (12.7%)	24,398,180
1960s	11,578,880 (48.4%)	5,803,276 (24.3%)	3,404,244 (14.2%)	3,141,582 (13.1%)	23,927,982
1970s	11,626,911 (48.9%)	5,755,537 (24.2%)	3,383,924 (14.2%)	3,002,933 (12.6%)	23,769,305
1980s	12,152,603 (48.3%)	5,804,320 (23.1%)	4,113,254 (16.3%)	3,108,775 (12.3%)	25,178,952
1990s	13,272,162 (47.6%)	7,440,305 (26.7%)	4,060,570 (14.6%)	3,104,303 (11.1%)	27,877,340
2000s	14,590,078 (49.5%)	7,678,830 (26.0%)	4,088,704 (13.9%)	3,121,839 (10.6%)	29,479,451
Σ	177,705,445	84,483,627	40,124,656	49,495,602	351,809,330

Variation of sentence length across time and genre

Influence on syntactic usage in English

Karolina Rudnicka
University of Freiburg

The goal of this paper is threefold: (i) to present some practical aspects of using the full-text version of the *Corpus of Historical American English* (COHA), the largest diachronic multi-genre corpus of the English language, in the investigation of a linguistic trend of change; (ii) to test a widely held assumption that sentence length in written English has been steadily decreasing over the past few centuries; (iii) to point to a possible link between changes in sentence length and changes in English syntactic usage. The empirical proof of concept for (iii) is provided by the decline in the frequency of the non-finite purpose subordinator *in order to*. Sentence length, genre and the likelihood of occurrence of *in order to* are shown to be interrelated.

1. Introduction

Sentence length, defined as the number of words that come between the opening word (starting with a capital letter) and the end punctuation sign such as full stop, question mark or exclamation mark is a subject that attracts a lot of interest in different contexts, such as cognitive linguistics, rhetoric and language teaching, to name just a few. There are handbooks of English suggesting language users and language learners use shorter sentences, as this would make their English appear more modern (McGregor 2002: 33), or advising them to keep their sentences within a given length range (Hult 2015: 105):

> In an age of tweets, text messages and *Facebook* posts, English sentences seem to be decreasing in length. Sentences in English vary from one word (*Help!*) to forty or more words. To keep your sentences both readable and meaningful, however, you will want to stay in the fifteen- to twenty-word range most of the time.

Another frequently investigated context is the relation between sentence length and syntactic complexity (e.g. Klare 1963: 170; Westin 2002: 81; Biber & Conrad 2009: 152; Fahnestock 2011: 169; Štajner & Mitkov 2012: 1578), which might

https://doi.org/10.1075/scl.85.10rud

directly and indirectly influence text complexity and text comprehension (Fries 2010: 21; Fahnestock 2011: 169–170). Sentence length is one of the main variables which are included in various readability formulas (equations yielding quantitative scores assessing text difficulty) such as the Gunning Fog Index and Lexile reading measure (Gross et al. 2002: 171; Fahnestock 2011: 170). One of the applications these measures have is, for instance, in the assessment of school textbooks for certain age groups. Similarly, the length of sentences produced by language users is used as a valuable index of language development in children (e.g. Davis 1937: 69).

From a diachronic perspective the relation between sentence length and syntactic complexity can tell us something about changes in the way in which authors use the resources offered by the linguistic system (Biber & Conrad 2009: 152). Or it could help us answer the question asked by Štajner and Mitkov (2012: 1577–1578) whether texts are becoming simpler and easier to read, as sentence length is frequently used as one of several factors of text complexity (the other ones are e.g. Automated Readability Index, sentence complexity and the use of passive voice).

The present paper looks at sentence length from a different angle and focuses on its evolution across the last two hundred years, offering insights into possible explanations for the observed decrease. Additionally, it points to a possible influence this decrease might have on the syntactic usage of English. The first section presents works focused on the evolution of sentence length across time and, in the same context, addresses the question of the influence of changing punctuation conventions. The second section presents the methodology of working with the full-text version of COHA. It contains a detailed description of the possible pitfalls of using textual versions of mega-corpora and introduces *Mathematica* as a programme of choice for the study.

In total, the lengths of more than nineteen million sentences extracted from the corpus are calculated. The results of the comprehensive analysis reveal a visible decrease in sentence length across time for all of the investigated genres. It is suggested that this decrease can have an influence on the syntactic usage of constructions that show correlating patterns of frequency development (decrease or increase in the frequency of use). The third part is devoted to the non-finite purpose subordinator *in order to* which serves as an example construction and which is shown to be decreasing in the frequency of use. This decrease, however, is different in the case of each of the investigated genres. Insights in the history of *in order to* and the genre-related distribution of *in order to* are then presented to justify the hypothesis that the general decrease in sentence length and the general decrease in frequency of use of *in order to* might be interrelated. It is likely that the described phenomenon is generalizable to more constructions from the network of English purpose subordinators such as *in order that*, *so as to*, *lest* (Rudnicka 2018, in preparation) or to other constructional networks of both English and other languages.

2. Sentence length in written English: The diachronic evolution across genres

More than one hundred years ago, Lewis (1894: 34) stated that "the English sentence has decreased in average length at least one-half in three hundred years". His work presents "a count of the average number of words to the sentence and to the paragraph, in representative authors since the middle of the fifteenth century". Although the main focus of Lewis's paper is on the structure of a paragraph in written English across centuries, he additionally identifies a visible trend towards a decrease in average sentence length. According to his calculations, the paragraph in the late 19th century has the same length as in the 16th century, but it contains twice as many sentences (1894: 170). On the other hand Säily, Vartiainen and Siirtola (2017), who look at sentence length evolution across the time period represented by the *Parsed Corpus of Early English Correspondence* (c.1410–1681), find that sentence length stayed roughly the same in the entire period studied. This might mean that the decrease in sentence length referred to by Lewis either starts a bit later or is only visible in literary works and not in private correspondence.

Biber and Conrad (2009: 151–153) compare linguistic characteristics of novels from the 18th and 20th century and call the change in syntactic complexity "perhaps the most important change" in the way that authors use the resources of the given linguistic system. Sentence length is one of the measures used to study the differences in syntactic complexity between the works of particular authors from the eighteenth and twentieth centuries. The comparison of sentence lengths in narratives across the two historical periods yields results pointing towards a rather steady decrease of the average sentence length (Biber & Conrad 2009: 152):

> While there is some variation at any given historical period, there is also a very steady progression from the extremely long sentences of Defoe to the short sentences of Vonnegut and Bellow.

The change in sentence length is said to be, to a large extent, a reflection of changing punctuation practices, such as "a much more extensive use of colons and semicolons in earlier historical periods" (Biber & Conrad 2009: 153).

The work of Gross, Harmon and Reidy (2002) focuses on scientific prose. Their results show "a definite shrinking in average sentence length over time, from 33 words in 1876–1900 to 30 words in 1901–1925 to 27 in 1976–2000"

(2002: 171), while the clausal density is said to have remained relatively stable. Additionally, their work offers an important observation suggesting that the decrease in sentence length might actually be a phenomenon generalizable to other languages. The comparison of average sentence lengths across languages in 19th century passages shows there is a decrease in English, German and French scientific texts (2002: 124). Gross et al. (2002: 171) identify two opposite trends which concern the readability of the scientific prose of today. According to their conclusions, on the one hand it is becoming more difficult to read because of the use of increasingly complex and compact noun phrases. On the other hand, it is becoming easier to read "because of its declining sentence length and number of clauses per sentence". The work of Dorgeloh (2005) on patterns of agentivity and narrativity in early science discourse adds a few more observations on the development of language of science from the 15th century to the present, and on the factors that shaped the way scientific texts look like today. In her work (2005: 92) she claims that "modern science texts tend to nominalise the experience and to impersonalise the argument", whereas in the past it was typical for the narrative and the argument to be based on personal reference. The observations concerning the more complex and compact noun phrases and the decrease in the number of clauses per sentence are further supported by the research of Hundt et al. (2012: 224), which also focuses on scientific discourse. Additionally, their research notes that the marked decrease in the frequency of relative clauses coincides with a marked decrease of sentence length occurring around the beginning of the twentieth century (2012: 225).

A large part of the recent research dealing with sentence length decrease focuses on newspaper language. Westin (2002) provides statistically significant evidence for this effect in a wide range of English newspapers, such as *The Times, The Guardian* And *The Daily Telegraph* for the period 1900–2000. Additionally to the sentence length itself, she also investigates sentence length distributions in the texts from particular years. One of the observations she makes is that "the shortest sentences, from 1 to 10 words, increased in number, from 5.7% in 1900 to 16.6% in 1993" (81). Westin concludes that if we assume there is a relation between sentence length and sentence complexity, "it is obvious that the complexity of the sentences in English upmarket newspaper editorials decreased considerably during the 20th century" (81). Likewise, focusing on newspaper language, Fries (2010) shows a decrease of approximately 10 words in sentence length during the 18th century in the *London Gazette*. Apart from the diachronic development of sentence length, he compares sentence length in different sections of the investigated journal and points to sections (such as foreign news) that, on average, seem to show longer sentences, and sections that tend to contain more short sentences, such as the advertising section. Schneider (2002: 98) reports that the sentence length in newspaper news has decreased by an average of 15 words since 1700. The reasons for this decrease given by Schneider (2002: 98ff.) include a necessity for greater comprehensibility for a mass readership and a trend in which newspaper language became more similar to the spoken word. Also Westin (2002: 161) concludes that authors of some of the studied newspapers might adjust the language of their editorials on purpose in order to attract a broader audience than competing newspapers.

2.1 Just a matter of punctuation conventions?

> English punctuation will apparently never stop causing division among scholars. There appears to be serious disagreement about its nature, functions and formal status. (Shou 2007: 195)

Whereas most authors of the empirical studies described in the section above claim that the decrease in sentence length in written English is, at least partly, a linguistic fact, other scholars, e.g. Fahnestock, express a view that the observed decrease might actually be "a by-product of changing punctuation conventions" (2011: 265) or a reflection of changing punctuation practices (Biber & Conrad 2009: 153). The full stop, exclamation mark and question mark are said to have taken over the work that was previously also done by colons, semicolons and commas. As Westin (2002: 79) points out, some historical linguists "consider the colon and semicolon as sentence delimiters". Support for this observation is provided by e.g. Miller who in his paper offers a first-hand view on punctuation practices at the beginning of the twentieth century (1908: 327):

> There are four structural equivalents of the period; namely, the semicolon (or colon), the structural connective, the series, and the balance.

This might suggest that the assumed decrease in sentence length would, to say the least, look differently if one used semicolons and colons as sentence delimiters, instead of only using full stops, exclamation marks and question marks. Westin (2002: 79) claims that in her earlier research she adopted both perspectives and the two approaches have shown to have almost identical patterns of development (1997: 16–17; quoted in Westin 2002: 79). Thus, in the paper in which she shows the universal decrease in sentence length in *The Times*, *The Guardian* And *The Daily Telegraph* for the period 1900–2000, she includes only sentences ending with a full stop, a question mark, or an exclamation mark. The question arises if the two approaches compared by Westin in her first work would still show similar patterns of development if she did her research also on some slightly earlier texts. Since super-long sentences, containing 60 or more words, seem to be much more typical for the nineteenth century than for the twentieth century (Biber & Conrad 2009:

152–153), the inclusion of earlier texts in the investigation could have changed the observed patterns of development.

The results from COHA seem to support this observation, as semicolons were much more frequently used in the beginning of the nineteenth century than they were in the twentieth century. Figure 1 shows the development of the frequency (normalised, per million words) of semicolons across time. The data used for the plot were obtained via the online interface of COHA.[1] We see a dramatic decrease in the frequency of use of semicolons happening between 1810 and 1900. From 1900 on, there still is a decrease, but the curve is much less steep, suggesting that the beginning of the twentieth century was a turning point for the use of semicolons.

Still, even though colons and semicolons were once considered sentence delimiters, the punctuation of present-day English follows different rules. If we accept the view that at least part of the decrease in sentence length can be attributed to a change in punctuation practices, does it mean that this decrease is only apparent? And, if the previous punctuation system was comprehensive, why did it not stay the same until today?

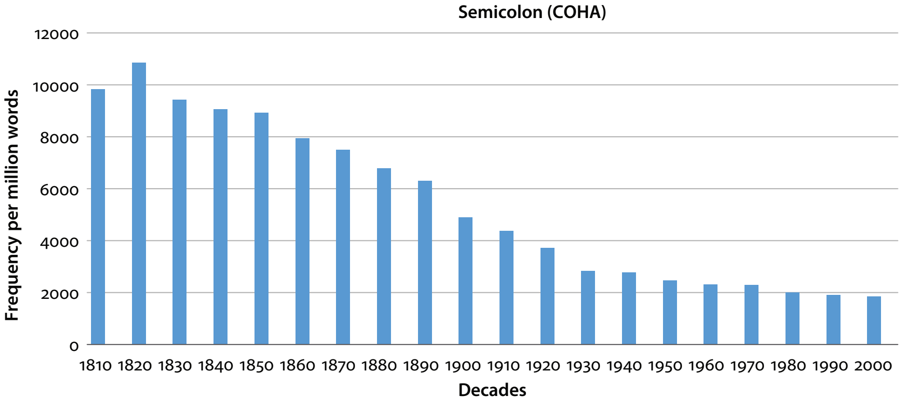


Figure 1. Frequency of semicolons across the period 1810–2009 in COHA[2]

The explanations for the decrease in sentence length offered by Schneider (2002: 98ff) and Westin (2002: 161) seem to make a lot of sense, also in the context of changing punctuation practices. Adding to their conclusions one could hypothesise that as the development of mass readership went on, newspaper editors

[1] <http://corpus.byu.edu/coha/old/>

[2] Data retrieved from COHA on May 13th 2017.

and authors might have adopted different punctuation conventions because of the need for greater text comprehensibility for the new society of readers. The dramatic decrease in the frequency of semicolon use (Figure 1) coincides with the achievement of mass literacy by American society and the invention of new printing technologies, since these developments took place around the mid- to late nineteenth century, as observed by e.g. Hames and Rae (1996: 227):

> All this changed with technological innovations in the mid- to late nineteenth century, and it is at this time that the rise of the independent mass circulation daily newspaper began. In addition to changes in printing technology that made production easier, the invention of telegraph and the advent of the wire services meant that news could be quickly transmitted to all areas of the country.

One of the possible explanations for the change in punctuation conventions might thus have been the development of mass readership. The change in punctuation conventions might have, in turn, led to a decrease of sentence length in written present-day English.

3. A comprehensive analysis of sentence length in the time period of 1800–2000

This work defines sentence length as the number of words that come between the opening word starting with a capital letter and the end punctuation mark, namely a full stop, exclamation mark and question mark. The main aim of the analysis is to detect and visualise the trend of change in the sentence length across the time period 1810–2009 and across different genres of American English. The novelty of this analysis lies in the application of the full-text COHA data. Since most of the research described in the section above is based on relatively small samples of texts, it is hoped that the use of a large collection of textual data available such as the COHA corpus sheds more light on the phenomenon and makes an interesting contribution to the discussion about sentence length in the English language. Also, the practical aspects and technical details of using the full-text COHA are expected to be of interest to scholars who would like to profit from the availability of big data in their research.

3.1 Design of the analysis and methodology

3.1.1 *Full-text COHA*

COHA is known as the largest diachronic multi-genre corpus of American English and it contains 400 million words of text from the time period 1810–2009 belonging to four genres, namely *fiction*, *magazine, newspaper* and *non-fiction*. For the purposes of this analysis the contents of the *fiction* genre are modified in order to decrease the level of bias, namely, the subgenre of *movie & play script* is removed from *fiction* and treated as an independent genre – a proxy for the spoken language. According to the terminology in Culpeper and Kytö (2010), the data constituting *movie & play script* section are speech-purposed, i.e. designed in order to imitate real-time spoken interaction.

COHA is available online via an interface at <https://corpus.byu.edu/coha/>. It is, however, impossible to conduct an in-depth analysis of the evolution of sentence length via the online interface, as the length of search strings in COHA cannot exceed the limit of fifteen words. Because of this fact, the paper uses the full-text offline version of the corpus. The full-text COHA has a form of twenty folders. Each of the folders encompasses one decade and contains various amounts of text files (txt format) representing different genres. The name of each file serves the purpose of genre and decade identification, e.g. mag_2007_387216.txt belongs to the *magazine* genre and is from the 2000 decade, while the last number in the file name may be used to identify its exact source in the sources_coha.xls table. All in all, the full-text COHA contains 116 615 text files.

3.1.2 *Genres in COHA*

On the webpage of COHA it can be read that the corpus is well-balanced in terms of genres, and the proportions of genres stay roughly the same from decade to decade. What are the contents of particular genres? For the purpose of clarity, a few words need to be said about the COHA genres themselves.

The genre accounting for the largest share of COHA (48–55% of the total in each decade) is *fiction*. According to the information on the webpage, *fiction* contains texts from digitized books from many different sources, such as *Project Gutenberg* (1810–1930) and *The Cornell University Library Making of America Collection* (1800–1900), a digital library of primary sources in American social history;[3] scanned books (1930–1990); movie and play scripts which can also be found in COCA,[4] another mega-corpus of the same family.

Magazine contains digitized journals from *The Cornell University Library Making of America Collection* (1810–1900), scanned magazines and the contents of the COCA genre *popular magazine* (1990–2009), which is described as "nearly 100 different magazines, with a good mix (overall, and by year) between specific domains (news, health, home and gardening, women, financial, religion, sports,

[3] <http://ebooks.library.cornell.edu/m/moa/>

[4] <http://corpus.byu.edu/coca/>

etc.)".[5] Some examples of the contemporary magazines include *Time*, *Men's Health*, *Good Housekeeping*, *Cosmopolitan*, *Fortune*, *Christian Century* and *Sports Illustrated.* According to a detailed list of COHA sources, which is attached to the full-text version of the corpus, among the magazines representing the early nineteenth century are: *The North American Review, The New England Magazine, The United States Democratic Review, New Englander* and *Yale Review.*

Newspaper genre contains various digitized texts from *The New York Times* (decades 1860–2009), *The Chicago Tribune* (1900–2009), *The Wall Street Journal* (1910–1989), *The Christian Science Monitor* (1930–2009), *The Boston Globe* (1980–2009), *USA Today* (1990–2009), *The San Francisco Chronicle* (1990–2009), *The Washington Post* (1990–2009), *The Atlanta Journal Constitution* (1990–2009), *The Houston Chronicle* (1990–2009), *The Associated Press* (1990–2009) and *The Denver Post* (1990–2009).

As one can read on the webpage of COHA, in the non-fiction genre there are ebooks from *Project Gutenberg* and www.archive.org (for the decades 1810–1900), scanned books (1900–1990) and the contents of COCA (1990–2009). The texts in the non-fiction genre are balanced across the Library of Congress classification system,[6] which in practice means that in the detailed list of all the text sources, there is a letter assigned for each source, which classifies it as belonging to one of the twenty one categories such as *A – general works; B – philosophy, psychology, religion; C – auxiliary sciences of history*; etc. The genre *non-fiction* is, however, not homogenous across the whole time span of COHA. Until the decade 1990 it is mostly composed of books, and starting from the decade 1990 texts from nearly a hundred different peer-reviewed journals such as *Ear, Nose & Throat Journal; The Physical Educator* and *Arab Studies Quarterly* are added to the corpus. On the COHA webpage these texts are referred to as the contents of COCA (in COCA they constitute *Academic Journals* genre). As one can see in the COHA online interface, the genre *non-fiction* is sometimes referred to as *non-fiction* and sometimes as *non-fiction books*. In the present work it is treated as the rough equivalent of the *learned* category from the Brown Corpus.[7]

3.1.3 *Sentence tokenisation: Methodology*

The present study uses *Mathematica* as the programme of choice to tokenise texts into individual sentences and provide the word counts. The version 10.4, which is

[5] <http://corpus.byu.edu/coca/old/>

[6] <http://www.loc.gov/catdir/cpso/lcco/>

[7] The standard corpus of present-day edited American English.

used here, contains tools designed especially to deal with linguistic data.[8] Below is the description of the sentence tokenisation procedure with some notes on the content recognition features of the programme:

1. The corpus data (text files) are imported and stored in the form of strings of text. Each text file becomes one string, so that it is possible to control the origin of the text (decade and genre). The tool TextSentences[9] parses strings of texts into individual sentences creating a very long list of sentences. The precision of TextSentences tool can be assessed as 93–98%, which means that it correctly tokenises between 93% and 98% of all the sentences. The exact precision depends to a large extent on the quality of a given text sample, which is in general worse for the earlier decades (1810–1890) and better for the later decades (1900–2000).
2. As explained on the COHA webpage, the texts included in the full-text version have been modified in order to avoid any economic impact on the holders of the copyright. In practice this means that out of every 200 words of the running text, ten words have been removed and replaced with "@". The modifications thus have a form of "@ @ @ @ @ @ @ @ @ @". The distribution of modifications is the same for each decade and genre; it is also independent of the text source. Keeping these strings in the dataset would make the analysis of sentence lengths meaningless, as we do not know, for instance, if there is a full stop somewhere in these modified part of text. All the sentences containing the string "@ @ @ @ @ @ @ @ @ @" are thus removed from our list of sentences. As explained in the COHA tutorial, due to the random distribution of the modifications, the removal of affected data should not have an influence on the statistical analysis.
3. The TextSentences tool recognizes common abbreviations such as "Mr.", "Mrs." or "Jr.", so we do not need to worry about a situation in which the full stops here would be treated as sentence delimiters.
4. On the other hand, cases of enumerations, used frequently to mark the beginning of a chapter / article / passage and containing a full stop exemplified by (2) are not detected by the programme and only a careful visual inspection of the data could keep them out of the dataset. Given the size of the corpus, the best solution is to remove all the one-word sentences from the final dataset.

(2) Art. 1 (1839, NF, American Fruit Garden, COHA)
Vol. II. p. 77. (1870, MAG, Atlantic, COHA)

5. Common tags such as "<P>" need to be removed from the data.

[8] <http://reference.wolfram.com/language/guide/LinguisticData.html>

[9] <https://reference.wolfram.com/language/ref/TextSentences.html>

6. Parsing is conducted for each text file of the corpus, leaving us with over nineteen million sentences (19,768,290 to be precise). Table 1 shows the exact numbers of sentences extracted for each decade and genre.

Table 1. Number of sentences extracted per decade and per genre

Decade	Magazine	Newspaper	Non-fiction	Fiction	Movie & play script
1810	2,242	NA	11,299	44,843	NA
1820	43,080	NA	42,524	150,977	NA
1830	78,267	NA	88,951	299,450	NA
1840	92,112	NA	109,403	368,834	NA
1850	116,018	NA	93,772	385,595	NA
1860	121,195	9,048	92,201	442,831	NA
1870	125,968	34,773	96,158	513,287	NA
1880	130,737	48,199	109,512	544,721	NA
1890	137,595	52,954	107,512	549,291	NA
1900	164,386	63,440	116,223	655,739	49
1910	206,582	62,465	126,706	705,553	5,720
1920	258,497	154,553	117,524	741,621	33,877
1930	254,573	138,671	113,920	709,298	70,225
1940	262,520	134,679	115,432	736,662	74,429
1950	266,879	145,524	117,991	740,315	68,921
1960	260,855	148,992	118,577	722,116	83,936
1970	253,159	141,422	113,998	756,625	68,288
1980	267,081	168,057	122,086	800,548	78,927
1990	335,913	191,881	115,085	871,569	82,290
2000	360,496	182,950	121,762	995,354	NA
Intotal	**3,738,155**	**1,677,608**	**2,050,636**	**11,735,229**	**566,662**

3.2 Results

A visual inspections of a sample of a data set from both the earlier and later decades led to an assumption that potential bias in any direction caused by the errors of the algorithm (a string of sentences not parsed, undetected short bogus sentences, the presence of undetected abbreviations etc.) is more or less the same for each decade. Thus, it is still possible to reliably study the diachronic evolution of the sentence length, even though one should be careful when providing "exact" values of variables such as the aforementioned mean sentence length.

WordCount tool is used to provide the word counts of each sentence in each genre. After running the routine we obtain a very long list of pairs (a, b), where "a" refers to the decade labelling of a particular sentence and "b" to its word length. The 19 768 290 data points (one for each sentence) are used to create a bigger picture of the evolution of sentence length for each of the genres across time. Figures 2 to 6 present visualisations of the trends in the form of box plots. The white line on each of the boxes indicates the median; the black line, which is also shorter than the white one, indicates the mean.

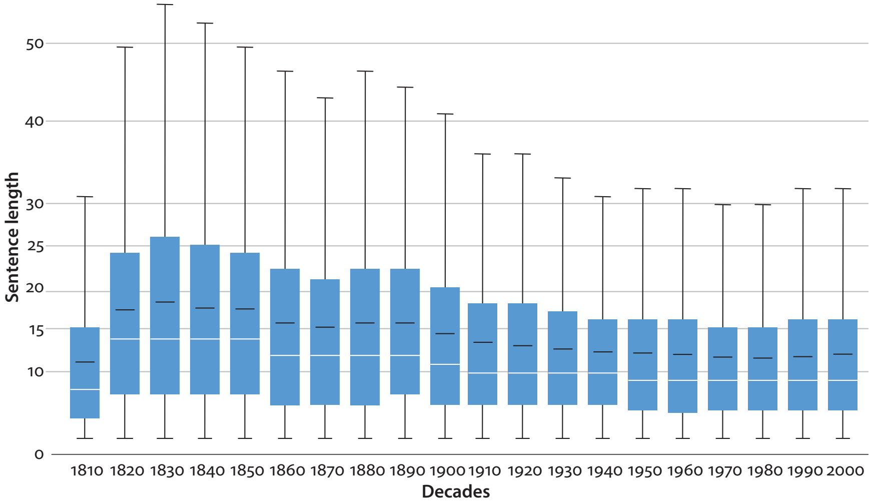


Figure 2. Sentence length across the time period 1810–2009 for the COHA genre *fiction*

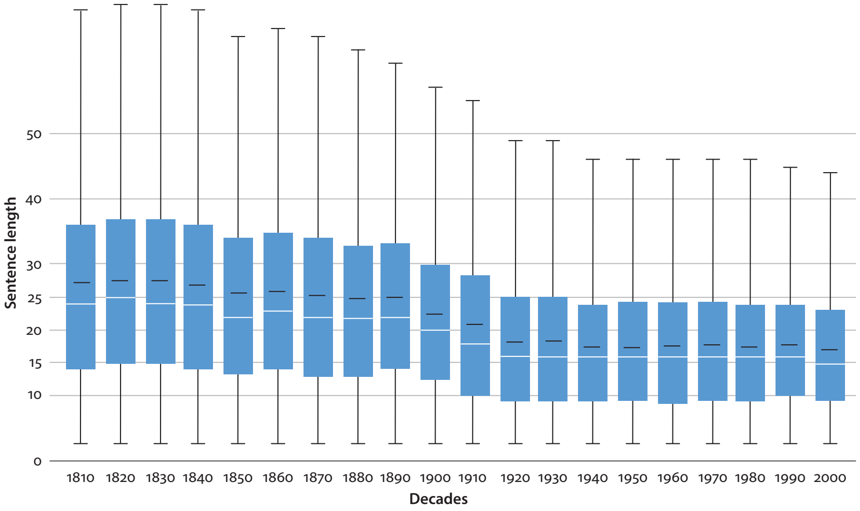


Figure 3. Sentence length across the time period 1810–2009 for the COHA genre *magazine*

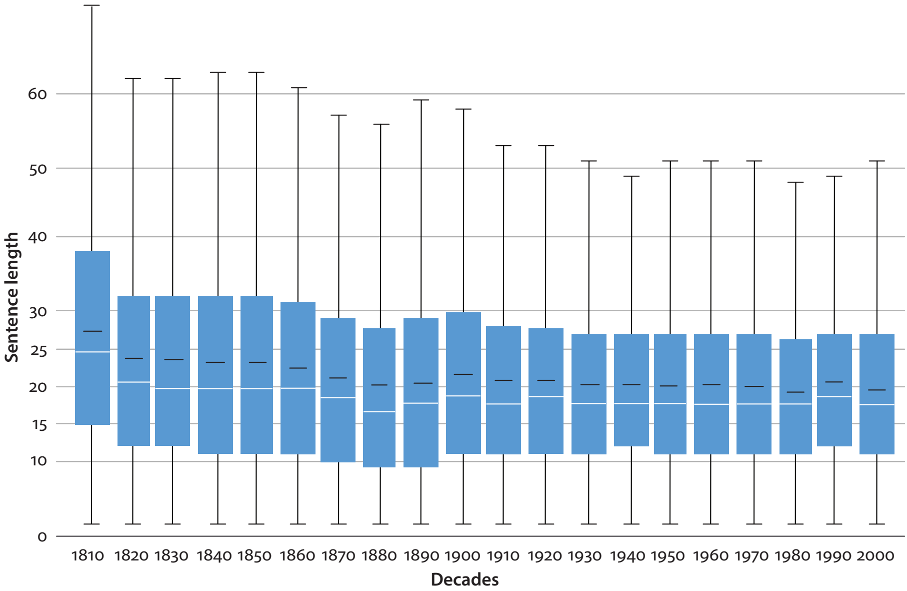


Figure 4. Sentence length across the time period 1810–2009 for the COHA genre *non-fiction*

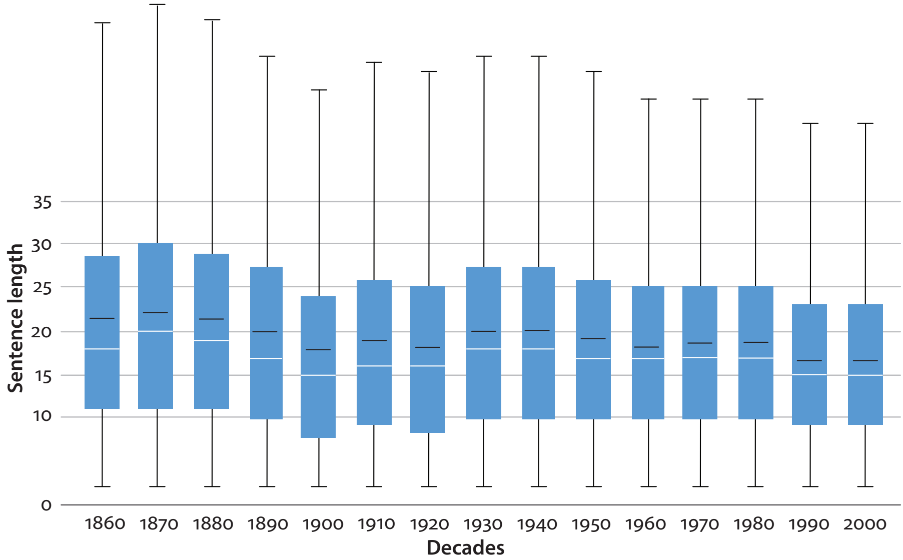


Figure 5. Sentence length across the time period 1860–2009 for the COHA genre *newspaper*

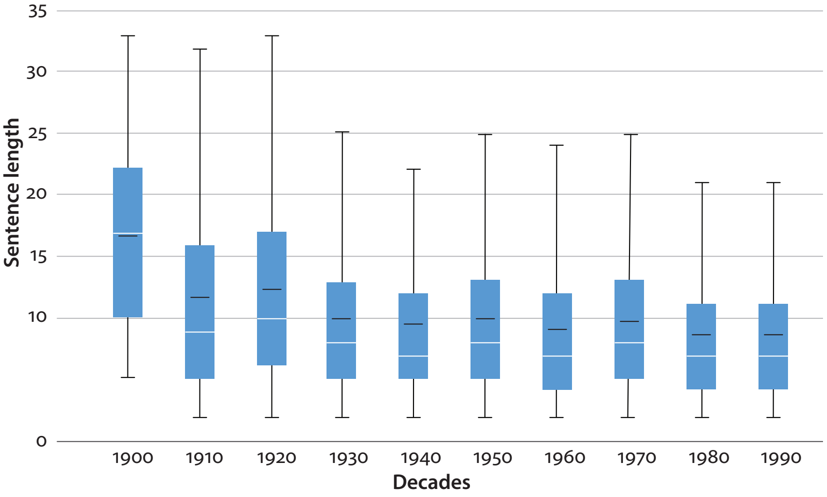


Figure 6. Sentence length across the time period 1860–2009 for the COHA genre *movie & play script*

The plots in Figures 2 to 6 reveal a decrease in the sentence length for each of the genres. Figure 2 presents the distribution of sentence lengths for *fiction*. We can see a visible downward trend starting from 1830. Also in the case of *magazine* (Figure 3) there is a visible decrease of sentence lengths across time. Sentence length in *non-fiction* (Figure 4) also decreases; however, the degree of decrease does not seem to be that visible as in the case of *fiction* and *magazine*. In the case of *newspaper* (Figure 5) there is a decrease as well; however, it is worth noting that the starting decade is 1860. COHA does not feature texts belonging to this genre from earlier decades. Similarly, for *movie & play script* (Figure 6) the starting decade is 1900. Even though we have less data representing this subgenre of *fiction*, we can assume that there is also a certain degree of sentence length decrease over time.

The value which will be used to compare the results between the studied genres is the mean value of sentence length. Table 2 gives an overview of the mean values. As we can see, if we use the mean value to draw conclusions, there is a visible decrease over time for each of the five studied genres. For *magazine*, the sentences in the onset of the twenty-first century are approximately 10 words shorter than in the beginning of the nineteen century, for *non-fiction* the difference amounts to 8 words. Texts in *newspaper* are almost 5 words shorter now than they were in the middle of the nineteenth century. In the case of *fiction*, there first seems to be an increase in the sentence length and then a visible decrease starting from the third decade (1830). The increase in sentence length in 1810 and 1820 might seem surprising, but a closer inspection of the dataset for 1810 and 1820 reveals a large amount of two-word sentences such as in (3):

(3) Eugenia Nothing. (COHA: 1817; FIC; "How to try a lover")
Prince Mentzkioff. (COHA: 1812; FIC; "Alexis, the Czarewitz")

The number of two-word sentences in *fiction* per million words is almost three times higher in 1810 than in 2000 decade. Neither *magazine*, nor *non-fiction* of 1810 come close to the amount found in *fiction*. It seems that the bias can be attributed to characteristics of texts included in the dataset for 1810 and, possibly, to some degree of 'pollution' of the texts.

Also the proxy for the spoken language – *movie & play script* – noted a decrease in the mean sentence length. It is, though, harder to asses even the approximate level of decrease, as for this genre we have the smallest dataset. Not surprisingly, it is also the genre with the shortest sentences nowadays – in the nineties its mean sentence length was 8.6 words, whereas for *magazine* and *newspaper* it was 17.78 and 16.27 words respectively. Long sentences seem to be the domain of *non-fiction*. The mean sentence length in this genre amounts to almost 20 words and, as we can see in Figure 4, the upper whiskers of the box plots representing recent decades reach values higher than in the case of any other genre.

Table 2. Mean sentence lengths in words in COHA across time and genre

Decade	Magazine	Newspaper	Non-fiction	Fiction	Movie & play script
1810	27.29	NA	27.96	11.12	NA
1820	27.77	NA	24.23	17.64	NA
1830	27.76	NA	24.06	18.51	NA
1840	27	NA	23.41	17.85	NA
1850	25.84	NA	23.58	17.69	NA
1860	26.02	21.57	22.77	15.93	NA
1870	25.44	22.13	21.52	15.47	NA
1880	24.96	21.39	20.5	15.9	NA
1890	25.02	19.9	20.9	15.9	NA
1900	22.65	17.83	21.97	14.58	16.86
1910	20.9	18.89	21.1	13.54	11.66
1920	18.34	18.14	21.25	13.2	12.46
1930	18.53	19.96	20.74	12.69	10.1
1940	17.54	20.2	20.56	12.37	9.49
1950	17.57	19.04	20.36	12.36	9.84
1960	17.72	18.2	20.53	12.09	9.12
1970	17.87	18.72	20.45	11.76	9.77
1980	17.65	18.77	19.65	11.73	8.63
1990	17.78	16.72	20.8	11.8	8.62
2000	17.14	16.7	19.94	12.07	NA

3.3 Discussion

The observed decrease in sentence length across time seems to be an outcome of many factors. Some, already mentioned in this work, range from the need for greater text comprehensibility accompanying the development of mass readership (Schneider 2002: 98ff) to competition for the potential reader (Westin 2002: 161) and change of punctuation conventions. Contrary to the idea that the decrease of sentence length is only a "by-product" of change in punctuation practices (Fahnestock 2011: 265), it did not stop once the punctuation became its present-day version. So there have to be other trends and phenomena at work. One of them might be the overall informalization of the language of the media – a trend described by Leech et al. (2009: 239). The decrease in sentence length is, however, not only typical for media language but, as this work aims to show, also for other genres such as *fiction* and *non-fiction*. Thus it might actually be one of the symptoms of a larger phenomenon concerning all the genres of the written language. A phenomenon encompassing this kind of change is the trend introduced by Mair as "colloquialization" (1998: 153), namely "a significant stylistic shift in twentieth century English" (Mair 2006: 187) due to which the written language becomes more similar to the spoken language and more tolerant to various degrees of informality. One of the clear examples of colloquialization is the increase in the frequency of use of contractions (negative contractions in particular) during the twentieth century, which has been quite dramatic in written American English (Mair 2006: 190). The results of the present analysis, however, also point towards a decrease in the sentence length in the *movie & play script* genre, which is used as a proxy for the spoken language. Genre conventions typical for scripts might account for the fact that *movie & play script* has the shortest sentences overall. They, however, do not explain the gradual decrease in sentence length. If the decrease of sentence length in written language might be a symptom of colloquialization, what is it a symptom of when happening in spoken language? The answer to this question might be of interest to not only linguists but also to sociologists and cognitive scientists. However, to be answered in a satisfactory way, it requires a separate study. For the *non-fiction* genre, part of the decrease might have to do with the trend associated with the growing lexical difficulty of the scientific prose, which make the word-load of the sentences considerably lower (Gross et al. 2002; Dorgeloh 2005; Hundt et al. 2012).

4. Sentence length and syntactic usage

According to the results of the previous analysis, sentence length in American English has decreased quite considerably during the last 200 years. But does this mean that the messages produced by the language users of today are less complex?

Probably not. On the other hand, a decrease in sentence length will, quite likely, have an influence on the contents of this new, shorter sentence.

An interesting case in this context is provided by a non-finite subordinator *in order to*. The first use of its present-day version dates back to the early seventeenth century (according to OED Online).[10] *In order to* came into existence as a kind of 'reinforcement' of the purposive meaning of the *to*-infinitive (Schmidtke-Bode 2009: 174). According to Los (2005: 28), "the function in which the *to*-infinitive first appeared was that of a purpose adjunct". As the scope of usage of the *to*-infinitive extended greatly during the Old and Middle English periods, the addition of *in order* in front of *to* was supposed to disambiguate the purposive meaning.

In order to is, even today, one of the main purposive subordinators. Figure 7 shows its frequency of use (normalised, per million words) in COHA. The frequency values look rather stable until the end of the nineteenth century. But then, just around the turn of century, we can see a visible decrease. This decrease follows a pattern very similar to the decrease of sentence length presented in Figures 2 to 6.

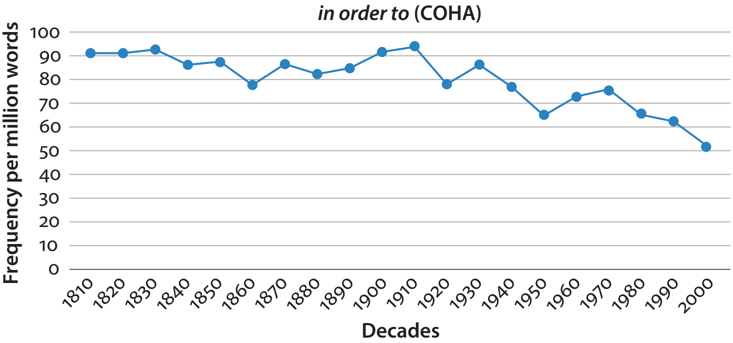


Figure 7. Frequency of *in order to* across the time period 1810 to 2009[11]

This correlation between the decrease of frequency of *in order to* and the decrease of sentence length could reflect a process opposite to what happened when *in order* was added in front of the purposive *to*-infinitive. Since sentences are shorter now, chances are that there are less sentences containing more than one *to*-infinitive, so the need for precision, to the fulfilment of which *in order to* was born, might just not be there anymore.

Another observation that supports the link between the decrease in sentence length and the decrease in the frequency of use of *in order to* is the fact that this

[10] OED Online, s.v. *in order to*, retrieved from <http://www.oed.com> (29 May 2016).

[11] Data retrieved from COHA on May 30th 2016.

decrease seems, at least to some extent, genre-dependent. Figure 8 presents the decrease in frequency of *in order to* across time and across different genres of COHA (normalised frequency, per million words).

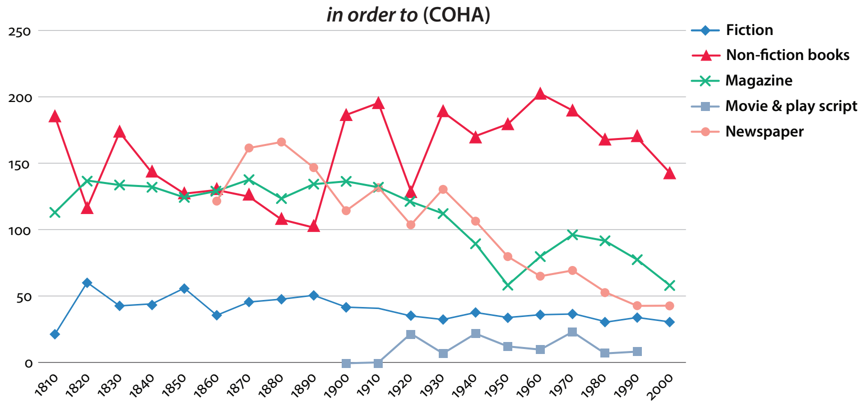


Figure 8. Frequency per million words of *in order to* across the time period 1810 to 2009[12] in *fiction*, *magazine*, *non-fiction*, *newspaper* and *movie & play script* (time period 1860 to 2009)

As we can see *non-fiction* is the only genre in which we can see some degree of increase in the frequency of *in order to*. Part of the explanation for this might be provided by the fact that, at present, it is the genre with the longest sentences among all the investigated genres (see Table 2). In *fiction*, *newspaper* and *magazine* there is a visible decrease in the frequency of *in order to*. The genre *movie & play script* is the genre with the lowest overall number of instances.

Another part of the explanation might be in the original function of *in order to*. Since its aim was to provide more clarity and precision, the genre *non-fiction* might actually prove to become its "natural environment" due to the rhetorical conventions and the need for clarity and unambiguousness typical of *learned* genres.

5. Conclusions

The results of the comprehensive analysis of sentence length across time (1810–2000) and genre show a decrease in sentence length for all the investigated genres. The degree of the decrease is slightly different for each genre; namely, in *fiction* sentences are approximately (on average) 6.5 words shorter now than they were

[12] Data retrieved from COHA on January 15th 2017.

in the beginning of the nineteenth century. For *magazine* the decrease amounts to 10.2 words, for *newspaper* 4.9, for *non-fiction* 8 and for *movie & play script* 8.3 words. The initial sentence lengths were, however, different for each genre, with *movie & play script* starting from the lowest sentence length and finishing at the lowest sentence length from all genres and *non-fiction* starting and staying the genre with longest sentences.

Various explanations have been offered to account for the observed decrease in sentence length across time and genre. Among the processes at work which might influence sentence length there are informalization of the language of the media (Leech et al. 2009: 239), colloquialization (Mair 1998: 153), a trend towards less explicitness in writing (noted by e.g. Biber & Gray 2010), trends accounting for changes in scientific discourse such as the increase in the use of compressed and complex noun phrases (Gross et al. 2002: 171; Dorgeloh 2005: 92; Hund et al. 2012: 236) and a decrease in the relative clause frequency (Gross et al. 2002: 171; Hundt et al. 2012: 225). In the *non-fiction* genre, the decrease happening between 1990 – 2009 might, at least partially, be attributed also to the addition of COCA contents (mostly texts from peer-reviewed academic journals).

The more marked decrease in sentence length observed around the beginning of the twentieth century might have to do with the development of mass readership (Schneider 2002, quoted in Fries 2010: 23), competition for potential readers (Westin 2002: 161) and a change in punctuation conventions (noted by e.g. Fahnestock 2011: 256 and Biber & Conrad 2009: 153), which as such might actually be the resultant of the two previous phenomena. The above mentioned processes and developments might account for the decrease in sentence length across particular written genres but they do not explain the observed decrease in the proxy for the spoken language – *movie & play script*. The question whether the observed decrease would also be visible in the spoken language and which processes could possibly account for it remains open.

This paper aimed to point out that the decrease in sentence length might be linked to changes in the syntactic usage of English. Non-finite purpose subordinator *in order to* is used to illustrate the potential influence that the decrease of sentence length might have on the constructional layer of the language. The observed decrease in sentence length is, according to the present work, an instantiation of a higher-order phenomenon. What are other examples of higher-order phenomena? According to Hilpert (2013: 14), processes and phenomena which affect multiple constructions at the same time represent higher-order phenomena, among the examples there are the development of a global, phonotactic constraint and deflexion (a general loss of inflectional morphological categories):

> [W]henever several members of grammatical paradigm are in demise or even whole groups of paradigms disappear, there are reasons to view the change as non-constructional, because a higher level of grammatical organization than the construction is concerned.

There are thus reasons to view the decrease of sentence length as a higher-order phenomenon at work, which, in turn, might – at least to some extent – be influencing syntactic usage. The shorter sentences of today might make the use of more explicit purpose subordinators such as *in order to* redundant. It is very likely that *in order to* is not the only construction affected, as the decrease in the frequency of use is generalizable to other constructions of the network of purpose subordinators, such as *so as to*, *lest* and *in order that* (Rudnicka in preparation). A scenario for the future of *in order to* could be a situation in which it becomes a marker of style, used mostly in learned texts or for rhetorical purposes.

Since the decrease in sentence length has been, at least in the language of science (Gross et al. 2002: 124), shown to be generalizable to other languages, it seems worthwhile to check if other genres across different languages and across time would show comparable degrees of sentence length decrease, and if the possible explanations for this decrease would be similar cross-linguistically.

Corpora

Davies, Mark. 2008–. *The Corpus of Contemporary American English (COCA): 520 million words, 1990–present*. <http://corpus.byu.edu/coca/>

Davies, Mark. 2010–. *The Corpus of Historical American English (COHA): 400 million words, 1810–2009*. <https://corpus.byu.edu/coha/>

References

Biber, Douglas & Conrad, Susan. 2009. *Register, Genre, and Style*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511814358

Biber, Douglas & Gray, Bethany. 2010. Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes* 9: 2–20. https://doi.org/10.1016/j.jeap.2010.01.001

Culpeper, Jonathan & Kytö, Merja. 2010. *Early Modern English Dialogues: Spoken Interaction as Writing*. Cambridge University Press.

Davis, Edith A. 1937. Mean sentence length compared with long and short sentences as a reliable measure of language development. *Child Development* 8(1): 69–79. https://doi.org/10.2307/1125824

Dorgeloh, Heidrun. 2005. Patterns of agentivity and narrativity in early science discourse. In *Opening Windows in Discourse and Texts from the Past* [Pragmatics & Beyond New Series

134], Janne Skaffari, Matti Peikola, Ruth Carroll, Risto Hiltunen & Brita Warvik (eds), 83–94. Amsterdam: John Benjamins. https://doi.org/10.1075/pbns.134.10dor

Fahnestock, Jeanne. 2011. *Rhetorical Style: The Uses Of Language In Persuasion*. Oxford: OUP. https://doi.org/10.1093/acprof:oso/9780199764129.001.0001

Fries, Udo. 2010. Sentence length, sentence complexity and the noun phrase in 18th-century news publications. In *Language Change and Variation from Old English to Late Modern English* [Linguistic Insights 114], Merja Kytö, John Scahill & Harumi Tanabe (eds), 21–33. Bern: Peter Lang.

Gross, Alan G., Harmon, Joseph E. & Reidy, Michael S. 2002. *Communicating Science: The Scientific Article from the 17th Century to the Present*. Oxford: OUP.

Hames, Tim & Rae, Nicol C. 1996. *Governing America: History, Culture, Institutions, Organisation, Policy*. Manchester: Manchester University Press.

Hilpert, Martin. 2013. *Constructional Change in English: Developments in Allomorphy, Word Formation, and Syntax*. Cambridge: CUP. https://doi.org/10.1017/CBO9781139004206

Hult, Christine A. 2015. *The Handy English Grammar Answer Book*. Canton MI: Visible Ink Press.

Hundt, Marianne, Denison, David & Schneider, Gerold. 2012. Relative complexity in scientific discourse. *English Language and Linguistics* 16(2): 209–240. https://doi.org/10.1017/S1360674312000032

Klare, George R. 1963. *The Measurement of Readability*. Ames, IA: Iowa State University Press.

Leech, Geoffrey, Hundt, Marianne, Mair, Christian & Smith, Nicholas. 2009. *Change in Contemporary English. A Grammatical Study*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511642210

Lewis, Edwin H. 1894. *The History of the English Paragraph*. Chicago IL: The University of Chicago Press.

Los, Bettelou. 2005. *The Rise of the To-Infinitive*. Oxford: OUP. https://doi.org/10.1093/acprof:oso/9780199274765.001.0001

Mair, Christian. 1998. Corpora and the study of major varieties in English: Issues and results. In *The Major Varieties of English: Papers from MAVEN 97*, Hans Lindquist, Staffan Klintborg, Magnus Levin & Maria Estling (eds), 139–157. Växjö: Acta Wexionensia.

Mair, Christian. 2006. *Twentieth-century English: History, Variation and Standardization*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511486951

McGregor, Gordon P. 2002. *English for Life? Teaching English as a Second Language in Sub-Saharan Africa with Special Reference to Uganda*. Kampala: Fountain Publishers.

Miller, Raymond D. 1908. Coordination and the comma. *PMLA* 23(2): 316–328. https://doi.org/10.2307/456693

Rudnicka, Karolina. In preparation. The Statistics of Obsolescence: Purpose Subordinators in Late Modern English. PhD dissertation, Albert-Ludwigs-Universität Freiburg.

Säily Tanja, Vartiainen Turo & Siirtola, Harri. 2017. Exploring part-of- speech frequencies in a sociohistorical corpus of English. In *Exploring Future Paths for Historical Sociolinguistics* [Advances in Historical Sociolinguistics 7], Tanja Säily, Arja Nurmi, Minna Palander-Collin & Anita Auer (eds). Amsterdam: John Benjamins. https://doi.org/10.1075/ahs.7

Schmidtke-Bode, Karsten. 2009. *A Typology of Purpose Clauses* [Typological Studies in Language 88]. Amsterdam: John Benjamins. https://doi.org/10.1075/tsl.88

Schneider, Kristina. 2002. The Development of Popular Journalism in England from 1700 to the Present, Corpus Compilation and Selective Stylistic Analysis. PhD dissertation, Universität Rostock.

Schou, Karsten. 2007. The syntactic status of English punctuation. *English Studies* 88(2): 195–216. https://doi.org/10.1080/00138380601042790

Štajner, Sanja & Mitkov, Ruslan. 2012. Diachronic changes in text complexity in 20th century English language: An NLP approach. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet U. Dogan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds), 1577–1584.

Westin, Ingrid. 2002. *Language Change in English Newspaper Editorials*. Amsterdam: Rodopi.

A comparison of multi-genre and single-genre corpora in the context of contact-induced change

Carola Trips & Achim Stein
University of Mannheim / University of Stuttgart

This chapter discusses results from a quantitative study of possible contact-induced change in Middle English in a multi-genre corpus (the *Penn-Helsinki Parsed Corpus of Middle English 2* (PPCME2), Kroch & Taylor 2000, and a single-genre corpus, the *Penn Corpus of Early English Correspondence* (PCEEC), Taylor et al. 2006). We claim that the data from the correspondence corpus are critical to understanding the rise of the recipient passive (more traditionally called the indirect passive) in late ME because they reflect the active competence of writers, much more than other genres do (at least in the PPCME2). More precisely, we argue that our data reflect a verb class specific passive construction that seems to be firmly established in the grammar of the writers. This construction is not calqued from the model language (Old French) but the result of interpreting the French dative as different from the English 'dative'.

1. Introduction

Investigating phenomena like language change and contact by means of corpus-based studies involves the process of selecting "suitable" corpora. Today a plethora of different types of corpora is available and sometimes only by comparing results of different corpora do we gain new insights into hitherto unexplained findings. In the case we are going to discuss here a comparison of a multi-genre and a single-genre corpus was undertaken because a previous study revealed that working with just one (multi-genre) corpus could not explain the empirical facts in a satisfying way.

The case at hand is the rise of a type of passive in English which Allen (1995) calls the recipient passive. Traditionally this type of passive is called the indirect passive of ditransitive verbs. The example in (1) illustrates the two main types of passive in present-day English:

(1) a. Tom gave Mary presents. active
b. **Presents** were given to Mary (by Tom). direct passive
c. **Mary** was given presents (by Tom). indirect/recipient passive

https://doi.org/10.1075/scl.85.11tri

When the direct object of a ditransitive verb is passivised as in (1b) we talk about the direct passive, and when the indirect object is passivised as in (1c), we talk about the indirect or recipient passive (the latter term emphasises that it is the recipient of the action expressed by the verb that is passivised).

In the literature on passives in English the recipient passive is a well-researched topic concerning its (formal) properties and analysis, but when it comes to its origin and development, not much can be found. In our research project Borrowing of Argument Structure in Contact Situations (BASICS) we investigate English-French contact in medieval times on the level of argument structure. The contact situation is well known, of course, triggered by the Norman Conquest of 1066, and resulting in a considerable number of borrowings. What has not been investigated so far is in how far the argument structure of English verbs was affected by the borrowing of Old French verbs. This also includes syntactic operations and changes affecting the verb (phrase). What we know about the development of the passive in English is that it was affected by the development of the English case system (cf. Allen 1995), that 'dative' is not a homogeneous category (it can be expressed by a NP, PP, structural case, inherent case, etc.), and that it is harder to acquire 'dative' in child bilingual language acquisition (cf. studies of German/Romance bilinguals, for example Schmitz 2006 and Scherger 2015, and the general observations concerning structural vs. inherent case in monolingual acquisition resumed in Eisenbeiss et al. 2009: 379–382). In the literature it has also been noted that dative case is more easily affected by language contact than nominative or accusative (see the numerous examples in e.g. Heine & Kuteva 2005 and Johanson 2009).

The goal of this chapter is to discuss the plausibility of the contact hypothesis, i.e., whether contact with French speakers/writers in Middle English times could have led to the rise of the recipient passive. We will provide some interesting results that could not have been gained without using different types of annotated corpora for Middle English (ME) and Old French (OF).

The chapter is organised as follows: Section 2 briefly sets the stage by discussing the case systems of both ME and OF as well as passive constructions occurring at this time. In Section 3 we present our assumptions about the rise of the recipient passive which are empirically based on corpus studies of two annotated corpora, the *Penn-Helsinki Parsed Corpus of Middle English 2* (PPCME2, Kroch & Taylor 2000), a multi-genre corpus, and the *Penn Corpus of Early English Correspondence* (PCEEC, Taylor et al. 2006), a single-genre corpus which spans Middle English and Early Modern English (EModE). Section 4 discusses the rise of this passive construction as an instance of contact-induced change. Section 5 concludes.

2. Passive and case

It is well-known that Old English (OE) displayed inflectional morphology on both nouns and verbs. Nouns showed a four-way case system (nominative, genitive, accusative, dative), and a three-way number (singular, plural, dual) and gender (masculine, feminine, neuter) system. Verbs inflected for tense, person, number, and mood. From about 1100 the development of this language system is one of continuous attrition. According to Allen (1995) and others, morphological marking of the dative case was lost in the Midlands by the middle of the 12th century, in the South by the end of the 13th century and in Kent by the middle of the 14th century. So by 1400, dative marked by the inflectional ending *-e* had vanished from the language.

This and further processes of attrition led to the loss of the accusative-dative distinction resulting in the objective case, the loss of the nominative-dative case distinction resulting in an ambiguity between preverbal nominal indirect object and subject, and the rise of new types of passives (cf. Allen 1995, Denison 1993).

In Old English two types of passive occurred (in (2) we give modernised examples). First, the impersonal passive (2a–c) where the objects retained case, the finite verb showed the 3-SG ending and there was no nominative subject. Second, the direct passive (2d–e) where the direct object of an active sentence was promoted to the subject position of a passive sentence. Monotransitives like *helpan* 'help' showed the first type, and ditransitives like *giefan* 'give' showed the impersonal or direct passive. The examples in (2c/e) show that dative fronting was also possible.

(2) a. Us (obj-DAT) **is** helped by God. impersonal
b. Presents (obj-ACC) **was** given her (obj-DAT). impersonal
c. Her (obj-DAT) **was** given presents (obj-ACC). impersonal
d. Presents (subj-NOM) **were** given her (obj-DAT). direct
e. Her (obj-DAT) were **given** presents (subj-NOM). direct
(see also Denison 1993: 104)

What is not found at that time is the indirect passive where the indirect object of an active sentence is promoted to the subject position of a passive sentence (Denison 1993; Mitchell 1985: 110ff).

In Middle English almost all of the OE constructions were lost (or could no longer be produced after the loss of case marking; see also Fischer (2010) on OE quirky subjects). At some point after the loss of the OE passive constructions the recipient passive must have emerged since today we have the following possibilities to build the passive, where (3c) is the recipient passive:

(3) a. Presents were given to Mary (by Tom).
b. %Presents were given her. [dialectal]
c. **Mary** was given presents (by Tom).

Before we present our investigation based on the assumption that the rise of the recipient passive may be seen as a case of contact-induced change with Old French, a look at the Old French case system and passive constructions is in order.

The Old French case system was subject to the attrition of inflectional morphology just like the Old English case system was. The inflection of nouns was reduced to two cases: nominative (< Latin nominative and vocative) and accusative (< other Latin cases). This case distinction gradually disappeared towards the end of the OF period (around 1300). Personal pronouns still distinguish three cases (nominative, accusative and dative) in OF (Buridant 2000: §326) as well as in Modern French (Grevisse 2011: Chapter 4):

(4) a. e.g. 'he': nom. *il* – acc. *le* – dat. *li*
b. cf. ModF: nom. *il* – acc. *le* – dat. *lui*

Generally, inflectional case marking is distributed etymologically in OF: so case endings (e.g. -*s*) can come from Latin nominative or accusative which leads to many ambiguities and intransparent paradigms (cf. Jensen 1990: 30ff).

Concerning the morphological marking of dative case, in Old French former Latin datives are rarely expressed as absolute 'dative', marked by oblique case (< accusative) as the following example shows:

(5) *il estoit Lancelot*
it was Lancelot$_{\text{OBL}}$
'it belonged to Lancelot'

For the most part, Latin datives were expressed as prepositional 'datives': the preposition *a* plus an oblique NP, as in Modern French (ModFr, cf. Jensen 1990: 30ff). The following example illustrates the use in an Old French text:

(6) *et dunet a hume graze et dulce parole*
and gives to man.OBL grace and sweet word
'and he gives the man grace and sweet word' (SRCMF Lapidf p. 102)

As in Old English, Old French exhibits a number of different types of passive constructions. In the most frequent case, the direct object of the active clause becomes the subject, as in (7). The other option is the impersonal passive, as in (8):

(7) *Quant le parole fu oïe si en furent tout li Franchois*
When the word was heard then by it were all the French

molt liés
very glad
'When the word was heard then all the French were very glad'
(SRCMF, clari_17_1291552991.41)

(8) *Ne vos sera mie celé/ Qui nos somes et de*
NEG you.DAT.2.PL be.FUT.3-SG ever hidden who we are and of
quell terre
which country
'It will never be hidden to you who you are and where from'
(SRCMF,YvainKu,pb:99,lb:5244)

Crucially, however, the indirect or recipient passive does not exist at this time: no ditransitive verbs like *doner* 'give' can be found in this construction. Variation between accusative and dative can only be observed with some monotransitive verbs like *aider* 'help' (cf. Troberg 2008), and with some ditransitives taking an object and a clausal complement, for example communication verbs like *comander* 'command' (in the SRCMF corpus 414 active occurrences, regardless of the verb):

(9) *Jo vos cumant qu' en Sarraguce algez*
I you.DAT.2PL command that to Zaragoza go.3PL
'I command you to go to Zaragoza' (SRCMF,roland,pb:198,lb:2673)

However, passivisation only occurs with verbs taking a direct object, so recipient passive constructions are neither part of Old French nor Modern French grammar. At this point the case is clear: if our assumption that the recipient passive emerged in English due to contact with Old French had been tenable, we should have found evidence for it in Old French as the model language in this contact situation. Yet, this is not the end of the story, as we will show in the next section.

3. The rise of the recipient passive in English

3.1 Allen's (1995) study

In her 1995 monograph *Case Marking and Reanalysis* Allen provides a comprehensive account of the recipient passive and other types of passive in Old and Middle English.[1] For her, clear cases of the recipient passive are passives formed from ditransitive verbs where the recipient is the subject and the subject either occurs

[1] For other accounts see Estival 1989: 25, Los 2009, Seoane 2006.

in the nominative or agrees with the finite verb. In her study which includes texts from different genres (homily, chronicle, history, sermon, lives of saints, etc.) written both in prose and verse,[2] Allen found that in ME recipient passives scarcely occur, and she states that this cannot be explained by assuming that it spread from a particular subclass of ditransitives. Rather, she states that 'bare objects directly following the verb were generally interpreted by the last quarter of the fourteenth century as direct objects, but these direct objects were not frequently passivized' (1995: 394–395). Further, she observes a gap in the development of ditransitives (1995: 384). Whereas the change from an indirect passive of monotransitives (Old English 'Us was helped') to a direct passive (Middle English 'We were helped') can be interpreted as reanalysis, this cannot be assumed for ditransitives. Dative fronting (Old English 'Us was given the book') was not reanalysed as recipient passive (Middle English 'We were given the book') because they did *not* co-occur. Evidence for this hypothesis is that dative fronting is limited to archaic texts, e.g. poetry, and that in prose texts from the fourteenth century neither dative-fronted passives nor recipient passives occur (e.g. in the works of Wycliffe). So clearly no replacement was taking place.

Allen makes another interesting observation: she gives two examples as possible cases of dative fronting from the *Ayenbite of Inwyt* (1340):

(10) a. *Ine þis heste* ***ous is uorbode*** *alle zenne of ulesse*
in this commandment us is forbidden all sins of flesh
'In this commandment, all sins of the flesh are forbidden to us'
(Ayenbite 9.9, in Allen 1995: 385)

b. *Ine þise heste* ***ous ys uorbode*** *þet we ne lyeȝe*
in this commandment us is forbidden that we not lie
'In this commandment we are forbidden to lie'
(Ayenbite 10.6, in Allen 1995: 385)

At first sight, these cases might be taken to exhibit residues of fronted datives but if we take into account that this text is a direct translation from the French text *Somme le Roi* (1279) things become less clear. It is well-known that the *Ayenbite*

[2] Note that Allen's study is not a corpus study in the sense we would use the term today, i.e. based on an (annotated) electronic corpus. At the time when Allen wrote her book, such corpora were not available for Old and Middle English. Many of the texts she investigated are part of the York-Helsinki Parsed Corpus of Old English Prose (YCOE) and the PPCME2 but Allen looked at the texts in full whereas the parsed corpora used in our studies provide samples of the texts only. So despite the fact that there is an overlap in the texts used, the results of Allen's studies and ours are comparable only to a limited degree.

of Inwyt 'contains some very un-English constituent order' (Allen 1995: 386).[3] A closer look at the French original shows that the two occurrences cited in (11) may indeed be calqued from Old French:

(11) a.

En	*ce*	*commandement*	***nous***	***est***	***deveé***	*tout*	*pechié*
in	this	commandment	us	is	forbidden	all	sin

de	*char*
of	flesh

'In this commandment, all sins of the flesh are forbidden to us'
(SOMME, Chapter 10, Para 57)

b.

En	*ce*	*commande-[b]-ment*	***nous***	***est***	***deveé***	*que*	*l'en*
in	this	commandment	us	is	forbidden	that	of-it

ne	*mente*
not	lie.3.SG

'In this commandment we are forbidden to lie'
(SOMME, Chapter 10, Para 71–72)

These cases clearly are direct translations from Old French into Middle English. Thus, they occur due to French influence and not because they were part of the English grammatical system at the time.

Concerning the rise of the recipient passive, Allen brings in French influence as well, albeit in an indirect fashion. She notes that not many examples can be found in the fourteenth and fifteenth centuries and she attributes the low frequency of this type of passive to language contact with Latin and French. Since many of the ME texts are based on Latin and French texts, and since these languages do not allow recipient passives, it is not surprising that they are rarely found in the ME 'translations'. For the same reason, 'the first examples of recipient passives are found in texts not aspiring to a polished literary style' (Allen 1995: 395). Allen further states:

> It is not difficult to find a reason why such passivization should be kept to a minimum. Most prose texts from this time are translations from Latin or French, and even original texts were affected by the **grammatical models of French** and Latin, which did not allow recipient passives. It seems most likely that recipient passives were **first used in speech** and only gradually gained acceptance in writing. (Allen 1995: 395, our emphasis)

Allen's comment addresses two points: first, the absence of a syntactic construction can be evidence for contact, and second the mode of speech (being related to

[3] See also the comments by Richard Morris in the preface of the EETS edition (O.S. 278. London: Oxford University Press).

genre) is crucial for tracing the development of the recipient passive. According to these hypotheses, which we are going to investigate in the following section, we will use two different types of corpora: a multi-genre corpus that includes many of the prose texts Allen included in her studies, and a single-genre corpus of letters which reflects authentic speech to a much higher degree than most of the texts in the multi-genre corpus, since letters are not copied or translated from other sources and normally motivated by a specific communicative goal rather than by textual traditions. We expect to find differences in frequency of occurrence of the recipient passive between the two corpora, i.e. more cases of the recipient passive in the corpus of correspondence than in the multi-genre corpus. In the next section we will present and discuss our results.

3.2 Comparing results from a multi-genre and a single-genre corpus study

As we mentioned in the previous sections, we used two syntactically annotated corpora for ME (and EModE), the PPCME2 as multi-genre corpus and the PCEEC as single-genre corpus. For Old and Middle French (MF) we used the *Syntactic Reference Corpus of Medieval French* (SRCMF, Prévost & Stein 2013). Table 1 gives the English and French periods as they are defined in the three corpora used:

Table 1. English and French periods in the three corpora used

English	time span	French	time span
OE	< 1150		
ME-M1	1150–1250	OF	842–ca.1320
ME-M2	1250–1350		
ME-M3	1350–1420	MF	ca.1320–1500
ME-M4	1420–1500		
Early Mod. English	1500–1710		

Based on what we said in Section 3 we hypothesised that first cases of recipient passives are found rather late in the ME period. The first 'genuine example of a recipient passive' (Allen 1995: 393) is quoted by Visser in a text from 1375:

(12) Item as for the Parke **she** is a lowyd Every yere a dere
'Item: as for the park, she is allowed a deer each year'
(AwardBlount p. 205 (1375), from Allen 1995: 393)

Allen further states that the recipient passive becomes more frequent in the course of the fifteenth century, and it is 'quite common' (395) by the early sixteenth century.

We queried our corpora by searching for passives of ditransitive verbs governing a subject *and* a direct object. In the passive, the subject occurs as the former dative argument. Our first results confirmed Visser's and Allen's observation that the first occurrences of true recipient passives are found in late ME (in the corpus in the subperiod M4, 1420–1500). A rather unexpected finding was, however, that recipient passives are much more frequent with verbs of French origin than with native verbs. This result was found in both corpora, but to a much higher degree in the PCEEC.

Let us first turn to the PPCME2. As a multi-genre corpus it is a representative corpus including text samples from the genres document, handbook, science, philosophy, homily, sermon, rule, religious treatise, prologues, epilogues, history, travelogue, biography, fiction, romance, and bible. It spans the time 1150–1500, contains 55 texts and overall comprises ca. 1.2 million words.

Only five examples of the recipient passive occurred in the whole corpus, and all of them in period M4. Four out of five cases exhibit verbs of French origin (*deliveren*, *banishen*, *serven*, *paien* 'pay'), one case exhibits a native verb (*smiten* 'cut').

Two examples are from Thomas Malory's *Morte Darthur*. Malory was a soldier born to a Warwickshire family around 1410. He probably wrote the *Morte Darthur* in prison and finished it around 1469. He died in 1471. Book V is based on the English *Alliterative Morte Arthure*, and the rest is from various French romances.[4] The two cases with recipient passive show the verbs *smyten* and *deliveren*:[5]

(13) And so hit befelle that **a man of kynge Evelakes** was smytten hys honde off
'And so it happened that a man of king Evelakes was cut off his hand'
(MALORY,641.3972)

(14) Than **Ulphuns and Brastias** were delyvirde three thousand men of armys
'Then Alfons and Brastias were delivered three thousand men of arms'
(MALORY,21.633)

Two further examples occur in *Gregory's Chronicle* which was written by William Gregory in 1451–52. He put together the chronicle, but only the part from the 19th to the 30th years of the reign of Henry VI can be attributed to him. The two

[4] The philological information about texts is gained from the PPCME2 corpus website.

[5] We quoted (2) here although it is not a typical instance of a recipient passive: *a man of kynge Evelakes* is not a recipient, but a pertinence dative, typically occurring with body parts. We are also aware that a thorough analysis of the examples quoted here would produce further differences, but they are not in the focus of the present contribution.

cases found in this text show the two French verbs *banishen* and *serven* in the recipient passive:

(15) … that ys to wete, **iij sowdyers** were banyschyde the towne of Caleys.
'… that is to say, three soldiers were denied the town of Calais'
(GREGOR,176.1112)

(16) and **they** were servyd nexte unto the quene every cours coveryde as the quene.
'and they were served next to the queen every course covered as the queen'
(GREGOR,139.584)

Finally, there is one example from *Capgrave's Chronicle*. The chronicle was written by John Capgrave who was born in Norfolk and became an Augustinian friar. He completed the chronicle, which is a holograph, between 1462 and 1463; it is partly influenced by the Latin source *Chronicon Pontificum et Imperatorum* of Martinus Polonus. The example shows the French verb *paien* used in the recipient passive.

(17) and **þei** þat took hem were treuly payed too þousand pound.
'and those who took them were truly paid two thousand pounds'
(CAPCHR,153.3587)

A clear contrast emerges now when we compare the findings from the PPCME2 whith those from the *Parsed Corpus of Early English Correspondence* (PCEEC, Taylor et al. 2006). The PCEEC consists of 84 letter collections, i.e. a total of 4970 letters (ca. 2.2 million words), from the time between 1410 and 1695. The corpus comprises letters by people of good social position from all dialectal regions like the Paston family (Norfolk gentry), the Bacon family (Norfolk gentry, Nathanial Bacon, justice of the peace in Norfolk), the Cely family (London, wool merchants, middle-class) the Plumpton family (Yorkshire), and the Stonor family (Oxfordshire, John Stonor, Chief justice of the common pleas). However, it also includes letters by the servants of these families as well as by individuals of high esteem like William Allen (Cardinal of Canterbury). From this information we can conclude that if the recipient passive occurs in letters by writers of all regions and social positions it is highly probable that it was an established construction in English grammar at that time.

Table 2 presents the overall quantitative results of the active and the recipient passive with the most frequent French and native verbs in the PCEEC. The most striking result is that despite the high token frequency of native verbs like *senden* and *yeven* (PDE *give*), the percentage of recipient passive occurrences is only 0.13 within this group of verbs. These figures contrast sharply with those for verbs of French origin like *paien* and *promisen*: their token frequencies are much lower, but they account for a disproportionately and significantly larger number of tokens of recipient passive (11.44%, $P < 0.001$).

Table 2. Active and recipient passive with ditransitive verbs in the PCEEC

French verbs				**Native verbs**			
Verb	**Active**	**RP**	**% RP**	**Verb**	**Active**	**RP**	**% RP**
paien	162	10		senden	2502	1	
promisen	96	7		yeven	1545	2	
offren[*]	55	12		tellen	446	2	
allouen	40	7		sheuen	223	1	
denien	31	7					
serven	2	1					
finen	1	6					
	387	50	11.44		4716	6	0.13

*From Latin *offerre*, reinforced by French *offrir*

Before we interpret this rather puzzling finding let us take a look at some examples:

(18) *and that* ***he*** *shuld be servid the same wythinne fewe dayes.*
and that he should be served the same within few days
'and that he should be served the same within a few days'
(M4-1461,PASTON,II,248.387.10097)

In this example, dated to 1461, the verb *serven* is used in the recipient passive in a letter by Thomas Playter, the family servant of the Paston family, to John Paston, the master of the house.

The example in (19) is also part of the Paston letters, but here John Paston wrote to his wife Margaret Paston in 1465. He used the recipient passive with the verb *paien*:

(19) *And seye þat* ***ye*** *will be paijd euerj pene,*
and say that you will be paid every penny
'… and say that you will be paid every penny'
(M4-1465,PASTON,I,133.035.765)

In the Example (20), which is one of the examples of the first Early Modern English subperiod in the corpus, Edward Bonner, one of the English ambassadors at the time, wrote to Thomas Cromwell, the chief minister to Henry VIII. The verb used with the recipient passive is *denien* (note that the direct object in the subordinate sentence is the demonstrative *that*):

(20) *And if* ***I*** *be denid that, … he shall neuer skape my handes.*
and if I be denied that he shall never escape my hands
'..And if I am denied that he shall never escape my hands'
(E1,1539,WYATT,108.016.493)

The next example which is from 1585 (E2 in the corpus) is from a letter by Robert Dudley, Earl of Leycester, to William Cecil. Dudley uses the recipient passive with the verb *promisen*:

(21) ***he** is promised ayd of men and gallyes from the pope and the*
he is promised aid of men and gallyes from the pope and the
dukes of Savoy and Florence
dukes of Savoy and Florence
'He is promised aid of men and galleys from the pope and the dukes of Savoy and Florence' (E2,1585,LEYCEST,42.012.346)

The last example discussed in detail here is dated to 1589 in the corpus (E3 period). The author is Bishop Brian Duppa who wrote to Sir Justinian Isham. He uses the recipient passive with the verb *deny*:

(22) *But since **the Jesuits themselves** have been denied the priviledge of*
but since the Jesuits themselves have been denied the privilege of
receaving confession by letters
receiving confession by letters
'But since the Jesuits themselves have been denied the privilege of receiving confession by letters' (E3,1650,DUPPA,18.011.245)

These examples illustrate that the recipient passive is frequently used, and our extensive corpus study corroborates this. It is found in letters from people of the gentry, both men and women (writing to each other), from servants of lords and ladies, from merchants, and from people of different regions.

To explain why English speakers were more likely to use the recipient passive when using verbs of French origin we will further explore the language contact situation between OF and ME. In the next section, we will discuss an analysis along these lines.

4. The language contact hypothesis

In the following we will apply Johanson's (2002) model of language contact to the contact situation under investigation. Johanson states that situations of contact are generally characterised by an asymmetrical dominance relation between a Basic Code A which is the sociolinguistically dominated (or weak) code, and a foreign Model Code B which is the sociolinguistically dominant and prestigious (or strong) code. From ca. 1066 to 1250 we can assume that this was the situation in our case, i.e., Middle English was the Basic Code dominated by the foreign Model Code of Old French. From ca. 1250 onwards speakers changed from being post-childhood learners to becoming bilinguals (Ingham 2012a,b) and therefore we can expect changes in the dynamics of the relation between the two languages (cf. Trips & Stein forthcoming).

Instead of using the rather problematic term (and concept) of borrowing, we use Johanson's term of copying. Further we adopt the distinction between global coyping and selective copying. Johanson defines global copying as the case when a unit is copied from a foreign Model Code as a whole, i.e. a block of different properties (material, semantic, combinational, frequential). In contrast, selective copying is the case when only some of these properties are copied. Although we believe that when verbs and their argument structures are copied from the foreign Model Code to the Basic Code they are predominantly cases of global coyping, at present we cannot exclude that selective copying also takes place.

In the previous section we showed that the occurrence of the recipient passive seems to be related to the use of French verbs. In our contact scenario we interpret this finding as a (syntactic) result of the global copying of French verbs to ME. More precisely, we assume that when French verbs were copied they brought along their argument structures, i.e. an indication of the number of arguments the verb takes as well as their syntactic expression and their semantic relation to the verb (cf. Levin 1993, Levin & Hovav 2005). On the level of the syntactic expression of the copied French verbs we identified three different cases: The first case is global copying of a verb where both the foreign Model Code and the Basic Code have identical structures and where therefore no changes on the level of syntax occur. This case is illustrated in (23):

(23) a. Old French: *(il) crëante [NP les detes] [PP a autres persones]*
b. English: *He granted [NP the debts] [PP* ***to other persons****]*
(Basic Code: He gave the debts to other persons.)

The Basic Code marks the direct object with a NP and the indirect object (recipient) with a PP. The foreign Model Code shows the same structure and therefore both codes match.

The second case is when there is a structural mismatch, i.e. a structure of the foreign Model Code is adopted and its frequency in the Basic Code increases:

(24) a. Old French: *plaisir [PP a Deu])*
b. English: preference for *please [PP to God]* over *please [NP God]*.

The OF verb *plaire* was copied as *plesen* to ME and brought along its preference to use a PP to mark the indirect object instead of a NP. One instance of this case is illustrated with the ME example in (25):

(25) a. *For God wasted þe bones of hem þat* ***plesen to men;***
for God wasted the petitions of them that please to men.Exp
'Because God destroyed the petitions of those who please men'
(Basic Code: He queems men.) (EARLPS,63.2771,M2)

In the semantic class of the verbs of pleasing in ME the native verbs generally mark the EXPERIENCER with a NP (Allen 1995, Trips & Stein forthcoming). In the process of copying, the OF verb *plesen* with a preference for marking the EXPERIENCER with a PP was added to this class and consequently the Basic Code and foreign Model Code mismatch. The new structure is adapted in the Basic Code, which can be evidenced by higher frequencies of this particular structure (with native verbs):

(26) *and makeþ ofte / lete þet guod to done: and do þet kuead /*
and makes often let that good to do and do that evil
uor to ***kueme*** *kueadliche* ***to þe wordle****.*
for to please wickedly to the world.Exp
'… and do evil to wickedly please the world' (AYENBI,26.403,M2)

French original
et fait mout de foiz lessier le bien a fere et fere le
and makes many of times let the good to do and do the
mal pour ***plere*** *mauvesement* ***au monde****.* (SOMME, ch32, par687)
evil to please wickedly to+the world

The third case is when a structure that had previously not existed in English is added due to contact, e.g. reflexive uses of native verbs in ME on the model of OF:

(27) a. Middle English
Men ***feeren hem*** *in al the toun* (reflexive)
'men were afraid in the whole town'

b. Anglo-Norman
Meis mult ***se dute*** *de la gueite.* (reflexive)
'But (he) was very afraid of the lookout' (example from Richard Ingham)

The crucial question is now whether the rise of the recipient passive can be attributed to one of these cases. More precisely, we must explain that the recipient passive appears only in late Middle English (>1375), that it is much more frequent with verbs of French origin, and that (Old) French does not have this type of passive. What we can exclude right away is that the recipient passive is an instance of grammatical replication, defined by Heine & Kuteva (2008: 59) as "a process whereby speakers of a language, called the replica language, create a new grammatical structure on the model of some structure of another language". Since the recipient passive is not part of the grammar of the foreign Model Code this is not an option. We can further exclude that the absence of the recipient passive in OF had an inhibiting influence because then, clearly, recipient passives should not appear with French verbs first. The same applies to the assumption that the recipient passive is of native origin, emerging in one verb class and spreading to others (e.g. object predicatives: *He was named king*). If this is what happened then the recipient passive should not appear with French verbs first.

The data which provide critical evidence come from the PCEEC. They show the recipient passive in the correspondence of people and thus reflect the active competence of the writers. In our opinion, these must be distinguished from instances of syntactic calquing (see again the examples in [10]), since they are part of the writers' grammars. The syntactic expression of the argument structure of French verbs was fundamentally different from the one of the native verbs. We saw in (25) and (26) above, that in (Old) French there is a strong preference to syntactically express the indirect object with a PP. In the literature it is assumed that this is an instance of structural case (Zaenen & Maling 1990). In OE indirect objects were expressed by morphologically marking a NP as dative which is analysed as inherent case (cf. Chomsky 1981, McFadden 2002). Whenever indirect objects were marked by the dative this was associated with the native verbs. At the time of the contact situation dative case was on the decline but there were still some traces of the OE system left. During contact with OF new verbs came into ME and brought along properties that were different from the properties of the existing native verbs. One such property was structural case (PP) of indirect objects. We assume that speakers/writers were sensitive to this difference which is to say that they interpreted native and non-native indirect objects differently. The native indirect objects may still have had a flavour of 'dativeness' (i.e., properties of the former inherent dative), even without systematic inflectional case marking. The 'French' indirect objects, however, were perceived as a different kind of dative, i.e., a structural dative.

Based on these assumptions we can now explain why the recipient passive first occurs with French verbs: the datives of French ditransitives were analysed as instances of structural case and this is why they could become the subject of passive clauses. At that time the datives of native verbs were not interpreted as structural datives (they were still associated with the OE four-way-case system and with inherent case) and therefore in ditransitives they could not occur as the subject of passive clauses. In our data this is reflected by the observation that first, the recipient passive occurs with a number of French verbs and only gradually also occurs with highly frequent ditransitive native verbs. In the course of time this difference vanished with the result that today all ditransitive verbs of English build the recipient passive.

The claim that speakers are sensitive to the native/non-native distinction can be found as early as in Pinker's seminal work on the acquisition of argument structure. Pinker (1989: Chapter 2.1) discusses the contrast between native verbs which can build the double object construction (28a/b) and non-native verbs which cannot (29a/b):

(28) a. He gave a painting to the museum.
b. He gave the museum a painting.

(29) a. He donated a painting to the museum.
b. *He donated the museum a painting.

Pinker (1989: 46) proposed that morphophonological properties like the second-syllable stress would allow children to acquire the particularities of 'Latinate' verbs. On this basis, they distinguish two sets of verbs: verbs that are copied from French and native verbs. This difference is reflected in their grammar, including their inventories of argument structures and in the way they acquire them. In a similar vein, recent studies by Ambridge and collaborators (Ambridge et al. 2012) dealing with dative argument-structure overgeneralisation assume a morphophonological constraint which states that speakers are sensitive to the syllable stress patterns of Latinate verbs. In a number of experiments the authors showed that for adults this constraint has a psychological reality whereas for older children (aged 9–10), this is true only to a limited degree (for already existing Latinate verbs but not for newly coined ones). These findings may be correlated with frequency and register: non-native verbs are more likely to be associated with a higher register, are often learned vocabulary, and are less frequent. The lower frequency of these verbs could explain that grammaticality restrictions are acquired only later (according to the pre-emption and entrenchment hypotheses, see Ambridge et al. 2014 for a discussion of these statistical approaches in this context), but it could not explain the rise of the recipient passive with French verbs.

Generally, it seems to be plausible that both phenomena, the rise of the recipient passive and the non-availability of the double-object construction with non-native verbs can be explained by a morphophonological constraint along the lines of Pinker and Ambridge and collaborators. However, this is only part of the story, and the question remains of how the recipient passive was licensed in the first place. We believe that the rise of this phenomenon is of a much more complex nature (than the acquisition of the double-object construction), involving the loss of morphological case in English and the coyping of structural case from French. In the following we will briefly elaborate on the differences we see here.

The psycholinguistic results cited above relate to the double-object construction, i.e., a construction that is and always was part of the speakers' grammar (an existing variant). The RP construction, on the other hand, is a new construction that was not part of the writers' grammar. Therefore the question is not how a restriction was acquired, but how a new construction was licensed. The fact that the new construction almost exclusively appeared with French verbs indicates that language contact is at play. The fact that French has never had a RP construction led us to assume that what was copied was only the particular property of the French dative, i.e. structural case. Structurally case-marked arguments can be licensed as subjects of passives. Thus, a property of the French grammar, introduced by the copying of French verbs, may have induced English speakers to *overcome* a restriction of English grammar, i.e. to structurally mark the dative that was inherently case-marked in the English grammar (morphological dative). This happened first in the context of French-origin verbs and later spread to native verbs at a point where the distinction between accusative and dative was lost in English due to the overall loss of morphological inflection. Thus, the property of French verbs to structurally mark dative case "surfaced" at the level of syntax not only in active constructions by exhibiting dative-PPs but also in passive constructions promoting structural dative to the subject position.

5. Conclusion

In this chapter we have presented data from studies of multi-genre and single-genre corpora which made it possible to find an explanation for the rise of the recipient passive in the history of English. We discussed the rather puzzling finding that the recipient passive was found first with verbs of French origin at the end of the ME period. Interestingly, there was a sharp contrast in frequency between the data from the multi-genre and the single-genre corpus. Since dialect or social status were not significant variables in the occurrence of the recipient passive in the latter corpus we assume that the passive construction was firmly established in the grammar of the writers. We have also shown that it cannot be analysed as a calque from the foreign Model Code (OF) but that it is the result of interpreting the French dative as different from the English 'dative'. More precisely, we assume that OE datives were inherent datives, and therefore excluded from the DAT-NOM alternation. This property was maintained in Middle English, even after the loss of morphological case marking. The datives of French ditransitives were analysed by speakers/writers as structural case. As a result they could become the subject of passive clauses. This analysis was possible because the French case system was not transparent anymore. A further piece of evidence are French impersonal passives with null subjects that could be reanalysed as indirect passives. In this (language contact) situation, the indirect object of native ditransitives was under double pressure: (i) there was almost no case marking left, which made the reanalysis of the first preverbal object as an internal argument possible, (ii) French verbs formed recipient passives. In the end these conditions led to building the recipient passives with all ditransitive verbs regardless of their origin.

If our analysis is on the right track then the rise of the recipient passive is an instance of contact-induced change. Without doubt, the contact situation between

ME and OF was intense enough to trigger syntactic change. Further, it seems to have been of the right 'quality':

> For example, an event such as the Norman Invasion would result in substantial changes in the sorts of sentences heard by language-learners as French speakers learned English and transported some French constructions into their English and native speakers of English imitated the speech of the influenctial French speakers. (Allen 1995: 13)

Further insights into language acquisition, bilingualism and second language learning will certainly lead to a better understanding of contact-induced change, both today and in the history of a language.

Acknowledgments

The project BASICS is funded by the *Deutsche Forschungsgemeinschaft*, DFG grant STE1188/TR555, 2015–2018.

References

Allen, Cynthia. 1995. *Case Marking and Reanalysis: Grammatical Relations from Old to Early Modern English*. Oxford: OUP.

Ambridge, Ben, Pine, Julian M., Rowland, Caroline F. & Chang, Franklin. 2012. The roles of verb semantics, entrenchment, and morphophonology in the retreat from dative argument-structure overgeneralization errors. *Language* 88(1): 45–81. https://doi.org/10.1353/lan.2012.0000

Ambridge, Ben, Pine, Julian M., Rowland, Caroline F., Freudenthal, Daniel & Chang, Franklin. 2014. Avoiding dative overgeneralisation errors: Semantics, statistics or both? *Language, Cognition and Neuroscience* 29(2): 218–243. https://doi.org/10.1080/01690965.2012.738300

Buridant, Claude. 2000. *Grammaire nouvelle de l'ancien français*. Paris: Sedes.

Chomsky, Noam. 1981. *Lectures on Government and Binding Studies in Generative Grammar*. Dordrecht: Foris.

Denison, David. 1993. *English Historical Syntax Longman Linguistics Library*. Harlow: Pearson Longman.

Eisenbeiss, Sonia, Narasimhan, Bhuvana & Voeikova, Maria. 2009. The acquisition of case. In *The Oxford Handbook of Case*, Andrej Malchukov & Andrew Spencer (eds), 369–383. Oxford: OUP.

Estival, Dominique. 1989. A diachronic study of the English passive. *Diachronica* 6(1): 23–54. https://doi.org/10.1075/dia.6.1.03est

Fischer, Susann. 2010. *Word-order Change as a Source of Grammaticalisation* [Linguistik Aktuell/Linguistics Today 157]. Amsterdam: John Benjamins. https://doi.org/10.1075/la.157

Grevisse, Maurice. 2011. *Le bon usage. Grammaire française, 15th edn*. Paris-Gembloux: Duculot.

Heine, Bernd & Kuteva, Tania (eds). 2005. *Language Contact and Grammatical Change*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511614132

Heine, Bernd & Kuteva, Tania. 2008. Constraints on contact-induced linguistic change. *Journal of Language Contact – THEMA* 2: 57–90. https://doi.org/10.1163/000000008792525363

Ingham, Richard. 2012a. Syntaxe et valeur discursive de la construction et VS en anglo-normand par rapport au français du continent. In Actes du 3e Congrès Mondial de Linguistique Française (CMLF), Lyon, 4–7 Juillet 2012, 177–186. Paris: Institut de Linguistique française. https://doi.org/10.1051/shsconf/20120100259

Ingham, Richard. 2012b. *The Transmission of Anglo-Norman: Language History and Language Acquisition* [Language Faculty and Beyond 9]. Amsterdam: John Benjamins. https://doi.org/10.1075/lfab.9

Jensen, Frede. 1990. *Old French and Comparative Gallo-Romance Syntax* [Zeitschrift für Romanische Philologie 2232]. Tübingen: Niemeyer. https://doi.org/10.1515/9783110938166

Johanson, Lars. 2002. Contact-induced change in a code-copying framework. In *Language Change: The Interplay of Internal, External and Extra-linguistic Factors*, Mari C. Jones & Edith Esch (eds), 285–313. Berlin: De Gruyter. https://doi.org/10.1515/9783110892598.285

Johanson, Lars. 2009. Copying case markers and case functions. In *The Oxford Handbook of Case*, Andrej Malchukov & Andrew Spencer (eds), 494–502. Oxford: OUP.

Kroch, Anthony & Taylor, Ann (eds). 2000. *The Penn-Helsinki Parsed Corpus of Middle English (PPCME2)*, 2nd edn. Philadelphia PA: University of Pennsylvania.

Levin, Beth. 1993. *English Verb Classes and Alternations. A Preliminary Investigation*. Chicago IL: University of Chicago Press.

Levin, Beth & Rappaport Hovav, Malka. 2005. *Argument Realization Research Surveys in Linguistics*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511610479

Los, Bettelou. 2009. The consequences of the loss of verb-second in English: Information structure and syntax in interaction. *English Language and Linguistics* 13: 97–125. https://doi.org/10.1017/S1360674308002876

McFadden, Thomas. 2002. The rise of the to-dative in Middle English. In *Syntactic Effects of Morphological Change*, David Lightfoot (ed.). Oxford: OUP. https://doi.org/10.1093/acprof:oso/9780199250691.003.0006

Mitchell, Bruce. 1985. *Old English Syntax*. Oxford: Clarendon. https://doi.org/10.1093/acprof:oso/9780198119357.001.0001

Pinker, Steven. 1989. *Learnability and Cognition. The Acquisition of Argument Structure*. Cambridge, MA: The MIT Press.

Prévost, Sophie & Stein, Achim (eds). 2013. *Syntactic Reference Corpus of Medieval French (SRCMF)*. Lyon/Stuttgart: ENS de Lyon; Lattice, Paris; Universität Stuttgart. <http://srcmf.org>

Scherger, Anna-Lena. 2015. *Schnittstelle zwischen Mehrsprachigkeit und Sprachentwicklungsstörung: Kasuserwerb deutsch-italienischer Kinder mit spezifischer Sprachentwicklungsstörung* [Schriftenreihe Philologia 200]. Hamburg: Kovač.

Schmitz, Katrin. 2006. Indirect objects and dative case in monolingual German and bilingual German/Romance language acquisition. In *Datives and Other Cases* [Studies in Language Companion Series 75], Daniel Hole, Andre Meinunger & Werner Abraham (eds), 239–268. Amsterdam: John Benjamins. https://doi.org/10.1075/slcs.75.11sch

Seoane, Elena. 2006. Information structure and word order change: The passive as an information-rearranging strategy in the history of English. In *The Handbook of the History of English*, Ans van Kemenade & Bettelou Los (eds), 360–391. Oxford: Blackwell. https://doi.org/10.1002/9780470757048.ch15

Taylor, Ann, Nurmi, Arja, Warner, Anthony, Pintzuk, Susan & Nevalainen, Terttu (eds). 2006. *Parsed Corpus of Early English Correspondence (PCEEC)*. York & Helsinki: Universities of York and Helsinki.

Trips, Carola & Stein, Achim. Forthcoming. Contact-induced changes in the argument structure of Middle English verbs on the model of Old French. *Journal of Language Contact*. Special Issue *Valency and Transitivity in Contact*, Eitan Grossman, Ilja Serzants & Alena Witzlack-Makarevich (eds).

Troberg, Michelle. 2008. Dynamic Two-place Indirect Verbs in French: A Synchronic and Diachronic Study in Variation and Change of Valence. PhD dissertation, University of Toronto. <https://tspace.library.utoronto.ca/bitstream/1807/17269/1/Troberg_Michelle_A_200811_PhD_thesis.pdf>

Zaenen, Annie & Maling, Joan. 1990. Unaccusative, passive and quirky case. In *Modern Icelandic Syntax*, Annie Zaenen & Joan Maling (eds), 137–152. New York NY: Academic Press.

Some methodological issues in the corpus-based study of morphosyntactic variation

The case of Old Spanish possessives

Andrés Enrique-Arias
University of the Balearic Islands

This paper has two main objectives: to point out some of the challenges surrounding the study of morphosyntactic variation and change through historical corpora, and to present some methodological alternatives to help alleviate these problems using the specific case study of the expression of possession in Old Spanish. The study highlights some methodological issues such as the definition of the context of occurrence of morphosyntactic variables and the problem of comparability of texts in diachronic corpora. In order to showcase some methodological alternatives the paper presents the results of a study on Old Spanish possessives that uses data from a parallel corpus of medieval Bible translations and a multinomial logistic regression analysis.

1. Introduction

This paper has two main objectives: to point out some of the challenges surrounding the study of morphosyntactic variation and change through historical corpora, and to present some methodological alternatives to help alleviate these problems. To attain these goals, the paper provides a critical discussion of a number of issues in the study of morphosyntactic variation using the specific case study of the expression of possession in Old Spanish. I begin by highlighting some of the methodological concerns which arise in the corpus-based study of morphosyntactic variation and change. Next, I present the expression of possession in Old Spanish as an example of variation phenomena subjected to a complex set of structural and contextual factors. Finally, in order to showcase some methodological alternatives, I present the results of a study on Old Spanish possessives that uses data from a parallel corpus of medieval Bible translations and a multinomial logistic regression analysis.

https://doi.org/10.1075/scl.85.12enr

2. Methodological issues in the study of morphosyntactic variation

Most theoretical models of historical linguistics conceive linguistic change as a three-phase process comprising: (a) a stage previous to the change, (b) a moment when a new structure replaces the older one and thus the change comes to completion, and (c) an intermediate period in which the original structure coexists with the innovative one that will eventually replace it. This intermediate period is the most interesting one for the researcher, as the examination of the factors that favor the encroachment of the innovative variant at the expense of the older one will give us the clues to understand the causes, chronology and diffusion channels of the change. Accordingly, searching through text databases to locate occurrences of the forms under study and then applying quantitative techniques – mainly coocurrence analyses of those linguistic variants that compete in the same contexts of occurrence – have become the essential procedures in corpus-based studies of language change. As Joseph (2008: 182) points out, two methodological techniques that support variationist studies have afforded some of the latest advances in diachronic linguistics: the development of electronic corpora and analytical tools have revolutionized data mining as it is now possible to search through millions of words of text in a fraction of a second, and the increased amount of data has facilitated the application of statistical methods developed in variationist sociolinguistics.

In most cases, linguists who embark on diachronic research will use what can be called *conventional corpora*. These 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 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 context and basic background information on the source text, such as title, author and date of composition. The advantages afforded by large conventional corpora are obvious, as they make it possible to search through multimillion-word databases in a fraction of a second. When corpora are lemmatized and richly annotated, it is possible to further observe subtle relationships between different elements, such as parts of speech, morphological markers, related vocabulary, and collocations, in a way that is virtually impossible with older, manual methods.

If, as I have mentioned, linguistic change results from changes in the distribution of alternative expressions with the same meaning and function, the linguist needs to carry out a careful quantitative description of the development of the constraints on the selection of one variant over the other. In carrying out their analyses, therefore, 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 to each other, and thus lend themselves to comparison. Despite the aforementioned methodological advances provided by corpora, the corpus-based study of diachronic variation still faces a number of challenges that have to do with the notion of comparability, whether it be comparability of texts, of contexts of occurrence or even of the very linguistic variants under study.

2.1 The problem of the comparability of texts

Any diachronic investigation confronts sample-related problems concerning which texts to compare. Is it methodologically sound to compare data extracted from diverse text types? For instance, a common procedure in historical studies of the evolution of Spanish morphosyntactic structures (cf. Lapesa 2000 [1970], or Company Company 2009, among many others) involves using textual corpora that contain a mix of text types, such as epic poetry (e.g. the 12th-century *Cantar de Mio Cid*), historiographical texts (e.g. the 13th-century *General estoria*), or a play (e.g. the 15th-century *Celestina*). Even though in all cases it can be argued that we are dealing with narrative discourse, these are works with very different textual conventions, and in which the distribution of narration, description and dialogue will differ considerably. An added difficulty involves identifying and controlling all the dimensions that potentially condition variation in older texts, such as text type and genre, poeticality, orality, dialect, and writer demographics. The reason is that historical texts – and more so medieval ones – often come down to us devoid of information on author and intended readership, as well as their social and geographic dialects or precise date of composition. As a consequence, differences that we may attribute to the structural evolution of the language may actually stem from geographical differences or the deliberate use of archaisms in certain registers. In sum, since we cannot control the factors that condition variation in each individual text, we face the risk of a partial or misleading analysis.

The methodological challenge of compiling a representative yet balanced historical corpus can be characterized in terms of a *comparability paradox* (Enrique-Arias 2012a: 97): a historical corpus has to be *diverse* because, assuming that we want to study variation of some sort, it must contain texts that represent different periods, genres, or dialects. At the same time, this corpus must be *uniform* in that the distribution of content type, genres, or dialects along the different chronological sections in the corpus must be as similar as possible so they can be compared. In quantitative studies of syntactic change, we therefore face sample-related problems concerning which texts to compare. Even if we restrict our sample to, let us say, narrative texts, we can never be completely sure whether, for each period represented in the corpus, we are characterizing states of language rather than mere text types.

2.2 The problem of the comparability of contexts of occurrence

Once a suitable textual corpus has been established, researchers are expected to count all places where the variable under study occurs and – crucially – all places where the variable could have occurred but did not, so as to be able to quantify the rate of occurrence for each variant (c.f. Labov 1982: 30 and Tagliamonte 2006: 72 for discussion of these methods). That is, the analyst must start by determining the set of environments in which the variable could possibly occur, i.e., the so-called *envelope of variation*. In synchronic variationist linguistics, this definition of the variable context is usually carried out by a combination of qualitative analysis and introspection. However, neither of these methods is satisfactory in historical linguistics because (a) we can never be sure that we have identified all of the constructions with a certain function from the analysis of isolated examples, and (b) direct introspection is not available for historical data. Second, variationist approaches to syntactic change face the problem of the comparability of contexts: in order to study diachronic change, we have to make sure that the individual examples that we draw from texts from different periods are indeed in a relation of equivalence to each other, a necessary condition for their comparison. In order to achieve this, it would be necessary to locate and examine a large number of occurrences of the *same* linguistic structure in versions that were produced at different time periods. But this is not easy to do with conventional corpora.

2.3 The problem of the comparability of variants of the same variable

In addition to the problems concerning the comparability of texts and contexts of occurrence, we face the problem of furnishing an exhaustive list of comparable variants. In corpus-based diachronic studies linguists carry out queries for the forms that are relevant for the research question they are investigating. This means that it is necessary to identify beforehand, via historical grammars, dictionaries, or previous studies, the possible expression units for the structure 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 research with conventional corpora typically requires searching for explicit markers; unless we have access to a richly annotated corpus, it is impossible to identify all the occurrences of linguistic phenomena that may be expressed in a great variety of ways, including zero-marking. 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 it 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.

3. Parallel texts versus conventional corpora

Researchers have long been aware of the comparability problems set forth in the previous sections. One response has been the development and use of *parallel corpora*, which today constitutes a well-established practice within both corpus linguistics (McEnery & Xiao 2007) and variationist sociolinguistics (Tagliamonte 2012: 162). Parallel corpora are collections 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 correspondences in the other languages. One such corpus, which I will be using to exemplify some of the features of parallel texts, is the *Biblia Medieval* corpus (Enrique-Arias & Pueyo Mena 2008–2017), which became available in 2009. Containing over 5 million words, *Biblia Medieval* (henceforth BM) is a freely accessible online tool that enables linguists to consult and compare the existing medieval Spanish versions of the Bible side-by-side next to their Hebrew or Latin sources. When the user enters a query for any of the parallel versions in the BM corpus, whether it is one of the 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.[1]

In what follows I will demonstrate the advantages of a parallel corpus such as *Biblia Medieval* in the study of morphosyntactic variation and change through an analysis of the expression of possession in Old Spanish.

3.1 The problem of the comparability of texts

Most research on Old Spanish possessives has focused on the variation in the use of the definite article preceding the possessive marker (*la mi casa* 'the my house', henceforth ART+POSS), a structure that is absent from present-day standard Spanish, as opposed to possessive alone (*mi casa* 'my house', henceforth POSS). Some authors have observed that ART+POSS is a structure that emphasizes possession and thus it is used with stylistic functions such as expressivity, solemnity,

[1] For a detailed description of the *Biblia Medieval* corpus refer to the project webpage at <www.bibliamedieval.es>.

poeticality, or reverence (Eberenz 2000: 265–319; Lapesa 2000 [1970]). Other authors have focused on the influence of structural factors in the distribution of the two variants; for instance, Wanner (2005: 39–40) pointed out that the first and second person, or singular possessors, as well as possessive structures embedded in prepositional phrases, favor the use of the ART+POSS construction. In addition, syntax-discourse factors have been taken into consideration: Company Company (2006) claimed that ART+POSS is relatively more frequent when the possessor or the possessed referent has been mentioned in the previous discourse (a property she calls *activation*) or when the possessed referent has a high degree of accessibility by virtue of a relationship of inherent possession. The appearance of the article before the possessive is thus conditioned by a considerable number of structural and contextual factors, as summarized in (1) (Rosemeyer & Enrique-Arias 2016):

(1) Environments that favor ART+POSS in Old Spanish[2]
  a. Stylistic factors
    – lyrical texts > narrative texts
    – direct speech > nondirect speech
    – possessor is a superior entity (God, the king) > normal entity
  b. Features of the possessor:
    – 1st and 2nd person > 3rd person
    – singular > plural
    – inanimate > animate
  c. Features of the possessed entity
    – inherent > noninherent
  d. Syntactic function of the NP
    – subject > other functions
  e. Discourse factors
    – activated entities > non activated

This means that the content of the texts (i.e. whether they include more direct speech in the form of dialogues, or the distribution of participant entities), and other features such as their degree of formality, the appearance of distinguished characters, such as God, saints, or kings, will greatly influence the distribution of these two variables. As a result, when comparing the percentage of article plus possessive in historical texts, it is rather complicated to control for all the possible factors that may be conditioning the variation observed in each text. Using a sample as uniform as possible will help to control for these factors; thus, using translation equivalents will afford advantages in this respect.

[2] The vector sign (>) represents that the category to the left is expressed with ART+POSS with greater frequency than the one to the right.

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 appropriate 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.

3.2 The problem of the comparability of contexts of occurrence

Direct comparability of concrete examples across different historical periods is an advantage of the parallel text method. Whereas 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, since 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 a phenomenon such as the expression of possession that exhibits covariation with a complex set of structural and contextual factors. 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 *Joel* 2:18 in the BM corpus in (2), which in the JPS Bible is translated as "Then the Lord was roused on behalf of his land and had compassion upon his people" (cf. Jewish Publication Society 1985). Here we get a good number of occurrences of possessive structures (the Spanish equivalents of Hebrew *ʾarṣō* 'his land' and *ʿammō* 'his people') embedded in very similar syntactic environments: [3]

(2) Translations of Joel 2:18
[Hebrew] wa-yəqannēʾ YHWH lə-ʾarṣō wa-yaḥmōl ʿal-ʿammō
[Vulgate] zelatus est Dominus terram suam et pepercit populo suo
[Fazienda] E receló el Señor por su tierra e ovo piedat de so pueblo
[E6] Receló Dios sobre su tierra e perdonó al so pueblo
[GE]Celó el Señor la su tierra y perdonó al su pueblo

[3] All passages quoted are from the *Biblia Medieval* corpus. I follow the standard practice of quoting Old Spanish biblical manuscripts from the library in the Escorial Monastery by using the letter E plus the final digit in the signature (thus Escorial I.i.6 is quoted as E6, Escorial I.i.4 is E4 and so forth). In addition we include the *Fazienda de Ultramar*, University of Salamanca ms. 1997 (quoted as *Fazienda*), the *General estoria* (*GE*), the *Santillana Bible*, Madrid, Biblioteca Nacional ms. 10288 (Sant.), and the *Arragel Bible*, Madrid, Liria Palace (Arr.). 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, see <www.bibliamedieval.es>.

[E3] E ovo embidia el Señor por su tierra e piadó sobre su pueblo
[E5] E ovo zelo el Señor por su tierra e apiadó sobre el su pueblo
[Santillana] E celó el Señor a su tierra e apiadó sobre su pueblo
[Arragel] E zeló el Señor por la su tierra e piadat del su pueblo ovo
'Then the Lord was roused on behalf of his land and had compassion upon his people.'

The parallel corpus allows us to observe structures with the same content in texts ranging from ca. 1200 to ca. 1450. If we look, for instance, at the occurrences of *(la) su tierra* 'his land' we can see that they are all inserted in identical contexts of occurrence and thus exhibit the same characteristics concerning contextual factors (narrative text, non direct discourse), features of the possessor, which is God (superior entity, third person, singular, animate), features of the possessed entity (non inherent), and same syntactic function of the NP containing the possessive structure (non subject / oblique). In comparing examples like the ones in (2), we can abstract away from the influence of contextual properties and focus instead on the diachronic evolution of the structural type under investigation.

3.3 The problem of the comparability of variants of the same variable

Linguistic structure is accessed in different ways by researchers depending on the material used. 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 – what is commonly known as *concordances* – from which its meaning and function can be observed. In this fashion the analysis proceeds from form to function, which carries a clear disadvantage: 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 two basic advantages.

The first one is the heuristic function of parallel texts, which has no equivalent in other data sources. For instance, if we want to use a conventional corpus to investigate the historical evolution of the elements used to express nonpredicative possession, first we need to consult reference materials and compile a list of elements that can express this function to include in our queries (i.e. possessive adjectives *mi* 'my', *tu* 'your', *su* 'his/her/their', etc. and genitive phrase with *de* followed by a third person pronoun: *de él* 'of him', *de ella* 'of her', etc.). Then we must 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 a parallel corpus like BM we can extract the passages that contain these elements by searching for all the relevant forms in the Latin original (*meus*, *mea*, *meum*, *eius*, *eorum*, etc.) and even any of the relevant words in any of the Spanish texts (*mi*, *tu*, *su*, etc.), and then observe the forms that are used in the same context and with the same functions in the parallel versions. This perspective, from particular textually embedded contents to form, facilitates the observation of elements that otherwise would have been overlooked. As a concrete example, consider the translations of Maccabees 1 5:5 in (3).

(3) *et incendit turres eorum*
'and he burned their towers' (MAC1 5:5)

a.	*e*	*quem-ó=les*	*las*	*torres*		
	and	burn-PST.PFV.3.SG=PRO.DAT	DET.F.PL	towers		(E6)
b.	*e*	*Puso*	*fuego*	*a*	*las*	*torres*
	and	put.PST.PFV.3.SG	Fire	to	DET.F.PL	towers (GE)
c.	*e*	*Encendió*	*las*	*torres*	*d'=ellos*	
	and	ignite.PST.PFV.3.SG	DET.F.PL	towers	of=them	(E4)

By looking up the Old Spanish equivalents of *turres eorum* 'their towers' we can observe without limitations what linguistic structures are used by the translators to convey the functions expressed by this phrase in the Latin model. Notice that the results are not restricted to explicit possessive markers: we find a dative pronoun and a noun phrase (NP) with a definite article in (3a), an NP with a definite article and no explicit possessive marker in (3b), and a genitive phrase ('the towers of them') in (3c).

This is, I would say, the other main advantage of the function-to-form perspective of the parallel text methodology: the possibility of searching for a wider inventory of ways of expressing a linguistic function. This feature 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. When using a conventional corpus it may be relatively easy to extract the examples that contain forms of the possessive adjective (*mi casa, su casa*, *la su casa*, *la casa suya*), but it is more complicated to search for other means of expressing possession. Locating the sequence with preposition *de* 'of' followed by a pronoun (i.e. *la casa de él* 'the house of him') will be problematic in a corpus that has non-normalized spelling and diacritics, which is often the case for medieval texts. Sequences such as *de el* and *del* will include many cases of the preposition *de* with definite article *el* 'of the', which we are not concerned with, mixed with genuine possessive examples of the genitive phrase involving a personal pronoun: *de él*, *d'él* 'of him'. At the same time, searching for those cases in which possession is expressed with a dative pronoun will be very arduous, as we will have to find all cases of *le*, *les* with their formal variants (*li*, apocopated *-l*, *ge*, non-reflexive dative *se*) and then, among the thousands of forms that we may collect, identify the few that do have a possessive interpretation. Finally, it will be virtually impossible to automatically extract those cases in which possession is expressed with just a definite article or with a bare noun.

4. New insights in the study of possession in Old Spanish

To demonstrate how a parallel corpus can afford new insights into the study of morphosyntactic phenomena, an investigation of the expression of nonpredicative possession in Old Spanish will now be presented. This work uses data from the aforementioned BM corpus in combination with a multinomial logistic regression analysis (for a complete discussion of data and methodology see Rosemeyer & Enrique-Arias 2016).

As we have already mentioned, research on Old Spanish possessives generally focuses on the variation in the use of the definite article preceding the possessive marker (ART+POSS) as opposed to possessive alone (POSS). There are, however, other constructions that also serve to express possession in Old Spanish, such as a genitive phrase with a personal pronoun (*la casa de él* 'the house of him', henceforth GEN), the *strong possessive adjective construction*, in which the possessive is postposed (*la casa suya* 'the house his'), or even a simple *determiner + noun construction* (ART/BARE) as in *levantó la mano* 'he raised the [=his] hand' or a dative pronoun as in *le cortó la cabeza* (3.SG=PRO.DAT CUT.PST.PFV.3.SG the head, 'he cut his head'). All of these structures can appear in contexts similar to those of the ART+POSS and POSS structures, and their appearance also correlates with a complex set of structural and contextual factors such as lyrical register, ambiguity of reference, or cognitive prominence (Company Company 1994; Eberenz 2000: 299; Enrique-Arias 2012b: 827–828).

Of course, there are practical reasons why previous studies tend to limit themselves to looking at the distribution of one of these variants relative to another one. First, this is because in working with conventional corpora there is no way to verify that all of the aforementioned structures are possible in a given context. Second, it is rather impractical to locate all instances of those structures that do not use an explicit possessive marker. It is not feasible to search for definite articles or dative pronouns and then check their context to see whether they have a possessive interpretation. But the standard procedure of looking at the distribution of just two variants amounts to rendering an incomplete analysis as it presupposes that speakers had only two choices to express a relation of possession when in fact they could use any of the possessive expressions mentioned above in exactly the same context of occurrence. Because in our method we look at the Old Spanish translational equivalents of the possessive structures in the source text, we are able to incorporate a wider inventory of expression units for nonpredicative possession rather than reducing this constructional network to a binary opposition between two members of that network.

We began the data collection procedure by entering the search string in (4) in the Latin version of certain Bible passages. In order to obtain language samples that reflect register variation, we selected three narrative passages: the episode of Daniel in the lions' den in *Daniel* 1–6, the story of Samson in *Judges* 13–16, and the story of David and Goliath in *Samuel 1* 17; and two lyrical passages, the entire *Song of Solomon* and the book of *Lamentations*. This query rendered a total of 905 tokens of possessive expressions in the Latin Vulgate.

(4) *meus* | *mea* | *meum* | *meam* | *mei* | *meae* | *meo* | *meos* | *meas* | *meis* | *meorum* | *tuus* | *tua* | *tuum* | *tui* | *tuae* | *tuorum* | *tuarum* | *tuo* | *tuis* | *tuam* | *tuos* | *tuas* | *tue* | *suus* | *sua* | *suum* | *sui* | *suae* | *suorum* | *suarum* | *suo* | *suis* | *suam* | *suos* | *suas* | *sue* | *nostr** | *vestr** | *eius* | *eorum* | *earum* | *illius* | *illorum* | *illarum* | *ipsius*

Then we selected all the Spanish passages that corresponded to the Latin possessives in six Old Spanish Bible manuscripts from the 13th and 15th centuries, namely the E6/8 and the *General estoria* (13th century), and the E3, E5/7, *Santillana*, and *Arragel* (15th century).[4] The few cases in which the possessive pronoun was not translated using a nominal construction in one of the translations, but paraphrased in a different way, were excluded from the analysis.[5] Also we did not consider the strong possessive adjective construction (*la casa suya* 'the house his') because the diffusion of this structure takes place during the transition to Modern Spanish and thus there were only three examples in the entire corpus. This yielded a total of 4,800 tokens. Table 1 illustrates the overall usage frequency and the diachronic development of the four types of possessive constructions in the corpus.

In both the 13th and the 15th centuries, the possessive constructions with the highest usage frequency are the POSS construction and the ART+POSS construction. They are almost evenly distributed in the 13th century, with a relative usage frequency of 44 percent (POSS) and 41 percent (ART+POSS). However, the 15th

[4] Additionally, we conducted extensive searches of Spanish possessive markers in the remaining Spanish passages. This way, we added to the database a few more possessive structures that did not correspond to possessive forms in the Latin original.

[5] For instance, in Song of Salomon 3:11, the passage that corresponds to *in die disponsionis illius* in the Latin Vulgate 'the day of his espousal' is translated *dia que fue novio* ('day that he became bridegroom') in E3.

Table 1. Overall usage frequency of possessive constructions in the Bible corpus by century

	13th century	15th century	TOTAL
POSS	44% (648)	59% (1,963)	2,611
ART+POSS	41% (599)	35% (1,148)	1,747
ART/BARE	9% (139)	5% (168)	307
GEN	6% (90)	1% (45)	135
TOTAL	**1,476**	**3,324**	**4,800**

century sees a marked increase of the usage frequency of POSS, to 59 percent, at the expense of the three other constructional types. Because the usage contexts are stable in all versions of the Bible, we can exclude the possibility that this variation is due to differences in these usage contexts.

We undertook a token-by-token annotation of the Spanish data for a dependent variable (the type of possessive construction employed by the translator) and a series of predictor variables selected on the basis of the results of the previous studies summarized in Section 3.1 above. We summarize these predictor variables in (5):

(5) Summary of predictor variables
  a. Features of the possessor:
    – Person
    – Number
    – Animacy
    – Status
    – Activation
  b. Features of the possessed entity:
    – Animacy
    – Activation
    – Inherent possession
  c. Features of syntactic context:
    – Syntactic function
    – Dative
  d. Features of text type:
    – Register
    – Direct speech

Now, our objective of studying the interaction between all four variants faces another problem. Standard statistical tools for logistic regression analysis typically measure the probability of use of a variant relative to another one (binomial analysis). This procedure is inadequate for measuring the probability of use of the four competing variables in our study. In order to solve this problem, we subjected the data to two multinomial logistic regression analyses using the function *multinom()* (Ripley & Venables 2015) in *R* (*R* Development Core Team 2015); one for the data from the 13th century and one for the data from the 15th century. We set the reference level of Type to POSS (possessive adjective cases, as in *su casa* 'his/her house'); the regression analysis thus calculates the probabilities of use of the other types of possessive constructions against the reference level POSS, which, being the most frequent, is taken as the default variant. This method aims to capture the speaker's competence in a more realistic way compared to the binomial analysis, since it reflects the reality that speakers can opt for more than two options when expressing a relation of possession.

The combination of the parallel translation corpus and the multinomial logistic regression analysis makes it possible to model the variation in the expression of possession in medieval Castilian, taking into account the interaction of the four variants considered according to twelve explanatory factors (see Appendix I for a summary of results). For each of the contexts studied, the analysis indicates which structure reaches a level of probability high enough to constitute a competitor of the possessive adjective without article (POSS). Our analysis confirms the relevance of factors already considered in the literature. ART+POSS is related to agentive features of the possessor. It is more frequent when the possessor is animate, singular, 1st and 2nd person, and when its syntactic function is that of a subject. Likewise ART+POSS serves a stylistic function and so it is more frequent in the lyrical genre, or when expressing reverence (when the possessor is God). In other words, the use of ART+POSS appears to be most likely in those contexts in which the reference of the possessor is less ambiguous; that is, this construction is not used to disambiguate a reference, but rather to express the emphatic functions of expressivity, solemnity, poeticality or reverence described by Eberenz (2000) and Lapesa (2000 [1970]), among others. In other contexts of use, particularly when the possessor is inanimate, when the clause contains a dative pronoun or when the possessed entity is classified as inherent possession, the structure that increases its probability to rival with that of the possessive alone (POSS) is the noun phrase without explicit possessive marker (ART/BARE).

As far as the contrast between the 13th and the 15th century data, the most salient difference is the decrease in the frequency of the ART+POSS in the 15th century, a development that is in line with the well-known fact that this structure disappears from the standard variety by the end of the Middle Ages. On the other hand, the probabilistic analysis shows that diachronically, the structural type factors lose weight, but the stylistic ones are maintained. For example, in the 15th century, the use of ART+POSS versus POSS is no longer more probable when the possessor is 1st or 2nd person, or when the possessor is singular. But at the same time the probability of use continues or even increases when the possessor is classified as belonging to a superior entity, when the noun phrase containing the possessive functions as vocative, and in lyrical passages and direct discourse versus, respectively, narrative texts and indirect speech. That is to say, the frequency of ART+POSS diminishes in general terms but it is maintained longer in contexts that we can describe as reverential, emphatic or marked stylistically. This observation agrees with recent models of morphosyntactic change that propose that the contexts that are more typical in the use of a construction are affected later by replacement processes, whereas the less typical contexts tend to adopt the new structure first (an effect called *remanence* in Rosemeyer 2014: 89–90). The fact that the article with possessive is increasingly restricted to emphatic uses causes this structure to be interpreted in terms of stylistic factors which in turn would explain why, in certain contexts, it maintains or even increases its frequency, in spite of being a disappearing structure.

5. Summary and conclusions

In the preceding pages I have aimed to show how using parallel corpora goes a long way towards solving a number of issues in the study of morphosyntactic variation in historical texts. Such corpora alleviate the problem of the definition of the variable context because translated texts can be understood as a collection of the translator's linguistic choices among variants. If the same construction in the source language – in this case Latin possessive pronouns – is translated with different constructions in the target language, we can assume that these constructions indeed have a similar function; it is therefore possible to define linguistic variables not on the basis of prior assumptions but on the basis of empirical observation. Likewise, using parallel corpora with a diachronic dimension eliminates the problem of the comparability of contexts. Given that the parallel corpus used in our study comprises translations from the same source text produced in the 13th and 15th centuries, it is possible to compare the expression of the same content in the language of different centuries. By using multinomial logistic regression analysis, we are able to exploit these analytical advantages. In contrast to standard statistical tools, multinomial logistic regression analysis calculates the influence of predictor variables on multiple constructions, rather than a set of two variants. In other words, the analysis is able to consider a wider inventory of Old Spanish nonpredicative possession structures rather than reduce this constructional network to a binary opposition between two members of that network.

Because in a parallel corpus of translation equivalents the contexts of occurrence of the structures under scrutiny do not change over time, it is safe to conclude that the changes in the distribution of the possessive constructions demonstrated in our analysis (such as the successive replacement of ART+POSS with POSS) are not likely to be due to differences regarding genres, registers, styles or the contents of the texts in the corpus; rather, they reflect an actual syntactic change. At the same time, because the Bible contains different text genres, it has been possible to analyze stylistic variation to show, for instance, that the use of ART+POSS is more probable in stylistically marked registers such as lyrical passages or direct speech, and that in the 15th century, stylistic parameters replace structural parameters as important predictors of the opposition between POSS and ART+POSS.

There are, nonetheless, some possible problems associated with the use of biblical translations in linguistic research. Because they are translated texts, they pose the risk of interference from the source language. As sacred texts, they may also exhibit stylistically marked language (i.e. deliberate archaisms). We must keep in mind, however, that the methodological reliability of using translated texts in linguistic research largely depends on the nature of the phenomenon to be studied. Variation between POSS, ART+POSS, and ART/BARE in Spanish can hardly be affected by features of the original texts, as Latin has no articles and Hebrew does not employ them in possessive structures.[6]

While the methodology outlined in this paper does not solve all the problems inherent to working with historical texts, it does enable the analyst to gain new insights into the historical evolution of structural phenomena in a manner that is not possible with conventional methods.

Acknowledgments

This research has benefitted from a grant from the Spanish Ministry of Innovation reference FFI2014-59135-R (AEI/FEDER, UE) and a grant from the BBVA Foundation for the project "Complete electronic edition of Medieval and Renaissance Spanish Bibles" funded within the program for Scientific Research Teams / Digital Humanities 2016. I am grateful to Malte Rosemeyer for carrying out the statistical analyses presented in this paper and to Francisco J. Pueyo for the development of the corpus databases and analytical tools used for data mining.

[6] Most Spanish medieval texts in current corpora come from translations (from Latin, Arabic, Hebrew, French, etc.) or are subject to the strong Latinate influence that is characteristic of 15th century Spanish writing; moreover, non-translated secular texts such as legal documents, or literature, especially poetry, may be highly artificial as well. In sum, biblical texts are not necessarily worse sources of data relative to other medieval text types. For detailed discussions of the methodological soundness of biblical texts as data sources in linguistic research see Kaiser (2005), de Vries (2007) or Enrique-Arias (2008, 2009, 2012b, 2013).

References

Company Company, Concepción. 1994. Semántica y sintaxis de los posesivos duplicados en el español de los siglos XV y XVI. *Romance Philology* 48(2): 111–135.

Company Company, Concepción. 2006. Persistencia referencial, accesibilidad y tópico: La semántica de la construcción artículo + posesivo + sustantivo en el español medieval. *Revista de Filologia Española* 86(1): 65–103. https://doi.org/10.3989/rfe.2006.v86.i1.3

Company Company, Concepción. 2009. Artículo + posesivo + sustantivo y estructuras afines. In *Sintaxis histórica de la lengua española. Segunda parte: La frase nominal*, Concepción Company Company (ed.), 759–880. Mexico City: Fondo de Cultura Económica and Universidad Nacional Autónoma de México.

Eberenz, Rolf. 2000. *El español en el otoño de la Edad Media. Sobre el artículo y los pronombres*. Madrid: Gredos.

Enrique-Arias, Andrés. 2008. Biblias romanceadas e historia de la lengua. In *Actas del VII Congreso Internacional de Historia de la Lengua Española*, Concepción Company Company & José G. Moreno de Alba (eds), 1781–1794. Madrid: Arco Libros.

Enrique-Arias, Andrés. 2009. Ventajas e inconvenientes del uso de Biblia Medieval (un corpus paralelo y alineado de textos bíblicos) para la investigación en lingüística histórica del español. In *Diacronía de las lenguas iberorrománicas: Nuevas aportaciones desde la lingüística de corpus*, Andrés Enrique-Arias (ed.), 269–283. Frankfurt: Vervuert; Madrid: Iberoamericana.

Enrique-Arias, Andrés. 2012a. Dos problemas en el uso de corpus diacrónicos del español: Perspectiva y comparabilidad. *Scriptum Digital* 1: 85–106.

Enrique-Arias, Andrés. 2012b. *Lingua eorum—la lengua d'ellos*: Sobre la suerte de un calco sintáctico del latín en la historia del español. *Bulletin of Hispanic Studies* 89: 813–829. https://doi.org/10.3828/bhs.2012.61

Enrique-Arias, Andrés. 2013. On the usefulness of using parallel texts in diachronic investigations: Insights from a parallel corpus of Spanish medieval Bible translations. In *New Methods in Historical Corpora*, Paul Bennett, Martin Durrell, Silke Scheible & Richard J. Whitt (eds), 105–115. Tübingen: Gunter Narr.

Enrique-Arias, Andrés & Pueyo Mena, F. Javier. 2008–2017. *Biblia Medieval*. <http://www.bibliamedieval.es> (26 November 2014).

Jewish Publication Society. 1985. *Tanakh: The Holy Scriptures. The New JPS Translation According to the Traditional Hebrew Text*. Philadelphia PA: The Jewish Publication Society.

Joseph, Brian. 2008. Historical linguistics in 2008. The state of the art. In *Unity and Diversity of Languages*, Piet van Sterkenburg (ed.), 175–188. Amsterdam: John Benjamins. https://doi.org/10.1075/z.141.16jos

Kaiser, Georg A. 2005. Bibelübersetzungen als Grundlage für empirische Sprachwandeluntersuchungen. In *Romance Corpus Linguistics, II: Corpora and Diachronic Linguistics*, Johannes Kabatek, Claus D. Pusch & Wolfgang Raible (eds), 71–83. Tübingen: Gunter Narr.

Labov, William. 1982. Building on empirical foundations. In *Perspectives on Historical Linguistics* [Current Issues in Linguistic Theory 24], Winfred P. Lehmann & Yakov Malkiel (eds), 17–92. Amsterdam: John Benjamins. https://doi.org/10.1075/cilt.24.06lab

Lapesa, Rafael. 2000 [1970]. Sobre el artículo ante posesivo en castellano antiguo. In *Estudios de morfosintaxis histórica del español*, Rafael Cano Aguilar & María Teresa Echenique Elizondo (eds), 413–435. Madrid: Gredos.

McEnery, Tony, & Xiao, Zhonghua. 2007. Parallel and comparable corpora: The state of play. In *Corpus-based Perspectives in Linguistics*, Yuji Kawaguchi, Toshihiro Takagaki, Nobuo Tomimori & Yoichiro Tsuruga (eds), 131–145. Amsterdam: John Benjamins.. https://doi.org/10.1075/ubli.6.11mce

R Development Core Team. 2015. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. <http://www.R-project.org> (21 September, 2015).

Ripley, Brian & Venables, William. 2015. *nnet: Software for Feed-forward Neural Networks with a Single Hidden Layer, and for Multinomial Log-linear Models* [R package version 7-3-11]. Vienna: R Foundation for Statistical Computing.

Rosemeyer, Malte. 2014. *Auxiliary Selection in Spanish: Gradience, Gradualness, and Conservation* [Studies in Language Companion Series 155]. Amsterdam: John Benjamins. https://doi.org/10.1075/slcs.155

Rosemeyer, Malte & Enrique-Arias, Andrés. 2016. A match made in heaven. Using parallel corpora and multinomial logistic regression to analyze the expression of possession in Old Spanish. *Language Variation and Change* 28(3): 307–334. https://doi.org/10.1017/S0954394516000120

Tagliamonte, Sali. 2006. *Analysing Sociolinguistic Variation*. Cambridge: CUP. https://doi.org/10.1017/CBO9780511801624

Tagliamonte, Sali. 2012. *Variationist Sociolinguistics: Change, Observation, Interpretation*. Malden MA: Wiley-Blackwell.

de Vries, Lourens. 2007. Some remarks on the use of Bible translations as parallel texts in linguistic research. In *Parallel Texts: Using Translational Equivalents in Linguistic Typology. Sprachtypologie und Universalienforschung (STUF)* 60: 95–99. Special issue ed. by Michael Cysow & Bernhard Wälchli.

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

Appendix I

The tables below illustrate the results from the multinomial regression analyses over the 13th and 15th-century data. They give the coefficient calculated for each of the levels of the predictor variables (see (5) above) for each of the four levels of the dependent variable (POSS, ART+POSS, ART/BARE, or GEN). Because the reference level of the dependent variable is set to POSS, the coefficients refer to the probability of use of one of the three other levels (ART+POSS, ART/BARE, or GEN) in comparison to POSS in these specific usage contexts. For instance, the coefficients regarding 1st person morphology in the 13th century data indicate that if the possessive construction has first-person morphology (PS_PERSON = 1st) instead of third-person morphology (PS_PERSON = 3rd), the likelihood of use of ART+POSS in comparison to the use of POSS increases by 1.274, whereas the likelihood of use of ART/BARE decreases by 1.576. Those effects that are statistically significant have been marked with an asterisk. For a full explanation of data and methodology see Tables 4 and 5 in Rosemeyer and Enrique-Arias (2016: 333–334).

Multinomial logistic regression analysis, 13th and 15th century (reference level of the dependent variable = POSS) (Adapted from Rosemeyer and Enrique-Arias 2016)

13th century

Variable	Level	ART+POSS	ART/BARE	GEN
PS_Person	3rd	*Reference level*		
	1st	1.274*	−1.576*	−92.674
	2nd	0.948*	−1.912*	−94.390
PS_Number	Singular	*Reference level*		
	Plural	−0.892*	−0.202	1.172*
PS_Animate	FALSE	*Reference level*		
	TRUE	−0.796*	−0.634	−1.384*
PS_Status	Other	*Reference level*		
	Upperclass	−0.299	−1.227*	−1.320
	God	1.383*	−1.247	−0.475
PS_Activation	FALSE	*Reference level*		
	TRUE	0.055	0.210	−0.224
PD_Animate	FALSE	*Reference level*		
	TRUE	−0.379*	−1.011*	−1.790*
PD_Activation	FALSE	*Reference level*		
	TRUE	−0.139	−0.101	0.273
PD_Inherent	FALSE	*Reference level*		
	TRUE	0.222	0.021	0.840*
Syntactic function	Subject	*Reference level*		
	Object	−0.767*	0.128	−1.469*
	Prep	−1.227*	−0.705*	−2.192*
	Vocative	−2.537*	0.450	−7.155*
	Apposition	−1.738*	−0.774	−224.547*
Dative	FALSE	*Reference level*		
	TRUE	−0.160	2.642*	0.208
Direct speech	FALSE	*Reference level*		
	TRUE	0.590*	0.375	−0.232
Narrative	FALSE	*Reference level*		
	TRUE	−0.885*	1.128*	−1.865*

15th century				
Variable	**Level**	**ART+POSS**	**ART/BARE**	**GEN**
PS_Person	3rd	*Reference level*		
	1st	0.319*	−0.550*	−16.058*
	2nd	0.177	−1.042*	−3.083*
PS_Number	Singular	*Reference level*		
	Plural	−0.011	0.090	0.622
PS_Animate	FALSE	*Reference level*		
	TRUE	−0.469*	−0.556*	−0.069
PS_Status	Other	*Reference level*		
	Upperclass	0.741*	−0.111	−15.756*
	God	0.689*	−0.496	−14.520*
PS_Activation	FALSE	*Reference level*		
	TRUE	0.133	0.217	−0.439
PD_Animate	FALSE	*Reference level*		
	TRUE	−0.216*	−0.227	1.019*
PD_Activation	FALSE	*Reference level*		
	TRUE	−0.144	0.552*	−1.665*
PD_Inherent	FALSE	*Reference level*		
	TRUE	−0.067	0.078	0.715
Fn_Syntax	Subject	*Reference level*		
	Object	−0.145	0.259	−0.738
	Prep	−0.313*	0.403*	0.178
	Vocative	−0.261	1.921*	−3.540*
	Apposition	−1.154*	0.608	−0.087
Dative	FALSE	*Reference level*		
	TRUE	0.207	2.329*	−11.470*
Direct	FALSE	*Reference level*		
	TRUE	0.310*	0.296	1.237*
Narrative	FALSE	*Reference level*		
	TRUE	−0.874*	0.395*	0.427

The interplay between genre variation and syntax in a historical Low German corpus

Melissa Farasyn[1], George Walkden[2], Sheila Watts[3] &
Anne Breitbarth[1]
[1]Ghent University / [2]University of Konstanz / [3]University of Cambridge

In this chapter, we focus on the choice of different genres in the Middle Low German part of the tagged and parsed *Corpus of Historical Low German* and its implications for syntax. We discuss how the inclusion or exclusion of genres has an impact on the study and the discovery of syntactic phenomena in Middle Low German, such as null referential subjects, resumptive pronouns, relative particles and gaps in coordinations. This interplay between genre and syntax also influences parsing decisions. Furthermore, we look at the influence of (sparse) punctuation on (automatically) tagging the corpus itself, and how a closer study of genre-specific syntactic elements contributes to the improvement of the accuracy of automatic classifiers.

1. Introduction

In the last few years, syntactically annotated and parsed corpora, facilitating large-scale comparative and diachronic morphological and syntactic research, have become increasingly important. Such corpora make specifically designed queries reproducible and comparable amongst languages and language stages, due to widely used annotation and parsing standards. The corpus we focus on in this study is the Middle Low German (MLG) subcorpus of the *Corpus of Historical Low German*.[1] The *Corpus of Historical Low German* is a syntactically annotated and parsed corpus, which is currently under construction, and which will be specifically designed to enable further research into the syntax of Old Saxon (OS; also known as Old Low German) and Middle Low German.[2] Although there have

[1] The *Corpus of Historical Low German* (CHLG) is currently funded by the Flemish Hercules foundation/FWO (grant AUGE 13/02 A parsed Corpus of Historical Low German).

[2] HeliPaD, the Old Saxon subpart of the CHLG, is already publicly available at <http://www.chlg.ac.uk/helipad/corpus.html>.

https://doi.org/10.1075/scl.85.13far

been a number of recent studies focusing on the syntax of Middle Low German, leading to interesting insights (e.g. Breitbarth 2014; Farasyn & Breitbarth 2016a; Mähl 2014; Petrova 2012), this is still an under-researched field and worth further exploration. So far, it has been described as having its very own syntactic features, setting it apart from Middle Dutch in the (South)west and Middle High German (MHG) in the Southeast in many aspects. The language, although very important and internationally used, especially during the heyday of the Hanseatic League, never became fully standardized and was never used as a national standard language. From the second half of the 16th century onwards, Early New High German (ENHG) became the language of writing in the area. As a consequence, Low German continued to exist only in the spoken dialects. However, thanks to its erstwhile importance, many written sources have been preserved. The vast majority of MLG texts are secular, ranging from chronicles and laws (city statutes, land property rights, inheritance rights) to charters and administrative texts (correspondences, bills, accounting books). Due to the number of texts, the extent of the area in which the language was spoken, and only partial standardization, compilers of the corpus of MLG need to consider many kinds of variation. Consequently, one of the goals in constructing the corpus is to recognize this variation and the effects it can have on the outcome of linguistic research, and, if possible, to give the researcher using the corpus the possibility to counter or manage these effects by building a corpus balanced for factors influencing variation, including for example the genre of the texts.

In this paper, we focus on the effects that the genre of a text can have on certain syntactic research outcomes, based on preliminary research conducted on the data from the MLG subcorpus of the *Corpus of Historical Low German*. We define genre in this paper as a specific type of text with a specific purpose. A distinction between text types can thus be made on the basis of the purpose of the text, which in the case of the *Corpus of Historical Low German* expresses itself for instance in history writing (chronicles), defining rights and privileges for areas such as trade and heritage (charters), and defining rules for the community (legal texts/city laws).

In order to make the effects of genre in the corpus more tangible, this paper will present a number of case studies on MLG syntax within the generative framework, from a minimalist perspective. These indicate that syntactic phenomena can be strongly influenced by the genre of the text investigated. Before that, we describe the *Corpus of Historical Low German*, its purpose and the texts in the corpus in Section 2. In Section 3, we present the case studies, in which we look at the distribution of relative clauses with the comparative particle *alse*, null pronominal arguments, null resumptive pronouns in relative clauses, and double agreement in V2 clauses with and without inversion of subject and finite verb. In this way we discuss properties of specific genres such as charters, religious texts and chronicles. These case studies show that, in some instances, the syntactic observations can be significantly different in different genres, even leading to false claims about continuity or change from OS to MLG. They therefore demonstrate that the variety of genres in the *Corpus of Historical Low German* is needed to counter these and similar effects. Section 3 starts with an example of how the study of a genre can help to improve the accuracy of an automatic classifier by looking at text-specific discourse markers.

2. A parsed corpus of Middle Low German

The MLG subcorpus of the *Corpus of Historical Low German* will be the first fully annotated and parsed corpus of MLG, covering the whole period in which the language was written (i.e. c. 1250–1600). In order to accomplish this goal, the corpus closely collaborates with the *Referenzkorpus Mittelniederdeutsch/Niederrheinisch* (ReN)[3] in sharing texts, transcription and part-of-speech and morphological mark-up. In contrast to ReN, the *Corpus of Historical Low German* also adds an extra layer of syntactic annotation. An important feature is that the *Corpus of Historical Low German*, like ReN, does not start from editions. All transcriptions are based on the original manuscripts; spelling, word breaks, syntactic structures etc. are thus original and not influenced by editorial decisions. The corpus currently consists of 722,000 words, which equals about 13 prose texts, text collections or books of charters. Based on manually developed gold standards, a tagger automatically assigning Part-of-Speech (POS) tags and additional morphological mark-up have been trained. These taggers use the HiNTS tagset, an adaptation of HiTS (Dipper et al. 2013), a fine-grained tagset especially adapted for Middle Low German, in order to make inter-interoperability with ReN possible. The syntactic annotation layer is currently under development. It will follow the Penn Treebank system, developed for the *Penn Parsed Corpora of Historical English* (Mitchell et al. 1993), to make the corpus comparable with many existing historical corpora using this annotation system, such as for instance the *Icelandic Parsed Historical Corpus* (IcePaHC; Wallenberg et al. 2011), the *Tycho Brahe Parsed Corpus of Historical Portuguese* (Galves & Faria 2010) and the *Parsed Linguistic Atlas of Early Middle English* (LAEME; Truswell et al. 2016). However, as every language has its own specific syntactic properties, some decisions about the exact tags are corpus-specific, and must thus still be made for the *Corpus of Historical Low German*.

[3] <https://vs1.corpora.uni-hamburg.de/ren/>

These decisions include for instance the addition of extended labels such as locative, directional or temporal extensions for adverb phrases. The MLG subpart of the *Corpus of Historical Low German* will follow the guidelines from the HeliPaD, the OS subpart of the corpus, where possible.[4] Middle Low German is the name of a group of related dialects spoken in northern Germany, for which the first written attestations date from the first half of the 13th century. These surface after a period of about 150 years in which only Latin had been written, making it difficult to make statements about continuity between Old Saxon (OS), the predecessor of MLG, and MLG itself. Recent syntactic analyses, e.g. on the distribution of (different types of) referential null subjects in OS and MLG (Walkden 2014; Farasyn & Breitbarth 2016a, respectively) seem to point at some degree of continuity, however, which further underlines the importance of syntactically annotated corpora. MLG can be divided into two main periods (Peters 2003: 437). The first lasts until about 1350; in this period there were many different scribal dialects, each belonging to a separate smaller region and closer to the spoken language. When the Hanseatic League gained importance, i.e. between about 1350 and 1550, these scribal dialects were partly standardized. They incorporated features of the surrounding dialect or adapted to influential chanceries like the one of Lübeck, in order to facilitate interregional and international correspondence (e.g. in the Baltic region and in the East). MLG lost its role as the leading written language in the area after a transition to ENHG between c. 1550 and 1600 (Peters 1973).

The MLG subcorpus of the *Corpus of Historical Low German* is meant to be balanced concerning different parameters that tend to influence variation.[5] The intention is to keep the corpus as representative as possible for the language as a whole. As a first measure to reach this goal, the corpus only consists of non-translated texts. Diachronic variation is covered by including texts from the whole period in which MLG was written. Only texts which are clearly dated are included in the corpus. The corpus also tries to offer a balanced picture of diatopic variation by including localized texts from the main MLG dialectal areas: Westphalian, Eastphalian (including Elbe-Eastphalian) and North Low German, both from the Saxon heartland (*Altland*, lit. 'old land') and the areas colonized from the 11th century onwards (*Neuland*, lit. 'new land'). Texts from areas that do not belong to present-day Germany, such as the Baltic and Low Prussian areas, are not included

[4] <http://www.chlg.ac.uk/helipad/corpus.html>; see Walkden (2016) for an overview.

[5] Because of the sparse attestation of OS – the vast majority being religious epic texts in alliterative verse – the OS part of the *Corpus of Historical Low German* obviously can never be balanced for genre, for instance.

in the corpus.[6] Texts answering all these criteria belong to the key text types in the language: Numerous charters and legal documents have been selected alongside narrative texts including chronicles, religious and medical prose texts. Table 1 gives an overview of the texts that have been included in the corpus so far.

Table 1. Texts included in the MLG subpart of the *Corpus of Historical Low German* so far, by scribal language, genre, period and place of origin

Place	Scribal language	Genre	Period	Name	Number of word tokens
Braunschweig	Eastphalian (Altland)	charters	13th–15th c.	*Urkundenbuch Braunschweig*	ca. 81000
Herford	Westphalian (Altland)	legal texts	1375	*Herforder Rechtsbuch*	16227
Lübeck	North Low Saxon (Neuland)	charters	13th–15th c.	*Urkundenbuch Lübeck*	ca. 179000
Magdeburg	Elbe-Eastphalian (Neuland)	charters	13th–15th c.	*Magdeburger Urkundenbuch*	ca. 39000
Magdeburg	Elbe-Eastphalian (Neuland)	medical prose	1483	*Promptuarium medicinae*	ca. 128000
Münster	Westphalian (Altland)	religious prose	1444	*Spieghel der leyen*	24505
Münster	Westphalian (Altland)	religious prose	1480	*Dat myrren bundeken*	ca. 91000
Münster	Westphalian (Altland)	charters	14th–15th c.	*Urkundenbuch Münster*	ca. 95500
Oldenburg	North Low Saxon (Altland)	charters	14th–15th c.	*Oldenburger Urkunden*	28241
Oldenburg	North Low Saxon (Altland)	legal texts	1336	*Oldenburger Sachsenspiegel*	24377
Rüthen	Westphalian (Altland)	legal texts	3 parts: c. 1300, c. 1350, 1460–1500	*Statuarrecht Rüthen*	6804
Soest	Westphalian (Altland)	legal texts	c. 1367	*Soester Schrae*	8241

[6] They are, however, included in ReN.

The text selection criteria for the corpus are intended to offer a representative picture of the language written from 1250 until 1600. However, the user of the corpus needs to keep in mind that the corpus data represent scribal languages, and not the spoken dialects. In these scribal languages, the writers/scribes did not try to represent the local dialect, and the difference between spoken and written language might well have been considerable (Fedders 1988).

3. Syntactic variation and the role of genre in the corpus

In-depth studies of the syntax of MLG are still only sparsely available, and only a small number of comparative studies on the effect of the text genre on syntactic phenomena in MLG have been performed. Two of these rare examples are Dreessen and Ihden (2015), for the effect of genre on the placement of the verb in subordinate clauses, and Farasyn et al. (2016) on different aspects of MLG syntax in psalms translated from Latin into MLG compared to these syntactic phenomena in authentic MLG text material. In the first case study presented below, we focus on how investigating genre-specific phenomena can inform the construction of the corpus, and the automatic tools developed for this purpose, such as the part-of-speech tagger or the parser. As an example of this, we will specifically focus on the study of discourse particles. With the other case studies, we want to show how the inclusion of different genres in the corpus can lead to new and diverging results when studying syntax based on the corpus.

3.1 Discourse markers

In order to train a high-performing part-of-speech tagger, it is crucial to have rich textual information. A part-of-speech tagger always relies in the first place on a set of custom features related to the token, such as word length, first *n* letters, last *n* letters, capitalization and punctuation for tagging tokens in a natural language. Based on these features and robust machine learning algorithms, the tagger learns how tokens can be divided into different categories: in other words, it learns which labels should be assigned to the token. One of the challenges of constructing automated tools to tag a historical corpus is dealing with (a lack of) punctuation, as it is often hard to see where sentences start or stop without having such information. For the construction of an automatic tagger or parser, it is however highly useful if information regarding coherent chunks of information is already (partly) encoded in the training data. In order to indicate finite clauses in the *Corpus of Historical Low German*, clause boundaries need to be inserted manually in the transcription or in the POS tagging phase of the text. These get included as additional features for training the parser. Larger chunks of information are harder to find. That is the reason why several corpus-specific features on which the POS tagger relies were included in the *Corpus of Historical Low German*.

The POS tagger was trained on gold standard data from three legal texts, which means that they all belong to the same text-genre. In a later stage, the features will be adapted to be more applicable on other genres as well, as first out-of-domain tests have shown that the accuracy of the POS-tagger drops about 10% when applied on *Spieghel der Leyen*, a religious prose text. An outstanding feature of the legal texts on which the tagger was trained is the use of discourse markers. A recurrent property of charters, for example, which are highly formulaic, is that new parts of the text are always introduced with the same word, *vortmer* ('furthermore'), as can be seen in Example (1).

(1) *UOrtmer . js eyner vrowen ere man doyt . wel sey dan nemen eynen anderen man . heuet sey mer kindere dan eyn . so sal sey nemen den derden deyl alles des ghudes . heuet sey nicht mehr dan eyn kint . so nemet sey den haluen deyl*
Furthermore is a.GEN woman.GEN her husband dead want she then take an other husband has she more children than one so will she take the third part all the. GEN good. GEN has she not more than one child so take she the half part
'Furthermore. If a woman's husband died and she wants to take another man, if she has more children than one, then she will take the third part of all goods. If she does not have more than one child, she takes half of it.'
(Soest, *Soester Schrae*, 1367)

This word was thus included as a corpus-specific feature, i.e. as a discourse marker, indicating the start of a new sentence or the start of a new chunk of information. The other corpus-specific features which were added to the corpus were brackets and paratext.[7] Testing all possible features and feature combinations separately to see which one(s) perform(s) better is computationally very intensive. A solution was to work with genetic algorithms, i.e. algorithms that are based on the idea of natural selection, which make it possible to look for optimal features or feature combinations much more efficiently. Koleva et al. (2017) give a more extensive description of how this has been done. When applying genetic algorithms on the

[7] Brackets are either part of the original manuscript or decisions made by the transcribers, according to the guidelines in Barteld et al. 2014. Curly brackets are used to resolve abbreviations, square brackets for text that is unreadable/hard to read. Manuscripts usually contain round brackets to indicate non-specified continuation of the text.

gold standard data,[8] the outcome shows that most features (i.e. bigrams (sequences of two tokens), trigrams (sequences of three tokens) and lowercase) related to the token are consistently needed to obtain a well-performing tagger. Corpus-specific features do play a role in all texts, although to a lesser extent: in all experiments on legal texts, *vortmer* is selected in between 50% and 70% of all cases, which means that the classifier partly relies on the information about *vortmer* in between 50% and 70% of all cases to be able to predict the right POS tag for the token. In *Soester Schrae* and the *Herforder Rechtsbuch*, the word was found in more than 70% of all cases selected.

In later experiments, in which the tagger will be optimized to be applied on out-of-domain texts, i.e. on texts from different cities/periods/genres, further discourse markers indicating new chunks of information can be added as a feature to optimize the tagger. For other genres, the role of other structuring elements that seem to function as chunk-introducing discourse markers in MLG still needs to be investigated more carefully. The word *vnde* ('and') for instance seems to have a more discourse structuring role as introducer of new informative chunks as well, rather than being a conjunction (Farasyn & Breitbarth 2016a). Example (2) shows how *Vn(de) se wolden* clearly is not used as a second conjunct, but rather introduces a new piece of information. This function is also emphasized by the capitalization of the *V*.

(2) *Des quemen des bioscopes ammetlude van Mynden in dat erfhus in eyn gheheghet richte un(de) anclaghede(n) mit eren vorspreken dessulven Johannes herwede unde sin erve, went he des stichtes vulschuldighe eghene man w(er) e un(de) horde in dat ammet to Hul-Horst.* ***Un(de)*** *se wolden ene vorbosmen un(de) vortughen, alze des ammetes recht is*
this.GEN came the bishop.GEN officials from Minden in the inheritance.
house in a limited court and filed.suit with their spokesmen the.same.GEN
John.GEN armour and his inheritance for he the.GEN convent.GEN serf were
and belonged in the authority of Hüllhorst and they wanted him to.claim.
as.serf and testify as the.GEN authority's right is
'Because of that, the officials of the bishop of Minden came in the house of the decedent in a limited court and filed suit with their spokesmen about the armour of this John and his inheritance, as he was (supposedly) a serf of the convent and belonged to the authority of Hüllhorst. **And** they wanted to claim him as a serf, as is the right of the authority'
(Herford, *Herforder Rechtsbuch*, 1375)

[8] The gold standard consists of two texts manually POS-tagged by two annotators, and one by one annotator. In order to reach full inter-annotator agreement on the doubly annotated texts, the double annotations were compared. For some cases, such as idioms or multi-word expressions, there was no tagging consistency between annotators, so consensual decisions were taken.

3.2 Null pronominal arguments

Much syntactic research that has been done on the basis of the *Corpus of Historical Low German* under construction concerns (referential) null pronominal arguments. In this section, we focus on four types of (referential) null arguments or structures containing these arguments, which have turned out to have a connection with the genre of the texts they occur in: referential null subjects, pronominal gaps in *alse*-clauses, (null) resumptives in non-restrictive relative clauses with a first or second person head and pronominal gaps in asymmetric coordinations.

3.2.1 *Referential null subjects*

The presence of referential null subjects (RNS) in MLG is a syntactic phenomenon that is strongly related to text type/genre. A sentence containing an RNS is interpreted as if it contains a referential subject, although this subject is not expressed. Based on results of the distribution of RNS in a corpus of twenty MLG texts, both of the MLG subpart of the *Corpus of Historical Low German* and of ReN, Farasyn and Breitbarth (2016b) show that MLG was a partial null subject language, i.e. that the language allowed null subjects under certain conditions (cf. Walkden 2014; Holmberg 2010). One of these conditions, which is a common property of many partial null subject languages, is the preference of the RNS to occur in the main clause. This is the case in clause-initial topic position (3) as well as in V2 clauses with another topic (4) and in V1 conditional clauses (5) (and other V1 clauses, e.g. interrogatives).[9] RNS occur about three times more often in main clauses (3.3%) than in subordinate clauses (0.9%). Besides clause type, person seems to be an additional important conditioning factor for the occurrence of RNS in MLG, as RNS have a preference to appear in the 2nd and the 3rd person in MLG. This is

[9] We can be fairly certain that the gap precedes the finite verb in (3), as there is evidence that MLG did not have inversion in second conjuncts (following *vnde* 'and'). First, we find examples (both from the same text as (3); *Griseldis*) with overt pronominal (i) and full nominal subjects (ii).

(i) *Vnde sze hudden syck nycht vor de boesze anlaghe des vaders*
and they protect REFL NEG for the evil accusation of.the father
'and they do not protect themselves from the evil accusation of the father'

(ii) *vnde de vrowe slot de dore na to*
and the lady locked the door after shut
'and the lady locked the door shut after [them]'

Second, evidence from double agreement (see Section 3.2.4 below) also indicates that there is no inversion in second conjuncts in MLG.

however a relaxation of the strong preference for 3rd person in predecessor Old Saxon (OS) and related Old High German (OHG).

(3) *Vnde* [pro] *hebbe* *dyne* *kyndere* *beyde* *seer* *wol* *bewaren* *lathen*
and [I] have your children both very well protect let
vnde *nycht* *ghedodeth*
and NEG killed
'And I have let both your children be very well protected and have not killed them' (Hamburg, *Griseldis*, 1502)

(4) *v(m)me* *vns* *to* *verlose(n)* *heuest* [pro] *willen* *anneme(n)*
for us to redeem have [you] want.IPP on-take
vnse *kranch(ei)t* […]
our disease
'in order to redeem us, you have wanted to take on our disease'
(Münster, *Dat myrren bundeken*, 1480)

(5) *heuet* [pro] *ene* *ane* *bůrghe* *ghelaten* *so* *mach* *hey* *dat*
has [he/one] him without bailsman left so may he that
selue *doyn*
himself do
'if he/one left him$_i$ without a bailsman, he$_i$ may do that himself.'
(Soest, *Soester Schrae*, 1367; Farasyn & Breitbarth 2016b)

In addition to the internal linguistic properties of null subjects in MLG, the extra-linguistic factors period (of writing) and scribal language turn out to be strong predictors of the expression of null referential subjects: RNS become more common between 1450 and 1550 and are used more often in the Eastphalian dialect, whereas they are almost absent in earlier texts or texts from the area of Lübeck. The strongest extra-linguistic factor predicting the presence of RNS in MLG, however, is genre/text type, as can be seen in Table 1, in which the results of a multiple logistic regression performed in Rbrul are displayed (cf. Farasyn & Breitbarth 2016b).

Table 2. Influence of the factor genre on the expression of a referential pronominal subject as null (cf. Farasyn & Breitbarth 2016b)

Genre	Log odds	Odds	Factor weight	N	% RNS
chronicle	1.475	0.075	0.814	425	7.53
letter	0.428	0.028	0.605	216	2.78
religious	0.208	0.022	0.552	1249	2.24
literature	0.074	0.020	0.518	1882	1.97
legal	−0.403	0.012	0.400	1709	1.30
charters	−1.782	0.003	0.144	320	0.69

Table 2 shows that, of the six different genres that have been evaluated, RNS in MLG are much more common in chronicles than in any other genre. It is probably the narrative character of chronicles that leads to this text genre displaying a remarkable 7.53% of RNS, whereas on average RNS only make up 2.12% of all pronominal subjects in the corpus (cf. Farasyn & Breitbarth 2016a). The study of genre in relation to the expression of null subjects shows that genre additionally is a very strong predictor of the type of RNS (SpecCP or the position after C) that is found in the text. An example is the Saxon chronicle (which is currently not in the *Corpus of Historical Low German*, but is part of ReN), in which 31 out of 33 are RNS occurring in SpecCP, as can be seen in Example (6).[10] In this example, we first encounter three cases of regular conjunction reduction. The last two gaps seem to be cases of conjunction reduction, but the subjects do not refer to the subject of the preceding referent (i.e. God, in 'he gave her to Adam as his wife'), as the first gap rather refers to Adam ('He was meant to live forever') and the second one to God ('He forbade him [Adam] to eat fruit from a certain tree'). The example therefore also shows MLG's tendency to use discourse antecedents which are introduced more implicitly, as the antecedent is neither structurally parallel nor c-commanding the gap.

(6) *Vnd in der ersten stunde des dages mackede got*$_i$ *Adame*$_j$ *van der erde na synem likenisse vnd* [Ø$_i$] *gaff ome gewalt over fee ouer voggel ouer fissche vnd* [Ø$_i$] *sande one*$_j$ *in dat Paradis dar mackede he Eua van Adames ribbe Jn der dridden stunde des dages die wile dat he*$_j$ *sleyp vnd* [Ø$_i$] *gaff eua adame*$_j$ *to wiue vnd* [pro$_j$] *scholde ewich leuen vnde* [pro$_i$] *vorbot one frucht an eynem bome to eten*

and in the first hour of-the day made god Adam from the earth to his image and gave him power over mammals over birds over fish and sent him in the paradise there made he Eve from Adam's rib in the third hour of-the day the while that he slept and gave Eve Adam to wife and [he] should forever live and [he] forbade him fruit from one tree to eat

'And in the first hour of the day, God created Adam from earth in his image, gave him power over mammals, birds and fish and sent him to paradise. There, he made Eve from Adam's rib in the third hour of the day, while he was asleep, and gave her to Adam as his wife. [He] was meant to live forever and [he] forbade him to eat fruit from a certain tree.'

(*Cronecken der Sassen*, 1492)

[10] See Farasyn & Breitbarth (2016a) for a distinction between two types of RNS in MLG, viz. null topics in SpecCP and null clitics in Wackernagel position. Note that our treatment of this position differs from that of Wackernagel (1892) in assuming it to be a syntactically-defined (rather than prosodically-defined) position high in the clause structure, but below C: see Lenerz (1977), Anagnostopoulou (2008) and many others for this interpretation.

The fact that narrative texts display a higher amount of RNS in SpecCP is probably due to a form of emerging discourse drop, as this is the position in which topics occur. This is highly visible in texts with a narrative character like these chronicles. This case study consequently underlines that the inclusion of different genres is highly important for the study of null subjects: a corpus entirely based on the most common type of texts in MLG, i.e. charters and legal documents, would definitely create a misleading image of the distribution of null subjects in this language. Thanks to the study of narrative texts, however, we can conclude that the type of null subjects in SpecCP show that MLG is in the transition to a topic-drop language of the modern V2-Germanic type, although it displays a certain continuity with Old Saxon in its preference for clause type and person.

3.2.2 *Pronominal gaps in alse-clauses*

In their earlier research on null subjects in MLG, Farasyn and Breitbarth (2016a) report on pronominal gaps in adverbial clauses introduced by the comparative particle *alse* '(just) as'. The gap is located right after *alse*, and hence is in the Wackernagel position, as are the second type of referential null subjects described in the previous subsection. Further evaluation of these cases shows that these sentences behave like relative clauses modifying the whole preceding situation. In Example (7), for example, the *alse*-clause refers to the whole action of claiming someone as a serf and testifying.

(7)

Un(de)	*se*	*wolden*	*ene*	*vorbosmen*	*un(de)*	*vortughen,*	*alze*
and	they	wanted	him	claim.as.serf	and	testify,	as [_]

des	*ammetes*	*recht*	*is*
the.GEN	authority's	right	is

'And they wanted to claim him as a serf and testify, as [it] is the authority's right'
(= '… which is the authority's right')

(Herford, *Herforder Rechtsbuch*, 1375)

The use of *(al)so/als(o)* as (approaching) a relative particle has already been described for MHG and ENHG (Paul ^{25}2007: 405/426; Ebert et al. 1993: 447/479). Indeed, in our MLG data, we find cases like (8), where *alse* is clearly used as a relative particle. Instead of referring to a whole preceding situation like (7), (8) has an object gap, and an overt subject (Farasyn & Breitbarth 2016a).

(8)

van	*wegen*	*eynes*	*huszes*	*alse*	*de*	*obg(ena)nte*	*Jacob*	*van*
Because	of	a	house	as	the	above-mentioned	Jacob	of

luebeke	*dem*	*vorb(enomed)en*	*Bernd*	*papke(n)*	*verkofft*	*hadde*	*vp*
Lübeck	the.DAT	aforementioned	Bernd	Papken	sold	had	on

passchen	*lest*	*vorleden*	*tobetale(n)de.*
Easter	last	past	to pay

'because of a house, which the above-mentioned Jacob of Lübeck had sold to the aforementioned Bernd Papken, to be paid this past Easter'
(Lübeck, *Urkunde 1500b*, 20/01/1500)

In cases like (7), there is no relative or resumptive pronoun overtly realising the subject of the adverbial/relative clause. Farasyn and Breitbarth (2016a) therefore propose to assume a null resumptive pronoun in the Wackernagel position in such cases, which also agrees with the verb. This analysis is further corroborated by null first and second person resumptives in MLG relative clauses, which we discuss in the next subsection.

In the context of the focus of the current article, it is furthermore crucial to note that relative(-like) *alse*-clauses with a subject gap/null resumptive pronoun are almost exclusively found in charters, a highly formalized genre. This again emphasizes the importance of including different genres in a historical corpus. Additionally, not finding these structures in certain genres may be taken as an indication of the fact that the construction is not representative of spoken MLG, but of more formal writing. This contrasts sharply with RNS, for instance, which tend to show up much more frequently in narrative texts, a genre that tends to be closer to the spoken language.

3.2.3 *Null resumptives in non-restrictive relative clauses*

A third syntactic phenomenon for which the inclusion of different genres in a representative MLG corpus is indispensable is the variation between overt and null resumptives in non-restrictive relative clauses (NRRCs) with a first or second person head in MLG. In order to understand the possible structures in MLG, a comparison with present-day German (PDG) is useful. In PDG, NRRCs with a first or second person head can in principle take three different forms, though they are not equally acceptable (Ito & Mester 2000; Trutkowski & Weiß 2016).[11] In the first type, there is agreement between the 1st or 2nd person antecedent (head) in the main clause and the verb in the relative clause, which shows a 1st or 2nd person ending (9a). This type of agreement pattern will thus be called head agreement (HA) in what follows. In the second type, agreement seems to be established between the (3rd person) relative pronoun and the finite verb in the relative clause (9b), leading to the term relative pronoun agreement (RPA). The third type looks

[11] Although Ito and Mester label the first structure as ungrammatical in PDG, Trutkowski and Weiß (2016) point out, based on a magnitude estimation experiment, that the three structures are all used, but have a varying level of acceptability.

like the first type in that it has 1st or 2nd person agreement on the verb in the relative clause, but there is an additional resumptive pronoun (9c). This structure can therefore be called resumptive pronoun agreement (ResPA).

(9) a. *Du, der mein Bruder bist,* (HA)
you, REL my brother are.2.SG
'You, who are my brother'

b. *Du, der mein Bruder ist, …* (RPA)
You, REL my brother are.3.SG

c. *Du, der du mein Bruder bist,…* (ResPA)
You, REL RESP.2.SG my brother are.2.SG

In MLG, the only agreement patterns that can be found are HA (10a) and ResPA (10b); no examples of RP have been found so far. HA is a puzzle insofar as noun-verb-agreement is normally clause-bound, and as the head noun is not necessarily a subject, but may be the object, a prepositional object, or a possessive pronoun in the matrix clause. In (10a) an example of a non-restrictive relative clause modifying an object (*dy*) is given, Example (10b) illustrates how the clause modifies the complement (*dy*) of a preposition (*van*) and (10c) shows a clause modifying the possessive *dyner*.

(10) a. *vp dat ick dy* ***de*** *dat ouerste gud bist v(m)me myne*
so that I you REL the highest good are.2.SG for my
eghene traecheit vn(de) vnuulherdicheit nicht en
own slowness and unpersistency NEG NEG
mote verlesen (HA)
must.1.SG lose
'so that I mustn't lose you, who are my highest good, because of my own slowness and lack of persistency'
(Münster, *Dat myrren bundeken*, 1480)

b. *meer warhen sal ick van dy vlein* ***de*** *du*
but where.to shall.1.SG I from you flee REL you
allerwegen Jegenwordich byst… (ResPA)
everywhere present are.2.SG
'but where will I flee from you, who are present everywhere…'
(Münster, *Ey(n) Jnnige clage to gode*, 1480)

c. *v(er)beide(n)de de behoerlike tijd* ***dyner*** *gheboerten de*
biding the appropriate time your.GEN birth REL
[ ] *na dyner godheit ghine tijd en heuest noch iare*
[ ] after your divinity no time NEG have.2.SG nor years
'biding the time appropriate for your birth, who has no time nor years due to your divinity' (Münster, *Dat myrren bundeken*, 1480)

It is noteworthy that every NRRC is introduced by *de*. While the relative pronoun in PDG (*der, die, das*) spells out gender and number features, this is not the case in MLG for *de*: it is invariable, even though the respective heads can be marked for the number and gender features in question (Farasyn 2017a). The fact that relative pronoun agreement is not attested in MLG raises the question whether the clause-introducing *de* is a relative pronoun at all, or whether it is an invariant relative particle (a syntactic head in C) instead. As a particle never contains φ-features (i.e. person, gender and number) and therefore cannot agree with the verb, the assumption of a relative particle could explain why 3rd person agreement is impossible. However, even though *de* can be found in free relative clauses as a relative particle in the Eastphalian dialects, additional evidence leads to the assumption that this is not the case in NRRCs. In Example (11), for instance, there are three non-restrictive relative clauses. All contain *de*, followed by either another *de*, or *dar*. We can take the second *de* in these combinations – if present – to be a relative particle *de*, located in C, which alternates with other particles in NRRCs, like *dar*. Semantically, the heads of all three relative clauses in (11) are 2nd person (plural); in the third, *gy* 'you' is even overt.[12] Therefore, Farasyn (2017a) claims that the left-peripheral invariable *de* that is always present is a relative pronoun, though underspecified for number, gender and person, i.e. for features relevant for agreement with the verb inside the relative clause.

(11) *Vrowet iu in deme heren alle* ***de de*** *enes guden leuendes mit ruwen be=gynnet vn(de) bewiset vtwendich de vroude iuwes herten alle* ***de dar*** *vort treden in enem guden leuende vn(de) beromet iu der ewighen ere alle gy* ***de=de*** *rechtes herten sint ane straffinghe iuwer samwitticheit*
rejoice you in the lord all REL REL a good life with remorse begins and prove outwardly the joy of-your heart all REL REL forth go in a good life and glory in the eternal glory all you REL=REL right heart are without punishment of-your conscience
'Rejoice in the lord, all who begin a good life with remorse, and outwardly show the joy of your heart, all who progress in a good life, and glory in eternal glory, all of you who are of the right heart without a guilty conscience'.
(Wolfenbüttel, Eastphalian psalms, 15th century)

Given that MLG shows either HN agreement or ResP agreement, and given further that *de* is invariant, and thus either a relative particle or an underspecified (non-agreeing) relative pronoun, Farasyn (2017a) argues that agreement is always established via a resumptive pronoun inside the relative clause. In cases that look like head noun agreement, Farasyn argues for a phonetically null resumptive in the

[12] In the first two clauses, the 2PL interpretation of *alle* is forced by the pronouns *iu* 'you.ACC' and *iuwes* 'you.GEN/POSS'.

Wackernagel position. This assumption is supported by the fact that the Wackernagel position, i.e. the position right after C, also contains other empty pronouns, which has been described above for null referential pronouns as well as for *alse*-clauses. The underspecified relative pronoun makes it possible to establish an agreement chain through relations of Checking and Matching, as it mediates between the head in the matrix clause and the (null) resumptive in the relative clause.

NRRCs with 1st and 2nd person heads are only sparsely attested in the texts of the *Corpus of Historical Low German*, in particular with null resumptives/apparent HA. Table 3 shows through multiple logistic regression analysis in Rbrul that such clauses are almost entirely restricted to religious prose texts.[13]

Table 3. Influence of the factor genre on NRRCs with 1st and 2nd person heads and HA

Genre	Log odds	Odds	Factor weight	N
religious	7.133	0.013	0.814	2221
letter	6.212	0.004	0.605	273
literature	4.995	0.001	0.552	3216
legal	3.724	0.000	0.518	2158
chronicle	−10.900	0.000	0.400	1002
charters	−11.165	0.000	0.144	1360

This means again that it is absolutely necessary to add all kinds of texts to the dataset to be incorporated into the corpus. It also means that religious texts are of high importance to decide on the possible labels and (empty) categories that have to be included in the syntactic mark-up of MLG texts, as they raise the question of whether fixed slots have to be reserved for (null) resumptives, just like it is common to reserve for instance slots for traces in the syntactic annotation layer, in order to make it possible for the researcher using the corpus to search for exactly these clauses.

3.2.4 *Pronominal gaps in asymmetric coordinations*

Like many other languages, the subject of a second conjunct in a coordination construction is almost always omitted if it is shared with and overtly expressed in the first conjunct. Such conjunction reduction can be seen in Example (12).

[13] N refers to the total number of finite clauses in the current (sub)corpus, the odds ratio indicates the presence of NRRCs with a 1st or 2nd person head (e.g., 0.013 ~ 1.3%).

(12)	*dey*	*sal*	*deme*	*Rayde*	*wedden*	*eyn*	*half*	*půnt*	*ande*	[ ]	*sal*
	that-one	will	the	council	pay	a	half	pound	and	[ ]	will

vte	*deme*	*gherichte*	*ewelike*	*wesen*	*vorwyset*
out	the	court	eternally	be	outlawed

'he will pay the council half a pound and [ ] will forever be outlawed'

(Soest, *Schrae im Statutenbuch*, 1367)

When conjunction reduction takes place in coordinations with subject-verb inversion in the first conjunct, something remarkable happens. As can be seen in Example (13), the first clause has subject-verb inversion because of the presence of the adverb *Vortmer.* The subject in the second conjunct is omitted, as it is clear from the content that both clauses share the same referent. This does however raise the question where the gap in the second clause should be located. In Example (12) for instance, it is reasonable to assume that both clauses are parallel, since there is no inversion: the overt subject in the first conjunct as well as the subject gap in the second are located in front of the verb. In (13a), however, we do not know if *Vortmer* also induces inversion in the second conjunct, with the subject gap following the verb, as in (13b), or if it does not, with an expected gap preceding the verb, as in (13c).

(13)	a.	*Vortmer,*	***bidde***	*wi*	*vnde*	***manen***	*alle*	*guode*	*lude,*
		furthermore,	pray	we	and	demand	all	good	people,

Houeman,	*vnde*	*husman*	*Dat*	*se*	*alle*	*mit*	*eneme*
nobleman,	and	peasant	that	they	all	with	a

schrichte	*volghen…*
complaint	follow

'Furthermore we pray and demand from every good man, nobleman and peasant, that they all sue with a complaint…'

b. *Vortmer,* ***bidde*** *wi vnde* ***manen*** *[wi] alle guode lude […]*
c. *Vortmer,* ***bidde*** *wi vnde [wi]* ***manen*** *alle guode lude […]*

(Lübeck, *Urkundenbuch Lübeck*, 1334)

Examples of coordinations with first person plural subjects like (13), which are particularly found in the opening formulas of charters, are invaluable for solving this puzzle. Note that the ending of the verb in the first conjunct is different from the one in the second conjunct. This is due to the fact that MLG has different agreement morphology on the verb in the 1st and 2nd person plural depending on the position of the pronoun relative to the verb. If the pronoun precedes the verb, the normal ending of the *Einheitsplural* ('unity plural'), which is *-n* or *-t* in MLG depending on period and region/scribal dialect, is used. However, if the pronoun follows the verb, the ending is different: in most cases, the *-n* or the *-t* is dropped. As the *-n* in Example (13) is dropped in the first clause, but not in the second one, this means that the gap in the second clause must be located before the verb and that these coordinations are therefore asymmetric (Farasyn 2017b). This conclusion is supported by research into asymmetric coordinations in High German (Reich 2009).

This case study shows once more how the study of a particular genre can have implications for what we can know about MLG word order. As dated and localized historical documents containing 1st and 2nd person plural are very hard to find, it shows us that the inclusion of charters in the corpus is paramount for shedding light on the nature of these structures. Once again, it is the in-depth syntactic study of a particular genre that solves a (syntactic) puzzle that could not have been solved without including this specific genre in the dataset. Furthermore, this study demonstrates the importance of the examination of a specific genre for parsing decisions, as in this case, the decision on where the gap in the clause must be located.

4. Summary and outlook

This paper examined the role of genre in the construction of a historical corpus of Middle Low German, incorporating all kinds of variation. A case study on the role of discourse markers made clear how future in-depth study of genre-specific discourse markers might lead to improvement of the accuracy of the POS-tagger when applied to texts from genres that were not included in the gold standard training data. Multiple syntactic case studies on the role of genre for null referential arguments showed among other things how genre significantly influences the amount of referential null subjects and how formulaic structures in the legal genre can lead to the discovery of *alse* as a relative particle. The study of religious texts led in its turn to the discovery of a (null) resumptive inside the non-restrictive relative clause and delivered insights into the word order of (asymmetric) conjuncts in MLG. All of these insights can be used to adapt and extend the labels in the layer of syntactic annotation.

References

Anagnostopoulou, Elena. 2008. Notes on the Person Case Constraint in Germanic (with special reference to German). In *Agreement Restrictions*, Roberta D'Alessandro, Susann Fischer & Gunnar Hrafn Hrafnbjargarson (eds), 15–48. Berlin: Mouton de Gruyter.

Barteld, Fabian, Dreessen, Katharina, Ihden, Sarah, Schröder, Ingrid, Glawe, Meike, Kleymann, Verena, Nagel, Norbert, Peters, Robert & Schilling, Elmar. 2014. Transkriptionshandbuch des Projekts Referenzkorpus Mittelniederdeutsch/ Niederrheinisch (1200–1650) (Version 1. August 2014, Auszug). <https://vs1.corpora.uni-hamburg.de/ren/media/pdf/transkriptionshb1.pdf>

Breitbarth, Anne. 2014. *The History of Low German Negation*. Oxford: OUP. https://doi.org/10.1093/acprof:oso/9780199687282.001.0001

Dipper, Stefanie, Donhauser, Karin, Klein, Thomas, Linde, Sonja, Müller, Stefan & Wegera, Klaus-Peter. 2013. HiTS: Ein Tagset für historische Sprachstufen des Deutschen. *JLCL* 28(1): 85–137.

Dreessen, Katharina & Ihden, Sarah. 2015. Korpuslinguistische Studien zur mittelniederdeutschen Syntax. *Jahrbuch für Germanistische Sprachgeschichte* 6(1): 249–275. https://doi.org/10.1515/jbgsg-2015-0016

Ebert, Robert Peter, Reichmann, Oskar, Solms, Hans-Joachim & Wegera, Klaus-Peter. 1993. *Frühneuhochdeutsche Grammatik*. Tübingen: Niemeyer. https://doi.org/10.1515/9783110920130

Farasyn, Melissa. 2017a. Kongruenzmuster in mittelniederdeutschen Relativsätzen: Eine Pilotstudie. In *Aktuelle Tendenzen in der Variationslinguistik (Kleine und regionale Sprachen)*, Line-Marie Hohenstein, Kathrin Weber, Heike Wermer, Meike Glawe & Stephanie Sauermilch (eds), 67–90. Hildesheim: Georg Olms.

Farasyn, Melissa. 2017b. Deletion in the verbal paradigm in Middle Low German: An interface phenomenon. Ms, Ghent University.

Farasyn, Melissa & Breitbarth, Anne. 2016a. Nullsubjekte im Mittelniederdeutschen. *Beiträge zur Geschichte der Deutschen Sprache und Literatur* 138(4): 524–559. https://doi.org/10.1515/bgsl-2016-0040

Farasyn, Melissa & Breitbarth, Anne. 2016b. Null Subjects in Middle Low German. Ms, Ghent University.

Farasyn, Melissa, Witzenhausen, Elisabeth & Breitbarth, Anne. 2016. Anmerkungen zur mittelniederdeutschen Syntax in Psalmenübersetzungen des (14. und) 15. Jahrhunderts. Ms, Ghent University.

Fedders, Wolfgang. 1988. Zur Erhebung historischer Schreibsprachdaten aus der Textsorte 'Urkunde'. *Niederdeutsches Wort* 28: 61–74.

Galves, Charlotte, Andrade, Aroldo Leal de, & Faria, Pablo. 2017. Tycho Brahe Parsed Corpus of Historical Portuguese. URL: http://www.tycho.iel.unicamp.br/~tycho/corpus/texts/psd.zip.

Holmberg, Anders. 2010. Null subject parameters. In *Parametric Variation: Null Subjects in Minimalist Theory*, Theresa Biberauer, Anders Holmberg, Ian Roberts & Michelle Sheehan (eds), 88–124. Cambridge: CUP.

Ito, Junko & Mester, Armin. 2000. Ich, der ich sechzig bin: An agreement puzzle. In *Jorge Hankamer WebFest*, Sandy Chung, Jim McCloskey & Nathan Sanders (eds). <http://ling.ucsc.edu/Jorge/ito_mester.html>

Koleva, Mariya, Farasyn, Melissa, Desmet, Bart, Breitbarth, Anne & Hoste, Veronique. 2017. An automatic part-of-speech tagger for Middle Low German. *International Journal of Corpus Linguistics* 22(1): 108–141. https://doi.org/10.1075/ijcl.22.1.05kol

Lenerz, Jürgen. 1977. *Zur Abfolge nominaler Satzglieder im Deutschen*. Tübingen: Narr.

Mähl, Stefan. 2014. *Mehrgliedrige Verbalkomplexe im Mittelniederdeutschen. Ein Beitrag zu einer historischen Syntax des Deutschen*. Köln: Böhlau.

Paul, Hemann. ^{25}2007. *Mittelhochdeutsche Grammatik*, neu bearb. von Thomas Klein, Hans-Joachim Solms & Klaus-Peter Wegera. Mit einer Syntax von Ingeborg Schöbler, neu bearb. von Heiz-Peter Prell. Tübingen: Niemeyer.

Peters, Robert. 1973. Mittelniederdeutsche Sprache. *Niederdeutsch. Sprache und Literatur* 1: 66–129.

Peters, Robert. 2003. Variation und Ausgleich in den mittelniederdeutschen Schreibsprachen. In *The Dawn of the Written Vernacular in Western Europe*, Michèle Goyens & Werner Verbeke (eds), 427–440. Leuven: Leuven University Press.

Petrova, Svetlana. 2012. Multiple XP-fronting in Middle Low German root clauses. *The Journal of Comparative Germanic Linguistics* 15(2): 157–188. https://doi.org/10.1007/s10828-012-9050-y

Reich, Ingo. 2009. '*Asymmetrische Koordination*' *im Deutschen*. Tübingen: Stauffenburg.

Truswell, Robert, Alcorn, Rhona & Donaldson, James. 2016. A parsed linguistic atlas of Early Middle English. Paper presented at the first Angus McIntosh Centre symposium, June 10, 2016. Slides available at <http://robtruswell.com/assets/pdfs/AMC_talk.pdf>

Trutkowski, Ewa & Weiß, Helmut. 2016. When personal pronouns compete with relative pronouns. In *The Impact of Pronominal Form on Interpretation*, Patrick Grosz & Pritty Patel-Grosz (eds), 135–166. Berlin: De Gruyter.

Wackernagel, Jacob. 1892. Über ein Gesetz der indogermanischen Wortstellung. *Indogermanische Forschungen* 1: 333–436. https://doi.org/10.1515/9783110242430.333

Walkden, George. 2014. *Syntactic Reconstruction and Proto-Germanic*. Oxford: OUP. https://doi.org/10.1093/acprof:oso/9780198712299.001.0001

Walkden, George. 2016. The HeliPaD: A parsed corpus of Old Saxon. *International Journal of Corpus Linguistics* 21(4): 559–571. https://doi.org/10.1075/ijcl.21.4.05wal

Wallenberg, Joel C., Ingason, Anton Karl, Sigurðsson, Einar Freyr & Rögnvaldsson, Eiríkur. 2011. *Icelandic Parsed Historical Corpus (IcePaHC)*, Version 0.9. <http://www.linguist.is/icelandic_treebank>

Genre influence on word formation (change)

A case study of German adjectival derivation

Luise Kempf
University of Mainz

Building on Kempf (2016), this paper analyzes genre influence on adjectival derivation in the history of German. The approach is empirical, based on data from two corpora of historical German (1350–1800). Quantitative measures of word-formation productivity reveal substantial differences between sermons, narrative prose, scientific texts, and newspapers. A long-term diachronic comparison highlights considerable changes in suffix distribution. Against this backdrop, the four genres are assessed as progressive or conservative. An analysis of the syntactic use of derived adjectives reveals stronger or weaker tendencies to condense information within noun phrases. This reflects differing affinities to oral or written language, as does the differing use of endophoric vs. exophoric reference (*obig* 'the above' vs. *gestrig* 'of yesterday'). Overall, genre greatly impacts word-formation and conversely, word-formation contributes to genre variation.

1. Introduction

This paper investigates the question of genre influence on word formation and on word formation change, using as an example adjectival derivation in Early Modern German. The literature on this topic is relatively scarce and generally offers no quantitative information on genre differences (cf. Section 2). This can be related to the fact that up until recent years, there was a shortage of digital diachronic corpora balanced for genre. For instance, the Bonn corpus of Early New High German (Solms & Wegera 1998) is balanced for regions and time periods, but not for genre.[1]

[1] The corpus, compiled in the 1970s, was a pioneer project in terms of freely available historical corpora. It is currently being updated and integrated into the *Historisches Referenzkorpus des Deutschen* (historical reference corpus of German).

https://doi.org/10.1075/scl.85.14kem

The present study uses data primarily from the German Manchester Corpus (cf. Section 3), which was released in 2012. The corpus covers five regions, three fifty-year periods (1650–1800), and eight genres.[2] Each slot defined by these features is represented with an equal amount of text, so that the corpus is fully balanced. With this corpus design, it is possible to disentangle the influence of the factors region, time, and genre. The results are striking: while there is only minor variation among the different regions and periods, the impact of genre is tremendous (cf. Kempf 2016: 106).

The question of genre impact on derivational morphology has numerous facets. In this paper, they will be approached from three perspectives: after an overview of the state of research in Section 2 and a discussion of methodology in Section 3, Section 4 analyzes differences in morphological productivity among the genres by scrutinizing a number of quantitative measures. Section 5 explores the composition of suffix inventories through time and across genres. The goal is to identify progressive vs. conservative suffix distributions as well as the suffixational patterns most characteristic of each genre. Adjectival derivation in general, and certain patterns in particular, can be used to increase informational density in texts. In connection to that, features of conceptual orality and writing are discussed. Section 6 explores functional, syntactic, and textual implications of the observed suffix distributions. In particular, the section examines syntactically motivated formations as well as the rate of attributive use through time. Finally, Section 7 sums up the results and draws important conclusions with respect to the role of genre for word formation and conversely, the role of word formation in genre profiles. Additionally, the section discusses genres with respect to their role in language change, revisiting the notion of progressive vs. conservative genres.

2. State of research

In their famous article on morphological productivity in spoken vs. written registers, Plag et al. (1999) pointed out that the role of derivational morphology had been neglected in research on genre variation and that conversely,

[2] The regions are: East and West Upper, East and West Central, North; the genres are: "drama, newspapers, sermons and personal letters (to represent orally-oriented registers) and narrative prose (fiction or non-fiction), scholarly (i.e. humanities), scientific and legal texts (to represent more print-oriented registers)" (Durrell et al. 2012). The present paper focuses on the genres narrative prose, newspapers, scientific texts, and sermons.

morphological research underestimated the impact of genre. In the research on German language history, there is a fair number of studies that explicitly take into account genre as an influential factor. However, word formation hardly plays any role in these works. Most of them are concerned with syntactic phenomena (e.g. Ebert 1988; Prell 2000; Jäger 2008; Pickl 2017; multiple contributions in Wich-Reif & Simmler 2005; Ziegler 2010; and Braun 2011) or other linguistic levels, e.g. inflection (Scheible et al. 2011; Durrell & Whitt 2013; Durrell 2016), lexis and typography (Hölscher 2011), and pragmatics (Linke 2001; Whitt 2018).

As for the literature on word formation, it can be (very roughly) grouped into two types: studies focusing on one specific genre and studies drawing on mixed-genre corpora. The first type is represented e.g. by three studies based on a 13th-century corpus of charters (Kronenberger 2002; Ring 2008; Ganslmayer 2012), by Brendel et al.'s (1997) survey of Early New High German specialized texts, and by Scherer's (2005) study based on an Early Modern German newspaper corpus.[3] All of these works look into nominal word formation, except for Ganslmayer's (2012) work, which concentrates on adjectival derivation. As a consequence of focusing on one genre, studies of this type tend to provide little or no comparison to other genres.[4] The second type, i.e. the mixed-genre studies, comprises works resulting from three large corpus projects that have been realized in the last decades. Several publications are based on the *Bonn Corpus of Early New High German* (henceforth: BonnC), including two studies on verbal derivation (Prell 1991; Prell & Schebben-Schmidt 1996), one on selected nominal suffixes (Doerfert 1994), and one on the adjectival suffix *-lich* (Winkler 1995). Three monographs draw on a mixed-genre corpus of Nuremberg ENHG from around the year 1500: Müller (1993), Habermann (1994), and Thomas (2002) (surveying nominal, verbal, and adjectival derivation, respectively). Finally, a comprehensive monograph on Middle High German word formation (Klein et al. 2009) draws on a corpus containing verse, prose, and charters.

The works of the second type typically mention the genre factor in passing when obvious genre effects occur (e.g. Wegera & Solms 2002: 165; Klein et al. 2009: 303, 328). The observations are usually expressed as tendencies, but no

[3] The term Early New High German (ENHG) generally refers to the period 1350–1650. Additionally, I will use the term *Early Modern German* to refer to the period 1650–1800, i.e. the first part of the New High German period (1650–present).

[4] By and large, the focus of these studies lies in exploring the word formation systems of specific periods, not in genre comparison. Brendel et al. 1997, however, include one prose text in their study and discuss how it differs from specialized texts.

specific quantification is provided. This holds as well for three studies that date back some thirty years but are still relevant when surveying adjectival derivation: Kühnhold et al.'s (1978) monograph on 20th-century adjectival derivation and Bentzinger's (1987) and (1992) articles on adjectival suffixes in the 16th–18th centuries. Prell (1991) does provide quantitative results for the very broadly composed genres 'biblical' vs. 'non-biblical' texts. In works of both types, one usually finds remarks or even chapters on diachronic comparison, either within or beyond the corpus examined. However, given the heterogeneous corpus situation, this usually amounts to comparing genre A in period X to genre B in period Y. This does not allow a definitive conclusion whether observed characteristics arise from the genre, the period, or both.

In recent years, the corpus situation has continued to improve, and three diachronic word formation studies have been released that make use of genre-balanced corpora. Kempf (2016) examines adjectival derivation in both BonnC and the *German Manchester Corpus* (henceforth: GerManC); Kopf (2018) analyzes nominal compounding in the Mainz corpus of (Early) New High German, and Hartmann (2016) uses both GerManC and the Mainz corpus to investigate nominalization patterns. All three studies offer some genre-related observations (e.g. Kempf 2016: 50, 71, 223f.) or even data analyses (Hartmann 2016: 129, 200–204, 261). The present article elaborates in detail on the genre-related results touched upon in Kempf (2016).

3. Approach, corpora, and methods

The present paper takes an empirical approach to the influence of genre on word formation. The corpus data are evaluated for tokens, types, newly attested types, and types that exclusively occur in only one of the four genres. The goal is to learn about the morphological productivity and the preferred patterns of each genre. The question of a progressive vs. conservative (or "agile" vs. "uptight", Hundt & Mair 1999) character of the genres will be approached via long term diachronic comparison of suffix use. Additionally, the syntactic behavior of derived adjectives will be examined in order to learn about their role in the structuring of information and the constitution of texts.

The data are taken from a large database compiled from both BonnC and GerManC. Together, both corpora provide seven periods of fifty years each, two of them overlapping (cf. Table 1). Since GerManC is roughly twice as large per period as BonnC, I included only half of it, so as to keep the amount of text constant. In order to achieve the best possible genre consistency with BonnC, the GerManC genres 'narrative prose', 'newspapers', 'scientific texts', and 'sermons' were selected.[5] From the selected corpus texts, all tokens of suffixed adjectives were extracted along with their lemma annotations, and, in the case of BonnC, also suffix annotations.[6] All annotations were then manually corrected, or added in the case of GerManC suffix annotations. Table 1 provides an overview of the corpus sizes as well as the type and token numbers of all suffixed adjectives for each period.

Table 1. Data sizes of the main data base

Period	**BonnC**					**GerManC**				**Total**
	1350–1400	**1450–1500**	**1550–1600**	**1650–1700**	**total**	**1650–1700**	**1700–1750**	**1750–1800**	**Total**	
Corpus words	123,952	138,314	125,163	127,525	514,954	125,318	126,172	123,848	375,338	890,292
Tokens suff, adj,	2,189	2,447	3,201	4,317	12,154	3,698	4,052	4,108	11,858	24,012
Types suff, adj,	335	407	517	689	1,024	726	794	904	1,431	1,885

4. Quantitative productivity measures

There are a number of quantitative measures of morphological productivity, none of which is without flaws (cf. Kempf 2016: 113–124 , Kempf forthcoming). Therefore, it is crucial to include several measurements and to consider carefully what they actually measure. This section works with three basic values: types, newly attested types (short: neo-attestations), and hapax legomena (short: hapaxes). The value of types (i.e. lexemes) indicates the vocabulary size and is referred to as "realized productivity" in the literature (e.g. Baayen 2009). The term neo-attestation in this paper refers to types that are attested for the first time within the corpus (previous studies using newly attested types include, among others, Flury 1964; Cowie 1999; Berg 2016; and Ganslmayer & Müller 2016). While a neo-attestation

[5] BonnC contains 40 texts (four periods times ten regions), which are unevenly distributed across six genres. Since 18 BonnC texts pertain to a religious genre, GerManC 'sermons' were selected; another twelve texts are chronicles or reports, so GerManC 'newspapers' were chosen as the closest match. The remainder of texts are fiction or specialized texts, so 'narrative prose' and 'scientific texts' were selected.

[6] In (Early) New High German, adjectives can occur in attributive, predicative, adverbial, and nominalized use. A word was counted as an adjective if it occurred at least once in attributive use within the corpus (cf. Thomas 2002: 17).

need not necessarily be a newly coined word, the value is still a good approximation to productivity: by way of excluding previously attested types, the measurement comes noticeably closer to active word formation than the type value does. Hapax legomena are words that occur only once in a corpus. Given that newly coined words tend to occur infrequently at first, hapaxes are generally considered a good indicator of productivity (Baayen & Lieber 1991; Plag et al. 1999; Baayen 2009). For the present data, the hapaxes yield the same basic tendencies as the neo-attestations (note that hapaxes are a subset of the latter).[7]

This section focuses on the GerManC data, where the genres are equally represented. The differences among the three fifty-year periods (1650– 1800) as well as among the five regions are not significant. Thus, for the present purpose of discerning productivity variation among genres, the total data for each genre are merged. Figure 1 shows the absolute measures of types, neo-attestations, and hapaxes of suffixed adjectives in the four GerManC genres sermons (SERM), narrative prose (NARR), scientific texts (SCIE), and newspapers (NEWS). Table 2 presents the underlying numbers (which form the basis of Figure 2 as well).

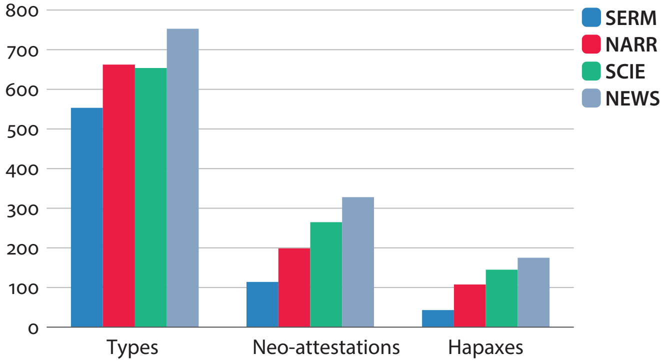


Figure 1. Values for types, neo-attestations, and hapaxes of suffixed adjectives in four GerManC genres

Figure 1 shows that newspapers are leading in all three measures, while sermons rank last. The absolute numbers, however, need to be put in perspective: the newspapers subcorpus contains the most tokens by far and thus has the best

[7] Both hapaxes and neo-attestations are calculated related to the entire data base, i.e. the data from both BonnC and GerManC. Neo-attestations are all those (and only those) tokens that constitute the first attestation (by year) of their type in the data base; for instance, if a type X is attested once in 1728 in genre A, and once in 1743 in genre B, it counts only once as a neo-attestation for genre A.

chances of reaching high values of types, neo-attestations and hapaxes. Figure 2 visualizes a number of relative measures, where types, neo-attestations, and hapaxes are related to token or type values.[8] However, even when looking at quotients, the results are not independent of input size: the number of types does not increase proportionally when more and more tokens are evaluated. Instead, the more tokens have been evaluated, the less likely it becomes to encounter a type that has not been sampled yet.[9] To gain more balanced data, I recalculated the values based on samples of equal token numbers (by way of randomly downsizing the data of sermons, science, and news to match the token size of narrative prose, which was the genre lowest in tokens). The results are presented in Table 3 and Figure 3.

Table 2. Productivity values for adjectival derivation in four GerManC genres[10]

	Tokens	**Types**	**Neo-attestations**	**Hapaxes**	**Types/ Tokens**	**Neo-attestations/ Tokens**	**Hapax/ Tokens**	**Neo-attestations/ Types**	**Hapax/ Types**
SERM	3158	553	114	43	0.18	0.04	0.01	0.21	0.08
NARR	2416	662	199	108	0.27	0.08	0.04	0.30	0.16
SCIE	2657	653	265	145	0.25	0.10	0.05	0.41	0.22
NEWS	3633	752	328	175	0.21	0.09	0.05	0.44	0.23
4 genres combined	11864	1431	906	471	0.12	0.08	0.04	0.63	0.33

[8] The measure types divided by tokens is discussed, e.g., by Klein et al. (2009: 12f). Hapaxes divided by tokens is a standard productivity measure (albeit much debated, e.g. Gaeta & Ricca 2006). It indirectly relates low frequency items (indicative of productive word formation) to high frequent types (indicative of established, potentially lexicalized formations). The measure neo-attestations divided by types has been discussed by Berg (2016). It offers the advantage of not being sensitive to the frequency of use (e.g. excessive usage of individual types due to genre conventions) while still relativizing the absolute numbers. The idea is to determine what proportion of the vocabulary is actually newly coined.

[9] Likewise, the score of neo-attestations will grow less and less the more tokens have been evaluated. This phenomenon is even more drastic with hapaxes: Their score will grow very fast and first, but it will eventually decrease, since more and more derivatives will occur for the second time and thus cease to be hapaxes (cf. Gaeta & Ricca 2006; Evert & Baroni 2007; Kempf 2016: 117–124, Kempf forthcoming).

[10] Note that the last line does not show the sum of the lines above, but rather the value calculated for the four genres taken together.

Table 3. Productivity values for adjectival derivation in four GerManC genres, recalculated for samples of 2416 tokens per genre

	Tokens	Types	Neo-attestations	Hapaxes	Types/Tokens	Neo-attestations/Tokens	Hapax/Tokens	Neo-attestations/Types	Hapax/Types
SERM	2416	501	82	29	0.21	0.03	0.01	0.16	0.06
NARR	2416	662	199	108	0.27	0.08	0.04	0.30	0.16
SCIE	2416	621	239	128	0.26	0.10	0.05	0.38	0.21
NEWS	2416	621	223	120	0.26	0.09	0.05	0.36	0.19
4 genres combined	9664	1333	743	385	0.14	0.08	0.04	0.56	0.29

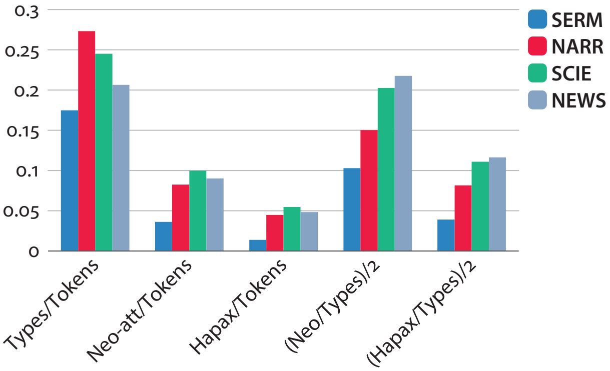


Figure 2. Productivity measures for derived adjectives in four GerManC genres

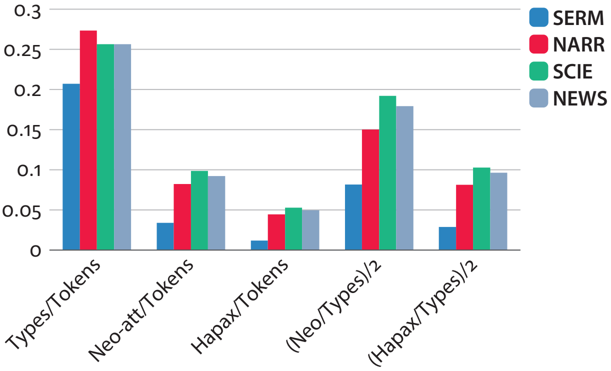


Figure 3. Productivity measures for derived adjectives in four GerManC genres, based on samples of 2416 tokens per genre

In Figure 2 and in Figure 3, narrative prose texts show the highest type-token ratio (i.e. they use the most diverse vocabulary) while sermons show the lowest. The size of vocabulary can be related to the diversity of topics. For instance, the scientific texts in GerManC include many subgenres, such as medicine, chemistry, and mining. The narrative prose texts are thematically open to virtually anything; they include, e.g. fairytales, biographic novels, utopian novels, travel literature, and much more. The texts in the sermons subcorpus cover different topics as well (mourning, holidays, even lottery addiction), but they do seem more homogenous than the other genres. The topical diversity may be ascribed, to some degree, to the corpus composition: if we were to limit science, for instance, to mining texts, the genre would probably attain lower type values. I would argue, however, that thematic diversity is, to a large degree, an intrinsic property of a genre. This becomes particularly clear with newspapers: their high topical diversity could not be avoided by selecting different texts since it becomes manifest *within* the individual corpus texts.

The differing degrees of topical diversity may have some influence on the productivity measures 'neo-attestations/tokens' and 'hapaxes/tokens', seeing that both neo-attestations and hapaxes are subsets of types. In the last two quotients in Figure 2 and in Figure 3, by contrast, the factor of topical diversity is leveled out, because the indicative values (neo-attestations and hapaxes) are viewed in relation to types. Interestingly, both groups of quotients (divided by tokens and divided by types) yield similar results: adjectival derivation appears to have been most productive in scientific texts and newspapers and least productive in sermons, while narrative prose takes an intermediate position. These findings correspond quite well to Hartmann's (2016: 129) results about *-ung*-nominalization in GerManC: measured by the hapax-token ratio, newspapers rank first, scientific and narrative texts second and third, while sermons take the last position.

A number of reasons described in the literature may account for the differing degrees of productivity in the four genres. One obvious factor would be the novelty of concepts discussed in both science and newspapers. Brendel et al. (1997: 595–605) as well as Polenz (2013: 374–395) discuss innovative tendencies in specialized texts of ENHG and in scientific texts of the 17th and 18th centuries, respectively. In both periods, the texts had to master the communicative task of expressing concepts in German that had previously been discussed in Latin only – resulting in borrowing, learned word formation, and native word formation. Polenz (2013: 402–403) describes that novel words of this kind are typically attested first in scientific texts, followed closely by newspapers, while literary texts show delays of years or decades. Furthermore, Polenz (2013: 397) observes that newspapers use more innovative language than sermons, albeit leaving open in what respect (e.g. word formation, syntax). This observation corresponds to socio-historical developments: in the 17th century, newspapers had just come into existence, while sermons were building on a long-standing discourse tradition. Martin Luther's work constitutes a major, if not the main influential factor:[11] Its great impact on the development of the German standard language in the 16th century and beyond is well known. More specifically, Prell (1991: 237–238, 245) observes a high rate of novel derived verbs in Luther's bible translation, whereas biblical texts of the subsequent (i.e. the 17th) century show no more evidence of productive (verbal) word formation. Given that Luther's oeuvre includes sermons as well, we can assume similar dynamics in this genre. Generally, the sixteenth century – the century of the Reformation – likely witnessed rather innovative language in sermons, while sermons of the subsequent centuries may have relied on the vocabulary inherited from this period.

Further explanation for differences in word formation productivity can be derived from characteristics of spoken vs. written registers: based on a survey of various registers in the BNC,[12] Plag et al. (1999) show that various suffixational patterns in English are more productive in writing than in speech. This tendency can be accounted for by the *parameter of time* (Ágel & Hennig 2007: 197–200). In writing, there is more planning time as well as more parsing time than in speech. This factor may encourage the formation and the use of complex words in writing and rule out structures too complex for online processing in speech. Among the four genres in the present study, sermons are closest to spoken language (cf. Durrell 2016: 219, 222). As they are written for oral delivery, the factor 'limited parsing time' may rule out the use of highly complex structures. Early newspapers, by contrast, are strongly influenced by ENHG chancery style. They are characterized by exceedingly long sentences and high informational density (Polenz 2013: 398–400). Complex adjectival attributes lend themselves well to condensing information, since they can incorporate information that would otherwise be expressed by a subordinate clause. This aspect will be addressed in more detail in Section 6. Section 5 will complement the quantitative productivity measures discussed so far by a qualitative analysis of suffixational patterns in the four genres.

5. Distribution of suffixational patterns

Adjectival derivation in German involves up to 25 different suffixational patterns (depending on the analysis). This section will first outline major diachronic trends and then present the specifics of each genre with respect to these trends. Figure 4

[11] I would like to thank Martin Durrell, who drew my attention to this aspect and also pointed out that religious language can aim to be deliberately archaic.

[12] *British National Corpus*, <http://www.natcorp.ox.ac.uk/>

gives an overview on the primary developments in a long-term diachronic perspective. It combines results from previous studies with results of the present study (BonnC and GerManC, highlighted by "B" or "M" prefixed to the period label). Each column indicates the composition of types in a given period. For simplification, the graph shows only the three most dominant suffixes separately and merges all other patterns into the groups "other" or "learned".[13] The numbers in the columns indicate the percentage of types that each suffix or group attains in a period.[14] Please note that the x-axis is not linear, but simply arranges the different studies in chronological order.

Even though the studies work with different genre compositions, the shares of the patterns appear to be roughly comparable (but cf. details below). Some essential long-term trends can be discerned: (1) *-lich* (as in *königlich* 'kingly') decreases continuously; (2) *-isch* (as in *Englåndisch* 'English') increases substantially but likely stagnates after 1800; (3) *-ig* (e.g. *steinig* 'stony') decreases slightly but remains stable after 1750; (4) learned suffixes (e.g. *-al* in *pyramidal*, *-abel* in *konsiderabel*) enter the system and increase after 1650.[15]

Figure 5 applies the method used in Figure 4 to the genres of GerManC (1650–1800) as well as to data for spoken 20th-century German (Gersbach & Graf 1985: 563 + 600 [for the addition of *-los* '-less']). The genre differences are quite striking: sermons show an overall distribution reminiscent of the time around 1500. Narrative prose texts display a distribution quite in line with the average of the period; scientific texts differ from that only in using more "other" and learned suffixes. Newspapers, finally, show an exuberant use of*-isch*-suffixation (this is linked to *-isch* frequently combining with proper names, see below). Based on these results, the four genres can be viewed on a scale from conservative to progressive in the order they are shown.

[13] The group "other" includes the suffixes *-artig*, *-bar*, *-e*, *-echtig*, *-er*, *-ern*, *-fach*, *-förmig*, *-haft*, *-haftig*, *-haltig*, *-icht*, *-los*, *-mäßig*, *-reich*, *-sam*, *-selig*, *-valt*; the group "learned" includes: *-abel/-ibel*, *-al*, *-ant*, *-ar/-är*, *-at*, *-ell*, *-esk*, *-(is)iert*, *-il*, *-iv*, *-os/-ös*, *-oid*. It is important to note that, unlike in English, word formation with learned elements in German constitutes a subsystem partially independent from native word formation (cf. Bergmann 1998, Kempf 2010).

[14] The measurement "share of types" was chosen because it is available in all the studies consulted and it is also fairly robust for comparison among differently sized corpora. As for the BonnC and the GerManC data, a number of other measures have been applied, resulting in the same essential trends as outlined in this section (see Kempf 2016).

[15] These findings essentially confirm and refine the tendencies described in previous accounts (see Thomas 2002: 559–565 for a comparison between ENHG and NHG, Ganslmayer 2012: 1080–1086 for a synopsis of MHG, ENHG, and NHG, and Polenz 2013: 314–319 for a concise account of word formation changes in the 17th and 18th centuries).

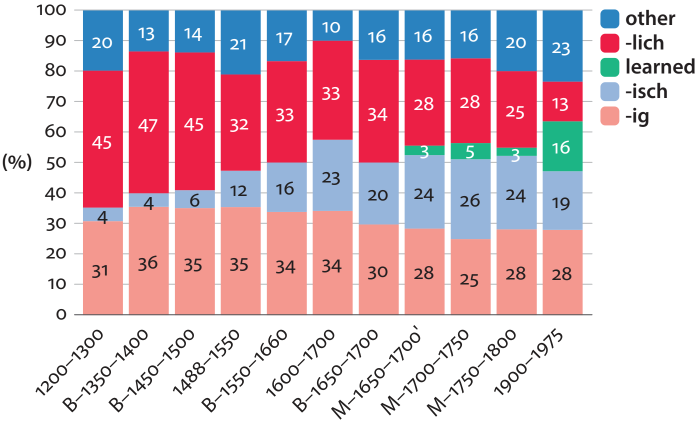


Figure 4. Major trends in adjectival suffixation measured in percentage of types per period[16]

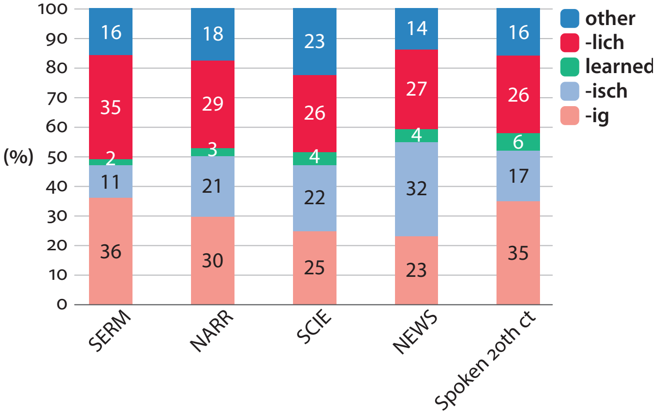


Figure 5. Genre and register differences in adjectival suffixation measured in percentage of types[17]

The 20th-century spoken data are included for comparison. All corpora underlying Figure 4 include charters, reports, chronicles, or newspapers (to a greater

[16] Numbers of types/sources: 1200–1300: 587/Ganslmayer (2012: 35); 1350–1400/1450–1500: 335/407/BonnC; 1488–1550: 561/Thomas (2002, recalculated, cf. Kempf 2016: 102–104); 1550–1660: 517/BonnC; 1600–1700: 1526/Schulz (2002, 2007: 193); 1650–1700: 689/BonnC; 1650–1700′/1700–1750/1750–1800: 726/794/904/GerManC; 1900–1975: 7189/Kühnhold et al. (1978: 100–117 + p. 444 for the addition of *-los* '-less'). The absolute numbers for all suffixational patterns are given in Tables 5 and 6 in the Appendix.

[17] The absolute numbers are given in Table 7 in the Appendix.

or lesser extent). Therefore, it could be questioned to what extent the essential long-term trends stated above hold for "the language" in general (Hundt & Mair 1999: 236). In the spoken data, we find these basic trends confirmed (decrease of -*lich*-derivation, expansion of learned suffixes, increase of -*isch*-derivation) – yet to a lesser degree than in writing. This could mean that we are looking at developments of written language that have spread to speech. If we were able to trace spoken texts throughout centuries, we might find different trends that eventually spread to written language (e.g. an upsurge of -*ig*-derivation). This thought casts a different light on the lagging distribution in sermons: the genre might appear conservative because it is conceptually closer to spoken language, where different diachronic trends prevail.

So far, this section has focused on distributions of types. In order to gain a more accurate picture of active word formation, Figure 6 shows the suffix distributions found in the four genres when counting neo-attestations only.

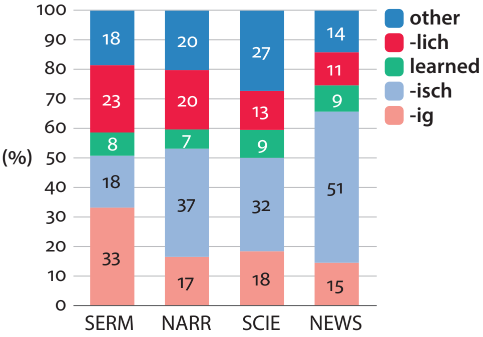


Figure 6. Genre differences in adjectival suffixation measured in percentage of neo-attestations[18]

The genres display roughly the same tendencies relative to one another as they do in Figure 5 (types). However, the overall proportions among the patterns have shifted: -*isch*-derivation as well as learned suffixes – i.e. the patterns that are diachronically on the rise – indeed attain a larger percentage in the neo-attestations. Conversely, the two more traditional patterns, -*lich* and -*ig*, show smaller percentages when only neo-attestations are counted. Three features, again, are particularly noticeable: -*isch*-derivation in newspapers, "other" patterns in scientific texts, and -*ig*-derivation in sermons. These affinities will be explored in the remainder of this section, where the most characteristic patterns of each genre are exemplified.

[18] The absolute numbers are given in Table 8 in the Appendix.

Specifically, the characteristics will be traced by taking into account neo-attestations as well as “exclusive formations”, i.e. words that occur exclusively in that genre. Figure 7 shows the composition of exclusive formations in each genre. The tendencies are similar to those in Figure 6 (neo-attestations), if not even more pronounced. The absolute numbers of exclusive formations are indicated below the genre labels.

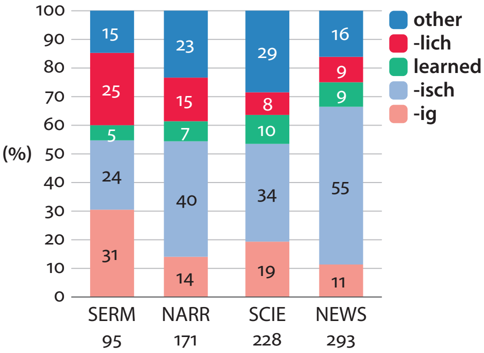


Figure 7. Genre differences in adjectival suffixation measured in percentage of genre-exclusive formations[19]

As shown above, sermons can be characterized by high proportions of -*ig*- and -*lich*-derivation. When looking into individual derivatives of these patterns (neo-attestations or exclusive formations), a functional tendency becomes transparent: very often, they conceptually involve actions or predispositions towards actions. This is reflected in numerous -*ig*-derivatives that can be analyzed as deverbal active formations or denominal possessive formations (or both), e.g. *allwissig*, lit. “all-know-y”, ‘all-knowing’/‘having all knowledge’; *bissig*, lit. “bite-y”, ‘snappish’. A large share of the -*lich*-derivatives are deverbal, e.g. *unausbleiblich* ‘not failing to appear’, *unauslöschlich* ‘inextinguishable’. This tendency also appears with “other” (*unzerstörbar* ‘indestructible’) and learned suffixes (*inexpugnabel*, *perseverant*).

Narrative prose texts display fairly average suffix distributions. Yet, from an onomasiological perspective there is a remarkable wealth of ‘comparative’ formations (i.e. formations that serve the function of comparing their base and the noun they modify, cf. the function ‘[[like]] X’ in Pounder 2000: 113f). This wealth stands out e.g. compared to the lack of comparative formations in Ganslmayer’s (2012: 1086) corpus of charters. In the neo-attestations, they are predominantly formed with -*isch*, e.g. *majestätisch* ‘majestic’, *stürmisch* ‘stormy, turbulent’, *kraftmännisch*,

[19] The absolute numbers are given in Table 9 in the Appendix.

lit. “bruiser-ish” ‘uncivil, brutal’. Another frequent group is formed by possessive formations (‘with X’). This class contains rather creative words derived with the rare suffix *-icht* (which is extinct by now), e.g. *mauseköpficht*, lit. “mouse-head-y”, ‘having the head of a mouse’; *wohlwangicht*, lit. “well-cheek-y”, ‘round-cheeked’, *grünhöslet* ‘green-trousered’. Both functional classes have in common that they denominate characteristic qualities of humans or other agents.

In the scientific texts, *-isch*, learned, and “other” suffixes play significant roles. The former two both bear witness to language contact having shaped this genre (indeed, most scientific texts were still written in Latin at the time, Durrell 2016: 222): *-isch* frequently combines with loan words, e.g. in an identificational function (*antiquitätisch*, *epidemisch* ‘being antique/epidemic’; see Kempf 2017, Kempf & Eitelmann 2018); learned suffixes are often used in technical terms, e.g. pertaining to geometry (*pyramidal*, *perpendikular*). In many or even in most cases, the adjectives with learned suffixes may have been borrowed as a whole. Nevertheless, they were included in the database, since their presence may have given rise to corresponding schemas with the potential of becoming productive (cf. Riehemann’s 1998, 2001*Type based derivational morphology*).

In the derivatives formed with “other” suffixes, there are two prominent themes: material and shape. Two old-established suffixes, *-e(r)n* and *-icht*, serve to express material composition, content or coveredness, e.g. *leimern* ‘of loam/adobe’, *salzicht* ‘salty’, *haaricht* ‘hairy’. Additionally, quite an interesting novel brand of suffixes appears in this domain: the complex suffix *-haltig* (roughly: “-contain-y”) expresses content, as in *kohlenstoffhaltig* ‘carbonic’, *zinnhaltig* ‘stannous’. Two other complex suffixes express general comparison (*-artig* ‘-like’) and comparison in shape (*-förmig* ‘-shaped’), e.g. in *tonartig* ‘clay-like’ and *eiförmig* ‘egg-shaped’. In GerManC, these novel suffixes occur only in scientific texts; in the 19th and 20th centuries, they have gained productivity and spread to other genres as well, cf. Herrmann (2017: 23, 37, 43). These complex derivational patterns are particularly interesting with respect to their informational density: they are more explicit than traditional suffixes (cf. e.g. *salzig* ‘salty’ vs. *salzhaltig* ‘saliferous’), yet shorter than corresponding relative clauses (*that contains salt*). With these properties, they are well suited for the needs of conceptually written language – including brevity and unambiguousness (cf. e.g. Koch & Oesterreicher 1985: 21– 22). The specifics of newspapers will be discussed in the next section, since they pertain to semantic, textual, and syntactic aspects.

6. Semantic, syntactic, and textual implications

Newspapers display two very interesting characteristics. First, they prove an excellent breeding ground for adjectives derived from proper names (short: *deonymic adjectives*): among the 168 *-isch*-derivatives that are newly attested in this genre, 120 are derived from a proper name (*Engländisch* 'English', *Crakowisch* 'of Krakow', *Glandenbergische Buchhandlung* 'Glandenberg's bookstore'). As for town names, the GerManC data witness the rise of the rival suffix *-er* (*Warschauer* 'of Warsaw', *Pariser* 'Parisian', 19 neo-attestations in newspapers), which became dominant around 1800 (cf. Kempf 2017). The affinity to proper names can be explained by frequent references to individual people or places, which are typical of newspapers, while e.g. sermons may tend to describe generic circumstances instead.

In fact, there is an essential functional difference between the adjectives discussed above (denoting predispositions, character traits, consistencies) and deonymic adjectives in their typical use: they fall on either side of the dichotomy 'descriptive' vs. 'classifying' adjectives, or *quality* vs. *relational adjectives* (cf. e.g. Frevel & Knobloch 2005). Quality adjectives (like *bissig* 'snappish') typically can undergo comparison, be modified, be used predicatively (*sie war sehr bissig* 'she was very snappish'), and in German, they can also be used adverbially (*sie reagierte bissig* 'she reacted snappishly'). Relational adjectives tend to occur in attributive use only. They classify the noun they refer to (as in 'Glandenberg's bookstore') by establishing a relationship with the entities denoted by their bases (here: relation of ownership between base entity *Glandenberg* and modified noun 'bookstore'). The majority of deonymic adjectives attested in GerManC newspapers are relational adjectives.

The dichotomy of quality vs. relational adjectives corresponds (at least to some degree) to the two major functions that are often attributed to word formation (cf. Kastovsky 1986; Plag et al. 1999: 225): the essentially lexical function of labeling, i.e. creating a name for segments of extralinguistic reality vs. the essentially syntactic function of condensing information in order to reference things previously mentioned in the discourse. The deonymic adjectives in the corpus rather serve syntactic purposes, given that they semantically correspond to inflectional or syntactic constructions (cf. *Glandenberg's bookstore*; Example (1))

(1)	*Vor-gestr-ig-e*	*Poln-isch-e*	*und*	*Warschau-er*	*Brief-e*
	before-yesterday-y-PL	Polish-PL	and	Warsow-ian	letter-PL

'letters from Poland and Warsaw from the day before yesterday'
(GerManC, NEWS_P1_NoD_1666_berlin1)

Example (1) also illustrates the second tendency characteristic of newspapers: more than any other genre, newspapers employ complex adjectives for endophoric or exophoric reference (Halliday & Hasan 1976: 31– 36). The formation *vorgestrig* in (1) is one of numerous derived adjectives in GerManC newspapers that serve referential purposes rather than characterizing their noun. They are thus relational rather than quality adjectives, and they serve syntactic rather than lexical purposes. Table 4 shows the most frequent examples along with their token frequencies in the four genres as well as an assessment of their endophoric or exophoric referential usage. For instance, *nächstkünftig* ‘next, coming’ is generally used to refer to the coming month, night, etc. with respect to the time of writing – thus making an exophoric or extratextual reference. By contrast, the formation *sonstig* ‘other’ generally marks the last item of an enumeration (cf. [2]), i.e. endophoric/text-internal reference.

(2) *Da unterschiedliche getreue Unterthanen des Koͤnigs* […] *geneigt seyn duͤrften, ihre desfallsige Zufriedenheit durch Beleuchtungen, und sonstige Freudensbezeugungen gefließentlich zu offenbaren* […]
‘Since several faithful subjects of the king […] will likely be inclined to eagerly display their enthusiasm about this (lit. “their this-case-y enthusiasm”) by illuminations or other expressions of delight…’
(GerManC, NEWS_P3_OOD_1780_wien)

The distributions in Table 4 are quite interesting: while scientific and narrative texts clearly prefer expressions of endophoric reference, newspapers eagerly use both kinds of phoric adjectives (and sermons do so reluctantly). This reflects different communicative situations: exophoric expressions like *hiesig* ‘local, present’ or *baldig* ‘coming, soon-to-be’ can only be understood with reference to the place and time of writing (thus, they are deictic elements). By way of refering to the writer’s local and temporal situation, they boost the authenticity and immediacy of the news (for a more differentiated account see Lefèvre 2008). By contrast, scientific writing seeks to present objective information, valid beyond the immediate experience of the author. Similar observations are reported by Brendel et al. (1997: 607): they find a lack of deictic expressions in ENHG specialized texts, which they trace back to the written character of the genre. Narrative texts behave relatively similar (cf. Table 4), even though “hiding” the voice of the (fictive or real) narrator is certainly less of a goal here. What may prevent a usage as frequent as in newspapers is the fact that deictic expressions like *nächstkünftig* ‘next, coming’ can only be used when still applicable at the time of writing (cf. *Back then, I was expecting to see him next week* ‘the following week’). In sermons, anchoring the narrative in the place and time of utterance may even be a volitional technique (of connecting with the audience, cf. *unsrig* ‘our’); yet, this can be achieved by means other than derived adjectives. The latter may pertain too much to written style by virtue of being too condensed.

Phoric or relational adjectives as in Table 4 bear an immense potential of textual densification. All examples in Table 4 can be – and almost exclusively are – used in the middle field of a noun phase, i.e. attributively. Like the complex patterns discussed above (*-haltig* “-contain-y”, *salzhaltig* ‘that contains salt’), they can assume the function of complex syntactic structures, in particular of prepositional

Table 4. Endo- and exophoric adjectives in GerManC arranged by tokens in NEWS

Lemma	Translation	Base	Translation (base)	Endophoric	Exophoric	SERM	NARR	SCIE	NEWS
selbig	'the same'	*selb-*	'same'	X		4	18	16	74
hiesig	'local, present'	*hier*	'here'		X	3		4	66
vorig	'previous'	*vor*	'before'	X	X	6	20	22	41
dasig	'local'	*da*	'here/there'		X		1	8	19
dortig	'local'	*dort*	'there'		X				18
unsrig	'our'	*wir/uns*	'we/us'		X	4	1		10
bisherig	'hitherto existing'	*bisher*	'to date, hitherto'	X	X	2	1	2	8
auswärtig	'external'	*auswärts*	'out of town, outwards'		X				7
obig	'above'	*oben*	'above'	X		1	1	6	6
gestrig	'of yesterday'	*gestern*	'yesterday'		X	2	3	1	5
nächstkünftig	'next, coming'	*nächst + kommen*	'next', 'come'		X				4
baldig	'coming, soon-to-be'	*bald*	'soon'		X		2		3
nunmehrig	'present, current'	*nun + mehr*	'now', 'more'		X				3
sonstig	'other'	*sonst*	'else'	X					3
damalig	'former'	*damals*	'back then, at the time'		X	2	3	4	
SUM						**24**	**50**	**63**	**267**

phrases (*vorgestrig* ‘from the day before yesterday‘), but also e.g. of complement clauses (cf. (3)).

(3) *Gewißheit/ des Auffbruchs und baldiger Uberschiffung*
lit. “intelligence of the departure and “soon-y” crossing”
‘the information that they had left and would soon be traversing’
NEWS_P1_NoD_1666_berlin1

Example (3) illustrates the distinctly nominal style realized by phoric adjectives, i.e. the information is coded within heavy noun phrases, so that verbal elements can be omitted. This preference for nominal vs. verbal expression has been shown to characterize written registers in contrast to speech (cf. Wells 1960; Biber & Finegan 1997; Plag et al. 1999; Halliday 2004; Hartmann 2016: 261). German is known to have drifted in its history towards the pole of conceptual writtenness (in writing, and by way of transmission also in speech). This tendency is evidenced predominantly by syntactic developments, such as a growing inventory of subjunctions (cf. Szczepaniak 2015). The remainder of this section examines whether this tendency can be confirmed by results pertaining to adjectival derivation. Figure 8 surveys all tokens of suffixed adjectives in BonnC and GerManC.

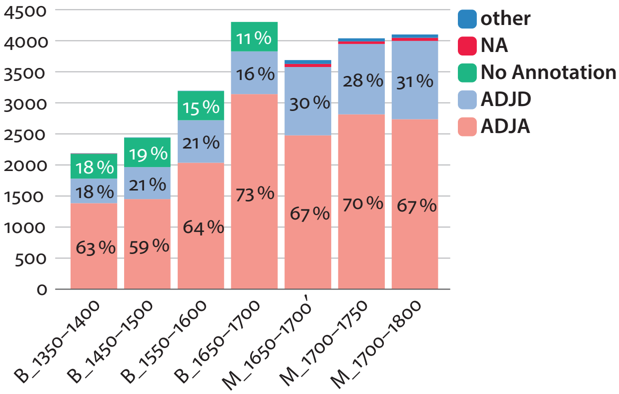


Figure 8. Frequency and syntactic functions of suffixed adjectives (tokens) in BonnC and GerManC[20]

[20] ADJA: attributive adjective; ADJD: adjective used adverbially or predicatively; NA: adjective used as noun; other: asyntactic use (e.g. in enumerations) or erroneous annotation; NoAnnotation: In BonnC, the annotation for ±adverbial use is sometimes missing; this annotation was added manually only for neo-attestations. These 160 cases contained 75% predicative, 20% adverbial, and 5% asyntactic, attributive or nominalized tokens. Extrapolating from that, “NoAnnotation” can be taken to predominantly contain “ADJD”.

The absolute numbers are given in Table 10 in the Appendix.

Interestingly, the total amount of tokens increases substantially throughout the diachronic subcopora – even though their size was kept approximately constant. Moreover, the share of attributive usage has increased. While the first tendency generally indicates a rise in morphological complexity, the second trend additionally points to noun phrases being incrementally loaded with information. Both developments taken together support the assumption of conceptual writtenness having advanced throughout the centuries.

This raises the question to what degree different genres have participated in this development. Figure 9 examines the four GerManC genres for the syntactic functions among all tokens (left panel) and among neo-attestations (right panel).

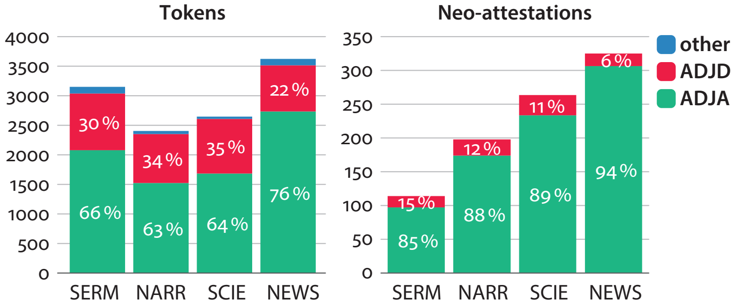


Figure 9. Frequency and syntactic functions of suffixed adjectives in four GerManC genres[21]

With respect to the number of tokens, newspapers are leading, followed by sermons. The high scores are partially caused by high-frequency words (that often are part of formulaic phrases): the items *königlich* 'kingly/of the king', *kaiserlich* 'imperial/of the emperor', and *französisch* 'French' (in newspapers) as well as *heilig* 'holy', *ewig* 'eternal', and *göttlich* 'divine' (in sermons) yield well over a hundred tokens each, while narrative and scientific texts do not contain any items that attain 70 tokens or more. In newspapers, the next items in terms of token frequency are the phoric elements *selbig* 'the same' and *hiesig* 'local, present' (cf. Table 4), while in sermons, many more "formulaic" items follow (i.e. items that typically occur as part of a formulaic phrase, e.g. the items *selig* 'blessed', *himmlisch* 'celestial', which occur in phrases like *selig entschlafen* 'to pass away peacefully' or *der himmlische Vater* 'heavenly father'). Thus, one may conclude that sermons do not actually participate in the written trend of nominal densificiation, despite their high token score. As concerns the syntactic use of adjectival derivatives, newspapers are distinguished by a higher proportion of attributive uses – which can largely be

[21] The absolute numbers are given in Table 11 and Table 12 in the appendix.

attributed to relational adjectives, i.e. phoric elements and derivatives from proper names, as discussed above.

Interestingly, attributives reach a dramatically higher share when only neo-attestations are counted (this holds for all four genres, but sermons lag behind, while newspapers are in the lead once more). Even though neo-attestations do not equal actual new coinages, they do represent a more recent layer of derivatives (than the unfiltered tokens). It is, of course, possible that a novel word is coined in predicative usage but attested first in an attributive position. Nevertheless, I take the results to be indicative, especially seeing that relational adjectives generally are restricted to attributive usage. If these assumptions are correct, the results in Figure 9 can be taken to indicate that attributive usage is indeed the key domain of new adjectival word formation. This, in turn, would support the view that adjectival derivation is increasingly used to accommodate information within noun phrases.

7. Discussion and conclusion

The empirical results about the four genres under investigation can be summarized as follows: based on quantitative productivity measures as well as on long term diachronic comparison with data from mixed-genre corpora, the genres ‘scientific texts’ and ‘newspapers’ stand out as being progressive. They exhibit most signs of productive word formation (which is intrinsically progressive in the sense that it involves novel words, but also conforms to the long term trend to make more use of word formation) and they display comparatively modern suffix distributions. Among those two genres, scientific texts stand out in the distribution of suffixes: they are ahead in using learned suffixes and are the first genre to exhibit complex suffixational patterns like *-haltig* ‘-containing’ – thus contributing to nominal densification. Newspapers, in turn, stand out in using relational adjectives and in contributing to nominal densification as measured by the rate of attributive tokens. The view of Early Modern German newspapers as a genre close to conceptual orality (cf. Durrell et al. 2012) thus has to be relativized. On the one hand, they can be said to display situationality or involved style – typical features of conceptual orality – by using deictic elements that point to the local and temporal situation of the writer; on the other hand, they use exceedingly long sentences (Polenz 2013: 398) and show strong tendencies to nominal style, as illustrated in (1)– (3). Writers of early newspapers were influenced by chancery style on the one hand and by the need to publish fast on the other hand. Presumably, newspapers in German (as in English, cf. Hundt & Mair 1999) have experienced colloquialization throughout the centuries, as the influence of chancery style was fading. This aspect is certainly worthy of further investigation.

Narrative prose texts display fairly average behavior in the measures applied here, but they do appear to use the most diverse vocabulary. This can be attributed to a large topical diversity but also to an elaborate choice of words, seeing that they use some interesting rare formations (*wohlwangicht* 'round-cheeked'). The status of sermons is somewhat ambivalent: their low type values can be accounted for by topical homogeneity; but their productivity values are low independent of this and their distribution of suffixational pattern appears antiquated compared to the general diachronic trends. To some degree, this can be accounted for by the semi-oral character of the genre. Another factor is the frequent usage of formulaic elements, which points towards a conservative character. This ambivalent status of sermons in Early Modern German contrasts with the results in Pickl's (2017) study on the development of negation in Middle High German, where sermons prove progressive compared to literary texts (2017: 3, 33, 36, 38). This contrast may be caused by different genres serving as a standard of comparison and/or by differences in the linguistic domains and periods under investigation. Certainly, the impact of Luther's work in the 16th century is a factor here (in the sense that sermon writers in subsequent centuries were able to build on that tradition and thus may have felt little need for innovation).

Overall, the results of this study show that genre constitutes a tremendously influential factor with respect to word formation and word formation change. This needs to be taken into account in studies of word formation as well as in the design of (diachronic) corpora. Conversely, word formation contributes strongly to genre variation and the evolution of genre profiles. Thus, word formation needs to be considered when advancing research of genre variation.

Further, the results of the present study challenge the idea that linguistic change happens in spoken rather than in written varieties, or that spoken varieties tend to be more progressive, while writing is more conservative. This may be true for some aspects, but does not hold for (adjectival) word formation (in German). The dichotomy of progressive vs. conservative (or "agile" vs. "uptight", Hundt & Mair 1999) has to be refined. Particular attention should be paid to different linguistic domains and to the idea that they need not necessarily align: a given genre can be agile in its syntactic behavior, yet uptight as regards word formation. Moreover, the standard of comparison should be reflected upon carefully: with respect to what other genre(s) do we classify a given genre as progressive or conservative, and is this comparison really valid? It may prove useful to distinguish explicitly between two types of progressive behavior: a broader notion of "progressive" could refer simply to a genre being prone to change – as opposed to remaining unaltered, irrespective of what happens in other genres. By contrast, a more narrow notion of progressive behavior could be used to refer to genres showing developments that later spread to other genres as well (as it was shown here for scientific writing, which featured learned and complex suffixes before other genres eventually joined in).

Acknowledgments

I would like to thank Richard J. Whitt as well as two anonymous reviewers for their helpful and constructive comments on earlier versions of this paper. I would also like to thank Mehmet Aydın for proofreading this paper. All remaining errors are, of course, mine.

References

Corpora and online sources

BonnC <http://www.korpora.org/Fnhd/>
GerManC <http://www.ota.ox.ac.uk/desc/2544>

References

Ágel, Vilmos & Hennig, Mathilde. 2007. Überlegungen zur Theorie und Praxis des Nähe- und Distanzsprechens. In *Zugänge zur Grammatik der gesprochenen Sprache*, Vilmos Ágel & Mathilde Hennig (eds), 179–214. Tübingen: Niemeyer.

Baayen, R. Harald. 2009. Corpus linguistics in morphology. Morphological productivity. In *Corpus Linguistics. An International Handbook*, Vol. 2, Anke Lüdeling & Merja Kytö (eds), 899–919. Berlin: De Gruyter. https://doi.org/10.1515/9783110213881.2.899

Baayen, R. Harald & Lieber, Rochelle. 1991. Productivity and English derivation: A corpus based study. *Linguistics* 29: 801–843. https://doi.org/10.1515/ling.1991.29.5.801

Bentzinger, Rudolf. 1987. Zur Verwendung von Adjektivsuffixen in Erfurter Frühdrucken. In *Zum Sprachwandel in der deutschen Literatursprache des 16. Jahrhunderts. Studien – Analysen – Probleme* [Bausteine zur Sprachgeschichte des Neuhochdeutschen 63], Joachim Schildt (ed.), 151–266. Berlin: Akademie-Verlag.

Bentzinger, Rudolf. 1992. Zur Verwendung von Adjektivsuffixen in der deutschen Literatursprache (1570–1730). In *Aspekte des Sprachwandels in der deutschen Literatursprache 1570–1730* [Bausteine zur Sprachgeschichte des Neuhochdeutschen 66], Joachim Schildt (ed.), 120–225. Berlin: Akademie-Verlag.

Berg, Kristian. 2016. Historische Produktivität: Die Lebensdauer von Wortbildungen. Paper presented at Historische Wortbildung. Theorie – Methoden – Perspektiven, University of Münster, 25–26 November.

Bergmann, Rolf. 1998. Autonomie und Isonomie der beiden Wortbildungssysteme im Deutschen. *Sprachwissenschaft* 23(2), 167–183.

Biber, Douglas & Finegan, Edward. 1997. Diachronic relations among speech-based and written registers in English. In *To Explain the Present Studies in the Changing English Language in Honour of Matti Rissanen*, Terttu Nevalainen & Leena Kahlas-Tarkka (eds), 253–275. Helsinki: Société Néophilologique.

Braun, Christian. 2011. *Kanzleisprache auf dem Weg zum Neuhochdeutschen* [Beiträge zur Kanzleisprachenforschung 7]. Wien: Praesens.

Brendel, Bettina, Frisch, Regina, Moser, Stephan & Wolf, Norbert Richard. 1997. *Wort- und Begriffsbildung in frühneuhochdeutscher Wissensliteratur. Substantivische Affixbildung* [Wissensliteratur im Mittelalter 26]. Wiesbaden: Reichert.

Cowie, Claire S. 1999. Diachronic Word Formation: A Corpus-based Study of Derived Nominalizations in the History of English. PhD thesis, Cambridge University.

Doerfert, Regina. 1994. *Die Substantivableitung mit -heit, -keit, -ida, -î im Frühneuhochdeutschen* [Studia Linguistica Germanica 34]. Berlin: De Gruyter. https://doi.org/10.1515/9783110877496

Durrell, Martin & Whitt, Richard J. 2013. Zum Abbau der regionalen Varianten im Standardisierungsprozess 1650–1800. Belege aus dem GerManC-Korpus. In *Akten des XII. internationalen Germanistenkongresses Warschau 2010: Vielheit und Einheit der Germanistik weltweit*, Vol. 17, Franciszek Grucza (ed.), 107–111. Frankfurt: Peter Lang.

Durrell, Martin, Bennett, Paul, Scheible, Silke & Whitt, Richard J. 2012. *The GerManC Corpus: A Representative, Multi-Genre Corpus of Early Modern German, 1650–1800*. University of Manchester. <http://ota.ox.ac.uk/desc/2544> (7 January 2017).

Durrell, Martin. 2016. Textsortenspezifische und regionale Unterschiede bei der Standardisierung der deutschen Sprache. In *PerspektivWechsel oder: Die Wiederentdeckung der Philologie, Vol. 1: Sprachdaten und Grundlagenforschung in der Historischen Linguistik*, Sarah Kwekkeboom & Sandra Waldenberger (eds), 211–231. Berlin: Schmidt.

Ebert, Robert Peter. 1988. Variation in the position of the attributive genitive in sixteenth-century German. *Monatshefte für deutschen Unterricht, Deutsche Sprache und Literatur* 80: 32–49.

Evert, Stefan & Baroni, Marco. 2007. zipfR: Word frequency distributions in R. In Proceedings of the *45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Sessions*, 29–32. Prag: Association for Computational Linguistics.

Flury, Robert. 1964. Struktur- und Bedeutungsgeschichte des Adjektiv-Suffixes -bar. PhD dissertation, University of Zürich.

Frevel, Claudia & Knobloch, Clemens. 2005. Das Relationsadjektiv. In *Wortarten und Grammatikalisierung: Perspektiven in System und Erwerb* [Linguistik – Impulse & Tendenzen 12], Clemens Knobloch & Burkhard Schaeder (eds), 151–175. Berlin: De Gruyter.

Gaeta, Livio & Ricca, Davide. 2006. Productivity in Italian word formation. A variable-corpus approach. *Linguistics* 44: 57–89. https://doi.org/10.1515/LING.2006.003

Ganslmayer, Christine. 2012. *Adjektivderivation in der Urkundensprache des 13. Jahrhunderts. Eine historisch-synchrone Untersuchung anhand der ältesten deutschsprachigen Originalurkunden* [Studia Linguistica Germanica 97]. Berlin: De Gruyter. https://doi.org/10.1515/9783110213614

Ganslmayer, Christine & Müller, Peter O. 2016. Historische Fremdwortbildung – Forschungsstand und Perspektiven. Paper presented at Historische Wortbildung. Theorie – Methoden – Perspektiven, University of Münster, 25–26 November.

Gersbach, Bernhard & Graf, Rainer. 1985. *Wortbildung in gesprochener Sprache. Die Substantiv-, Verb- und Adjektiv-Zusammensetzungen und Ableitungen im "Häufigkeitswörterbuch gesprochener Sprache"*, Bd. II. Tübingen: Niemeyer.

Habermann, Mechthild. 1994. *Verbale Wortbildung um 1500. Eine historisch-synchrone Untersuchung anhand von Texten Albrecht Dürers, Heinrich Deichslers und Veit Dietrichs* [Wortbildung des Nürnberger Frühneuhochdeutsch 2]. Berlin: De Gruyter.

Halliday, Michael A. K. 2004. *The Language of Science. Collected Works of M. A. K. Halliday*, Vol. 5, Jonathan J. Webster (ed.). London: Continuum.

Halliday, Michael A. K. & Hasan, Ruqaiya. 1976. *Cohesion in English* [English Language 9]. London: Longman.

Hartmann, Stefan. 2016. *Wortbildungswandel. Eine diachrone Studie zu deutschen Nominalisierungsmustern* [Studia Linguistica Germanica 125]. Berlin: De Gruyter. https://doi.org/10.1515/9783110471809

Herrmann, Maria. 2017. Produktivitätsentwicklung reihenbildender relationaler Zweitglieder als Adjektivsuffixe im 19. und 20. Jahrhundert. MA thesis, University of Mainz.

Hölscher, Sandra. 2011. *Familienanzeigen. Zur Geschichte der Textsorten Geburts-, Verbindungs- und Todesanzeige, ihrer Varianten und Strukturen in ausgewählten regionalen und überregionalen Tageszeitungen von 1790–2002* [Berliner sprachwissenschaftliche Studien 23]. Berlin: Weidler.

Hundt, Marianne & Mair, Christian. 1999. "Agile" and "uptight" genres: The corpus-based approach to language change in progress. *International Journal of Corpus Linguistics* 4(2): 221–242. https://doi.org/10.1075/ijcl.4.2.02hun

Jäger, Agnes. 2008. *History of German Negation* [Linguistik Aktuell/Linguistics Today 118]. Amsterdam: John Benjamins. https://doi.org/10.1075/la.118

Kastovsky, Dieter. 1986. The problem of productivity in word formation. *Linguistics* 24: 585–600. https://doi.org/10.1515/ling.1986.24.3.585

Kempf, Luise. 2010. Warum die Unterscheidung fremd–nativ in der deutschen Wortbildung nicht obsolet ist. In Carmen Scherer & Anke Holler (eds.), *Strategien der Integration und Isolation nicht-nativer Einheiten und Strukturen* [Linguistische Arbeiten 253], 123–142. Berlin/New York: De Gruyter.

Kempf, Luise. 2016. *Adjektivsuffixe in Konkurrenz: Wortbildungswandel vom Frühneuhochdeutschen zum Neuhochdeutschen* [Studia Linguistica Germanica 126]. Berlin: De Gruyter.

Kempf, Luise. 2017. Englåndisch, Hamburgisch, Lutherisch – Degrees of onymicity reflected in the history of German *-isch*-derivation. *Folia Linguistia* 51(2): 391–417.

Kempf, Luise. forthcoming. Methoden der Produktivitätsmessung in diachronen Korpusstudien. In *Historische Wortbildung. Theorie – Methoden – Perspektiven* [Germanistische Linguistik], Christian Schwarz & Christine Ganslmayer (eds.). Berlin: De Gruyter.

Kempf, Luise & Matthias Eitelmann. 2018. Von diutisk zu dynamisch, von englisc zu anythingish. -is(c)h kontrastiv diachron. *Zeitschrift für Wortbildung/Journal of Word Formation (ZWJW)* 2018(1): 93–134.

Kopf, Kristin. 2018. *Fugenelemente diachron: Eine Korpusuntersuchung zu Entstehung und Ausbreitung der verfugenden N+N-Komposita* [Studia Linguistica Germanica 133]. Berlin, New York: De Gruyter.

Klein, Thomas, Solms, Hans-Joachim & Wegera, Klaus-Peter. 2009 *Mittelhochdeutsche Grammatik, 3: Wortbildung*. Tübingen: Niemeyer.

Koch, Peter & Oesterreicher, Wulf. 1985. Sprache der Nähe - Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch* 36: 15–43.

Kopf, Kristin. 2016. Von der Syntax in die Wortbildung. Zur Diachronie der verfugenden N+N-Komposition. PhD dissertation, University of Mainz.

Kronenberger, Kerstin. 2002. Die Substantivableitung mit -e, -ede und -heit in der Urkundensprache des 13. Jahrhunderts. In *Historische Wortbildung des Deutschen* [Reihe Germanistische Linguistik 232], Mechtild Habermann, Peter O. Müller & Horst Haider Munske (eds), 193–210. Tübingen: Niemeyer. https://doi.org/10.1515/9783110940756.193

Kühnhold, Ingeburg, Putzer, Oskar & Wellmann, Hans. 1978. *Deutsche Wortbildung. Typen und Tendenzen in der Gegenwartssprache, Teil 3: Das Adjektiv. Eine Bestandsaufnahme* *des Instituts für Deutsche Sprache, Forschungsstelle Innsbruck* [Sprache der Gegenwart 43]. Berlin: De Gruyter.

Lefèvre, Michel. 2008. Kontrastive Untersuchung zu (*d-*)*selb*(*ig*)- und anderen Einheiten des Wiederaufgreifens im 17. Jahrhundert. Ein systemischer Ansatz. In *Die Formen der Wiederaufnahme im älteren Deutsch* [Berliner sprachwissenschaftliche Studien 10], Yvonne Desportes (ed.), 289–306. Berlin: Weidler.

Linke, Angelika. 2001. Trauer, Öffentlichkeit und Intimität. Zum Wandel der Textsorte "Todesanzeige" in der zweiten Hälfte des 20. Jahrhunderts. In *Zur Kulturspezifik von Textsorten* [Textsorten 3], Ulla Fix, Stephan Habscheid & Josef Klein (eds), 195–223. Tübingen: Stauffenburg.

Müller, Peter O. 1993. *Substantiv-Derivation in den Schriften Albrecht Dürers. Ein Beitrag zur Methodik historisch-synchroner Wortbildungsanalysen* [Wortbildung des Nürnberger Frühneuhochdeutsch 1]. Berlin: De Gruyter. https://doi.org/10.1515/9783110859072

Pickl, Simon. 2017. Neues zur Entwicklung der Negation im Mittelhochdeutschen. Grammatikalisierung und Variation in oberdeutschen Predigten. *Beiträge zur Geschichte der deutschen Sprache und Literatur (PBB)* 139(1): 1–46. https://doi.org/10.1515/bgsl-2017-0001

Plag, Ingo, Dalton-Puffer, Christiane & Baayen, Harald. 1999. Morphological productivity across speech and writing. *English Language and Linguistics* 3(2): 209–228. https://doi.org/10.1017/S1360674399000222

von Polenz, Peter. 2013. *Deutsche Sprachgeschichte vom Spätmittelalter bis zur Gegenwart, Band II: 17. und 18. Jahrhundert*, 2. Auflage, bearbeitet von Claudine Moulin unter Mitarbeit von Dominic Harion. [De Gruyter Studium]. Berlin: De Gruyter.

Pounder, Amanda. 2000. *Process and Paradigms in Word Formation Morphology*. Berlin & New York: De Gruyter.

Prell, Heinz-Peter & Schebben-Schmidt, Marietheres. 1996. *Die Verbableitung im Frühneuhochdeutschen*. Berlin: De Gruyter. https://doi.org/10.1515/9783110817218

Prell, Heinz-Peter. 1991. *Die Ableitung von Verben aus Substantiven in biblischen und nichtbiblischen Texten des Frühneuhochdeutschen* [Europäische Hochschulschriften 1274]. Frankfurt: Peter Lang.

Prell, Heinz-Peter. 2000. Die Stellung des attributiven Genitivs im Mittelhochdeutschen. Zur Notwendigkeit einer Syntax mittelhochdeutscher Prosa. *Beiträge zur Geschichte der deutschen Sprache und Literatur* 122(1): 23–39. https://doi.org/10.1515/bgsl.2000.122.1.23

Riehemann, Susanne Z. 1998. Type-based derivational morphology. *The Journal of Comparative Germanic Linguistics* 2(1): 49–77. https://doi.org/10.1023/A:1009746617055

Riehemann, Susanne Z. 2001. A Constructional Approach to Idioms and Word Formation. PhD dissertation, Stanford University.

Ring, Uli. 2008. *Substantivderivation in der Urkundensprache des 13. Jahrhunderts. Eine historisch-synchrone Untersuchung anhand der ältesten deutschsprachigen Originalurkunden* [Studia Linguistica Germanica 96]. Berlin: De Gruyter. https://doi.org/10.1515/9783110212433

Scheible, Silke, Whitt, Richard J., Durrell, Martin & Benett, Paul. 2011. Investigating diachronic grammatical variation in Early Modern German. Evidence from the GerManC corpus. In *Grammatik und Korpora* 2009 [Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache 1], Marek Konopka, Jaqueline Kubczak, Christian Mair, František Štícha & Ulrich H. Waßner (eds), 539–549. Tübingen: Narr.

Scherer, Carmen. 2005. *Wortbildungswandel und Produktivität. Eine empirische Studie zur nominalen '-er'-Derivation im Deutschen* [Linguistische Arbeiten 497]. Tübingen: Niemeyer. https://doi.org/10.1515/9783110914887

Schulz, Matthias. 2002. Wortbildung in Wörterbüchern und Texten des 17. Jahrhunderts. In *Historische Wortbildung des Deutschen* [Reihe Germanistische Linguistik 232], Mechthild Habermann, Peter Müller & Horst Haider Munske (eds), 269–288. Tübingen: Niemeyer. https://doi.org/10.1515/9783110940756.269

Schulz, Matthias. 2007. *Deutscher Wortschatz im 17. Jahrhundert. Methodologische Studien zu Korpustheorie, Lexikologie und Lexikographie von historischem Wortschatz* [Reihe Germanistische Linguistik 278]. Tübingen: Niemeyer.

Solms, Hans-Joachim & Wegera, Klaus-Peter. 1998. Das Bonner Frühneuhochdeutsch-Korpus. Rückblick und Perspektiven. In *Probleme der Textauswahl für einen elektronischen Thesaurus. Beiträge zum ersten Göttinger Arbeitsgespräch zur historischen deutschen Wortforschung, 1. und 2. November 1996, Stuttgart*, Rolf Bergmann (ed.), 22–39. Leipzig, Hirzel.

Szczepaniak, Renata. 2015. Syntaktische Einheitenbildung ‐ typologisch und diachron betrachtet. In *Handbuch Satz, Äußerung, Schema* [Handbücher Sprachwissen (HSW) 4], Christa Dürscheid & Jan Georg Schneider (eds), 104–124. Berlin: De Gruyter. https://doi.org/10.1515/9783110296037-006

Thomas, Barbara. 2002. *Adjektivderivation im Nürnberger Frühneuhochdeutsch um 1500. Eine historisch-synchrone Analyse anhand von Texten Albrecht Dürers, Veit Dietrichs und Heinrich Deichslers* [Wortbildung des Nürnberger Frühneuhochdeutsch 3]. Berlin: De Gruyter. https://doi.org/10.1515/9783110896176

Wegera, Klaus-Peter & Solms, Hans-Joachim 2002. Wortbildung des Mittelhochdeutschen. Zur Methode und zum Stand ihrer Erforschung, dargestellt am Beispiel der Diminutive. In *Historische Wortbildung des Deutschen* [Reihe Germanistische Linguistik 232], Mechthild Habermann, Peter O. Müller & Horst Haider Munske (eds), 159–170. Tübingen: Niemeyer. https://doi.org/10.1515/9783110940756.159

Wells, Rulon. 1960. Nominal and verbal style. In Style in language, Thomas A. Sebeok (ed.), 213–220. Cambridge MA: Technology Press; New York NY: Wiley.

Whitt, Richard J. 2018. Evidentiality and propositional scope in early modern German. *Journal of Historical Pragmatics* 19(1): 123–150. https://doi.org/10.1075/jhp.00013.whi

Wich-Reif, Claudia & Simmler, Franz. 2005. *Syntax, Althochdeutsch – Mittelhochdeutsch: eine Gegenüberstellung von Metrik und Prosa. Akten zum internationalen Kongress an der Freien Universität Berlin 26. bis 29. Mai 2004* [Berliner sprachwissenschaftliche Studien 7]. Berlin: Weidler.

Winkler, Gertraud. 1995. *Die Wortbildung mit -lich im Alt-, Mittel- und Frühneuhochdeutschen*. Heidelberg: Winter.

Ziegler, Arne. 2010. *Historische Textgrammatik und Historische Syntax des Deutschen. Traditionen, Innovationen, Perspektiven, 2 Vols: 1. Diachronie, Althochdeutsch, Mittelhochdeutsch; 2. Frühneuhochdeutsch, Neuhochdeutsch*. Berlin: De Gruyter. https://doi.org/10.1515/9783110219944

Appendix

Table 5. Type measures of suffixational patterns in different studies (Part I)

Suffix	Category	Ganslmayer (2012)	Kempf (2016)	Kempf (2016)	Thomas (2002)	Kempf (2016)	Schulz (2002, 2007)
		1200–1300	1350–1400	1450–1500	1488–1550	1550–1660	1600–1700
-abel	learned						
-al	learned						
-ant	learned						
-är	learned						
-artig	other						
-at	learned						
-bar	other	23	8	12	10	16	38
-ell	learned						
-e	other	13					
-echtig	other				1		
-er	other			1		2	
-ern	other	23	10	13	17	14	
-esk	learned						
-fach	other			1	9	1	
-valt	other	2					
-förmig	other						
-haft	other	26	9	7	12	10	26
-haftig	other				2		
-haltig	other						
-icht	other	3	5	11	42	17	87
-(is)iert	learned						
-ig	-ig	181	119	143	199	175	522
-il	learned						
-isch	-isch	26	15	24	67	84	356
-iv	learned						
-lich	-lich	264	156	184	177	172	497
-los	other	14	4	1	5	6	
-mäßig	other	2					

Table 5. (*Continued*)

Suffix	Category	Ganslmayer (2012)	Kempf (2016)	Kempf (2016)	Thomas (2002)	Kempf (2016)	Schulz (2002, 2007)
		1200–1300	**1350–1400**	**1450–1500**	**1488–1550**	**1550–1660**	**1600–1700**
-ös	learned						
-oid	learned						
-reich	other				5		
-sam	other	10	9	9	12	14	
-selig	other			1	3	6	
sum		**587**	**335**	**407**	**561**	**517**	**1526**

Table 6. Type measures of suffixational patterns in different studies (Part II)

Suffix	Category	Kempf (2016)	Kempf (2016)	Kempf (2016)	Kempf (2016)	Kühnhold et al. (1978)
		1650–1700	**1650–1700′**	**1700–1750**	**1750–1800**	**1900–1975**
-abel	learned		5	10	4	73
-al	learned		4	4	3	
-ant	learned		4	13	4	96
-är	learned		3	5	2	124
-artig	other				7	
-at	learned		2	3	3	
-bar	other	25	17	19	25	389
-ell	learned			1	4	403
-e	other					
-echtig	other					
-er	other	7	9	7	13	
-ern	other	13	19	24	14	159
-esk	learned					28
-fach	other	2	2	2	5	
-valt	other					
-förmig	other				6	
-haft	other	17	18	17	24	429
-haftig	other					

(*Continued*)

Table 6. (*Continued*)

-haltig	other				3	
-icht	other	10	17	25	27	
-(is)iert	learned					29
-ig	-ig	205	206	198	254	2008
-il	learned		1	2	1	
-isch	-isch	140	175	208	218	1387
-iv	learned		2	2	3	251
-lich	-lich	232	205	221	227	940
-los	other	14	5	2	19	413
-mäßig	other	1	7	4	8	206
-ös	learned		2	2	1	131
-oid	learned					40
-reich	other					
-sam	other	16	15	17	21	64
-selig	other	7	8	8	8	19
sum		**689**	**726**	**794**	**904**	**7189**

Table 7. Distribution of suffixational patterns (types) in four GerManC genres and spoken 20th century German

	ig	**isch**	**learned**	**lich**	**other**	**sum**
SERM	200	60	11	195	87	**553**
NARR	196	136	18	195	117	**662**
SCIE	162	145	29	170	147	**653**
NEWS	174	239	32	202	105	**752**
spoken 20th ct	184	89	31	134	84	**522**

Table 8. Distribution of suffixational patterns across neo-attestations in four GerManC genres

	ig	**isch**	**learned**	**lich**	**other**	**sum**
SERM	38	20	9	26	21	**114**
NARR	33	73	13	40	40	**199**
SCIE	49	84	25	35	72	**265**
NEWS	48	168	29	37	46	**328**
sum	**168**	**345**	**76**	**138**	**179**	**906**

Table 9. Distribution of suffixational patterns across exclusive formations (types) in four GerManC genres

	ig	isch	learned	lich	other	sum
SERM	29	23	5	24	14	95
NARR	24	69	12	26	40	**171**
SCIE	44	78	23	18	65	**228**
NEWS	33	161	25	26	47	**292**

Table 10. Syntactic use in BonnC and GerManC measured in tokens

	ADJA	ADJD	NoAnnotation	NA	other	sum
B_1350–1400	1386	396	404	2	1	**2189**
B_1450–1500	1452	519	474		2	**2447**
B_1550–1600	2039	688	473		1	**3201**
B_1650–1700	3150	688	479			**4317**
M_1650–1700′	2484	1103		49	62	**3698**
M_1700–1750	2823	1136		42	51	**4052**
M_1750–1800	2745	1265		47	57	**4114**
sum	**16079**	**5795**	**1830**	**140**	**174**	**24018**

Table 11. Syntactic use of suffixed adjectives in four GerManC genres (all tokens)

	ADJA	ADJD	other	sum
SERM	2085	963	110	**3158**
NARR	1530	832	52	**2414**
SCIE	1693	924	40	**2657**
NEWS	2744	783	106	**3633**
sum	**8052**	**3502**	**308**	**11862**

Table 12. Syntactic use of suffixed adjectives in four GerManC genres (neo-attestations only)

	ADJA	ADJD	sum
SERM	97	17	**114**
NARR	174	24	**198**
SCIE	234	30	**264**
NEWS	307	19	**326**
sum	**812**	**90**	**902**

Index

This volume provides a state-of-the-art overview of the intersecting fields of corpus linguistics, historical linguistics, and genre-based studies of language usage. Papers in this collection are devoted to presenting relevant methods pertinent to corpus-based studies of the connection between genre and language change, linguistic changes that occur in particular genres, and specific diachronic phenomena that are influenced by genre factors to greater and lesser degrees. Data are drawn from a number of languages, and the scope of the studies presented here is both short- and long-term, covering cases of recent change as well as more long-term alterations.

ISBN 978 90 272 0148 5

9 789027 201485

JOHN BENJAMINS PUBLISHING COMPANY