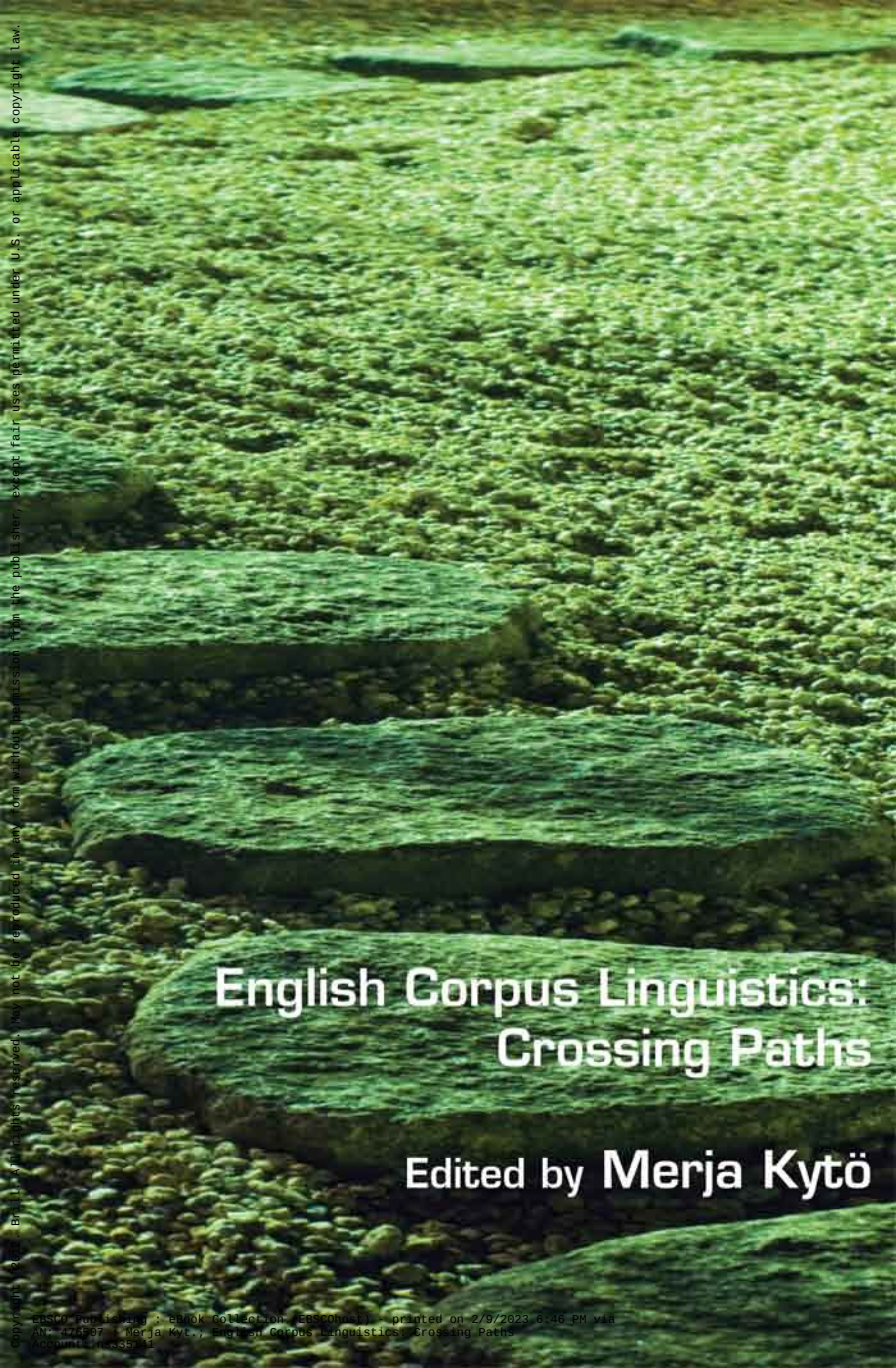


English Corpus Linguistics: Crossing Paths

LANGUAGE AND COMPUTERS: STUDIES IN PRACTICAL LINGUISTICS

No 76

edited by
Christian Mair
Charles F. Meyer
Nelleke Oostdijk

English Corpus Linguistics: Crossing Paths

Edited by
Merja Kytö

Amsterdam - New York, NY 2012

Cover image: www.dreamstime.com

Cover design: Inge Baeten

The paper on which this book is printed meets the requirements of "ISO 9706:1994, Information and documentation - Paper for documents - Requirements for permanence".

ISBN: 978-90-420-3518-8
E-Book ISBN: 978-94-012-0793-5

Printed in The Netherlands

To the memory of Stig Johansson,
a founding father of English corpus linguistics

Contents

Introduction

Merja Kytö

Uppsala University

It is probably not an exaggeration to say that corpus linguistics is a methodology that enjoys an ever increasing popularity world-wide today. English corpus linguistics has been an influential area in this respect, showing the way since the 1970s and 1980s, when technological advances had begun to enable researchers to process vast amounts of text stored in electronic form with speed and efficiency (Johansson 2008: 33). Moreover, it is not only linguists that find the approach increasingly attractive: corpus linguistic methodology and language analyses are nowadays applied to fields beyond linguistics proper. Professionals profiting from techniques developed in corpus linguistics include historians, experts in law, literary critics, computer scientists and language teachers. Individuals representing a variety of research interests, institutional backgrounds and disciplinary affiliations find it useful to search for evidence in electronic text collections, for instance, in the form of collocations or co-occurrence patterns (Wynne 2010: 425).

The contributions to the present book reflect aspects of the dynamic development in English corpus linguistics today. They highlight some of the fundamental issues in the corpus linguistic approach, and also throw light on patterns in Present-day English from the cross-linguistic perspective, on stylistic trends in recent English, and on aspects of the rich variation and long-term change characteristic of early English.

Two issues have received special attention across the chapters. Firstly, the volume and diversity of digitized material available for English corpus linguists today is impressive, and the contributors were encouraged to comment on the characteristics and potential of the corpus (corpora) or database(s) that they used for their studies, as well as on the significance of such and similar sources to corpus linguistics and/or literary computing methodology. These two related areas share many affinities but also differ in how they approach the notion of a corpus (architecture, annotation issues, representativeness, exploitation strategies, etc.).

Secondly, much still remains to be done to communicate the benefits of the corpus linguistic approach to those working in disciplines other than linguistics. There is an urgent call for more communication and collaboration across subjects and research areas (for discussion, see Curzan in this volume). In many of the contributions, explicit mention is made of the potential that the advances made in English corpus linguistics have to other disciplines and of the ways in which it would be possible to increase interdisciplinary effects by making the work done in the field more approachable to those working within other disciplines.

The book is divided into three parts. Part I is devoted to methodological issues (Curzan; Meyer), Part II turns to studies of Present-day and recent English (Johansson; Leech, Smith and Rayson), and Part III highlights work carried out on early English (Brinton, Claridge, Kohnen, Lancashire, Rissanen, and Traugott).

In Part I, Anne Curzan looks into the various ways in which corpus linguists might promote the use of corpora in sub-disciplines of English literature and language study, and also across other disciplines. The starting-point for her discussion is intriguing: if Toni Morrison, the Nobel Prize and Pulitzer Prize-winning American novelist, editor, and professor, were a corpus linguist, how might she profit from corpus linguistic techniques when wishing to establish how an Africanist presence and persona were created in American literature? That texts, literary and nonliterary, are increasingly available in electronic form has brought completely new challenges to those working on English literary and language studies, and the chapter takes stock of the options available.

In his chapter, Charles F. Meyer contrasts automated methods of data collection and analysis with close analyses of example data familiar to us from the great grammarians' work in the pre-electronic era. The danger of us being able to collect vast amounts of data from a variety of electronic sources may be that no attention is paid to close analysis of data when, ideally, today's corpus linguist should both benefit from the unprecedented access to diverse data sources at various times in the history of English and also from careful close analyses. An illustration in empirical terms is provided in a study of gapping phenomena (a type of coordination ellipsis).

The two contributions included in Part II turn to the contrastive perspective (Johansson) and recent change in English (Leech, Smith and Rayson). In his chapter, Stig Johansson shows the power of multilingual corpora. Translation corpora help the researcher make meanings visible in a new systematic way, by profiting from the bilingual intuition of translators: the corpus allows one to investigate the forms and expressions in the target text which can be found to correspond to their counterparts in the source text, and vice versa. The three case studies turn to close cognates that in the light of corpus evidence differ significantly in use across languages: English *here* vs. Norwegian *her*, expressions of possibility in English and Norwegian, and expressions of habituality in English, Norwegian and German.

In their chapter, Leech, Smith and Rayson investigate variation and change in stylistic norms in the twentieth century from the perspective of text comparison and quantitative techniques. In the first part of their study, they trace developments in recent English over a sixty-year period by investigating data drawn from three matching corpora of a million words each, the B-LOB (c. 1931), LOB (1961) and F-LOB (1991) corpora, with further material drawn from a fourth corpus from c. 1901 (underway). They trace the frequency distributions of the use of *not*-contraction, the passive voice, pied piping, *upon*, noun-noun sequences, and the *s*-genitive, and turn to stylistic trends such as *colloquialization* (movement towards spoken norms of usage) and *densification* (movement

towards denser or more compact expression of meaning) to account for the changes observed. In the second part of their study, they use the corpora as a reference norm against which they explore, in statistically sophisticated terms, the stylistic features of a single text, Virginia Woolf's 'The Mark on the Wall'. Their investigations not only provide insights into stylistic trends characteristic of twentieth-century English but also throw light on the ways in which corpus linguistic techniques could be of profit to literary studies.

Part III, devoted to the diachronic perspective, comprises six studies, reflecting the huge interest in historical corpus linguistics since the ground-breaking *Helsinki Corpus of English Texts*, which celebrated its twentieth anniversary in 2011. In his address at the 16th ICAME (International Computer Archive of Modern and Medieval English) conference in New College, the University of Toronto (Canada), in May 1995, Stig Johansson predicted that of the central areas in English corpus linguistics, the one to develop beyond any other area in the years to come would be English historical corpus linguistics. This prediction finds support in Matti Rissanen's survey of the forty-year story of English diachronic corpora included in the present volume. According to Rissanen, without the advent of electronic diachronic corpora, evidence-based historical linguistics might not have survived, let alone experienced the Renaissance it did. An interdisciplinary approach is important: an end-user of a multi-genre or specialized single-genre historical corpus needs to consider cultural, political, social, geographical and other factors of extralinguistic nature when analyzing the data. To illustrate the ways in which the materials can be used, Rissanen concludes the chapter by tracing the development of the adverbial subordinator *provided (that)* from Middle to Present-day English using, in addition to the Helsinki Corpus, a number of multi-genre or specialized historical, recent and Present-day English corpora.

Of the other five chapters in this section, those by Laurel J. Brinton and Thomas Kohnen combine corpus analyses with historical pragmatics and genre studies. In her chapter, Brinton surveys diachronic corpora that provide material for studies within the historical pragmatics framework and also probes into the problems that historical pragmaticians tend to experience when using such corpora. Among these are the dearth of oral discourse data from past periods and the necessity to try to access 'spoken' interaction of the past via written records. Linguistic features of interest to historical pragmaticians are often multi-functional, and results of computerized searches mostly require time-consuming manual screening to pin down the meanings in a wider context. However, search programs do not always allow one to inspect examples in a sufficiently wide context, and having to look up further context manually in printed sources is usually too time-consuming. These problems inherent in the historical corpus linguistics setting – along with its obvious advantages brought by automated searches and access to stratified and other ready-made text collections – are demonstrated in the case study of the comment clause *(as) you say* across the history of English.

In his contribution, Thomas Kohnen turns to prayers, a neglected genre in the history of English, and examines a number of typical (text-)linguistic and discourse-functional features in them, among them personal pronouns, performative formulas and patterns of address. The data is drawn from a corpus of prayers intended to be part of the *Corpus of English Religious Prose* (underway). Few vernacular prayer collections survive from the Middle English period, so most material comes from the Early Modern and subsequent periods. In the light of the results obtained, prayers emerge as an interactive genre with affinities to a wide range of registers including conversation, written records of 'spoken' interaction, and oral and formulaic usage. They also display remarkable stability across the centuries.

The chapters by Claudia Claridge, Ian Lancashire and Elizabeth Traugott focus on semantic change. In her study, Claridge explores the interface between two data sources, early dictionaries and historical corpora, with the aim of checking the treatment of transferred senses of body part terms such as *head, face, eye, leg,* and *foot* recorded in four eighteenth-century dictionaries against the occurrence of corresponding terms attested in three English historical corpora. The dictionaries included in the study are based on more literary and specialized sources than the corpora, which contain texts representative of more private, colloquial, spoken-like and utilitarian registers. Differences in the treatment and occurrence of the terms in the two sources can be expected to show whether or to what extent the sources used by the dictionary compilers have biased the picture given of usage when compared with the picture of everyday language use that transpires from the corpora. The study shows that there is considerable overlap in the evidence gained of different word senses in the dictionaries and the corpora, the different dictionaries displaying variation as to the amount of the overlap. Interestingly, the corpora included in the study end up yielding richer and more varied evidence of collocations, fixed expressions and idioms than do the dictionaries. This seems to point to the usefulness of even small-size corpora for the study of historical phraseology.

In his chapter, Ian Lancashire illustrates the potential of his *Lexicons of Early Modern English* (*LEME*) database for the study of semantic derivation in Shakespeare's language. *LEME* includes some 588,000 word-entries from 176 bilingual, monolingual and polyglot dictionaries and glossaries dating from 1470 to 1700 (figures from October, 2011). In addition to English, 36 other languages are represented in the material. This electronic resource is highly useful in the study of semantic deviation, a well-known problem in the history of the English lexicon. When searched, *LEME* provides access to historical word profiles that often enrich the picture given of the usage in the *Oxford English Dictionary*. Regarding the interdisciplinary perspective, most use of *LEME* is made not by linguists but by literary and historical scholars.

In her chapter, Elizabeth Traugott explores the role played by the 'bridging contexts' in morphosyntactic change and grammaticalization processes. A number of models have been presented on the nature of these contexts in the literature over the years. After a careful examination of the issues involved,

Traugott sets out to look for evidence of such contexts and their status in historical corpora. Two topics are addressed, the *be going to* construction that grammaticalized from a propositional expression of motion into an auxiliary of future tense, and the development of a sub-set of ALL- and WH-pseudo-clefts (e.g. *all one had to do was to listen to it)*. The study reveals a great deal of evidence of a stage of pragmatic, semantic and structural ambiguity before *be going to* uses grammaticalized but only scanty evidence of such ambiguity in the grammaticalization of the pseudo-clefts considered. Both studies make it clear that the analyst needs to take into consideration a sufficiently wide context when sifting readings, especially in the case of ambiguous examples.

To conclude, the idea for the present book emerged at the conference of the International Association of University Professors of English in Lund in 2007, and the draft chapters were submitted and revised over a number of rounds in 2008–2011. Heartfelt thanks go to the contributors and, regarding advice on the production of the camera-ready copy of the manuscript, to Docent Erik Smitterberg at the Department of English, Uppsala University.

Finally, the book is dedicated to the memory of Stig Johansson, a founding father and major contributor to English corpus linguistics, whose chapter in this volume remained one of his last works. Stig Johansson submitted and revised his manuscript prior to passing away on 22 April, 2010, and his former student and subsequent colleague at the University of Oslo, Professor Hilde Hasselgård, kindly attended to the final editorial stages of the text.

References

Corpora

Helsinki Corpus = *The Helsinki Corpus of English Texts* (1991). Department of English, University of Helsinki. Compiled by Matti Rissanen (Project leader), Merja Kytö (Project secretary); Leena Kahlas-Tarkka, Matti Kilpiö (Old English); Saara Nevanlinna, Irma Taavitsainen (Middle English); Terttu Nevalainen, Helena Raumolin-Brunberg (Early Modern English). See http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/index.html.

Secondary sources

Johansson, Stig (2008), ‘Some aspects of the development of corpus linguistics in the 1970s and 1980s’, in: Anke Lüdeling and Merja Kytö (eds.) *Corpus linguistics: an international handbook* (Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science 29.1–2). Berlin and New York: Walter de Gruyter, 33–53.

Wynne, Martin (2010), 'Interdisciplinary relationships', *International Journal of Corpus Linguistics*, 15 (3): 425–427.

This page intentionally left blank

I Setting the scene

The electronic life of texts: insights from corpus linguistics for all fields of English

Anne Curzan

University of Michigan

Abstract

More English literary and nonliterary texts "go electronic" and often online every day, from literary projects like EEBO (Early English Books Online) *to linguistics projects like ARCHER* (A Representative Corpus of Historical English Registers)*, from lexicographic projects like the* Oxford English Dictionary Online *to projects so ambitious they are almost uncategorizable, like Google's digitization of entire university libraries. How should researchers and teachers of English best exploit these new electronic riches? Scholars in English corpus linguistics have been pushing the boundaries and addressing the challenges of working with collections of electronic texts for decades, in ways that can usefully inform all sub-disciplines of English literature and language study. This chapter focuses on the new research opportunities and lines of questioning that electronic text collections open in a variety of fields, on the wisdom gained in corpus linguistics on best practices for working with electronic texts, and on much-needed conversations between scholars in all sub-disciplines of English for how best to build electronic text collections so they can answer the questions we want to ask.*

1. English corpus linguistics and points of contact across disciplines

New developments in electronic text databases hit the headlines of *The New York Times* in December 2010: "In 500 Billion Words, New Window on Culture" (Cohen 2010). *The New York Times* was previewing a forthcoming article in *Science* (Michel et al. 2011) about "culturomics," or quantitative, computational analysis of a massive text corpus to investigate human culture. A team of researchers, primarily at Harvard, exploited a corpus of over five million books (from the over 15 million books digitized by Google), which contained 361 billion words of English as well as billions of words in six other languages—the largest text database ever compiled for humanities and social science research. Through frequency studies of words (1-grams) and set phrases (n-grams), the authors provide examples of what they argue can be learned about developments in the English language (e.g., the regularization over the past two hundred years of verbs such as *burn* and *thrive*) and cultural phenomena (e.g, the relative rapidity of the achievement and loss of famous people's fame since 1800). *The New York Times* article simultaneously celebrates the exciting possibilities such

database searching offers, and it quotes scholars such as Louis Menand explaining their concerns about the lack of collaboration with humanists on the project and about the limits of quantitative data for understanding culture (see also Nunberg [2010] for a thoughtful response to the *Science* article and its impact on research in the humanities).

Both the *Science* and *The New York Times* articles mention the word *corpus*; neither mentions the field corpus linguistics. In this piece, I propose that corpus linguistics can provide bridges in the conversation among scholars from various disciplines interested in exploiting large electronic text databases. I focus primarily on the ways those working in all sub-disciplines of English literature and language study would profit from advances made in English corpus linguistics, but I also gesture to a broader transdisciplinary conversation about the use of corpora.

When I presented a version of this argument at an international conference, I began by asking the audience to imagine Toni Morrison as a corpus linguist, or at least to imagine her employing corpus linguistic methodologies. This rhetorical ploy was certainly in part meant to be attention-getting, but more importantly, the image it conjured was meant to capture a much needed scholarly, cross-disciplinary conversation. If the image was jarring to some, we need to figure out why. What could it mean for literary and linguistic studies for an author like Toni Morrison to have the tools of corpus linguistics at her disposal? At many of our academic institutions, "digital humanities" has become a buzzword, circulating in discussions focused on the future of the humanities. Scholars with experience in English corpus linguistics – experience in the creation, manipulation, and analysis of electronic texts – have much to offer colleagues in a range of disciplines, and much to learn from them, if we explore the many potential points of contact. Which brings me back to the example of Toni Morrison.

In her ground-breaking book *Playing in the dark: Whiteness and the literary imagination* (1992), Morrison argues for extending the study of American literature to examine how American literature has been fundamentally shaped by responses to a "dark and abiding" Africanist presence (1992: 46). She asserts:

> Reading *and charting* the emergence of an Africanist persona in the development of a national literature is both a fascinating project and an urgent one, if the history and criticism of our literature is to become accurate. (1992: 48, emphasis added)

Morrison frames this as a literary project, but it is arguably a linguistic one as well. She asks:

> How did the founding writers of young America engage, imagine, employ, and create an Africanist presence and persona? (1992: 51)

Morrison herself catalogues six linguistic strategies authors employed to engage the consequences of this Africanist presence, and she closely examines individual texts by Hemingway, Poe, and Cather to reveal their racial subtexts. These brilliant readings stand on their own, fully supported and not in need of other tools. But how might corpus linguistic methodologies help with the larger project that Morrison envisions, the reading and charting of an Africanist presence and persona?

Corpus linguistic methods could potentially facilitate and enrich this endeavor in at least three ways: they could help identify texts worthy of investigation now that so much is online; they could help provide systematicity to the cataloguing of linguistic features; and they could potentially reveal new patterns of co-occurrence. For example, Morrison discusses images of "impenetrable whiteness" that appear almost always in conjunction with representations of black or Africanist people who are dead, impotent, or enslaved (1992: 33). This observation is based on close reading and intuition, and Morrison shows us the literary payoff: the juxtaposition is a key to understanding constructions of whiteness. And I, as a linguist, am left wondering whether there are other noteworthy clusters of words/concepts that are equally frequent but might be below our conscious radar. Morrison also provides a list of collocations for blackness, including strangeness, desire, irrationality, and the thrill of evil; and she argues that these metaphorical oppositions have allowed writers to explore their fears about, for example, freedom and lack of restraint. As a linguist, I ask: what might a corpus-based study of collocations also reveal? The task would remain for the literary scholar to analyze what these findings mean for our understanding of American literature, but it would potentially be rich material indeed.

This speculative essay addresses the relatively broad question of how methods and insights from English corpus linguistics can usefully inform all sub-disciplines of English literature and language study. And from the reverse perspective, it makes an argument to corpus linguists about how specific methodological approaches can benefit from the insights of literary studies, as well as how corpus linguists could more richly conceive of the role of corpus linguistics generally.

Corpus linguistics is typically defined as the systematic study of language based on examples of "real life" language use (McEnery and Wilson 1996; see also Stubbs 1996, Biber, Conrad, and Reppen 1998), and many scholars pursuing this research have defined corpus linguistics as more a methodology than a "field" per se. As a methodology, corpus linguistics can be applied to the various subfields of linguistics (e.g., morphology, syntax, discourse) – and as this essay argues, to fields far beyond linguistics. Corpus linguistics aims to assess the extent to which patterns of language use are found in a given body of texts (spoken or written) and to analyze the contextual factors that influence language variation in the texts. Furthermore, corpus linguistics is generally characterized as making extensive use of computers and electronic collections of texts. In the narrowest definition of corpus linguistics, these electronic collections, or corpora,

need to be principled compilations of texts that aim to be balanced and representative. But over at least the past two decades, corpus linguists have been having extensive and productive conversations about how to think about and use the unprincipled, often much larger electronic collections of texts that are now readily available (see Curzan 2008a, 2008b). As perhaps a telling case in point, the definition of "corpus" on AskOxford is any electronic collection of text.

As more English literary and nonliterary texts "go electronic" every day and often immediately go online, the field of English faces challenges and possibilities we as scholars could not even have conceived of two or three decades ago. It is exciting and, honestly, overwhelming to contemplate the electronic resources now at our fingertips – from literary projects like EEBO (*Early English Books Online*) to linguistic projects like ARCHER (*A Representative Corpus of Historical English Registers*), from lexicographic projects like the *Oxford English Dictionary Online* to projects so ambitious they are almost uncategorizable, like Google's digitization of entire university libraries. How should teachers and researchers of English best exploit these new electronic riches? How do we know what to read if we can suddenly read everything? How can corpus linguistics help?

2. Conversation? What conversation?

Scholars who pursue work in English corpus linguistics may be housed in English departments or in Linguistics departments. In either one, colleagues may not immediately see their work as in conversation with work in corpus linguistics. Corpus linguists have been concerned about this lack of conversation, and they have focused almost exclusively on their relationship with colleagues in linguistics, particularly those in the Chomskyan tradition, strongly defending the value of real spoken and written data for the field of linguistics (see, for example, Fillmore 1992, Biber and Finegan 1991). While I have been known to be overly optimistic, I think significant progress has been made in terms of the broader field's attention to actual language use and corpus-based methodologies. For example, Ray Jackendoff (2007: 255) writes:

> My own position is that each source of evidence is valuable for certain purposes, that each must be used with care, and that we need all the tools we can get. … [Our first challenge is] [g]etting people to pay attention to other frameworks, to address the phenomena that other frameworks take as central, and to engage in conversation with a willingness to uncover and possibly even relinquish their own deeply held beliefs.

This passage appeared in Jackendoff's contribution to a special issue of the *Journal of English Linguistics* (2007) entitled "Directions for linguistics in the 21st century," in which almost all the authors, including those within the mentalist generative program, called for more collaboration and more recognition

of the work going on across all sub-disciplines of language study, often mentioning specifically corpus-based research.

Less noted but of equal concern should be the lack of conversation between corpus linguists and literary scholars within English departments. In 1992, Charles Fillmore drew pointed caricatures of the corpus linguist and the rational linguist struggling to find each other's work compelling or convincing. To transfer the trope to this other disciplinary context, the caricature of a literary scholar faced with corpus linguistic studies might wonder, "How exactly do the details about speakers' variable use of *–th* vs. *–s* endings (e.g., *sayeth* vs. *says*) help me read literature more insightfully?" And the caricature of the corpus linguist faced with an analysis of conceptualizations of the body in Victorian literature might wonder, "What exactly is the systematic basis of your interpretation?" This caricature, as Fillmore's original did as well, makes the gap between the two fields feel very wide, and there is much more common ground than typically recognized. However, that common ground can be exploited only through sustained conversation, and there is much to be learned by all involved.

Literary scholars and corpus linguists share an abiding interest in the workings of language, textual analysis, intertextuality, and "close reading." While the definition and practice of these terms can differ dramatically in the two fields, literary scholars can benefit from the work and tools of corpus linguists, and corpus linguistic methodologies and research would be richer from being informed by the questions being asked in literary studies. William Kretzschmar (2009) describes how the sheer volume of and easy access to electronic texts has brought some aspects of corpus linguistics into literary studies "through the back door" – for example searching multiple texts for all occurrences of a word to understand nuances of meaning. He writes: "The technology and the availability of texts have become so mainstream that now, finally, the basic findings of corpus linguistics are also becoming mainstream, because we replicate them whenever we try to solve the problems that concern us" (90). This essay presents an initial exploration of some of the ways this mutual conversation could be developed more explicitly ("through the front door," so to speak), focusing specifically on four areas: reading the linguistic features of literary narratives; exploring intertextual connections; analyzing collocations; and exploiting keyness, a concept developed in corpus linguistics.

To begin, I would argue that at the most fundamental level, literary scholars interested in the historical context of the literature they read should consider language history part of that context, and some of the most interesting and innovative work on the history of English at this point is corpus-based. While courses in the history and structure of English used to be required components of Ph.D. programs, this is no longer the case at most U.S. institutions, and the details and implications of linguistic study for literary scholars can seem fairly foreign. But a focus on specific linguistic features opens up new ways to think about textual analysis and close reading.[1] For example, in literature of the eighteenth and nineteenth century, in which authors play with linguistic differentiation based on class and region as a way to develop characters, the richer picture of linguistic

variation in these periods that we have gleaned from corpus-based studies allows us to understand better how authors are using these features – without assuming we know which features were standard and which were stigmatized in that period. Taryn Hakala (2010), for example, examines how selected British authors in this period manipulate specific non-standard features, such as the *–th* ending, to shape the status of – and readers' sympathies toward – female characters of different classes. To take another example, post-colonial literary studies could be enriched by linguistic studies of the English spoken in these postcolonial contexts, some of which are based on the *International Corpus of English* (ICE). These studies provide a more detailed picture of how different varieties of English – from the most standard to the most local – distribute in real use in real contexts, from newspapers to literature to the spoken language. Within this context, literary scholars can situate and analyze the linguistic choices authors are making for their narrative and for the depiction of particular characters. Corpus-based literary stylistics has focused on developments of particular authors' style (e.g., Binongo and Smith 1999), faithfulness of discourse representation (Short, Semino and Wynne 2002), and literary attribution (see Hoover 2001, 2002, 2003).

Many scholars in literary studies already turn to resources such as the electronic *OED* to learn more about the history and use of particular words in the literary texts with which they work. But it is a much richer world now for this kind of research. For example, Ian Lancashire (1997) demonstrates how the corpus of Renaissance dictionaries that he has compiled can enrich literary studies. In perusing these dictionaries, he and his students hit upon a new possibility for making sense of Aaron's name in Shakespeare's *Titus Andronicus*: the definitions of *aron* in these early dictionaries suggest that Shakespeare may have been playing with the name of the Moor to refer to a common English plant sometimes called "the devil" with a spotted black body and a bitter tongue (cf. Lancashire in this volume, p. 186ff.). And it is now possible for each and every one of us to do more extensive corpus-searching, in minutes, than was ever possible for the first or second edition of the *OED*. (This assertion relies on the broader definition of a corpus as any body of electronic texts, such as EEBO or the *English Poetry Database*.) For any linguistic oddity, scholars can now explore the extent to which it is an anomaly or a feature that appears in other texts. And if it does, what does that reoccurrence, in light of other shared or disparate features, reveal about forms of intertextuality?

That question serves as a useful segue to the second point of conversation: what corpus linguistics can offer for the exploration of intertextual connections. For instance, literary scholars could find interesting the work that has been done on the development of registers or genres (e.g., Biber and Finegan 1989, 1997; Atkinson 1992, 1999); these studies situate the formal features of literary texts, as well as scientific texts, journalism, and other genres, along a historical continuum. As one example, they provide details on the increasing linguistic formality of scientific and medical texts and the growing informality of political and other forms of public discourse. It is a very different take on studying intertextuality and could usefully inform literary scholars interested in exploring

intertextual relationships between, for example, nineteenth-century literature and published developments in science. Literary scholars are already working on how astronomical, psychological, and medical discourses influence literature of this period, but they often focus on *discourse* only in the sense of intellectual content. The conversation I am suggesting here would allow simultaneous interrogation of similarities in content, rhetorical traditions and structures, register features, and specific language usage (features of *discourse* as the term is used in linguistics) – potentially a much richer picture of intertextuality than we gain when each discipline works independently.[2] To take a second example, Costas Gabrielatos and Paul Baker's (2008) work on contemporary discursive constructions of refugees and asylum seekers in the British press using corpus linguistics and Critical Discourse Analysis, if framed with a broad audience in mind, could be of significant interest to scholars in cultural studies, communications/media studies, political science, and public policy, as well as literary scholars seeking to detail the historical context of recent British immigrant narratives. How do the narratives in the press and in literature compete and overlap, with what implications? To make this conversation happen, however, corpus linguists need as a rule to make more effort to explore the relevant questions and invoke the frameworks in these fields as well as to theorize the transdisciplinary implications of more empirical linguistic studies.[3]

The third point of potential conversation involves collocations. The study of collocations (i.e., the words that tend to co-occur with a given word) has long been a focus of corpus linguistic research, and has often been criticized as demonstrating the lack of significance of this research for other fields. Erin McKean wrote an entertaining column for *The New York Times Sunday Magazine* in the summer of 2007 about how the collocations lexicographers can locate with corpora help create better dictionaries (e.g., how the fact that we chide ourselves more than others could affect the definition of *chide*), but the application of studying collocations does not extend to the literary. Here I offer a tentative, exploratory example of how collocations could be of interest to literary scholars. I am intrigued by what it could mean for the study of poetry to be able to more systematically examine collocations, at any historical moment. One of the rich features of poetry is the meaningful juxtapositions it creates. Corpus-based studies could give scholars of poetry much more context to understand how poetry accepts, stretches, and breaks typical word patterns.[4] Also, I can imagine poets playing in innovative ways with corpus linguists' findings about collocations (as they do with found items such as newspaper clippings or advertising slogans) – to exploit and challenge our expectations for how language works and means (e.g., to migrate north rather than south). The earlier example of Toni Morrison has already demonstrated how potentially powerful the study of collocations could be for reading literature.

Toni Morrison's challenge to the field of literary studies to examine the Africanist presence in American literature usefully gestures toward the fourth point of a scholarly intersection: the concept of keyness. In taking up Morrison's call, it is not realistic for a literary scholar simply to search the library's electronic

texts for a term, like *blackness* or *whiteness*. The results will overwhelm even the most ambitious scholar given the large unprincipled collections now available. But in corpus linguistics, scholars have developed the notion of keyness, or the occurrence of a word or words in higher proportion in a given text or set of texts than would be the expected frequency across all texts (see Baker 2004). This technique could help literary scholars narrow their focus to literary and nonliterary texts of particular interest given their question. To find these key terms, corpus linguistic methodologies could be used on texts already known to be of interest for a literary question – to analyze the linguistic patterns of these texts and then use those patterns and/or word clusters to identify other texts of potential interest.

When discussing potential intersections between literature and linguistics, some colleagues in literary studies could quickly imagine how they could use such corpus-based methodologies to study citizenship and race in American literature, or justice and gender – both to understand better texts they already knew and to find texts they should know. The key point here is that it is a mutual enterprise: for literary scholars to show corpus linguists what collocations or lexical fields or combinations of lexical fields are of interest, and for corpus linguists to develop methodologies that exploit those focal points to provide literary scholars even better avenues of inquiry and data. In the process, all of us stand to gain important insights about connections among texts and text types.

In this way, corpus linguistic methodologies would provide tools for analyzing linguistic patterns in texts (e.g., Morrison's observations about collocations of blackness) and navigational tools for the immense electronic database of texts to which we are gaining access. One exciting possibility is that these methodologies could identify both literary and nonliterary texts of interest that a scholar would not intuitively look to, or even know about. Even if texts are electronic and available at a click, they do us little good if we do not know we might want to examine them.

It could be useful for corpus linguists to consider statistics as one possible analogy for the field/methodology that is corpus linguistics. Corpus linguistics has traditionally been focused on aiding in research on linguistic questions, and I do not want to downplay the importance of this enterprise. But as a methodology, corpus linguistics can serve a wide range of disciplines, beyond linguistics and even literary studies. The kinds of collaboration described in this essay could be of interest to political scientists, historians, sociologists, public health scholars, history of science scholars, and others. Corpus linguistic methodologies provide scholars in all these disciplines new ways to categorize and analyze texts. And as scholars invested in both corpus linguistics and critical discourse analysis are proving, this kind of close, systematic study of language reveals much about how topics are being framed and what ideologies surface. Scholars in a variety of disciplines will exploit these findings differently, but we all share the goal of more careful attention to the language informing scholarship in all relevant disciplines. To accomplish this goal, corpus linguists must take seriously the

burden of making our work and our tools accessible and interesting to scholars in other fields.

3. Building better databases

The developing world of electronic texts also demands collaboration among scholars in corpus linguistics, literary studies, and other fields along another avenue. As more and more texts gain electronic lives, we as researchers and teachers urgently need to speak up about how electronic texts can be most useful to us. At the University of Michigan, where the entire library is being digitized, the issue feels particularly pressing. Of course, it is a luxury in and of itself to be able to access books electronically, but let's be selfish: what else do we want to be able to do? My concern is that the technology could drive the questions we are able to ask, rather than our ensuring that the questions we want to ask drive the technology.

To take EEBO as an example,[5] this collection allows scholars to search the database by word or phrase, and it is designed to account for a wide range of spelling variation, which is important and not true of many electronic databases. In a subset of the collection, the texts have been keyed to allow for Boolean, proximity, and similar searches – a key resource for scholars interested in the language. However, EEBO does not provide ready access to collocations, or the words that typically appear on either side of a given word. The results of a given search are presented with the heading of the work, and it requires following a link to retrieve the text of any one example. The collection allows searches by categories such as poems, letters, drama, and notes. What other categories would we as scholars ideally want to be able to distinguish? And what do we want to be able to search for? Right now, the standard search parameters in electronic databases are authors, works, dates, and words and phrases, sometimes with the option of doing proximity or Boolean searches. But let's think outside the box, as it becomes possible to link various resources. Would we want to be able to search for a set of semantically related words? Corpus-based work on lexical fields could be useful here. And how do we want various editions of one work handled? If all the editions are electronic, we should build databases that facilitate electronic comparison of them; otherwise, all we have really done is save ourselves a walk to the library.

These questions seem all the more imperative not only as we build more literary databases but also as Google makes entire libraries electronic. How do we make such an enormous resource – some 7 million volumes in some cases – manageable so that we can actually use the data we are suddenly able to collect? We need finer-grained categories. Corpus linguists' experience with text tagging can be a starting point, but we need many voices in this conversation to know what we want the tagging to achieve. Unless we tell the database builders how we think about textual categories, the relationship of textual editions, and the ways

we want to search and manipulate electronic texts, we will be left to make do with what is built for us without our input.

We should not be overwhelmed by the available electronic resources or be passive recipients of them. We are well situated to develop the tools and methodologies to help linguists, literary scholars, and a range of other researchers to better navigate and analyze the linguistic and textual electronic world now open to us. And while we should proceed with capacious ambitions, corpus linguists have also learned hard lessons about the importance of exercising caution with electronic resources. Computers are invaluable in searching large amounts of text, but they make mistakes that can only be caught by human eyes, and only we can interpret the material they systematically gather. In the end, critical interpretation is what makes the data interesting. It is our challenge to help computers gather in seconds exactly the data we want them to gather and then to collaborate methodologically and theoretically to interpret what we find. It is simultaneously scary and exhilarating to know that computers can do a dissertation's worth of research in a long eye-blink, and as experienced researchers we need to ensure it is the research we need. Corpus linguistics has much to offer in terms of knowledge about how texts and the language of texts work, as well as how computers and electronic text collections help us gain that knowledge.

Notes

1 Historical language study has traditionally characterized medieval literary studies, and resources such as the *Middle English Compendium* allow scholars to examine nuances of Middle English beyond material available in the *Middle English Dictionary*.

2 Corpus linguists are also helping to answer the question of what it means, specifically, for many genres of the written language to become "more informal" over time, which would seem of interest to scholars of twentieth and twenty first-century literature.

3 Corpus linguists, who rely on literary texts among other genres for data, have a responsibility to ensure their research is informed by work in literary studies. Each and every piece of data comes from an individual manuscript, and often to understand variation, we need to turn to qualitative analysis that examines the features of and historical context of that particular manuscript. As I have argued elsewhere (Curzan and Palmer 2006), quantitative analysis must be balanced with the qualitative, and for the latter, scholars in literary studies are valuable guides. For example, to understand the linguistic variation in literary texts, may well depend on understanding the literary practices of a specific author or

genre. And certainly studies of medieval texts must take into account the work by medievalists on the transmission history of specific manuscripts.

4 As an interesting side note here (but one that leads me to the second part of my argument), Clai Rice has argued that in order to truly understand how modern words work, our databases must include ephemera, such as advertising slogans and supermarket signs; otherwise, any analysis of real-life examples is going to miss, for example, how *pledge* and *lemon* are integrally related in modern American English.

5 This database includes the works listed in Pollard & Redgrave's *Short-Title Catalogue (1475–1640)* and Wing's *Short-Title Catalogue (1641–1700)* and their revised editions, as well as the *Thomason Tracts (1640–1661)* collection and the *Early English Books Tract Supplement* (http://eebo.chadwyck.com/home).

Electronic and online sources

ARCHER Corpus. *A Representative Corpus of Historical English Registers 1650–1990*. http://llc.stage.manchester.ac.uk/research/projects/archer/.

Corpus of Middle English Prose and Verse. 2006. Available through the *Middle English Compendium*. http://quod.lib.umich.edu/c/cme/.

Davies, Mark. 2011–. *Google Books: American English* corpus (155 billion words, 1810–2009). Available online at http://googlebooks.byu.edu/.

Early English Books Online (EEBO). Chadwyck-Healey Ltd. http://eebo.chadwyck.com/home.

English Poetry Database. Chadwyck-Healey Ltd.

Google Books Ngram Viewer. http://ngrams.googlelabs.com/.

International Corpus of English [ICE]. http://ice-corpora.net/ice/index.htm.

Middle English Dictionary [MED]. 2001. Electronic version available through the *Middle English Compendium*. http://quod.lib.umich.edu/m/med/.

Oxford English Dictionary [OED]. 2000–. 3rd ed. online (in progress). Oxford: Oxford University Press. http://www.oed.com/.

References

Atkinson, Dwight (1992), 'The evolution of medical research writing from 1735 to 1985: the case of the *Edinburgh Medical Journal*', *Applied Linguistics*, 13: 337–374.

Atkinson, Dwight (1999), *Scientific discourse in sociohistorical context: the* Philosophical Transactions *of the Royal Society of London, 1675–1975*. Mahwah, NJ: Lawrence Erlbaum.

Baker, Paul (2004), 'Querying keywords: questions of difference, frequency, and sense in keywords analysis', *Journal of English Linguistics*, 32 (4): 346–359.

Biber, Douglas, Susan Conrad and Randi Reppen (1998), *Corpus linguistics: investigating language structure and use*. Cambridge: Cambridge UP.

Biber, Douglas and Edward Finegan (1989), 'Drift and evolution of English style: a history of three genres', *Language*, 65: 487–517.

Biber, Douglas and Edward Finegan (1991), 'On the exploitation of computerized corpora in variation studies', in: Karin Aijmer and Bengt Altenberg (eds.) *English corpus linguistics: studies in Honour of Jan Svartvik*. London, NY: Longman, 204–220.

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

Binongo, José Nilo G. and M.W.A. Smith (1999), 'A bridge between statistics and literature: the graphs of Oscar Wilde's literary genres', *Journal of Applied Statistics*, 26 (7): 781–787.

Cohen, Patricia (2010), 'In 500 billion words, new window on culture', *The New York Times* (December 16).

Curzan, Anne (2008a), 'Corpus-based linguistic approaches to the history of English', in: Haruko Momma and Michael Matto (eds.) *A companion to the history of the English language.* Malden, MA: Wiley-Blackwell, 596–607.

Curzan, Anne (2008b), 'Historical corpus linguistics and evidence of language change', in: Anke Lüdeling and Merja Kytö (eds.) *Corpus linguistics: an international handbook*. Berlin and New York: Walter de Gruyter, 1091–1109.

Curzan, Anne and Chris C. Palmer (2006), 'The importance of historical corpora, reliability, and reading', in: Roberta Facchinetti and Matti Rissanen (eds.) *Corpus-based studies in diachronic English*. Bern: Peter Lang, 17–34.

Fillmore, Charles J. (1992), '"Corpus linguistics" or "Computer-aided armchair linguistics"', in: Jan Svartvik (ed.) *Directions in corpus linguistics. Proceedings of Nobel Symposium 82*. Berlin and New York: Mouton, 35–60.

Gabrielatos, Costas and Paul Baker (2008), 'Fleeing, sneaking, flooding: a corpus analysis of discursive constructions of refugees and asylum seekers in the UK Press 1996–2005', *Journal of English Linguistics*, 36 (1): 5–38.

Hakala, Taryn (2010), Working dialect: nonstandard voices in Victorian literature. Ph.D. dissertation. Ann Arbor: University of Michigan.

Hoover, D. (2001), 'Statistical stylistics and authorship attribution: an empirical investigation', *Literary and Linguistic Computing*, 16 (4): 421–444.
Hoover, D. (2002), 'Frequent word sequences and statistical stylistics', *Literary and Linguistic Computing*, 17 (2): 157–180.
Hoover, D. (2003), 'Frequent collocations and authorial style', *Literary and Linguistic Computing*, 18 (3): 261–286.
Jackendoff, Ray (2007), 'A whole lot of challenges for linguistics', *Journal of English Linguistics*, 35 (3): 253–262.
Kretzschmar, William A. (2009), 'Habeas corpus?', *Journal of English Linguistics*, 37 (1): 88–92.
Lancashire, Ian (1997), 'Understanding Shakespeare's *Titus Andronicus* and the EMEDD', *Early Modern Literary Studies Special Issue*, 1: 6.1–20. Available at http://extra.shu.ac.uk/emls/si-01/si-01lancashire.html.
McEnery, Tony and Andrew Wilson (1996), *Corpus linguistics*. Edinburgh: Edinburgh UP.
McKean, Erin (2007), 'Corpus', *The New York Times Sunday Magazine*, (July 27): 14.
Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden et al. (2011), 'Quantitative analysis of culture using millions of digitized books', *Science* 331 (January 10): 176–182.
Morrison, Toni (1992), *Playing in the dark: whiteness and the literary imagination*. Cambridge, MA: Harvard UP.
Nunberg, Geoffrey (2010), 'Counting on Google books', *Chronicle of Higher Education* (December 16). Available at http://chronicle.com/article/Counting-on-Google-Books/125735/.
Short, Mick, Elena Semino and Martin Wynne (2002), 'Revisiting the notion of faithfulness in discourse presentation using a corpus approach', *Language and Literature*, 11 (4): 325–355.
Stubbs, Michael (1996), *Text and corpus analysis*. Oxford: Blackwell.

This page intentionally left blank

Textual analysis: from philology to corpus linguistics

Charles F. Meyer

University of Massachusetts Boston

Abstract

In the pre-electronic era, textual analysis was largely a matter of analyzing "static" texts, i.e., texts produced by writers at a given point in time. For instance, Otto Jespersen's (1909–1949) seven volume A Modern English Grammar on Historical Principles *is based on a large collection of written texts (e.g. novels, essays, newspaper articles) that Jespersen used for examples and as the basis of generalizations that he made about the structure of the English language. Technology, however, has greatly changed how we view the text: it is no longer an isolated entity existing only in printed form and accessible only through tedious manual analysis. Instead, when a text is encoded in computer-readable form and becomes part of an electronic corpus, it can be subjected to many different kinds of linguistic analysis. In my chapter, I will focus on how the field of corpus linguistics has greatly changed the potential for textual analysis, moving us beyond traditional philological analyses of "dead" texts to analyses that highlight the dynamic nature of language. For instance, traditional research in historical linguistics has focused mainly on structures found in a series of largely canonical written texts. However, because the* Corpus of Early English Correspondence *contains texts (e.g. personal letters) written between 1440–1800 that are very close to spoken language, analyses of this corpus can give us a sense of what the spoken language of this period might have been like, and additionally, can help us track changes in the language as they occurred in correspondence written by males and females belonging to different social classes. More synchronically oriented corpora permit similar kinds of analyses, primarily because they contain spoken as well as written language, and are encoded in a manner permitting easy access to the information in them. In short, current linguistic corpora are giving us unprecedented views into the structure of English used in diverse contexts at various times in its history by differing speakers and writers.*

1. Introduction

In the pre-electronic era, textual analysis was largely a matter of analyzing "static" texts: written texts existing only in printed form that had to be analyzed by hand. For instance, early 20th century grammarians such as Otto Jespersen, Hendrik Poutsma, and Etsko Kruisinga based the grammars of English they wrote on analyses of primarily canonical written texts (e.g. novels) that they used for

examples and as the basis of generalizations that they made about the structure of the English language. Technology, however, has greatly changed how we view the text: it is no longer an isolated entity existing only in printed form and accessible only through sometimes tedious manual analysis. Instead, when a text is encoded in computer-readable form and becomes part of an electronic corpus, it can be annotated with linguistic information (e.g. all words can be assigned a part of speech tag) and subjected to many different kinds of linguistic analysis. The existence of electronic corpora has greatly changed the potential for textual analysis, moving us beyond traditional philological analyses of "dead" texts to analyses that highlight the dynamic nature of language. For instance, traditional research in historical linguistics has focused mainly on structures found in a series of largely canonical written texts. However, because the *Corpus of Early English Correspondence* contains texts (e.g. personal letters) written between 1440–1800 that are very close to spoken language, analyses of this corpus can give us a sense of what the spoken language of this period might have been like, and additionally, can help us track changes in the language as they occurred in correspondence written by males and females belonging to different social classes (see http://www.helsinki.fi/varieng/domains/CEEC.html for further information about the corpus). More synchronically oriented corpora permit similar kinds of analyses, primarily because they contain spoken as well as written language, and are encoded in a manner permitting easy access to the information in them. In short, current linguistic corpora are giving us unprecedented views into the structure of English used in diverse contexts at various times in its history by differing speakers and writers.

But while electronic corpora and the software tools used to analyze them can generate results in a matter of seconds, such analyses can potentially take us too far away from the texts being examined and, more importantly, instill a false sense of security in the accuracy of the results that are ultimately obtained. For this reason, I argue in this chapter that corpus linguists need to go "back to the future" and complement automated analyses of corpora with the more philologically analyses conducted by earlier grammarians. To demonstrate the importance of this kind of approach to corpus analysis, I focus on the work of one member of "The Great Tradition" (Aarts 1975): Otto Jespersen. I examine his method of corpus analysis as reflected in his seven volume grammar *A Modern English Grammar on Historical Principles* (1909–1949; hereafter MEG), demonstrating that the kind of qualitative analyses that he conducted are important models for textual analysis. I then provide an overview of annotated corpora – a major innovation in corpus compilation and a key difference between modern-day corpora and the kinds of pre-electronic corpora with which grammarians such as Jespersen worked. To describe the strengths and limitations of annotated corpora and the tools used to analyze them, I review the results of three corpus analyses of such corpora that I have conducted. This review demonstrates that although the annotation of corpora greatly automates their analysis, corpus linguists still need to actually examine the texts upon which their analyses are based because many annotation schemes may be based on

preconceptions with which the corpus linguist may not be familiar. I conclude by arguing that the ideal corpus analysis combines the efficiency of an automated search with the kind of careful analysis done by grammarians from the Great Tradition.

2. The Great Tradition

The Great Tradition is a term that Aarts (1975) coined to describe the kind of linguistic analysis conducted by a certain group of grammarians writing grammars of English in the first part of the 20th century. Of these grammarians, three made use of what are commonly referred to as pre-electronic corpora: collections of written texts available only in printed form and upon which linguistic analyses were conducted and relevant examples selected. In addition to Jespersen, both Hendrik Poutsma (*A Grammar of Late Modern English*, 1904–1926) and Etsko Kruisinga (*A Handbook of Present-day English*, 5th ed., 1931–1932) made extensive use of pre-electronic corpora that each personally compiled to aid in the writing of their respective grammars. These grammarians were revolutionary in the sense that their aims were to write descriptive rather than prescriptive grammars of English. In addition, they began a tradition of basing grammatical analyses on actual texts – a key tenet of corpus linguistics.

Compared with modern-day corpus linguists, however, grammarians working with pre-electronic corpora were at a disadvantage in terms of the texts they were able to analyze (only written), the prevailing attitudes towards which texts were "worthy" of study (primarily canonical works of literature), and the time it took to pour through texts manually to find relevant examples. These limitations have led many to criticize the nature of the grammatical descriptions that the grammarians of the Great Tradition produced. Because they did not analyze any spoken texts, Quirk (1974: 167) comments that "their generally eclectic use of [written] source materials too often leaves unclear the distinction between normal and relatively abnormal structures and the conditions for selecting the latter." Mönnink (2000: 1–2) notes that their over-reliance on older literary texts skewed analyses towards one rather specialized type of English; that the dated nature of the texts often produced "descriptions of rather obsolete use"; and that because grammarians of this era lacked a clearly defined methodology for analyzing texts,

> They only describe the frequent, more basic patterns and structures together with some unusual patterns which happened to attract their attention. As a consequence, some features were necessarily missed out and their grammars are not quite complete.

While all of these criticisms are valid, in one sense they oversimplify the shortcomings of the grammatical descriptions that grammarians of this period produced. For instance, the corpus that Jespersen used to write MEG is large and

fairly varied (at least by the standards of his time). The bibliography of texts included in the corpus covers nearly 40 pages in MEG and contains ca. 1000 different sources. Below is a list of some of the registers into which the texts can be classified:

- Literature (fiction, poetry, drama)
- Literary criticism
- Biography
- Science
- History
- Philosophy
- Linguistics (e.g. the journal *American Speech* from 1925)
- Press reportage

Because Jespersen merely lists the sources in his corpus, it is not possible to quantify the word count for each of the registers above. Clearly, the corpus contains a considerable amount of literature. But by including various types of non-fictional writing (e.g. science, history, and press reportage), Jespersen does exhibit an awareness of register variation.

More interesting than the corpus itself, however, are Jespersen's analyses of the grammatical constructions on which he focuses in MEG. Although Jespersen is not, as Mönnink (2000) notes, working from any well established linguistic methodology for conducting textual analysis, arguably he and other members of the Great Tradition originated the kind of qualitative linguistic analysis found not just in modern-day reference grammars (such as Quirk et al. 1985) but in other books and articles providing more descriptively-oriented discussions of corpus data.

Entries in MEG typically open with some general commentary by Jespersen that is illustrated with a few invented sentences. This overview is then followed by extensive lists of examples taken from Jespersen's corpus with the purpose of providing a comprehensive description of the grammatical category under discussion. For instance, in an entry discussing the use of plural *they* or *their* to refer back to a singular indefinite pronoun such as *anybody* or *none*, Jespersen (vol. II, p. 137) comments that these types of number disagreements result from "the lack of a common-number (and common-sex) form in the third-personal pronoun…." He follows this point with a quote from an earlier work of his, *Progress in language* (published in 1894), in which he claimed that using generic *he* in a tag question such as *Nobody prevents you, does he?* "is too definite, and *does he or she?* too clumsy." The use of a plural pronoun in constructions of this type, he notes, is "not wholly illogical; for *everybody* is much the same thing as *all men.*" However, he qualifies this statement by saying that for all instances of such usages, "this explanation will not hold good" (p. 138). He then provides numerous examples of how common this usage is, including examples such as those below:

> God send *euery one their* harts desire (Shakespeare, *Much Ado About Nothing* III 4.60, 1623)
> *Each* had *their* favourite (Jane Austen, *Mansfield Park*, 1814)
> If *anyone* desires to know…*they* need only impartially reflect (Percy Bysshe Shelley, *Essays and Letters*, 1912)
> Now, *nobody* does anything well that *they* cannot help doing (John Ruskin, *The Crown of Wild Olive*, 1866)

Jespersen even includes sentences containing plural pronouns with singular noun phrases as antecedents. He claims that these types of noun phrases often have "generic meaning" (Vol. II, p. 495):

> Unless *a person* takes a deal of exercise, *they* may soon eat more than does them good (Herbert Spencer, *Autobiography*, 1904)
> As for *a doctor* – that would be sinful waste, and besides, what use were *they* except to tell you what you knew? (John Galsworthy, *Caravan*, 1925)

Commenting on Jespersen's discussion of plural pronouns with singular antecedents, Curzan (2003: 70–73) notes that these constructions date back to Old English, and were especially common when the antecedent contained two nouns conjoined by *or* (e.g. Modern English *If a man or a woman want to get married, they must get a marriage license*). She indicates (pp. 73–79) that grammarians of Jespersen's era treated the construction primarily from a prescriptive perspective, preferring generic *he* over *they*, or insisting that *they* be restricted to highly informal contexts. In critiquing Jespersen, Curzan (2003: 76) is correct to note that there is "a hint of prescriptivism" in Jespersen's discussion when he states that using a plural pronoun to refer to a singular antecedent "will not hold good" in all instances. Overall, though, Jespersen's discussion foreshadows the perspective taken in many modern-day corpus analyses: what one finds in a corpus directly affects the resultant grammatical description.

The most obvious shortcoming of the kind of corpus analysis done during the pre-electronic period is that it involved a tremendous amount of manual analysis of printed texts – analyses that a scholar such as Jespersen had an entire lifetime to conduct. The creation of computerized corpora, however, has greatly automated the linguistic analysis of texts, particularly if the corpus being analyzed contains linguistic annotation. In an annotated corpus, it is relatively easy to retrieve abstract grammatical constructions (e.g. noun phrases or adverbial clauses) in a matter of seconds. But one unfortunate consequence of this type of automation is that it takes the analyst one step away from the texts being analyzed, and in many cases reduces the constructions being analyzed to a series of unrelated concordance lines extracted from disparate parts of a corpus. As I will demonstrate in the next section, not being familiar with the annotation scheme used in a corpus and exactly what is being retrieved can greatly diminish the accuracy of the results that are obtained.

3. Annotated corpora

Annotation provides various kinds of linguistic information about a text:

Structural information
- Identification of how a text is structured (e.g. Speaker IDs and boundaries of overlapping speech in spoken texts; paragraph boundaries and special fonts in written texts; etc.)

Lexical information
- Identification of which words are nouns, verbs, adjectives, etc.

Grammatical information
- Identification of noun phrases, subordinate clauses, subjects, objects, etc.

Ethnographic information
- Age, gender, social class, etc. of speakers/writers who contributed texts to a corpus

Annotation can either be placed directly in the corpus itself, or stored in a database linked to specific sections of the corpus. Earlier lexical corpora contained annotation placed directly in the text. For instance, the excerpt below contains structural information taken from a spoken dialogue in ICE-New Zealand:

<ICE-NZ:S1A-011#1:1:B>
the thing with this stuff is that when you rub it in

<ICE-NZ:S1A-011#2:1:B>
you know how you rub in liniment

<ICE-NZ:S1A-011#3:1:U>
mm

<ICE-NZ:S1A-011#4:1:B>
and your skin goes bright red <{><[>like</[> it's burning

<ICE-NZ:S1A-011#5:1:B>
well this doesn't do that <,,> so it must really soak in <,,>
<&>10</&>

<ICE-NZ:S1A-011#6:1:U>
<[>yeah</[></{>

<ICE-NZ:S1A-011#7:1:U>
it's on the <,,> the <,> pelvic bone

Because the markup is SGML-conformant, it is included within braces < >. The conversation is divided into text units, which are preceded by markup indicating, for instance, the sample number from which conversation was taken (S1A-011), the number of the text unit, and the particular individual who is speaker (B or U). In the text itself, short and long pauses are marked by <,> and <,,>, respectively; the markup around *like* in text unit 4:1 indicates that B's utterance of this word overlaps with U's uttering of *yeah* in 6:1.

As more and more annotation is added to a text, it becomes increasingly difficult to work with. As a result, newer corpora have created various kinds of interfaces allowing easy access to the information in a corpus. The *British National Corpus* (BNC) is bundled with a search program called Xaira that can search for strings and lexical tags (http://www.natcorp.ox.ac.uk/tools/index.xml). Davies (2005) has created a search interface linked to corpora set up as relational databases, including the BNC (http://corpus.byu.edu/bnc/) and a newer corpus that Davies created himself: the *Corpus of Contemporary American English* (http://www.americancorpus.org/). The British component of ICE (ICE-GB) can be searched with ICECUP, a program that can retrieve full or partial parse trees.

But even though increasingly sophisticated annotation schemes and search interfaces continue to be created, the status of annotation among many corpus linguists remains controversial. Proponents of annotated corpora, such as Aarts (1992: 181), argue that annotation is essential because without it "…the comparison of corpora containing just raw text cannot go beyond linguistically rather trivial observations." Others feel that annotation introduces bias into any corpus analysis – a preconceived notion of how the text is structured. Thus, they believe, as Sinclair (1992: 384) argues, that a corpus should be "in raw form and analyse[d]… fresh each time some analysis is required."

It is certainly true that a tagged and parsed corpus does present a particular view of language. However, some annotated corpora (e.g. BNC, ICE-GB) have interfaces allowing for annotation to be turned off or on, making this objection somewhat moot. In addition, a well annotated corpus can greatly expand the amount of text that can be analyzed well beyond what is feasible with manual analysis – provided, as I will demonstrate later, that the analyst understands the grammar underlying the annotation scheme being used in the corpus being analyzed. To illustrate these points, I will describe analyses I conducted based on two annotated corpora: ICE-GB and the *Michigan Corpus of Academic Spoken English* (MICASE).

4. Gapping in ICE-GB

Gapping is a type of coordination ellipsis involving omission of one or more elements in the middle of the 2nd conjunct under identity with the same elements in the 1st conjunct. For instance, in the example below (taken from the Brown Corpus), the copula *is* in the second conjunct (marked by brackets) is deleted under identity with *is* following *memory* in the first conjunct:

(1) The long-settled areas of states like Virginia and South Carolina developed the ante-bellum culture to its richest flowering, and there the memory **is** more precious, and the consciousness of loss [ ] the greater. (Brown G01 1040-60)

In Meyer (1995), I report the results of a manual analysis of gapping and other types of coordination ellipsis in a 96,000 word corpus consisting of samples of speech and writing taken from the Brown Corpus and ICE-USA. To conduct this analysis, I read through the entire corpus, identifying all instances of coordination ellipsis and noting various linguistic characteristic of each which were then entered into a database. I recorded, for instance, not only how many instances of coordination ellipsis I discovered but the particular form of the constructions that were ellipted. In this corpus, I found only 22 instances of gapping, with the majority of instances (12 of 22) involving the gapping of function words, such as the auxiliary *is*:

(2) This type of borrowing can be reduced to a minimum if quarterly installment payment of taxes **is** instituted and the first payment [ ] placed near the opening of the fiscal year. (Brown H07 760-90)

Fewer examples involved gapping of the entire verb phrase, as is the case with *is directed* below:

(3) Related to micelle formation is the technologically important ability of detergent actives to congregate at oil-water interfaces in such a manner that the polar (or ionized) end of the molecule **is directed** towards the aqueous phase and the hydrocarbon chain [ ] towards the oily phase. (J05 1460-90)

Because this analysis was done manually, it took months to complete and could only be done with a very small corpus.

More recently, Hongyin Tao and I examined gapped coordinations in ICE-GB (Tao and Meyer 2006), a million word corpus containing various types of spoken and written British English. Because ICE-GB is fully parsed, we were able to automatically retrieve instances of gapping by constructing two FTFs (fuzzy tree fragments) that searched all the parse trees in ICE-GB. For instance, Figure 1 contains an FTF that searched all trees containing an *–ed* participle in the second conjunct (abbreviated as *edp*) but no accompanying auxiliary verb.

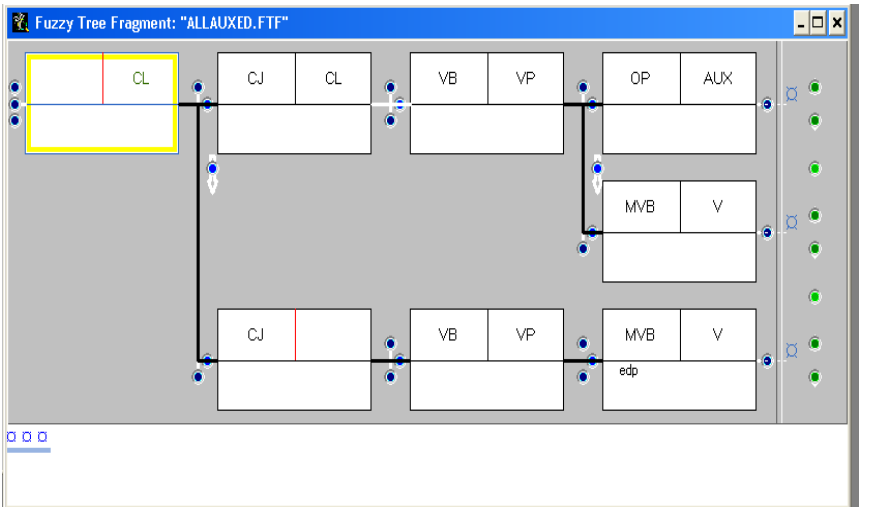


Figure 1: FTF for gapped auxiliaries

This FTF retrieved examples such as the one below in which the auxiliaries *have* and *been* are ellipted before *injured* in the second conjunct:

(4) It says three hundred and twenty civilians **have been** killed and more than four hundred [ ] injured. (S2B-037 #88:1:A)

The second FTF we created was much more general (see Figure 2, p. 32). It was devised to search for second conjuncts containing a clause (abbreviated as *CL*) but no verb phrase (abbreviated as *–v*).

This FTF retrieved quite a few false positives, since especially in the spoken sections of ICE-GB there were a number of conjuncts that contained incomplete structures. However, this FTF also located relevant structures such as the example below in which the verb phrase *had gone* is ellipted in the second conjunct, a clause containing only a subject (*the interviews*) and subject complement (*fine*):

(5) The documentary **had gone** well and the interviews [ ] fine. (W2B-001 #99:1)

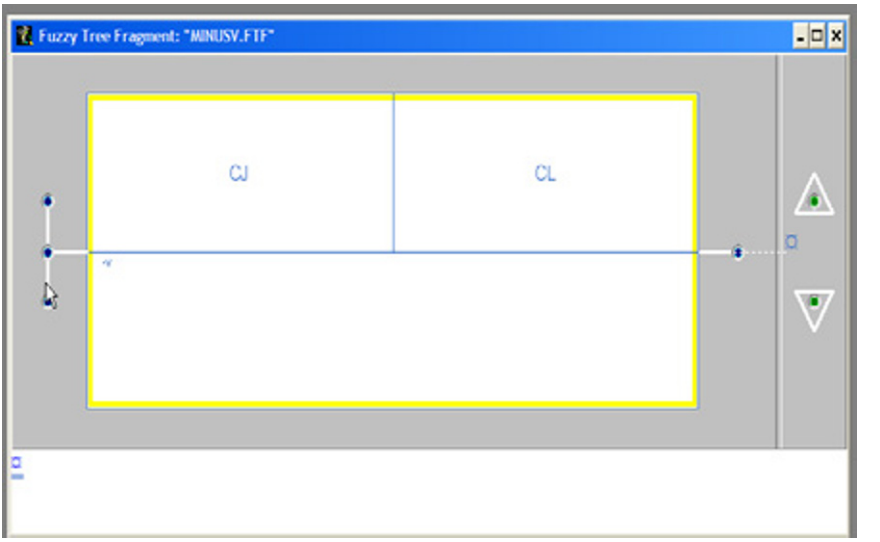


Figure 2: FTF for gapped verb phrases

After running both FTFs, we manually inspected all gapped structures to note additional grammatical features of them that could not be automatically generated – a procedure that did take some time. But the time expended was nothing compared to what a manual analysis would have involved, and the results shed more detailed information about gapping than was possible in my earlier study of gapping:

(i) Gapping is rare: only 120 examples were found in the entirety of ICE-GB.

(ii) Gapping is found mainly in spoken monologues and written registers, and only rarely in interactive speech.

(iii) As Meyer (1995) found, gapping favors coordinated clauses with short and non-complex structures and low content verbs (e.g. auxiliaries, copulas)

(iv) Gapping sometimes serves a stylistic function, and can enhance rhythm and parallelism. In the example below (which was taken from a broadcast news report), the omission of *is* in the second conjunct reinforces the parallelism of the subjects and subject complements in each clause:

The main post office **is** a burnt-out shell its telecommunication tower [ ] a twisted heap. (ICE-GB S2B-005 #65:1:E)

(v) In press reportage, gapping has become formulaic and often exhibits a negative prosody. For instance, we found examples such as the one below in which gapping occurred with verbs of violence, such as *killed, injured, destroyed, damaged*, or *shot down*:

Baghdad Radio has reported that nine planes **were** shot down over the city and another five [ ] destroyed as they were attacking Basra (ICE-GB S2B-008 #82:1:G)

One issue that is always relevant when doing an automated analysis such as this is whether the search algorithm (in this case, our FTFs) has retrieved all instances of the construction being studied so that any statistical analyses being conducted are based on all possible tokens. Answering this question as it relates to our study is difficult, since it requires a manual inspection of all of ICE-GB. We did consult Greenbaum's (1996) *Oxford English Grammar*, which uses examples from ICE-GB to illustrate various points of grammar, and indeed the four examples of gapping he cites (p. 313) were retrieved by our FTFs.

Ultimately, however, one may never know exactly whether all relevant cases for a particular corpus analysis were successfully retrieved. Consequently, any corpus analysis must allow for a certain margin of error. But this margin of error can be minimized if the analyst examines as much as is possible precisely what has been retrieved in a given search. To illustrate the importance of this notion, I will describe two analyses I recently conducted: one where understanding the underlying grammar used to annotate a particular construction helped clarify the results of the search, and another where investigating the sampling procedure used to create the corpus that was being used led to a more tentative reading of the results.

5. Object complements in ICE-GB

In a recent discussion of clause functions (Meyer 2009), I used ICE-GB to locate examples of clauses containing object complements. As Nelson, Wallis, and Aarts (2002: 51) state in their overview of ICE-GB, "Object complements occur with complex transitive verbs" and they provide as examples the two clauses below, both of which were taken from ICE-GB:

(6) a. Leave that battery *alone*. (ICE-GB S1A-007 #184)

b. *What* do they call it? (ICE-GB S1A-006 #16)

Both of these examples contain verbs – *leave* and *call* – that require a direct object and either an adjective phrase (*alone*) or noun phrase (*what*) that stands in a copular relationship to the direct object: *that battery is alone* and *it is what*.

Although object complements typically follow the direct object, because *What* is part of a *wh*-question, it is fronted in the clause.

While examining some of the examples that a search for object complements retrieved, I came across the example listed below, which contained the noun phrase *Jennifer* labeled as an object complement:

(7) Oh, she's called Jennifer (ICE-GB #122)

The parse tree for this example (see Figure 3) reveals that *Jennifer* has been parsed as an object complement (CO), even though the clause contains no overt direct object.

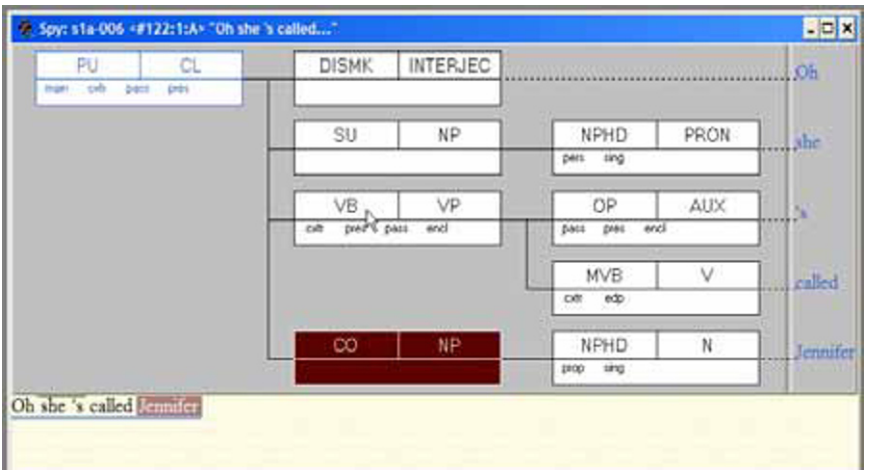


Figure 3: Parse tree for object complement with no overt direct object

Because this is an agentless passive, the agent (which would have functioned as subject in the active equivalent) is omitted, and the original object (*her*) has become subject in the passive. If the active equivalent of the clause is reconstructed, *Jennifer* is more clearly seen as an object complement:

(8) Oh, someone called her Jennifer.

What the parse tree in Figure 1 suggests is that in the ICE-GB grammar, clause functions are defined semantically as well as syntactically: in the agentless passive, *Jennifer* is semantically related to the overt agent.

To test whether other clause functions were defined semantically and syntactically, I searched for instances of indirect objects following the direct object and functioning as objects of the prepositions *to* or *for* to see how they were parsed:

(9) I'll save it for you (ICE-GB S1A-094 #2)

However, as the parse tree for this clause illustrates (see Figure 4), the prepositional phrase *for you* is parsed as an Adverbial (A), not an indirect object.

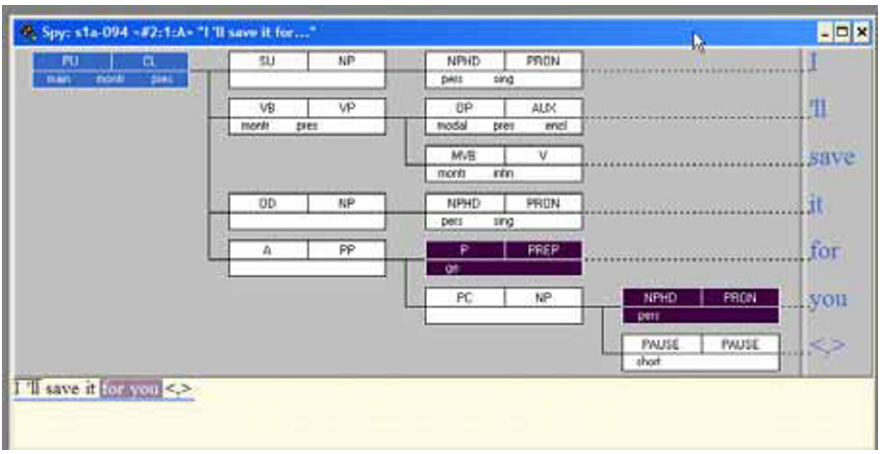


Figure 4: Parse tree for indirect objects following prepositions

Because the direct object is a pronoun in this example, an indirect object is not possible:

(10) ?I'll save you it.

Instead, the pronoun *it* has to be placed in prepositional phrase headed by *for* following the direct object, making the prepositional phrase semantically an indirect object.

This example illustrates that the ICE-GB grammar does not consistently use both syntactic and semantic criteria in parsing clause functions. Viewed from a larger perspective, this inconsistency is not unexpected, since it simply is not possible to tag and parse a large body of text without some level of error. The more important point, though, is that analysts need to be aware of the possibility of error, and to always double-check the results of searches. This awareness is even more important when analysts work with corpora allowing for results to be grouped by genre or by their use by males vs. females, for instance, and other demographic variables. These variables, as I will show in the next section, depend crucially on how a corpus has been created: the extent to which it is balanced and representative of the population of speakers and writers included in it.

6. Modal verbs of politeness in MICASE

The *Michigan Corpus of Academic Spoken English* (MICASE) contains various kinds of academic spoken English (e.g. class lectures, office hours, advising sessions) recorded and transcribed at the University of Michigan. The corpus consists of 152 transcripts totaling 1,848, 364 words (see http://lw.lsa.umich.edu/eli/micase/index.htm for details). Lexical searches of the corpus can be conducted online, and searches can be narrowed to speaker attributes (gender, age, academic role, native/non-native speaker of English, first language) and transcript attributes (speech event type, academic division/discipline, participant level, and interactivity rating). MICASE is therefore a very useful corpus to study the extent to which variables such as age and gender affect language use in various communicative contexts.

In Meyer (2006), I used MICASE to study the extent to which speakers adhere to Leech's (1983) notion of 'Tact' when using two directives – *you should* and *you might want to/wanna* – in two communicative contexts: advising sessions and office hours. Specifically, I wanted to determine how gender and power relationships (as defined by academic role) influenced the use of these two forms.

Tact is related to Leech's (1983: 109) work on politeness in English, and has two polarities:

> *Negative*: Minimize the cost to *h* [hearer]
> *Positive*: Maximize the benefit to *h*

In general, the directives *you should* and *you might want to/wanna* differ in terms of the extent to which they promote tact. If I say *You should do X* to someone, I am being more direct and potentially less tactful than if I had uttered *You might want to do X*, a directive that is less direct, more "mitigated", and potentially more tactful than *you should*. However, how these forms are interpreted will vary by context. For instance, research has shown that females are more likely to use mitigated forms than males, and that the particular power relationship existing between speakers will affect the level of politeness too: an instructor telling a student how to write a paper during office hours would be more likely to use *you should write your paper this way* than *you might want to write your paper this way*; an individual in an advising session with a student might be more likely to say *You might want to take this course* than *you should take this course*. However, these trends are not absolutes, and a corpus such as MICASE provides an opportunity to explore how the contexts in which the trends are reversed.

Because MICASE permits only lexical searches, I searched for three different strings: *you should, you might want to*, and also *you might wanna*, since in MICASE *want to* and *wanna* are transcribed differently depending upon pronunciation. Tables 1 and 2 contain the frequency with which these forms occurred in MICASE by gender and academic role.

Table 1: Modal breakdown by gender

	Advising and office hours	
Form	**Male**	**Female**
You might want to/wanna	1	51
You should	15	59
Total	16	110

Table 2: Modal breakdown by academic role

Role	*You might want to/wanna*	*You should*	**Total**
Graduate Student	31 (54%)	22 (46%)	53
Staff	19 (37%)	31 (63%)	50
Faculty	2 (5%)	12 (95%)	14
Student	0	5	5
Researcher	0	4	4

The results in these tables raise several issues that need to be addressed before any conclusions can be drawn, and that require a careful examination of the data upon which the frequencies are based.

First of all, after examining the concordance lines containing the individual examples, I noticed that while most of the instances of *you should* were deontic, some were epistemic. For instance, in the example below, the speaker is not using *you should* to get someone to do something but rather to suggest that if normal curves are sampled, a certain result is likely to be obtained – a clearly epistemic use of *should*.

(11) That's what *you should* get when you sample from normal curves. (Transcript # OFC575MU046)

Second, as Table 1 notes, the gender distributions were quite skewed. A breakdown by communicative context revealed that 70% of speakers in advising sessions were female. There was more balance in office hours (59% female, 41% male), but still females predominated. These distributions do not reveal any defect in the design of MICASE, but rather that individuals at American universities doing advising, for instance, tend to typically be female.

Table 1 does seem to suggest that while females may use *you might want/to* more than males, it does not mean that they use *you should* less than males. As Table 2 shows, academic role plays an important role too, with faculty using *you might want to/wanna* far more infrequently than academic staff and graduate students. But these results need to be qualified too because after I

examined exactly who used *you might want to/wanna*, I discovered that 26 (50%) of the instances of *I want to/wanna* were uttered by a single graduate student who was holding office hours with five students in one, rather lengthy sample that was 29,635 words in length.

Do complications such as the ones I have detailed invalidate the statistical information in Tables 1 and 2? Not necessarily, especially if some kind of qualitative analysis can be used to interpret the complexity of the results that are obtained. For instance, there are obvious pragmatic reasons why faculty advising students during office hours would be more likely to use *you should* than *you might want to/wanna*: faculty have the authority to use *you should* without appearing impolite; additionally *you should* is far clearer than the heavily mitigated *you might want to/wanna*. In the examples below, the use of *you should* tells students precisely what they need to do:

(12) a. in general **you should** do this throughout the paper too. (OFC115SU060)

b. so probably actually what **you should** do is go back and actually just, um, read this one more time, go back and read the Crawford one more time, and see if there're any sort of other sort of arguments that help you out. (OFC115SU060)

c. Don't interpret your confidence interval level, with just one interval. **you should** interpret it as being looking at many intervals. (OFC575MU046)

Using *you might want to/wanna* might give students the erroneous impression that the instructor is providing an option rather than a mandate. In the examples below, for instance, are instructors insisting, or merely suggesting, that students make changes?

(13) a. it doesn't change your argument necessarily, but **you might wanna** qualify it in that kinda way (OFC115SU060)

b. so you **you might wanna** say that, in order to understand um, the, programs and sort of missions of, these two organizations, **[S5:** mhm **]** we have to understand them within the context of, the, you know political and economic situations in these two cities. right? (OFC115SU060)

The intent of the speaker is not entirely clear.

On the other hand, as the examples below indicate, advisors in an advising session are often exploring options with advisees, and as a result, might be more inclined to use *you might want to/wanna* as a means of more diplomatically exploring choices with students:

(14) a. well **you might wanna** major in English (ADV700JU047)

b. okay so that sounds like **you might want to** take a mathematics class next semester, do you remember what math you placed into? (ADV700JU047)

In addition, at many American universities, staff interacting with students are often in a customer client relationship, and as a consequence of this power imbalance, must give advice to students in a more mitigated manner – a communicative need that a form such as *you might want to/wanna* satisfies quite well.

But advisors do use *you should*, particularly when students really do not have a choice:

(15) a. **you should** do your senior audit next fall (ADV700JU047)

b. **you should** take Intro Comp next semester. (ADV700JU047)

Perhaps the student being addressed in the first example does not have to do his/her senior audit in the following fall, but if the student is close to graduation, using a more direct form is likely to stress to the student the importance of following the advisor's advice.

What the analysis in this section shows is that statistical information is only a starting point for any investigation of language use: the examination of frequency trends is only a gateway into a closer investigation of actual examples and the functions that they serve in the contexts in which they occur. Furthermore, it is imperative that analysts examine exactly what their statistics are based on. In my analysis, if I had not discovered that 50% of the instances of *you might want to/wanna* were used by one individual, my discussion of gender trends would have been very inaccurate.

7. Conclusions

I have argued in this chapter that corpus linguists need to in a sense go "back to the future": they should complement fully automated methods of data collection and analysis with the kinds of close analysis conducted by grammarians of the "Great Tradition". Failure to strike this balance, I have demonstrated, can lead to results and claims that are not fully accurate.

But as corpora increase in size, many might question whether it is realistic (or even possible) to inspect the results of analyses based on multi-million word corpora. For instance, a search of *you should* in the 425 million word Corpus of Contemporary English (COCA) yielded 23,865 hits. Obviously, if an individual wishes to study the deontic uses of this construction, it would not be possible to examine each instance individually to exclude epistemic uses from

consideration. Restricting the search to the spoken section of the corpus (104 million words) resulted in 6,526 hits – still too many. But restricting it even further to the period 2010–11 narrowed the hits to a much more manageable number of 413. Of course, the ease with which one can easily focus in on specific parts of a corpus under analysis depends crucially upon the interface that is used to search, retrieve, and sort examples.

Even though a corpus may be large, it is still important to consider what kinds of texts the corpus has and the effects their composition may have on the types of constructions that are retrieved. The spoken part of COCA contains samples of speech taken from transcripts of broadcast television shows. Thus, the results will reflect how *you should* is used in a public forum – a forum quite different than, say, the private exchanges that are characteristic of casual conversations between friends. In addition, because the transcripts were provided by the broadcast companies themselves, there is no way of knowing how accurately the exchanges between speakers have been transcribed. These limitations do not necessarily invalidate a study of *you should* in the spoken section of COCA: any corpus will have its limitations. But knowing these limitations, as I have stressed throughout this chapter, is crucial to an accurate interpretation of the data upon which generalizations about structure and usage are based.

References

Corpora

British National Corpus (BNC). See http://www.natcorp.ox.ac.uk/.

Brown Corpus = *A Standard Corpus of Present-day Edited American English, for use with Digital Computers (Brown)*. 1964, 1971, 1979. Compiled by W. Nelson Francis and Henry Kučera. Brown University. Providence, Rhode Island. See http://www.helsinki.fi/varieng/CoRD/corpora/BROWN/index.html.

Corpus of Contemporary American English (COCA). See http://www.americancorpus.org/.

Corpus of Early English Correspondence (1998). Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Jukka Keränen, Minna Nevala, Arja Nurmi and Minna Palander-Collin at the Department of English, University of Helsinki. See http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/index.html.

ICE-GB = The British component of the *International Corpus of English*. See http://www.ucl.ac.uk/english-usage/projects/ice-gb/.

ICE-New Zealand = The New Zealand component of the *International Corpus of English*. See http://ice-corpora.net/ice/icenz.htm.

ICE-USA = The American component of the *International Corpus of English*.

International Corpus of English (ICE). See http://ice-corpora.net/ice/index.htm.
Michigan Corpus of Academic Spoken English (MICASE). See http://lw.lsa.umich.edu/eli/micase/index.htm.

Secondary sources

Aarts, Flor (1975), 'The great tradition of grammars and Quirk's grammar, *Dutch Quarterly Review of Anglo-American Lettersk*, 5: 98–126.
Aarts, Jan (1992), 'Comments', in: Jan Svartvik (ed.) *Directions in corpus linguistics. Proceedings of Nobel Symposium* 82. Stockholm, 4–8 August 1991. Berlin: Mouton de Gruyter, 180–183.
Curzan, Anne (2003), *Gender shifts in the history of English*. Cambridge: Cambridge University Press.
Davies, Mark (2005), 'The advantage of using relational databases for large corpora: speed, advanced queries, and unlimited annotation', *International Journal of Corpus Linguistics*, 10: 301–328.
Greenbaum, Sidney (1996), *The Oxford English grammar*. Oxford: Oxford University Press.
Huddleston, Rodney and Geoffrey Pullum (2002), *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.
Jespersen, Otto (1909–1949), *A modern English grammar on historical principles*. London: George Allen and Unwin LTD.
Kruisinga, Etsko (1931–1932), *A handbook of Present-day English*, 5th ed. Groningen: Noordhoff.
Leech, Geoffrey (1983), *Principles of pragmatics*. London: Longman.
Meyer, Charles F. (1995), 'Coordination ellipsis in spoken and written American English', *Language Sciences*, 17: 241–269.
Meyer, Charles F. (2006), 'Corpus linguistics, the World Wide Web, and language teaching', *Iberica*, 12: 9–21.
Meyer, Charles F. (2008), 'Pre-electronic corpora', in: Anke Lüdeling and Merja Kytö (eds.) *Corpus linguistics: an international handbook*. Walter de Gruyter, 1–14.
Meyer, Charles F. (2009), *Introducing English linguistics*. Cambridge: Cambridge University Press.
Mönnink, Inge de (2000), *The mobility of constituents in the English noun phrase: a multi-method approach*. Amsterdam/Atlanta, GA: Rodopi.
Nelson, Gerald, Sean Wallis and Bas Aarts (2002), *Exploring natural language: working with the British component of the International Corpus of English*. Amsterdam and Philadelphia: John Benjamins.
Poutsma, Hendrik (1904–1926), *A grammar of Late Modern English*. Groningen: Noordhoff.
Quirk, Randolph (1974), *The linguist and the English language*. London: Edward Arnold.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik (1985), *A grammar of contemporary English*. London: Longman.

Sinclair, John M. (1992), 'The automatic analysis of corpora', in: Jan Svartvik (ed.) *Directions in corpus linguistics*. *Proceedings of Nobel Symposium* 82. Stockholm, 4–8 August 1991. Berlin: Mouton de Gruyter, 379–397.

Tao, Hongyin and Charles F. Meyer (2006), 'Gapped coordinations in English: form, usage, and implications for linguistic theory', *Corpus Linguistics and Linguistic Theory*, 2: 129–163.

This page intentionally left blank

II Focus on Present-day and recent English

Cross-linguistic perspectives

Stig Johansson

University of Oslo

Abstract

The Oslo Multilingual Corpus *(OMC) is a collection of electronic text corpora comprising original texts and translations in several languages: English-Norwegian, German-Norwegian, French-Norwegian, English-German-Norwegian, Norwegian-English-German-French, etc. The OMC provides unique research material for use in contrastive studies and translation studies, as well as in theoretical and applied linguistics. The study reveals what is general and what is language specific and is therefore important both for the understanding of language in general and for the study of the individual languages compared. Current work on the OMC includes: studies of lexis, modality, coordination vs. subordination, explicit vs. implicit information, discourse markers. In this chapter I will give examples from some of these areas.*

1. Introduction

In recent years there has been a rapidly increasing interest in multilingual corpora. In the 1990s we built the *English-Norwegian Parallel Corpus* (ENPC), a bidirectional translation corpus consisting of English original texts and their translations into Norwegian and of Norwegian original texts and their translations into English. We were fortunate to cooperate with researchers at Lund University and Göteborg University, who compiled a similar corpus for English and Swedish, the *English-Swedish Parallel Corpus* (ESPC). Because of the way these corpora are structured, they can be used both as translation corpora and as comparable corpora of original texts, allowing us to ask questions both on language relationships and on translation (Johansson 1998). In Oslo we have expanded our corpus work to include other languages, in particular German and French. The umbrella term for the various subcorpora is the *Oslo Multilingual Corpus* (OMC).[1]

The corpora have been used for a wide range of studies on lexis, syntax, and discourse. Our colleagues in German linguistics have done very interesting contrastive work on topics such as coordination vs. subordination, information structure, and explicit vs. implicit information in discourse.[2] In my chapter I will focus on recent English-related work done at the University of Oslo. Three topics will be singled out for special mention: spatial linking, expressions of possibility, and expressions of habituality. But before going into the individual topics I would like to briefly discuss the possibilities of multilingual corpus research.

2. Contrastive linguistics in a new key

There are three characteristics of our research which warrant the description 'contrastive linguistics in a new key':

- the focus on immediate applications is toned down;
- the contrastive study is text-based rather than a comparison of systems in the abstract;
- the study draws on electronic corpora and the use of computational tools.

It is only the combination that is new, or relatively new. We have had a lot of language comparison before which has been primarily descriptive-theoretical. We have had text-based comparison before. There was even a contrastive project in the early days of corpus studies, planned about forty years ago as part of the Serbo-Croatian – English Contrastive Project (Filipović 1969). But it is only since the mid-1990s that contrastive studies related to multilingual corpora have started to come to fruition.

What is there to be gained by using multilingual corpora? In the first place, we can make sure that there is a sound empirical foundation for the studies. Secondly, our research can be made more efficient through the use of computational tools. But the motivation is much more fundamental. To see this, let's turn to what one of the members of the Serbo-Croatian research team had to say:

> […] similarity between languages is not necessarily limited to similarity between elements belonging to corresponding levels in the languages concerned, and […] is not necessarily limited to similarity between elements belonging to corresponding classes or ranks in the languages concerned. (Spalatin 1969: 26)

For example, if we are interested in studying modality across languages, it is not sufficient to compare the use of the modal auxiliaries, because modal meanings can be expressed by many other means (including lexical verbs, adverbs, noun constructions, and adjective constructions). Given an appropriate corpus structure, we can discover the different means, as I will show in one of my examples later (Section 4).

One of the most fascinating aspects of a translation corpus is that it provides a means of making meaning visible. To take an example, Dirk Noël argues that "translators, through the linguistic choices they make, inadvertently supply evidence of the meanings of the forms they are receiving and producing" (Noël 2003: 757). On the basis of translations in the *Canadian Hansard Corpus* (English-French), he shows that forms like *BE said to* and *BE reported to* are turning into evidential auxiliaries. To put it more generally, we can regard the use of a translation corpus as the systematic exploitation of the bilingual intuition of

translators, as it is reflected in the pairing of source and target language expressions in the corpus texts.

Given a multilingual corpus we can examine paradigms of correspondences, i.e. the set of forms in the target text which are found to correspond to particular words or constructions in the source text; or the other way around: the set of forms in the source text which are found to correspond to particular words or constructions in the target text. Figure 1 defines some major correspondence types.

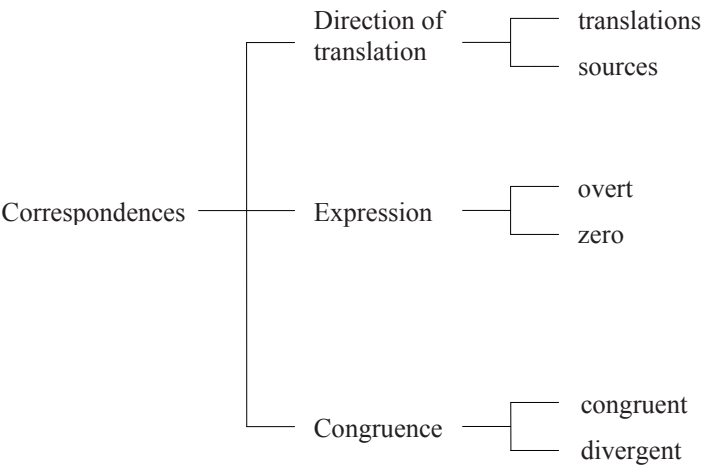


Figure 1: Classification of correspondences

Correspondences may differ depending upon the direction of translation (translations vs. sources), they may be overt or there may be no formal correspondence at all (overt vs. zero), and they may involve similar or different form types (congruent vs. divergent); see further Johansson (2007: 23ff.).

To take an example, correspondences of the English discourse particle *well* have been studied in relation to Swedish and Dutch by Aijmer and Simon-Vandenbergen (2003) and by myself (Johansson 2006) in relation to Norwegian and German. In both cases, we find a wide range of correspondences which serve to illuminate both English *well* itself and its relationships across languages. There is a lot of zero correspondence, i.e. *well* is often omitted in translation. It may also be added, so to speak out of the blue, in translation from other languages. About every fifth instance of *well* in English texts translated from Norwegian have no clearly identifiable source.

Formal similarity is no guarantee that there is identity of use. English *well* and its Norwegian cognate *vel* have the same core meaning and partly overlap in use, and yet they differ greatly (see Johansson 2007: 280ff.). We turn now to the first study which I will deal with in a bit more detail. This is another case where

close cognates can be shown to differ greatly in use across languages: English *here* vs. Norwegian *her*.

3. Spatial linking

In a series of studies based on the ENPC, my colleague Hilde Hasselgård has examined sentence openings in English vs. Norwegian, with special reference to thematic choice, i.e. the choice of opening element.[3] One of these studies deals with spatial linking, with special reference to the two deictic adverbs *here* and *her* (Hasselgård 2004b). These are dictionary equivalents, and at the outset one would expect a fairly straightforward relationship, apart from the well-known fact that English *here* corresponds both to the place adverb *her* (cf. German *hier*) and to the adverb of direction *hit* (cf. German *hierher*):[4]

(1) The service *here* was excellent. (AT1)
Servicen *her* var utmerket.

(2) "Butt, come *here*," Sam called. (GN1)
"Butt, kom *hit*," ropte Sam.

However, differences extend far beyond such examples, and it is these other differences which are in focus in Hasselgård's study.

Using the ENPC browser we can easily find all the cases where *here* and *her* do not correspond as well as those where they do correspond. Differences were found particularly in initial position. These are some of the main findings:

- Initial *her* was observed to be much more common than initial *here* (on average, 5.1 vs. 1.2 occurrences per text).[5]

- While English *here* (regardless of position) was found to correspond to *her/hit* in about 80% of the cases, initial *her* corresponded to English initial *here* only in about a third of the cases (cf. Table 2).

- There are a great number of other correspondence types for initial *her*: zero, non-initial *here*, other space adverbials, demonstrative pronouns/determiners, etc.

Tables 1 and 2 give a more detailed survey of the distribution of, and correspondence between, the two words. Table 1 shows that initial *her* is especially common in Norwegian non-fiction. In Table 2 we note that initial *her* corresponds to non-initial *here* in about 15% of the cases in translation from Norwegian into English, i.e. the adverb is fairly often moved away from initial position. But the most striking finding is that in about half of the instances overall initial *her* does not correspond to *here* at all.

Table 1: Occurrences of initial HERE (representing *her/here*) in the English and Norwegian original texts of the ENPC (quoted from Hasselgård 2004b)

Language	Text type	Total no. of *HERE*	No. of initial *HERE*	Proportion of HERE in initial position	Occurrences of initial HERE per text	Occurrences of initial HERE per text
English	fiction	376	35	9.3%	1.2	1.2
	non-fiction	111	24	21.6%	1.2	
Norwegian	fiction	547	102	18.6%	3.4	5.1
	non-fiction	431	155	36.0%	7.8	

Table 2: English correspondences of Norwegian initial *her*: both directions of translation, including both fiction and non-fiction (quoted from Hasselgård 2004b)

	English → Norwegian		Norwegian → English		Total	
	N	%	N	%	N	%
Initial *her* = initial *here*	46	32.9	91	35.4	137	34.5
Initial *her*= non-initial *here*	11	7.9	40	15.6	51	12.8
her = *here* in sentence fragm.	3	2.1	7	2.7	10	2.5
Initial *her* ≠ *here*	80	57.1	119	46.3	199	50.1
	140	100	257	100	397	100

Hasselgård gives a full survey of correspondences. Some examples from the material are:

(3) Den myten som er best kjent i Norge, kjenner vi fra diktet *Trymskvida*. *Her* hører vi at Tor lå og sov, og da han våknet, var hammeren hans borte. (JG1)
The myth that is best known in the Nordic countries comes from the Eddic poem "The Lay of Thrym." *It* tells how Thor, rising from sleep, finds that his hammer is gone.

(4) *Her* kunne de snakke sammen uten å bli ropt inn for å gå i melkebutikken eller til bakeren. (BV1)
They could talk *here* without being called in to go and buy milk or bread.

(5) *Her* på Bayer'n har jeg begynt å skrive. (JM1)
I've begun to write at Bayer jail.

(6) Det var som om huset lå ved verdens ende, for bak hagen hennes var ingen andre hus. *Her* begynte den dype skogen. (JG1)
There were no other houses beyond her garden, which made it seem as if her house lay at the end of the world. *This* was where the woods began.

In (3) the space adverbial is left out, in (4) it is postponed; in both cases the English sentence opens with a personal pronoun. In (5) the translation of *her* is omitted and the second space adverbial is moved to the end. In (6) we find an initial demonstrative link.[6]

The interpretation of the findings is that English *here* is not as clearly anaphoric as Norwegian *her*, which has a more textual, connecting function. According to Hasselgård, "It may be claimed that Norwegian *her* has acquired a grammaticalized usage as a discourse connector." Extending the study to initial prepositional phrases, she finds similar tendencies, although there is more congruence than for *her/here*. To sum up, there are more initial space adverbials in Norwegian, supporting the hypothesis that spatial linking is more common in Norwegian than in English. English instead tends to prefer participant continuity; cf. the opening of the translated sentences above (see the italicised forms).

4. Expressions of possibility

My second example concerns expressions of possibility in English and Norwegian, a topic dealt with in a PhD thesis from the University of Oslo (Løken 2007). Due to the complexity of the area of modality, I will not go into details of analysis, but will chiefly illustrate the methodology used. As pointed out in Section 2 above, it is not sufficient to examine modal auxiliaries. How do we find the relevant forms?

To begin with, let's examine the method used by Bengt Altenberg (1999) in a contrastive study of adverbial connectors in English and Swedish. Altenberg examines the mutual correspondence (MC), or intertranslatability, between forms and semantic subcategories in the two languages. Figure 2 summarises the relationships between English and Swedish contrastive conjuncts.[7]

Mutual correspondence is a good measure which can be used to relate not just individual forms but also semantic categories and subsystems across languages. With reference to his study of contrastive conjuncts, Altenberg stresses that the findings are independent of any preconceived classification:

> Even if the items […] had not been classified from the start as 'contrastive', their MC values would have brought them together and forced us to consider them as cross-linguistically related systems. Provided that the material is large enough, MC values are thus a useful

means of establishing semantic paradigms in contrasted languages, as well as of refining or 'correcting' existing classifications. (Altenberg 1999: 266)

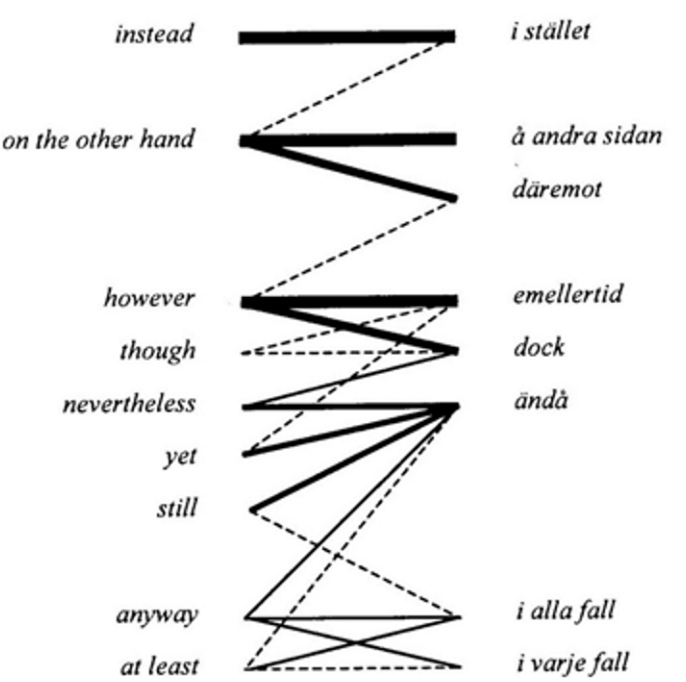


Figure 2: Cross-linguistic clustering of contrastive conjuncts: English-Swedish (from Altenberg 1999: 265)

Thus the study gives insight not just into cross-linguistic relationships, but may throw new light on the individual languages compared.

In defining the material for her study Løken starts from the modal auxiliaries, since "the modals are the central expressions of modality and expressing modality is the central use of the modals in the two languages" (Løken 2007: 13). Non-auxiliary expressions are established by navigating back and forth between the languages:

> The first step in the process is to list all correspondences of the modals expressing a particular meaning, what Dyvik (1998: 59) refers to as "first t-image". This list includes items chosen by one translator only. These are disregarded on the grounds of representativeness. There will also be items that are conditioned by context and do not necessarily express the relevant meaning on their own. These are also disregarded. The next step is to

> identify the correspondences of the remaining items, what Dyvik refers to as the "inverse t-image" (1998: 59). Again unique and context-dependent items are disregarded. Uni-directional translation relations are also disregarded, since these are likely to be instances of translationese. (Løken 2007: 14)

To take an example, Table 3 lists the expressions of low-value obligation (permission) which were included in the study.

Table 3: Expressions of low-value obligation (quoted from Løken 2007: 84)

	English	Norwegian
Modals	*can* *could* *may* *might*	*KUNNE* *FÅ*
Lexical verbs	*let* *allow* *permit*	*la* *tillate*
Constructions	*be allowed to* *be permitted to* verb + *permission to*	*være tillatt* verb + *tillatelse til* verb + *lov (til)*

After a close examination of the use of these expressions within each language, and their correspondences across the languages, Løken summarises the functions of the expressions as shown in Table 4.

In both languages the modal auxiliaries are the main means of requesting and giving permission, whereas other means are preferred to report permission. These are some examples of the three functions: requesting permission (7), giving permission (8), reporting permission (9, 10):

(7) "*May* we open the windows, Mr Barker?" asked the Queen. (ST1)
"*Kan* vi *få* åpne vinduene, Mr Barker?" spurte dronningen.

(8) "Very well. You *may* go out with him." (TH1)
"Det er greit. Du *kan få* gå ut med ham."

Table 4: Summing up the functions of expressions of low-value obligation: requesting, giving, and reporting permission (quoted from Løken 2007: 136)[8]

Function / Expression	request	give	report	
			receiver	source and receiver
may	+	++		
can		+		
kan	+	++		
kunne	+	+	+	
kan få	+	+		
få			++	
allow/permit				+
tillate				+
let				++
la				++
allowed/permitted	-		++	
tillatt/lov			-	-
permission		-		++
tillatelse	-	-		++
lov (noun)			++	

(9) He's my oldest friend, so I'*m allowed to* call him a great pedant. (JB1)
Han er min eldste venn, så jeg *har lov til å* kalle ham for pedant.

(10) To use the drug, you would have to *get permission* from this hospital and the next of kin. (AH1)
For å bruke medisinen, må De *ha tillatelse* av sykehusets ledelse og av nærmeste pårørende.

In (9) the receiver of permission is expressed, in (10) both the receiver and the source of permission.

This brief account can only give an indication of Løken's work. As shown in Table 4, the study goes beyond the modal auxiliaries and illuminates both intra- and interlingual relationships.

5. Expressions of habituality

My last example relates to some work I have done myself on expressions of habituality in English, Norwegian, and German on the basis of the ENPC and the English-German-Norwegian material of the OMC (Johansson 2005, Johansson 2007: 139ff.). A similar study has been carried out for English and Swedish

(Altenberg 2007), albeit restricted to the expression of past habit. Interestingly, the author ends his paper by calling for a broader cross-linguistic comparison:

> To arrive at a general language-independent definition of habituality we need to compare a wide range of languages. Then we might be able to define some prototypical notion of habituality that is shared by a number of languages and regard deviations from this notion as extensions or restrictions of this prototypical idea. (Altenberg 2007: 127)

My work widens the number of languages somewhat and extends the study to include expressions of present as well as past habituality.

5.1 Norwegian *pleie* and its English correspondences

Norwegian *pleie* (cognate of German *pflegen*) is a natural starting-point, as it is a common verb which can be used to mark both present and past habituality. The Norwegian verb overlaps with English *used to*, but the conditions of use of the two forms are quite different. The correspondences for the present-tense form *pleier* and the past-tense and past-participle forms *pleide/pleid* differ in a number of ways; see Table 5:

Table 5: Correspondences of *pleier* and *pleide/pleid* in ENPC fiction (based on Bjerga 1998)

E form type	*Pleier*		*Pleide/Pleid*	
	E translation	E source	E translation	E source
Simple present	12	4	-	-
Simple past	-	-	6	7
Used to	-	-	21	33
Would	-	-	6	22
Adverbial constr.	19	4	20	8
Be in the habit of	-	-	1	3
Other	6	2	1	16
Total	37	10	55	89

Judging by the overall distribution, habituality in Norwegian is more commonly marked in the past than in the present tense.[9] The frequency is strikingly low for the present-tense form *pleier* in translation from English, presumably because English has no grammaticalised marker of habituality in the present tense and the translator is therefore less likely to choose an explicit marker. On the other hand, the frequency in translation is high in the case of the past tense, where English has verb forms for the expression of habituality (see below).

Pleier most commonly corresponds to an adverbial combining with a simple present-tense verb form:

(11) Hun *pleier å gjøre* sånt når håret hans blir for langt og han skal til frisøren. (LSC1)
She *usually does* this when his hair gets too long and he's going to go to the barber.

(12) Jeg gikk fort, *som jeg pleier* [lit. 'I walked fast, as I usually-do'], uten å se meg om, gikk og så ned i fortauet, og plutselig var hun der, like foran meg, kom rett imot meg med barnevognen. (EHA1)
As usual I was walking quickly, my eyes on the sidewalk, and suddenly there she was right in front of me; she came right towards me with the baby carriage.

Usually is most frequent by far. Note the change of construction in (12), where there is a progressive verb form. *As usual* here refers to how the speaker normally walks, while *was walking* describes the particular situation.

The second most frequent correspondence type is a simple present-tense verb form, as in:

(13) Moren finner plass til dem i hjørnet, og det er stille som i graven, som bestefaren *pleier å si* når han forteller om bestemoren. (LSC1)
Mother finds room for them in the corner, and it is as quiet as the grave, as Grandfather *says* when he talks about Grandmother.

(14) "*Vi pleier ikke drasse på* [lit. 'we usually-do not cart'] håndveskene våre," forklarte fru Olsen. (EG1)
"*Nobody here carts* a handbag around with them all day," explained Mrs Johnsen.

(15) Søndag morgen, mens moren lager frokost, *pleier hun å krype* opp i sengen til faren, kryper inntil varmen hans, snuser på ham, moren sier hun er for stor til det. (BV2)
On Sunday mornings, whilst mother is making breakfast, *she slips* into bed with father, snuggles up to his warmth and nuzzles at him. Mother says she is too old for that.

The simple present tense works well, since it commonly refers to something habitual, in contrast to the present progressive. In (15) habituality is in addition marked by the initial adverbial.

For the past-tense form *pleide* we also sometimes find simple verb forms, as in:

(16) Det var på slike steder chokonene *pleide slå seg ned*. (SH1)
Those were the sort of sites the Chokonen *chose for their camps*.

(17) Faren *pleide aldri være borte* mer enn tre dager om gangen. (KAL1)
Father *never stayed away* for more than three days at a time.

In (16) the context makes it clear that the reference is to a repeated situation rather than a single past event. In (17) the presence of the adverb *never* makes additional marking of habituality redundant.

As in the case of the present-tense form *pleier*, we commonly find adverbial constructions as correspondences, in this case combining with simple past-tense verb forms. Again *usually* is most frequent, but other forms occur as well, as in:

(18) Han var ikke så pratsom som ellers, han arbeidet fortere enn han *pleide* og fortere enn han likte. (BV1)
He was not as talkative as usual, he worked more swiftly than he *normally did* and more swiftly than he liked to.

However, the most striking correspondences for the past-tense form *pleide* are *used to* and *would*, as in:

(19) Jeg *pleide å kjøpe* mat til'n. (LSC2)
I *used to buy* food for him.

(20) Mor til Magda *pleide synge* gamle viser når hun satt ved rokken og spant. (PEJ1)
Magda's mother *used to sing* old folk songs when she sat at the spinning wheel and spun wool.

(21) "Du finner aldri en ektemann, så gal som du er," *pleide han å si*. (SL1)
"You'll never find a husband, the way you go on," *he would say*.

(22) Jeg *pleide å kjøpe* med et par middagsaviser fra tobakkshandelen ved siden av, fant et lite bord borte langs en av veggene og ble sittende for meg selv. (GS1)
I *would buy* a couple of evening papers at the tobacconist's next door and take them in with me to read, find a small table against one of the walls at the back and sit there on my own.

Though the two forms often work in the same context, *would* is more limited in distribution and is typically used in narrative style to describe characteristic behaviour (cf. Quirk et al. 1985: 228). *Used to* is less formal than *would* and can also be applied to a past state, as in:

(23) Det var henne jeg *pleide å like* best av barna mine, da hun var liten sa hun ofte at jeg var den beste faren i verden. (KA1)
I *used to like* her the best of all my children, and when she was small she often said I was the best father in the whole world.

Using *would* in this situation, with the stative verb *like*, would result in quite a different meaning.[10]

The contrast between *used to* and *would* is discussed more fully in Altenberg (2007). A particularly interesting point is the behaviour of the two forms in sequences of habitual events, where *would* is typically non-initial:

> […] once the habitual nature of the sequence has been established by *used to* and/or some other information in the context, the following events can be expressed by *would* (Altenberg 2007: 123)

As an example, consider the context preceding (22) above:

(24) Det hendte at jeg spiste middag der, men som oftest drakk jeg bare et glass øl eller to. Jeg *pleide* å kjøpe med et par middagsaviser fra tobakkshandelen ved siden av, fant et lite bord borte langs en av veggene og ble sittende for meg selv. (GS1)
I ate lunch there occasionally but most often I made do with a glass or two of beer. I *would* buy a couple of evening papers at the tobacconist's next door and take them in with me to read, find a small table against one of the walls at the back and sit there on my own.

Would further lacks the notion of 'discontinuity' which is characteristic of *used to* and which suggests "that the past situation is 'discontinued', i.e. that it no longer applies and therefore contrasts with the moment of speaking" (Altenberg 2007: 126). This aspect of *used to* comes out clearly in its Norwegian and German correspondence patterns.

5.2 English *used to* and its Norwegian correspondences

At this point, let's reverse the perspective and examine Norwegian correspondences of *used to*. The mutual correspondence between *used to* and *pleide* is surprisingly low; *used to* corresponds to *pleide* in as little as a third of the material in the ENPC; in addition, we occasionally find the verb *brukte* (meaning literally 'used', cf. also Swedish *brukade*). Most often there is an adverbial combining with a past-tense verb form, as in:

(25) He *used to have* narrow gray slits of eyes; now they were wide and startled. (AT1)
Vanligvis hadde ('usually had') han trange, grå sprekker til øyne, men nå var de store og skremte.

(26) He *used to read* to us at night, Baby and me, whenever there were no meetings. (NG1)
Ofte leste ('often read') han høyt for oss, for Baby og meg, på kvelder når det ikke var møter.

(27) […] he is *acting*. Performing what he *used to* be. (NG1)
Han spiller en rolle. Spiller den han *før var* ('before was').

(28) It *used to* be the public washing square, and was known still to all the locals as the Soap Garden. (JC1)
En gang var ('one time was') dette den offentlige vaskeplassen, og stedet var fremdeles kjent for alle de fastboende som Såpehagen.

The Norwegian translations use either a frequency adverbial, as in (25) and (26), or an adverbial referring to a time in the past, as in (27) and (28). Frequency adverbials in the material include:

> *alltid* ('always'), *av og til* ('on and off'), *ofte* ('often'), *stadig* ('constantly'), *vanligvis* ('usually')

The meaning expressed in the translation may vary from 'always' to 'on and off', but all these adverbials express the notion of repeated occurrence. Though high-frequency forms are predominant, the exact frequency seems to be open to interpretation. Time adverbials in the material include:

> *den gang(en)* ('that time'), *en gang* ('once'), *en gang i tiden* ('once', lit. 'one time in the time'), *før* ('before'), *tidligere* ('earlier').

The adverbials in the second set of correspondences bring out the notion of discontinuity which is characteristic of *used to* (cf. the end of Section 5.1).

If there is already an adverbial in the context, or some other indication that the reference is to a habit or state in the past, there may be no need to add another adverbial corresponding to *used to*, as in:

(29) In the winter business was quieter, and Arthur *used to like* to spend Tuesdays and Thursdays from November through to March […] with whoever it was it happened to be. (FW1)
Om vinteren var det mer stille i butikken, og Arthur *likte* ['liked'] å tilbringe tirsdager og torsdager fra november til mars […] med hvem det nå kunne være.

(30) In his boyhood he *used to look to* the Queen for inspiration. (ST1)
Som barn *hadde han sett opp til* [lit. 'as child had he looked up to'] dronningen som et ideal.

(31) It came with cream, just the way it *used to* at his grandmother's house. (AT1)
Macon tok en honningkake med krem, akkurat slik *han hadde fått den* ['he had got it'] hjemme hos bestemor.

Note in the last two examples that the Norwegian translator has opted for the past perfect, thereby making explicit the notion of discontinuity associated with *used to*.

Judging by the correspondences, *used to* places a situation in the past and clarifies that the reference is not to a single event. *Used to* has a wider distribution than *pleide*. It can refer both to recurring events and to continuous states in the past. *Pleide*, on the other hand, is rarely found in combination with state verbs. Most important, unlike *used to*, the Norwegian verb *pleie* is not restricted to the past tense. The present-tense form *pleier* is best conveyed in English by a simple verb form or an adverbial construction.

5.3 German expressions of habituality

Since German *pflegen* and Norwegian *pleie* are closely related both in origin and meaning, we might at the outset have expected them to behave in the same way. This is not at all the case. The mutual correspondence (MC) is low:[11]

	Norw > German	**German > Norw**	**MC**
pleie vs. ***pflegen***	$\frac{7 \times 100}{57} = 12\%$	$\frac{20 \times 100}{24} = 83\%$	$\frac{(7 + 20) \times 100}{57 + 24} = 33\%$

What this means is that *pleie* is rarely translated into *pflegen*, whereas *pflegen* is translated into *pleie* in the great majority of cases. The correspondence is also low in relation to English *used to*; out of 70 examples in the English-German translations of the OMC, only four had a form of *pflegen*. The verb *pflegen* in the habituality sense seems to be formal, perhaps even old-fashioned.[12] So how is habituality expressed in German?

Judging by the material examined, German commonly resorts to adverbial markers. Here are some examples of translations of the present-tense form *pleier* and the past-tense form *pleide*:[13]

(32) Hun *pleier å gjøre* sånt når håret hans blir for langt og han skal til frisøren. (LSC1)
Das *macht* sie *immer*, wenn seine Haare zu lang werden und er zum Friseur muß.
She *usually does* this when his hair gets too long and he's going to go to the barber.

(33) Faren *pleide å gi* ham en bok […]. Moren *pleide å lese* i den for ham. (EFH1)
Der Vater *gab* ihm *immer* ein Buch […]. Die Mutter *las* ihm *immer* daraus vor.
His father *used to give* him a book […].His mother *used to read* to him out of it.

(34) Hun hengte fra seg klærne på gangveggen og lukket døra varsomt etter seg som hun *pleide.* (HW1)
Sie hängte ihren Mantel in den Flur und schloß sorgsam die Tür hinter sich, wie sie es *immer tat*.
She hung her coat up on the hook in the entryway and closed the door gently behind her as she *usually did*.

(35) Han *pleide å holde til* et bord eller to bortenfor meg. (GS1)
Gewöhnlich saß er ein oder zwei Tische von mir entfernt.
He *usually sat* a table or two away from me.

In German we typically find verb forms combining with frequency adverbials, most often *immer* ('always').

Sometimes the Norwegian original contains *pleier* in combination with *alltid* ('always'), i.e. habituality is overtly marked both by a verb and by an adverbial:

(36) Det er noe Herman alltid har lurt på, om latteren egentlig er en sykdom, for moren *pleier alltid å si* at latteren smitter. (LSC1)
Das wollte Herman schon immer gern wissen, ob Lachen eigentlich eine Krankheit ist, denn Mutter *sagt immer*, daß Lachen ansteckt.
That's something Herman has always wondered about, if laughter really is a sickness, because Mother *always says* that laughter is contagious.

(37) "[…] Jeg *pleier alltid å vite* hvor han er." (OEL1)
"[…] Ich *weiß* eigentlich *immer*, wo er ist."
"[…] I *always know* where he is."

Here the translations only preserve the adverbial in combination with a present-tense verb form.

If we turn to German translations of English *used to*, we generally find adverbial forms, as shown in:

(38) Another thing he *used to do*, like going straight to the fridge for a glass of water, he *used to call*, Aila? Aila? if she wasn't in the first room he entered. (NG1)
Noch etwas *hat* er *immer getan*, so wie er immer schnurstracks zum Kühlschrank gegangen ist, um sich ein Glas Wasser zu holen; *immer hat* er Aila? Aila? *gerufen*, wenn sie nicht in dem ersten Raum war, den er betrat.
Og noe annet han *pleide å gjøre*, som å gå rett til kjøleskapet etter et glass vann. Han *pleide å rope* Aila? Aila? hvis han ikke så henne med det samme han kom inn.

(39) I'm trying to reach a John Daggett, who *used to live* in this area. (SG1)
Ich versuche, einen John Daggett zu erreichen, der *früher* hier in der Gegend *gewohnt hat*.
Hallo, jeg prøver å få tak i en som heter John Daggett og som *bodde her før* ('lived here before').

The frequency adverbials represented are: *immer* (common), *oft*, *oft genug*, *manchmal*; i.e. adverbials with the meanings 'always', 'often (enough)' and 'sometimes'. Past time adverbials include *einmal*, *mal*, *früher* (most common by far), *früher mal*, *früher immer* (!), *vor gar nicht so lange Zeit*, meaning roughly either 'once' or 'before'. These closely match the two sets of adverbial forms found in Norwegian translations (see Section 5.2). Both sets co-occur with verb forms referring to past time, and the reference is to repeated events or states in the past.

As in the case of the Norwegian translations, there may be no explicit marker corresponding to *used to*:

(40) I *used to save* my money for opera when I had a free weekend. (ABR1)
Ich *sparte* mir mein Geld für die Oper auf, wenn ich mal ein freies Wochenende hatte.
Jeg *pleide å spare* pengene mine til operaen når jeg fikk en frihelg.

(41) The name of the gallery is Sub-Versions, one of those puns *that used to delight me* before they became so fashionable. (MA1)
Die Galerie heißt Sub-Versions, eines jener Wortspiele, *an denen ich Spaß hatte*, solange sie noch nicht derart in Mode waren.
Galleriet heter Sub-Versjoner, en type ordspill *som pleide å more meg* før det gikk inflasjon i dem.

(42) "I *used to admire* you. Now I despise you. I *used to find* you amusing. Now you bore me. I *used to love* you. Now I just feel sorry for you." She smiled apologetically. (MW1)
"*Früher habe* ich dich *bewundert*. Jetzt verachte ich dich. Ich *fand* dich amüsant. Jetzt langweilst du mich. Ich *liebte* dich. Jetzt tust du mir nur leid." Sie lächelte entschuldigend.
"Jeg *beundret* deg *den gangen* ['that time']. Nå avskyr jeg deg. *Den gangen syntes* jeg du var morsom. Nå kjeder du meg. *Den gangen elsket* jeg deg. Nå synes jeg bare synd på deg." Hun smilte unnskyldende.

In (40) and (41) the Norwegian translators opted for *pleide*, whereas the German translators just picked verb forms with past-time reference. In (42) an adverbial is used only in the first sentence of the German version; in the Norwegian version, the adverbial is repeated. The stylistic means are different, but the same message comes across.

A notable German correspondence is *sonst*, which is found both in rendering *pleide* and *pleier*:

(43) Den ødelagte skulderen hang enda mere enn den *pleide*. (HW1)
Die verstümmelte Schulter hing noch mehr herunter als *sonst*.
His crippled shoulder drooped even more than it *usually* did.

(44) I dag gjør jeg noe jeg sjelden *pleier* å gjøre […]. (KF1)
Heute tue ich etwas, was ich *sonst* selten tue […].
Today I do something I don't *usually* do very often […].

(45) Og han må ta på gråbuksene som stikker og skjorten han bare *pleier* å bruke søttende mai og julaften. (LSC1)
Und er muß die grauen Hosen anziehen, die kratzen, und das Hemd, das er *sonst* nur am Nationalfeiertag und zu Weihnachten anzieht.
And he has to put on the gray trousers that pinch and the shirt that he *usually* wears on Independence Day and Christmas Eve.

(46) – Jeg *pleier* ikke drikke noe, sier han. (TB1)
"Ich trinke *sonst nie*", sagt er.
"I don't *usually* drink," he says.

(47) Selv Rachel som *alltid pleier* å være full av prat og latter, går uten å si noe. (TB1)
Auch Rachel, die *sonst immer* plaudert und lacht, sagt nichts.
Even Rachel, *usually* laughing and talkative, goes along without a word.

German *sonst* is similar to English *otherwise* and Norwegian *ellers*, but the latter did not turn up as correspondences of expressions of habituality in the material examined here.[14]

The use of *sonst* as an expression of habituality is noted in dictionaries I have consulted: *Oxford-Duden German Dictionary*, and *Das große Wörterbuch der deutschen Sprache* (Duden). What seems to be happening is that the preferred German marker of habituality, *immer*, would clash with other elements in (43) to (45):

(43') *Die verstümmelte Schulter hing noch mehr herunter als *immer*.

(44') *Heute tue ich etwas, was ich *immer* selten tue […].

(45') *Und er muß die grauen Hosen anziehen, die kratzen, und das Hemd, das er *immer* nur am Nationalfeiertag und zu Weihnachten anzieht.

Sonst is appropriate because it sets up a contrast between what happens in a particular situation and what happens under other circumstances, i.e. usually. Interestingly, *sonst* may combine with frequency adverbials: *selten* in (44), *nie* in (46), and *immer* in (47).

6. Summing up and interpretation

I have taken up three studies showing how we can use a multilingual corpus. In the corpus we observe correspondences. These must be interpreted, however. English *here* and Norwegian *her* turn out to be surprisingly different in use. Hasselgård interprets the results as revealing differences in patterns of cohesion between the two languages. Løken outlines a methodology for identifying modality expressions. The example I quoted provides an instance of functional interpretation of corpus findings.

In my study of expressions of habituality we see both similarities and differences between English, German, and Norwegian. A variety of formal means are used, though preferences differ. All three languages use frequency adverbials to mark habituality. English stands out by having explicit verbal markers for past habituality (*used to* and *would*) and by commonly using simple verb forms without any overt formal marker. In addition to adverbials, Norwegian has a special verb (*pleie*) which is often used both in the present and the past tense. Although German has a cognate verb (*pflegen*), it appears to be more marginal and translators generally opt for adverbials, the most notable of which is *immer*. It remains to be explained how *immer* has developed into the preferred marker of habituality in German.

Pleie stands out as the preferred marker of habituality in Norwegian. There are indications that this verb is becoming grammaticalised in the habituality sense. It is only used together with a main verb (expressed or ellipted). Although it typically combines with action verbs, it is also found with state verbs, as in (23) and (37) above. It sometimes drops the infinitive marker *å* before the following

verb, as in (16) and (17). It easily combines with frequency adverbials, as in (36) and (37), possibly an indication of semantic bleaching.

Used to differs from *pleie* not only in being restricted to past-time reference, but also in combining freely with state verbs and in having two sets of adverbial correspondences: frequency adverbials and past-time adverbials. We can take this to mean that it is a marker of a somewhat different kind than *pleie*. It can refer both to recurring events and to continuous states in the past, contrasting them with the present; (42) above is a striking example. Altenberg (2007: 127) suggests that "[f]rom a Swedish point of view *used to* must be regarded as polysemous". The same conclusion can be drawn from our comparison with German and Norwegian. In other words, the contrastive perspective throws new light on English *used to*.

There is a need for further detailed work on expressions of habituality, exploring aspects such as:

- the range of expressions of habituality and their conditions of use;
- the degree of overt marking and the conditions of zero correspondence;
- combinations of markers of habituality;
- similarities and differences across a wider range of languages, including preferred ways of expressing habituality;
- diachronic changes in the expression of habituality.

With respect to the last point, it is interesting to note that the restriction to the past tense of English *used to* did not apply in Middle English and Early Modern English (see the *Oxford English Dictionary*, *use* v., 21). How can we account for the restriction in use in later English? This is a matter for further investigation.

7. The way forward

The studies reported here show the potential of exploiting multilingual corpora for language comparison and for throwing special features of the languages compared into relief, including preferred ways of expressing similar meanings. Using multilingual corpora we can perceive the characteristics of each language in a new way. This is why we might talk about 'contrastive linguistics in a new key'. There are important applications in lexicography, language teaching, and the training of translators.

Studies of multilingual corpora are still in their infancy, and we have only just started to exploit the potential of these resources. Some challenges that lie ahead are (see also Johansson 2007: 301ff.): we need to widen the range of languages, including the variety of texts. We need multi-register corpora. We need corpora with annotation of features which cannot be easily found in raw, unannotated text. Above all, we need to learn more about how we can best exploit multilingual corpora.

If used with care and imagination, multilingual corpora lead us beyond what we knew or did not see so clearly. This is the essence of the cross-linguistic perspective. To my mind, using multilingual corpora is a good means of doing language research and an important direction of research for the future.

Notes

1 See the websites listed at the end of the paper. For more information on the corpus models developed at the University of Oslo, see Johansson (2007: 9ff.).

2 See the website for the SPRIK (Språk i kontrast 'Languages in contrast') project: http://www.hf.uio.no/ilos/forskning/prosjekter/sprik/english/

3 See Hasselgård (1997, 1998, 2000, 2004a, 2004b, 2005).

4 In quotations the original version is generally listed first and is accompanied by a reference code. For an explanation of the reference codes, see the websites listed at the end of the paper.

5 The texts in the ENPC are extracts of 10–15 thousand words, 30 original fiction texts for each language and 20 original non-fiction texts, in all 200 texts, or about 2.7 million words.

6 For a full survey of correspondences, with examples, see Hasselgård (2004b).

7 The thickness of the lines reflects the strength of correspondence.

8 ++ indicates typical use; + indicates frequent use; - indicates use non-existent in the material examined. Due to the low frequencies, *could* and *might* are not included.

9 Because of the way the corpus was set up, i.e. with equal amounts of text in both languages, we can compare raw frequency figures. (See Johansson (2007: 14) for further details.)

10 Note that *pleie* with a stative verb is somewhat unusual (see Section 5.2), but (23) is an attested example from a text by a highly acclaimed Norwegian author (Kjell Askildsen).

11 For the calculation of mutual correspondence, see Altenberg (1999).

12 The Swedish cognate verb *pläga* (as opposed to the common habituality marker *bruka*) seems to be even more unusual. Not a single example was found in the *English-Swedish Parallel Corpus*.

13 The English translations are included for comparison.

14 These findings suggest that these expressions in German, English, and Norwegian deserve further investigation. For a comparison of English *otherwise* and Norwegian *ellers*, see Fretheim (2004).

Websites

English-Norwegian Parallel Corpus (ENPC):
http://www.hf.uio.no/ilos/english/services/omc/enpc/

English-Swedish Parallel Corpus (ESPC):
http://www.sol.lu.se/engelska/corpus/corpus/espc.html

Oslo Multilingual Corpus (OMC):
http://www.hf.uio.no/ilos/english/services/omc/

References

Aijmer, Karin and Anne-Marie Simon-Vandenbergen (2003), 'The discourse particle *well* and its equivalents in Swedish and Dutch', *Linguistics*, 41: 1123–1161.

Altenberg, Bengt (1999), 'Adverbial connectors in English and Swedish: semantic and lexical correspondences', in: Hilde Hasselgård and Signe Oksefjell (eds.) *Out of corpora. Studies in honour of Stig Johansson*. Amsterdam & Atlanta, GA: Rodopi, 249–268.

Altenberg, Bengt (2007), 'Expressing past habit in English and Swedish: a corpus-based contrastive study', in: Christopher S. Butler, Raquel Hidalgo Downing and Julia Lavid (eds.) *Functional perspectives on grammar and discourse. In honour of Angela Downing*. Amsterdam & Philadelphia: Benjamins, 97–128.

Bjerga, Trude Davidsen (1998), Continuative and habitual aspect in English and Norwegian, with special reference to the English verb *keep* and the Norwegian verb *pleie*. Unpublished *hovedfag* thesis. Department of British and American Studies, University of Oslo.

Dyvik, Helge (1998), 'A translational basis for semantics', in: Stig Johansson and Signe Oksefjell (eds), *Corpora and cross-linguistic research: theory, method, and case studies*. Amsterdam & Atlanta, GA: Rodopi, 51–86.

Filipović, Rudolf (1969), 'The choice of the corpus for the contrastive analysis of Serbo-Croatian and English', in: *The Yugoslav Serbo-Croatian – English contrastive project B. Studies 1*. Institute of Linguistics, University of Zagreb, 37–46.

Fretheim, Thorstein (2004), '"Switch-polarity" anaphora in English and Norwegian', *Special issue on contrastive lexical pragmatics. Working Papers ISK*, 1/2004, 45–67.

Hasselgård, Hilde (1997), 'Sentence openings in English and Norwegian', in: Magnus Ljung (ed.) *Corpus-based studies in English. Papers from the Seventeenth International Conference on English Language Research on Computerized Corpora*. Amsterdam: Rodopi, 3–20.

Hasselgård, Hilde (1998), 'Thematic structure in translation between English and Norwegian', in: Stig Johansson and Signe Oksefjell (eds), *Corpora and cross-linguistic research: theory, method, and case studies*. Amsterdam & Atlanta, GA: Rodopi, 145–168.

Hasselgård, Hilde (2000), 'English multiple themes in translation', in: Alex Klinge (ed.) *Contrastive studies in syntax* (Copenhagen Studies in Language 25). Frederiksberg: Samfundslitteratur, 11–38.

Hasselgård, Hilde (2004a), 'Thematic choice in English and Norwegian', *Functions of Language*, 11 (2): 187–212.

Hasselgård, Hilde (2004b), 'Spatial linking in English and Norwegian', in: Karin Aijmer and Hilde Hasselgård (eds.) *Translation and corpora*. Göteborg: Acta Universitatis Gothoburgensis, 163–188.

Hasselgård, Hilde (2005), 'Theme in Norwegian', in: Kjell Lars Berge and Eva Maagerø (eds.) *Semiotics from the North. Nordic approaches to systemic functional linguistics*. Oslo: Novus, 35–47.

Johansson, Stig (1998), 'On the role of corpora in cross-linguistic research', in: Stig Johansson and Signe Oksefjell (eds.), *Corpora and cross-linguistic research: theory, method, and case studies*. Amsterdam & Atlanta, GA: Rodopi, 3–24.

Johansson, Stig (2005), 'Some aspects of usuality in English and Norwegian', in: Kjell Lars Berge and Eva Maagerø (eds.) *Semiotics from the North. Nordic approaches to systemic functional linguistics*. Oslo: Novus, 69–86.

Johansson, Stig (2006), 'How well can *well* be translated? On the English discourse particle *well* and its correspondences in Norwegian and German', in: Karin Aijmer and Anne-Marie Simon-Vandenbergen (eds.) *Pragmatic markers in contrast*. Amsterdam: Elsevier, 115–137.

Johansson, Stig (2007), *Seeing through multilingual corpora: on the use of corpora in contrastive linguistics*. Amsterdam & Philadelphia: Benjamins.

Løken, Berit H. (2007), *Beyond modals: a corpus-based study of English and Norwegian expressions of possibility* (Acta Humaniora 296). Faculty of Humanities, University of Oslo.

Noël, Dirk (2003), 'Translations as evidence for semantics: an illustration', *Linguistics*, 41: 757–785.

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

Spalatin, Leonardo (1969), 'Approach to contrastive analysis', in: *The Yugoslav Serbo-Croatian–English contrastive project B. Studies 1*. Institute of Linguistics, University of Zagreb, 26–36.

English style on the move: variation and change in stylistic norms in the twentieth century

Geoffrey Leech, Nicholas Smith and Paul Rayson

University of Lancaster, University of Salford and University of Lancaster

Abstract

This paper has two related purposes. First, our goal is to explain the results of recent research on twentieth century British (as well as American) English, using equivalent corpora of general written (published) English known as the 'Brown Family' of corpora. Limiting our attention to British corpora, the 'Brown Family' contains three matching corpora of a million words each, the B-LOB, LOB and F-LOB corpora, sampled at roughly thirty-year intervals (1931±3[1] years, 1961 and 1991). (A fourth corpus from 1901±3 is under development, and one-third of it will be used in the latter part of this paper.) These enable us to trace the changing history of written (published) British English over a sixty-year period. Through changes in frequency in grammatical categories and constructions across a variety of genres, we observe largely consistent patterns of change which lend themselves to explanations in terms of what may be called general stylistic trends. To these trends we give such names as colloquialization *(movement towards spoken norms of usage),* densification *(movement towards denser or more compact expression of meaning) and* democratization *(the trend towards avoidance of discrimination or inequality in the linguistic treatment of individuals). Only the first two of these trends will be explored in this paper.*

In the second part of the paper, we show how general stylistic norms, such as are provided by the 'Brown Family' corpora, can be used as a reference norm *against which statistical deviations identify some of the characteristic features of style of an individual author or an individual text. For this we make use of Rayson's Wmatrix software (http://ucrel.lancs.ac.uk/wmatrix/) for comparing (groups of) texts in terms of lexical, grammatical and semantic characteristics. Although the comparison is in some respects lacking in accuracy, it identifies typical style markers of an individual text, ordering them in terms of their differentness from the reference norm. It remains to be seen how far this computational technique can place the elusive notion of authorial style on an objective footing, but results so far are promising.*

1. Style in terms of frequency

There are many definitions of style (see Enkvist 1973, Wales 2001: 370–372, Leech and Short 2007: 34–57), but in an everyday sense, a style is understood to

be a particular way of using the language, or a particular way of expressing meanings. These definitions fit the traditional literary concept of authorial style (the Miltonic style, Johnsonese, Woolf's prose style and so on), but can also be used outside the literary domain, to refer to the style of newspaper headlines, of email messages, of TV weather forecasts, and the like. In fact, style overlaps with terms like register and genre, and typically concerns language variation within the standard language, in this case English. Variation, in turn, implies differentness – the way linguistic choices pattern in X as opposed to Y (where X and Y are varieties defined by their own set of situational parameters). In fact, there is an implicit comparison of varieties in any discussion of style. The 'Miltonic' or 'Johnsonian' or 'internet advertising' style only makes sense against the background of some norm of comparison.

From a textual point of view, style is strongly associated with frequency. If it is claimed, for example, that Henry James is fond of abstract nouns, or that D. H. Lawrence is fond of adjectives, this (rather simplistic) stylistic claim can only be tested by comparing the frequency of that linguistic characteristic in the author's works with its frequency in some **reference corpus**, such as a representative set of texts (a corpus) of the author's period, used as a standard of comparison.[2] The promising side of such definitions is that they make style, in principle, a measurable commodity. In practice, however, there are difficulties. How, exactly, do we select a 'representative set of texts' for comparison with the works of James or Lawrence? Another question is: how do we decide on the set of features to be compared?

Two further issues are easier to deal with, at least at the present day, but nevertheless deserve mention. A difficulty which was insuperable in the 1950s when Bloch (see note 2) was writing – that of calculating frequencies in large bodies of texts – has now been substantially overcome by electronic text storage and text processing and the techniques of corpus linguistics. An additional question 'How do we compare frequencies in corpora of different sizes?' is relatively easy to answer: we make the two corpora 'as if of equivalent size' by comparing them in terms of relative frequency – say, occurrences per million words. (A common alternative procedure for measuring relative frequency is to consider a feature as a variant of a variable, and to calculate percentages of occurrence within the variant field. For an example of this, see Section 2.)

Let us think, then, of the X and Y being compared as corpora. In the case where an author's novels are being compared with a reference corpus of other writers' novels, then the domain of the 'authorial corpus' is less general than that of the reference corpus. But another type of comparison is between two corpora of equal generality: for example, a comparison between the Brown Corpus (of American English published in 1961) and the LOB Corpus (of British English published in 1961); or between the LOB Corpus and the F-LOB Corpus (of British English published in 1991). Here we are comparing equivalent samples of language use, differing only in the time/place of their origin. In this chapter, we will explore stylistic comparisons of both these types.

It is important that, if possible, there should be only one parameter on which the selection criteria for texts differ. In the case of authorial style, the authorship of the texts making up the corpus is the crucial criterion. In the case of the Brown and LOB corpora, the contrasting dimension is place (the country of the texts' origin being the US v. the UK), and in the case of the LOB and F-LOB corpora, the contrasting dimension is time: the year of the texts' composition (1961 v. 1991). We can also compare more specific varieties. It is meaningful to ask 'In what ways does the style of British government documents differ from that of American government documents of the same year (say 1961)?'. It is also meaningful to ask 'In what ways does the style of British newspaper editorials of 1961 differ from those of 1991?'. The diachronic comparison is the one on which we concentrate in the next part of this paper.

2. Diachronic studies of style: B-LOB, LOB and F-LOB

To demonstrate diachronic studies of style, we make use of the three equivalent corpora of published British English (BrE): the Lancaster-Oslo/Bergen (LOB) Corpus (sampled from texts first published in 1961), the Freiburg-LOB (F-LOB) Corpus (sampled from texts first published in 1991) and the Before-LOB (B-LOB) Corpus (sampled from texts first published in 1931±3).[3] These corpora, having virtually the same sampling design as the American Brown and Frown corpora, belong to the set of corpora known as the Brown Family (named after the original corpus, the Brown Corpus). The corpora each consist of c. 1,000,000 words from 500 text samples (each of c. 2000 words), classified in the 15 text categories. They are as close as is feasible to the ideal of **comparable corpora**, differing only in that their dates of publication differ (in this case by a generation-gap of 30 years) from their chronologically neighbouring corpora.

For diachronic comparisons of this kind, while we want the corpora to be comparable in their make-up, we also want them to be (as far as possible) representative. This means, if the sampling frame is of published English, that we would like them to be sampled from a broad cross-section of published text types (or genres), so that whatever we find in the corpora is likely to be generalizable, within limits of approximation, to the published written language as a whole. This was broadly the intention behind the design of the Brown Family of corpora, although more 'marginal' text types, such as poetry, dramatic texts and advertisements, were excluded from their make-up.[4]

For most purposes, it is convenient to subdivide the Brown Family corpora into four subcorpora, to which the fifteen text categories are allotted, as follows (see Table 1, p.72).

In what follows, we will minimize the presentation of numerical statistics, and will instead show the changing patterns of style using line charts. The changes represented are in most cases of high statistical significance. As the corpora and subcorpora vary slightly in their word counts,[5] instead of raw word counts, the frequency data are normalized to occurrences per million words (pmw). As two introductory examples of the line charts we will use, we now give

the changing occurrence of negative contractions (reductions of negative *not* to *n't*) as represented in the three comparable corpora; see Figures 1 and 2 (below).

Table 1: The make-up of Brown Family corpora in terms of subcorpora and numbers of text samples (*n*)

Subcorpora	N	Including the following text categories
Press	88	Press: A. reportage B. editorial C. reviews
General prose	206	D. religion. E. skills, trades and hobbies. F. popular lore. G. belles lettres, biography, memoirs, etc. H. miscellaneous (largely government documents)
Learned	80	J. Learned (academic)
Fiction	126	Fiction: K. general. L. mystery and detective. M. science fiction. N. adventure and western. P. romance and love story. R. humour
Total	500	

Figure 1 shows frequencies per million words – which, of course, is close to the raw frequency count. Figure 2 (p. 73) shows frequencies in comparison with frequencies of the corresponding full forms: that is, occurrences of *hasn't*, *didn't*, etc. are shown as a percentage of cases of *potential* contraction (including *has not*, *did not*, etc. as well as their contracted forms).

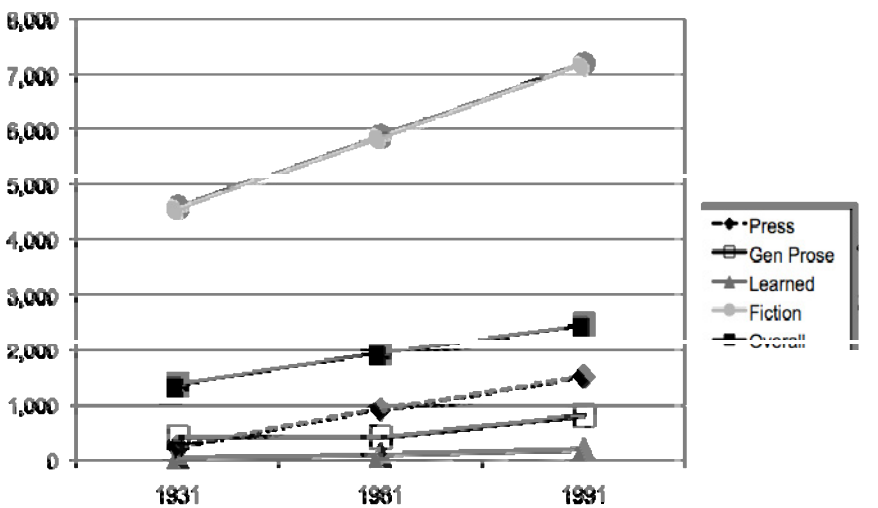


Figure 1: *Not*-contractions in twentieth-century BrE: frequencies per million words (pmw)

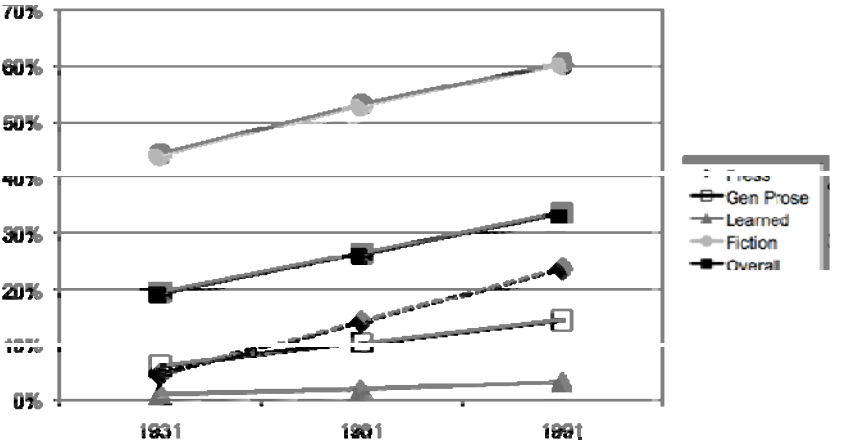


Figure 2: *Not*-contractions as a proportion of all *not*-negations in twentieth-century BrE

The reason for showing these two pictures of the same data is this: our two definitions of style at the beginning of the chapter – style as [1] a chosen way of using language and [2] as a chosen way of expressing meanings – correspond to the two different ways of representing the frequency of occurrence of a particular feature (in this case contractions). Figure 1 gives us frequencies in their most simplistic form (normalised to pmw), while Figure 2 gives frequencies in relation to another choice which keeps the meaning constant. That is, using style in sense [2], we can say that *I don't know* and *I do not know* differ 'merely in style', whereas, for example, *I don't know* and *I do know* differ more radically – in terms of meaning. It is often assumed that the measurement of frequency in language variation and change should be restricted to the latter, as the more linguistically relevant measure. However, it is certainly more difficult to measure change in a corpus in sense [2]: there is indeed a serious difficulty in deciding what options should be counted as meaning-preserving;[6] and in some cases there is no obvious alternative expression for the one chosen. Even in the present case, in some contexts contractions cannot substitute for full forms (try putting *n't* at the start of a sentence), and in other contexts full forms can scarcely substitute for contractions (in tag questions like *isn't it?*, for example).

The important point to make here, however, is that the two charts tell very much the same story. They show that the use of contractions has been increasing steadily between 1931 and 1991, not only in the corpora as a whole, but in each of the four subcorpora. They also show not unexpected differences between the

subcorpora: the frequency of contractions is by far the highest in Fiction writing, which also shows a dramatic increase.[7] At the other extreme, the Learned subcorpus has a very low incidence of contractions, and only a very small increase. The other subcorpora – Press and General Prose – are intermediate between these extremes, but over the sixty-year period, contractions in the Press (which often shows innovative and trend-setting tendencies) have overtaken those of General Prose in frequency.[8]

Having demonstrated how the two ways of measuring the increase of contractions tell essentially the same story, we will feel free in what follows to use the simpler way of measuring change of frequency – the measurement of occurrences per million words (pmw for short).

2.1 Colloquialization (including de-formalization)

The increasing use of contractions just discussed exemplifies a rather general stylistic trend observed taking place in the twentieth century, which can be termed colloquialization: this is the tendency for written language to move closer to the characteristics of spoken language.[9] Varieties of spoken and written English can be placed on a scale of distance from the most colloquial extreme – extempore speech in private contexts – to the most formal extreme – which, in the Brown Family corpora, is represented by the Learned subcorpus.[10] Of the remaining three subcorpora, Fiction, although it contains a great variety of styles, is in general the subcorpus that is closest to conversation – as is clearly shown in the case of contractions. We can therefore see colloquialization as a process of progressively moving towards the conversational pole of the scale, and away from the formal/literate end (see Figure 3).

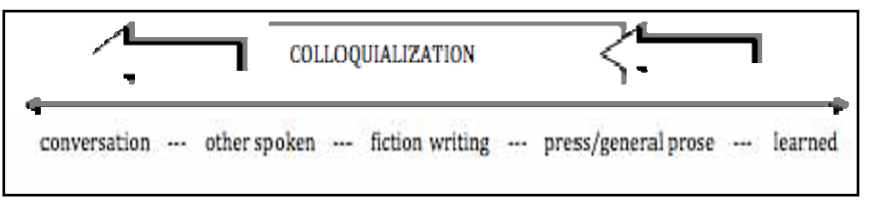


Figure 3: A rough representation of the formality scale with colloquialization

In the context of language development as a whole, colloquialization can be seen as a manifestation of the trend for innovation to arise in the spoken language and to spread to the written language, rather than vice versa. However, in the process of change, our conception of the scale itself is not immobile – the norm of what makes a colloquial style, what makes a formal style, also changes over time. When a passage such as the following (from *An Introduction to Report Writing*,

by William Lumb, Pitman, 1933) is read by a present-day reader, it strikes us, in historical retrospect, as distinctly formal in relation to the text category of 'popular lore' (part of the General Prose subcorpus) to which it belongs. (Relevant features of formal written style are italicized.)

(1) Special training in law, accountancy, social investigation, etc., *is required* in order to deal efficiently with documents, records, and books of account and to extract from them information *upon which* a report may *be based*. [B-LOB, F04]

Three grammatical features of stylistic interest in (1) are the passive,[11] the preposition *upon* and the so-called pied-piping construction (relativization by preposition + relative pronoun). To show that this extract is no mere oddity, the following sentence from a B-LOB text in the same 'popular lore' category (*How to Appeal against Your Rates*, by A. Stanley Eamer, Pitman 1930) gives a similar instance of formality:

(2) The grounds *upon which* the ratepayer is enabled to exercise his powers *are* statutorily *prescribed* to be those of incorrectness or unfairness, wrong insertion or omission from the Valuation List, or the valuation as a single hereditament of a building, or portion of a building, occupied in parts. (B-LOB, F06)

As we see from Figures 4–6 (pp. 76–77), these three features have all been declining since 1931, but the decline of *upon* is more marked than that of pied-piping, which in turn is more marked than that of the passive.

The passive and pied-piping charts both show an inverse pattern, compared with Figures 1 and 2, of lowest frequency in Fiction and highest frequency in Learned, which is what we expect if these are to be examples of de-formalization (the negative side of colloquialization). However, the picture in the case of *upon* is slightly different – the lowest frequency in LOB and F-LOB is found in Press, with Fiction the second lowest. This may represent a somewhat more complex case, where the reasons for the decline could be a combination of colloquialization and information compression (cf. 'densification', below). Journalists conscious of the need to save space could easily, in most cases, substitute *on* for *upon*, which might be part of the reason for the particularly steep decline in this subcorpus.

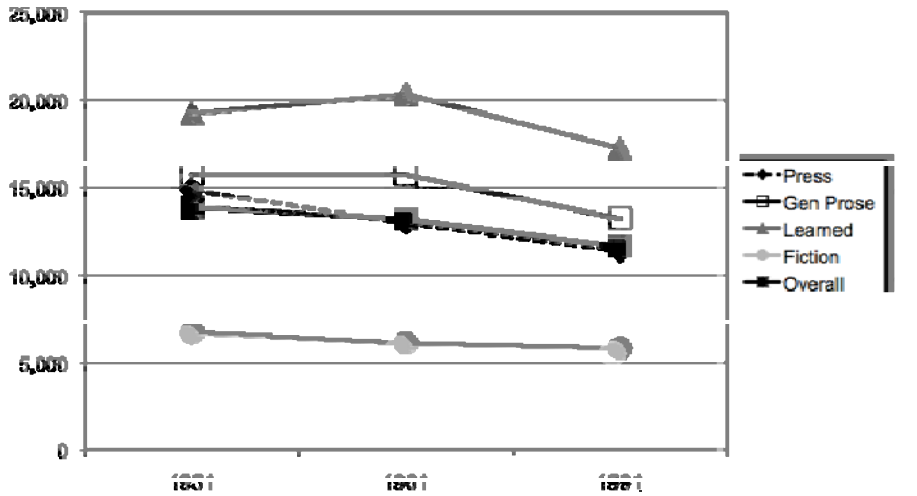


Figure 4: Passive voice in twentieth-century BrE: frequencies pmw

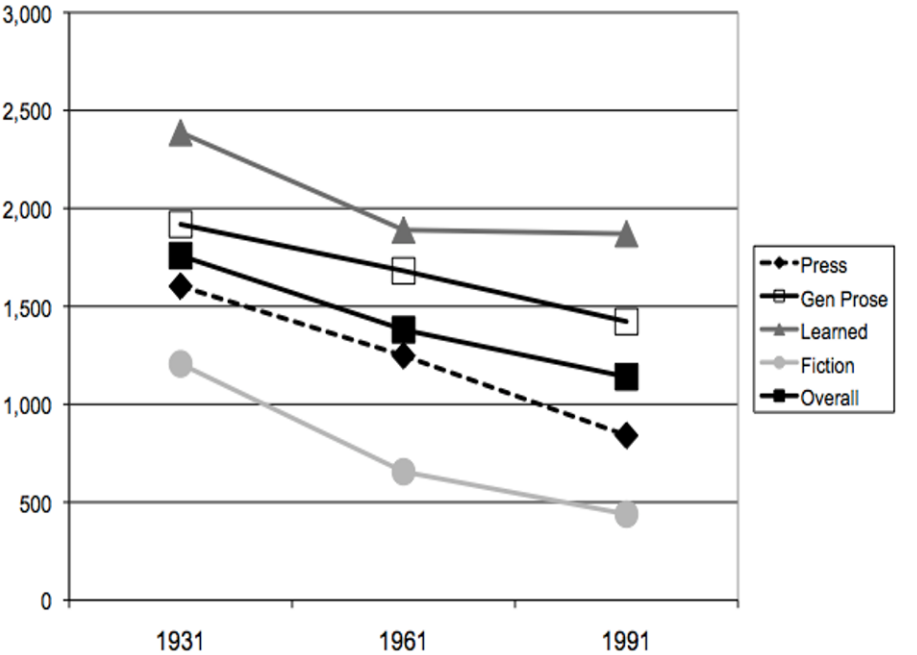


Figure 5: Pied-piping in twentieth-century BrE: frequencies pmw

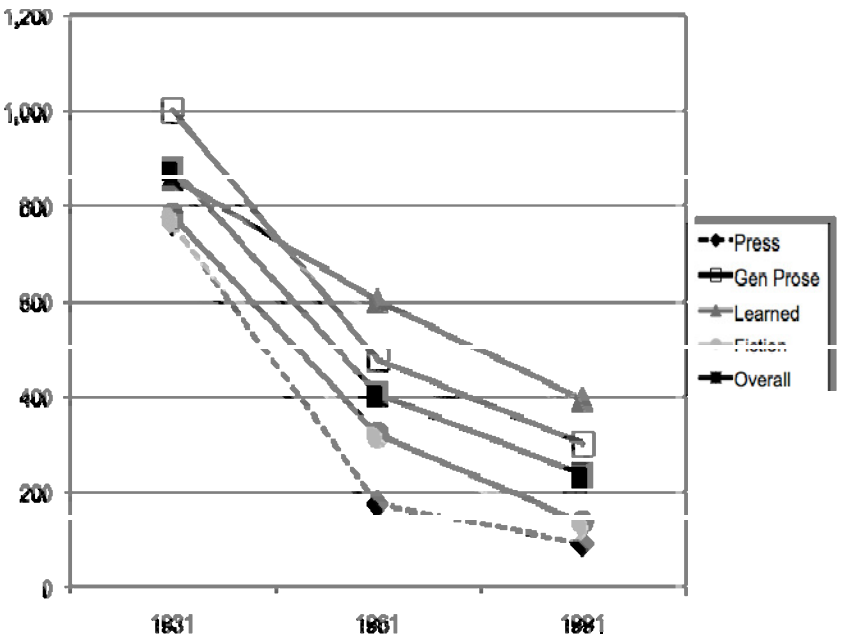


Figure 6: *Upon* in twentieth-century BrE: frequencies pmw

Another possible reason for the somewhat different profile of *upon* is that the Fiction subcorpus is actually a mixture of different styles. We have characterized it, simplistically, as the subcorpus that approximately most closely to spoken language – which, in broad generality, is true. Yet Fiction contains, in its descriptive and narrative passages, some more literary and conservative styles – especially since Fiction can depict life in earlier periods of history.[12] A check on the occurrences of *isn't it* (as an example of contraction) and *upon* in LOB showed that *isn't it* occurs mainly in speech quotation (64 out of 113 instances), whereas speech quotation accounted for only 1 out of 238 instances of *upon*. In brief – *upon* has no connection with the most speech-like parts of the Fiction subcorpus, and what we see in Figure 6 is consistent with the hypothesis that the decrease of *upon* is an instance of de-formalization.

If we measure decrease over the 60 years as a percentage of the original 1931 frequencies pmw, the decline of the passive is 15.9%, that of pied-piping is 35.2%, and that of *upon* is 73.2%. Like the increase in contractions, these are all very significant changes. But we have examined only four instances of colloquialization. Other instances could have been cited – for instance:

Further plausible cases of colloquialization

increasing use of semi-modals such as *have to, want to* and *need* to (Leech et al. 2009: 98–105)
increasing use of the progressive construction (Leech et al. 2009: 124–127)
increasing use of *that*-relativization (Leech et al. 2009: 229–231)
increasing use of questions: especially verbless questions (Leech et al. 2009: 242–243)
increasing use of preposition stranding (Leech et al. 2009: 231–233)
decreasing use of *wh*-relativization (Leech et al. 2009: 228–229)
decreasing use of 'no negation', as contrasted with *not* negation (Leech et al. 2009: 241–242; e.g. *We saw no one*, as contrasted with *We didn't see anyone*.)

If we accept the premise that colloquial features in the Brown Family corpora will be most frequent in the Fiction subcorpus and that formal features will be most frequent in Learned, then all the above (as discussed in Leech et al. 2009: 239–245) show features characteristic of speech on the increase or features uncharacteristic of speech on the decrease – both trends indicative of colloquialization.[13]

2.2 Densification

It would be wrong, however, to give the impression that colloquialization affects all aspects of the language. For example, modal auxiliaries as a class are significantly more frequent in speech than in writing – but their use in the written language has not been on the increase: in fact, it has been decreasing. Moreover, nouns – word classes which are significantly more frequent in the more formal, written registers than in speech, have been increasing (rather than decreasing, as they should in accordance with the colloquialization hypothesis) over our sixty-year period in the written language.

What is the explanation for this increased 'nouniness' of written language? Since nouns are key elements of the noun phrase, it means that nouns, as key parts of the noun phrase, have been taking a bigger role in written syntax. However, not all elements of the noun phrase have been increasing – prepositions, for example, particularly *of*, have been on the decline. The motivating force behind this change in the direction of 'nouniness' appears to be a need to compress more semantic content into fewer words, particularly through combinations such as Noun + Noun sequences and Noun's + Noun (*s*-genitive) sequences. Consequently, increasing use has been made of single-word modifying elements preceding the noun head of the noun phrase, rather than of phrasal elements following it. In certain contexts, the phrases labelled *a, b, c* in (3) and (4) can be seen as stylistic alternatives:

(3) *a*. the fruit of the coconut palm [Brown F34] - N_1 of N_2
b. the coconut palm's fruit - N_2's N_1
c. coconut palm fruit - N_2 N_1

(4) *a*. the behavior of a patient - N_1 of N_2
b. a patient's behavior [Brown J34] - N_2's N_1
c. patient behavior - N_2 N_1

As these examples show, choice of option *b*. (*s*-genitive) or *c*. (juxtaposition of nouns) can be a way of reducing the number of words used to express a given meaning, as compared with the more explicit form *a*. – acting, in effect, as a means of textual compression. This trend that we call **densification** has been attributed to the increasing complexity of modern society, the 'information explosion', and the need for more efficient and specialized information transfer (see Biber 2003).

This runs counter to colloquialization, because in spoken language meaning tends to be more diffusely expressed: nouns are uncommon, and pronouns common, compared with other varieties. Thus Biber et al. (1999: 65) show that lexical density[14] is lower in the conversation register (at c. 35%) than in written registers, and reaches its peak in News writing, at c. 54%. An increase in lexical density of 2.6 percentage points between LOB and F-LOB, therefore, is a manifestation of what we may call densification. Figures 7 (p. 80) and 8 (p. 81) show that the steady increase of Noun + Noun sequences[15] and of *s*-genitives has been truly remarkable between 1931 and 1991.

These diagrams show patterns of increase rather different from those in the colloquialization charts shown earlier. In Figure 7, the lowest frequency and increase of Noun + Noun sequences is found in Fiction, the inverse of what we found in Figure 2 for contractions. The highest frequency is found, however, not in Learned writing but in the Press, reflecting the common suspicion that this densification in style in the twentieth century has been spearheaded by journalism.[16] The same 'leadership' role of journalism is suggested for the *s*-genitive in Figure 8, where the Press subcorpus shows the greatest frequency and the sharpest rise, although the lowest frequency in this case is found in the Learned subcorpus – a matter we address briefly in 2.3 below.

2.3 Conclusions regarding diachronic style change

We have explored two contrasting types of diachronic style change – colloquialization and densification – as manifested in the use of a selection of grammatical features of the language. There are, however, one or two loose ends in this discussion that need to be addressed. One question to be explored is: can we explain the coexistence of these apparently antagonistic trends? If we look again at Figure 6 (*upon*) and Figure 8 (the *s*-genitive), we may be able to see a sign of their symbiosis.

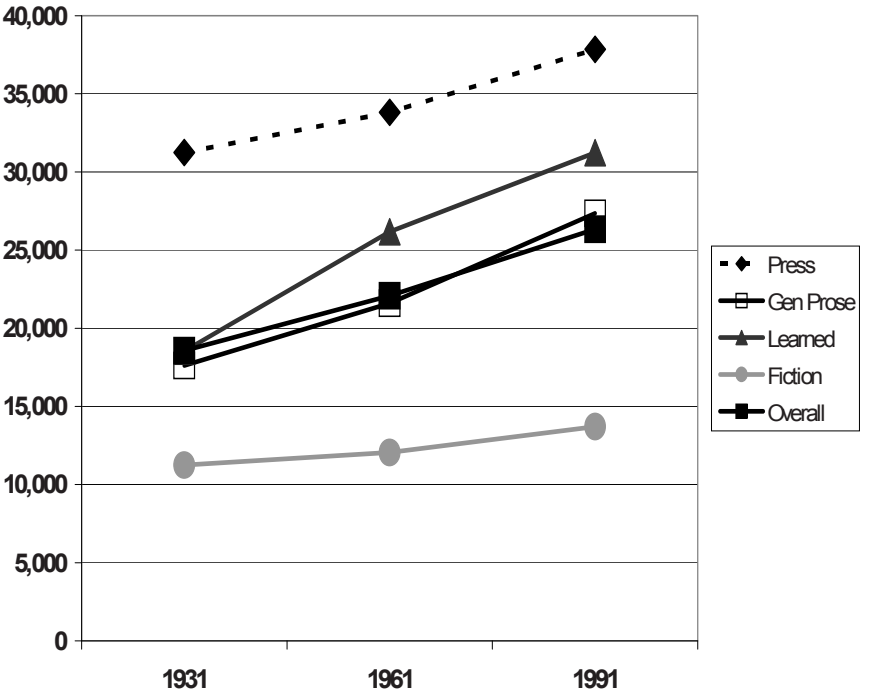


Figure 7: Noun-noun sequences in twentieth-century BrE: frequencies pmw

In Figure 6, the particularly precipitous decline of *upon* can possibly be explained as combining colloquialization (in its negative form of de-formalization) and densification. *Upon*, being commonest in Learned writing and General Prose, is associated with formal registers, and so colloquialization provides one possible reason for its decline. Another reason is provided by densification: *upon* is a two-syllable preposition which can normally be replaced by the single-syllable preposition *on*, with a consequent gain in density.

In Figure 8, this phenomenon of the two trends working *with* rather than *against* one another can perhaps also explain the especially sharp increase of *s*-genitives. On the one hand, as already suggested, the *s*-genitive compresses information into a smaller compass than its habitual rival the *of*-genitive. (See (3) and (4) above: the *s*-genitive generally saves one, or two, function words – here the preposition *of* and an article.)[17] On the other hand, the *s*-genitive is less formal than the *of*-genitive. Figure 8 shows that the genitive, in 1931±3, was as frequent in Fiction as in Press, although they diverged greatly after that. But nouns are in any case relatively infrequent in Fiction. This is revealed by Table 2 (p. 81), which indicates that the frequency of *s*-genitives relative to the frequency of nouns in the four subcorpora was as high in Fiction as in Press. That is, if we

consider the frequency of the *s*-genitive in relation to the opportunities for using genitives – for which nouns are obviously required – Fiction, along with Press, shows the highest in frequency of genitives.[18]

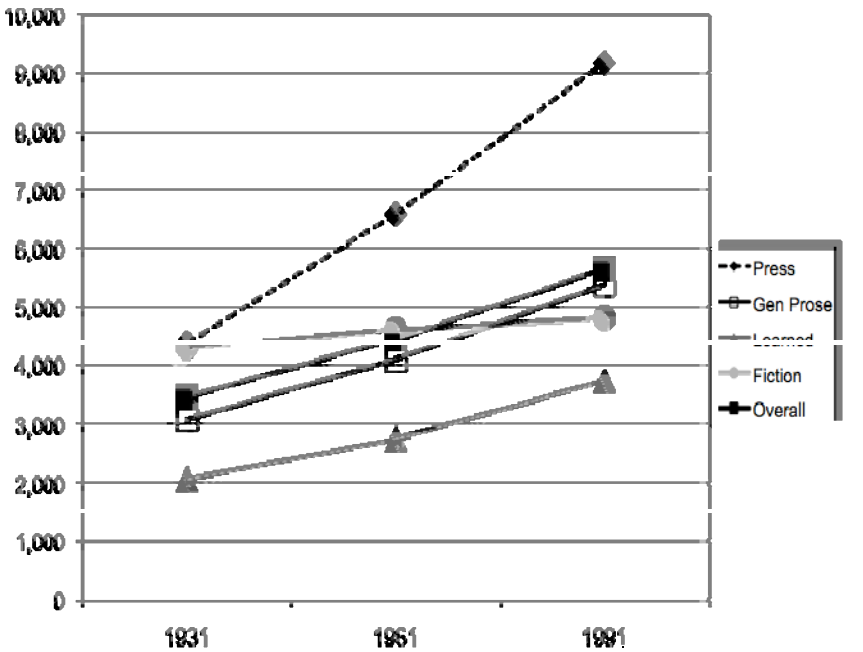


Figure 8: *S*-genitive in twentieth-century BrE: frequencies pmw

Table 2: Frequency of *s*-genitives as a percentage of frequency of nouns in the LOB Corpus

	Press	General prose	Learned	Fiction
a. frequency of *s*-genitives (pmw)	7,278	4,646	3,082	4,921
b. frequency of nouns (pmw)	296,198	259,943	261,802	200,212
a as a percentage of b	2.46%	1.79%	1.18%	2.46%

Most genitives, of course, require two nouns: the genitive noun itself, and the head noun. If the frequency of nouns is low, the frequency of genitives is expected to be correspondingly low. But this is not always the case with the *s*-genitive: given that the opportunities for using the genitive are of low frequency

in Fiction, its incidence is high. Compared with the formal/literate nature of the *of*-genitive, the *s*-genitive is the more colloquial option. It is plausible, then, that the rise in the *s*-genitive is due to two factors in combination – both the densifying effect and the colloquializing effect of this construction.

The relation between colloquialization and densification is therefore a mixture of opposition and cooperation. Biber (2003) has talked of the 'competing demands of popularization vs. economy' in the use of language in modern mass media – which may explain this mixture of trends. A style which meets these media-driven demands is likely to be one that manages to combine colloquialization with densification.

3. Style as a comparison between a focal text and a reference corpus

We turn now to our second, synchronic, exploration of style. For this, we focus on a short text – a short story by Virginia Woolf entitled 'The Mark on the Wall' (see Leech 2008: 162–178 for an earlier and lengthier treatment of this topic).

3.1 Virginia Woolf's 'The Mark on the Wall': our focal text

'The Mark on the Wall', written in 1917, might be described as a story in which nothing happens – where nothing happens, that is, except in the mind of the narrator. (We use the term 'narrator' here, although it is the *inner* voice of the narrator that we experience throughout the story.) The narrator, sitting down after tea, notices a mark on the wall. Her mind explores in a myriad ways the significance of that mark – what it might be, and where it came from. This train of thought leads her by digressions of memory and imagination to such topics as the preceding occupants of the house – the nature of life – life after death – the oddities of experience – the mysteries of existence – always following the stream of the narrator's consciousness. Every so often, however, the narrator's attention comes back to the mark on the wall – and at last, she learns what it is. To give the flavour of the text, here are its opening paragraph and the final few lines:

> *Opening paragraph:*
> Perhaps it was the middle of January in the present year that I first looked up and saw the mark on the wall. In order to fix a date it is necessary to remember what one saw. So now I think of the fire; the steady film of yellow light upon the page of my book; the three chrysanthemums in the round glass bowl on the mantelpiece. Yes, it must have been the winter time, and we had just finished our tea, for I remember that I was smoking a cigarette when I looked up and saw the mark on the wall for the first time. I looked up through the smoke of my cigarette and my eye lodged for a moment upon the burning coals, and that old fancy of the crimson flag flapping from the castle tower came into my mind, and I thought of the cavalcade of red knights riding up the side of the black rock. Rather to

my relief the sight of the mark interrupted the fancy, for it is an old fancy, an automatic fancy, made as a child perhaps. The mark was a small round mark, black upon the white wall, about six or seven inches above the mantelpiece.

Ending:

... – but something is getting in the way ... Where was I? What has it all been about? A tree? A river? The Downs? Whitaker's Almanack? The fields of asphodel? I can't remember a thing. Everything's moving, falling, slipping, vanishing ... There is a vast upheaval of matter. Someone is standing over me and saying:

'I'm going out to buy a newspaper.'

'Yes?'

'Though it's no good buying newspapers. Nothing ever happens. Curse this war; God damn this war! ... All the same, I don't see why we should have a snail on our wall.'

Ah, the mark on the wall! It was a snail.

3.2 Comparing the focal text and a reference corpus

We introduced in Section 1 the idea that stylistic analysis is essentially a comparative process. An automatic method of comparing bodies of text in order to characterize their 'differentness' is provided by the Wmatrix software developed by Paul Rayson (for details, see Rayson 2008; also http://ucrel.lancs.ac.uk/wmatrix/). For our purposes, as we are interested here in the stylistic analysis of a single text, the comparison will be between a single text (**the focal text**) and a corpus (**the reference corpus**).

The focal text, 'The Mark on the Wall', will be compared quantitatively with a reference corpus which should be representative to some degree of the variety from which the text is taken. However, there are obviously different degrees of generality in defining the language variety meant to act as a reference standard. We have decided to use three different 'reference varieties' (the choice being determined, obviously, by the availability of suitable texts in electronic form):

(A) A rather specific variety, resembling the focal text in three ways: it consists of (1) fiction writing (2) by women writers (3) published in 1917. On the other hand, this reference corpus is limited in representativeness, as it contains only three novels, the work of three authors.[19]

(B) A more general corpus of fiction, consisting of category K (General Fiction) in the Fiction subcorpus of B-LOB. This is more widely representative than (A), as it contains 29 text samples by different authors. However, it is less closely matched than (A) in time of publication, as the samples date from 1928–1934.

(C) A very general corpus, sampled from the written (published) English of roughly the same period and national variety (British English of the beginning of the twentieth century) as the focal text. For this we used a third of the as yet incomplete 1901±3 corpus of the Brown Family, covering all four of the subcorpora Press, General Prose, Learned and Fiction.[20] The corpus is not closely matched with 'The Mark on the Wall' temporally – indeed it is a worse match than (B), but may be considered more broadly representative than the other two of the written prose of the period, containing 166 text samples across a wide range of fiction and non-fiction writing.[21]

In practice, none of our reference corpora are ideal; and one of the interests of this study was to discover how far the differences between the three reference corpora of increasing generality would produce different results.[22] So, what is the method of comparison?

The methodology employed by Wmatrix is broadly definable as an extraction from the data of **keywords**, or rather **key features**: that is, words or other features of the text which stand out or deviate, in a statistical sense, from the frequencies of the reference corpus. The statistical concept of **keywords** has become familiar in corpus linguistics since it was built into the popular corpus software package WordSmith Tools (Scott 2004), and has since been the basis of a considerable body of published research.[23] In the case of Wmatrix, however, this method has been extended further to grammatical word classes (parts of speech) and to semantic domains, as will be shortly explained. In other words, the comparison is not purely lexical.

To begin with keywords: by 'keyness' here is meant the words which are most distinctive of that text, as contrasted with the reference corpus. Keyness so understood is of variable strength, so that the output of this process of keyword extraction is a list, in which words are listed in order of keyness. Similar lists can be obtained for any other features of language automatically identifiable in the textual data. The general set of procedures involved in a research project of this kind can be listed as the four stages below:[24]

1. Building the data: corpus design and compilation (in the case of our Wmatrix investigation, this has already been sufficiently described in terms of a focal text and reference corpora).

2. Annotating the data: analysing the corpus linguistically, using particular annotation tools: in the case of Wmatrix, the two annotation tools used are

(a) the CLAWS part-of-speech (POS) tagger, and

(b) the USAS semantic domain tagger.

Details of these tools are to be found on the UCREL (Lancaster) website at: http://ucrel.lancs.ac.uk/claws/ and http://ucrel.lancs.ac.uk/usas/.[25]

3. Retrieving: extracting from the text data some analytic results, which may be displayed in a variety of formats for inspection or further processing. In the

Wmatrix analysis, we are interested in three more or less standard listing formats:

(a) concordances, which list the occurrences of a particular word (or other feature) in their contexts of occurrence,

(b) frequency lists, which list words (or other features) in order of their frequency in a particular body of text data, and

(c) keyness lists, which list words (or other features) in order of their keyness in a given textual comparison.

4. Interpreting: This is the only stage of the process which is essentially non-automatic ('manual'), although it can be aided by automatic procedures such as using the 'Sort' and 'Collocation' facilities of corpus software. Whereas stage 3 above is essentially automatic and quantitative, stage 4 is qualitative: it makes use of the human ability to interpret texts and to explain the phenomena observed in them. In the case of the Wmatrix investigation, we may be interested here in examining the textual material more carefully, using especially the concordance displays, in order to explain the stylistic phenomena observed in the analysis.

We now have to focus on the third, 'Retrieving' stage above, in order to explain in a little more detail what the software does. At the same time, we will avoid going into technical detail, which can be studied in Rayson (2008) and on the UCREL webpages already cited.

To take the most basic case, the list of keywords is arrived at as follows:

i) Two word frequency lists are compiled: a list for the focal text ('List X'), and a list for the reference corpus ('List Y').
ii) List X and List Y are compared. This means that each word in List X is measured in terms of *comparative frequency* with the same word in List Y.[26] 'Comparative frequency' means that the raw count of a word's frequency is adjusted to a standard measure relative to corpus size, which in Wmatrix is the number of occurrences of the word as a percentage of all occurrences of words in the text/corpus.
iii) Each word's keyness in the focal text is measured by a statistical formula, which calculates the degree to which the word is either 'over-represented' or 'under-represented' in this text, as measured against the reference corpus. The normal understanding of keyness is that the word is *over*-represented, that is, is relatively more frequent in the focal text than in the reference corpus, to a certain high degree of statistical significance.[27]
iv) The words in List X are re-ordered in order of keyness. This means that the words at the top of the list are most distinctive of that text.

Concordance, frequency and key-feature lists of POS tags and semantic tags are extracted in the same way as the word lists described in 3(a)–(c) above. There are no particular difficulties in this, as the annotation (tagging) has meant that each word in each text is accompanied by label giving its grammatical and semantic classification.

3.3 Results: keywords, key POS tags, and key semantic domain tags

To begin with, Table 3 shows the top 12 keywords, in order, when 'The Mark on the Wall' is compared with each of the reference corpora.

Table 3: Keywords: words of abnormally high frequency in 'The Mark on the Wall'

A. compared with three 1917 novels by women writers	
1. mark	7. worshipping
2. is	8. thoughts
3. one	9. of
4. Whitaker	10. tree
5. wall	11. Precedency
6. tablecloths	12. chancellor

B. compared with 1931 general fiction (category K of B-LOB)	
1. mark	7. of
2. is	8. nail
3. wall	9. reality
4. thoughts	10. tablecloths
5. Whitaker	11. worshipping
6. one	12. tree

C. compared with the ⅓ 1901±3 Brown-family corpus	
1. mark	7. one
2. wall	8. I
3. Whitaker	9. Precedency
4. thoughts	10. mantelpiece
5. tablecloths	11. nail
6. worshipping	12. tree

Note: double underlining marks the words which are in the top 12 for all three comparisons; single underlining marks the words which are in the top 12 for two of the three comparisons.

Perhaps the most striking result is the amount of agreement that the three reference corpora show, in spite of their very different composition. Comparisons with A and B share all of their top 10 key words (out of 12); A and C share 9 of the 12; and B and C share 11. Perhaps this is a mild reflection of the degree of generality of the corpora. It seems that the keyword methodology is robust in

showing up the 'differentness' of a text without respect to the exact make-up of the reference corpus.

It is not surprising that *mark* is the 'keyest' of the keywords: it represents the theme of the story, as to a lesser extent does *wall*. These are words that, as we might imagine, occur relatively rarely in the reference corpora, and therefore their repeated use in 'The Mark' is salient, both statistically and thematically. Of the other words which occur in all three comparisons, *one* (typically used in the generic human sense) is perhaps a personal stylistic favourite of Virginia Woolf, representing as it does the objectification of the narrator's personal experiences, as illustrated in the following passage:

> because *one* will never see them again, never know what happened next ... as *one* is torn from the old lady about to pour out tea and the young man about to hit the tennis ball in the back garden of the suburban villa as *one* rushes past in the train.

We will not dwell on the items in this list, some of them uncommon words, like *Precedency*, which gain idiosyncratic prominence in Woolf's narrative – see Leech (2008: 168–171) for further discussion. But there are some interesting points to observe about the similarities and differences between the lists. For example, *is* is very much overrepresented when compared with the fictional reference corpora (but not with the more general reference corpus C), and this is probably because Woolf, in capturing the immediacy of the interior monologue, tells much of her story in the historic present, instead of using the past tense narrative convention of the majority of fictional writers. This choice of the present tense is understandably not so salient when compared with the full range of written texts (scientific, journalistic, etc.) in the 1901±3 corpus. On the other hand, the pronoun *I*, frequent in Woolf's first-person narrative, stands out as over-represented when compared with the cross-section of written texts in 1901±3, but is less salient in the two fiction reference corpora, where first person reference occurs frequently, for example in dialogue.

We move on now to the lists of key part-of-speech tags, reflecting the different grammatical choices made by Virginia Woolf as compared with the writers in the other reference corpora; see Table 4 (p. 88).

The amount of shared 'key tags' between the comparisons here is the same: nine tags are shared by the top twelve in A, B and C. What brings A and B closer together, however, is the fact that the top four tags are the same and in the same order. As mentioned above, the present tense (represented in the keyness of the *s*-form of lexical verbs VVZ as well as of VBZ and VV0), is a distinctive feature of 'The Mark', as opposed to fiction written in the more conventional past-tense narrative.

Table 4: The most 'key' parts of speech in 'The Mark on the Wall'

compared with three 1917 novels		compared with 1931 general fiction		compared with 1901 Brown Family corpus (⅓)	
1. VVZ	7. DDQ	1. VVZ	7. VV0	1. PN1	7. NN2
2. NN2	8. PPIS1	2. NN2	8. AT	2. PPIS1	8. PNX1
3. PN1	9. PNX1	3. PN1	9. PNX1	3. VVZ	9. RGQ
4. VBZ	10. NPD1	4. VBZ	10. RPK	4. VVG	10. DDQ
5. IO	11. RPK	5. IO	11. RGQ	5. RPK	11. AT1
6. AT1	12. RGQ	6. DDQ	12. NPD1	6. VV0	12 PPH1

Note: as in Table 3, double underlining marks the tags which are in the top 12 for all three comparisons; single underlining marks the tags which are in the top 12 for two of the three comparisons.

Key: AT – article neutral for number; chiefly the definite article *the*
AT1 – singular article; chiefly the indefinite article *a/an*
DDQ – *wh*-determiner or *wh*-pronoun (e.g. *what, which*)
IO – the preposition *of*
NN2 – plural common noun (e.g. *tables, women, thoughts*)
NPD1 – singular weekday noun (e.g. *Sunday, Monday*)
PN1 – singular indefinite pronouns (e.g. *one, anything, nobody*)
PNX1 – indefinite reflexive pronoun (i.e. *oneself*)
PPH1 – third person personal pronoun *it*
PPIS1 – the first person subject pronoun *I*
RGQ – *wh*-adverb of degree (*how* when modifying another word)
RPK – *about* used in the expression *be about to*.
VBZ – present tense *–s* form of the verb to be (i.e. *is*)
VVG – *ing*-form of lexical verb (e.g. *saying, wishing*)
VVZ – present tense lexical verb ending in *–s* (e.g. *says, wishes*)
VV0 – present tense lexical verb not ending in *–s* (e.g. *say, find*)

More difficult to explain is the second-keyest tag, the plural noun tag NN2; however, the following passage illustrates how Woolf's style may favour plural nouns in describing the multitudinous particularity of her experiential world:

let me just count over a few of the *things* lost in one lifetime, beginning, for that seems always the most mysterious of *losses* – what cat would gnaw, what rat would nibble – three pale blue *canisters* of book-binding *tools*? Then there were the bird *cages*, the iron *hoops*, the steel *skates*, the Queen Anne coal-scuttle, the bagatelle board, the hand organ – all gone, and *jewels*, too. *Opals* and *emeralds*, they lie about the *roots* of *turnips*.

It is striking, also, that this passage contains four examples of another key tag, IO (representing the preposition *of* in the tagging system). The word, of course, has many functions – but its main function, in the most general terms, is to signal the interconnectedness of things. It is noticeable in this list that IO stands out as a key tag in relation to the fictional reference corpora A and B, but not in relation to the most general reference corpus C, which is predominantly non-fictional. Elaboration of noun phrases by means of *of* is likely to be a characteristic of informational texts, which oddly here seem to be more akin to Woolf's own elaborative style. Of the other key tags, we will comment only on PN1, PNX1 and RGQ. PN1 chiefly represents the pronoun *one* already noted as favoured in 'The Mark'; and PNX1, normally a very rare tag (representing the word *oneself*) stands out in this text even though there are only two occurrences of it. RGQ represents the adverb *How* as a modifier, in this text especially associated with exclamations:

> How readily our thoughts swarm...
> How shocking, and yet how wonderful it was to discover...
> How peaceful it is down here.

This construction may, indeed be another authorial favourite of Virginia Woolf, indicative of the narrator's (or a character's) characteristic emotional involvement in her subject matter.[28]

The third level of analysis, that of semantic tagging, produces lists of key semantic domains as shown in Table 5 (pp. 90–91).

Key semantic domains tell us about the 'aboutness' of texts, rather than about their stylistic characteristics in the strict sense. They are therefore less relevant to most of our proceeding discussion of style, and there is less agreement between the different reference corpus comparisons: only half of the key semantic domains listed are shared by all three lists. On the other hand, there are some features which are salient not so much in style as in the authorial world view. The domain of colour is high on the list of key domains in all three comparisons, as are the domains relating to the natural world: 'Plants' and 'Living creatures'. Readers of Virginia Woolf will probably agree that these traits have a 'key' role in her writing. Other, more abstract domains are more difficult to interpret, but arguably reflect her exploration of the nature of reality and the ontological concerns of her writing. At the other extreme, the domain of 'Smoking' must be regarded as incidental to the text, in that it results from the semantic tagging of four words only: one of the drawbacks of choosing such a short focal text for analysis that such haphazard results can occur. Here is another excerpt, which contains a reference to smoking, but is also relevant to some other key features:

Even so, life isn't done with: there are a million patient watchful lives for a tree, all over the world, in bedrooms, in ships, on the pavement, lining rooms, where men and women sit after tea, smoking cigarettes. It is full of peaceful thoughts, happy thoughts, this tree.

This passage illustrates representation of some of the key features high on the list above: Plants (*tree*), Life and living things (*life, lives*), Mental object; conceptual (*thoughts*), Parts of buildings (*bedrooms, rooms*). Obviously there is much more to be said about this story, and the extent to which the 'key' analysis succeeds in highlighting stylistically important features. But the main point of this section of the paper has been to illustrate the potential of such analyses, using a chosen text and three alternative reference corpora of different generality.

Table 5: The most 'key' semantic domains in 'The Mark on the Wall'

compared with three 1917 novels	compared with 1931 general fiction	compared with 1901 Brown-family corpus (⅓)
1. General & abstract (*thing, things*)	1. Evaluation: authentic (*real, reality, really*)	1. General & abstract (*thing, things*)
2. Evaluation: authentic (*real, reality, really*	2. Plants (*tree, roots, stalk, flower*)	2. Colours & colour patterns (*blue, light, colour*)
3. Plants (*tree, roots, stalk, flower*)	3. Solid materials (*coals, glass, iron, emeralds*)	3. Evaluation: authentic (*real, reality, really*)
4. Life and living things (*life, lives*)	4. Colours & colour patterns (*blue, light, colour*)	4. Plants (*tree, roots, stalk, flower*)
5. Colours & colour patterns (*blue, light, colour*)	5. General appearance & physical properties (*mark*)	5. Life and living things (*life, lives*)
6. Mental object; conceptual (*thought, thoughts, ideas*)	6. General & abstract (*thing, things*)	6. Parts of buildings (*wall, room, door*)
7. Smoking and non-medical drugs (*cigarette(s)*)	7. Mental object; conceptual (*thought, thoughts, ideas*)	7. Furniture and household fittings (*chair, table*)
		Cont.

8. Living creatures: animals, birds (*cat, snail*)	8. Living creatures: animals, birds (*cat, snail*)	8. Smoking and non-medical drugs (*cigarette(s)*)
9. Solid materials (*coals, glass, iron, emeralds*)	9. Objects generally (*bowl, rock, hoops*)	9. Thought, belief (*think, believe, imagine*)
10. No kin (*illegitimate*)	10. Strong obligation & necessity (*must, should*)	10. The universe (*world, moon*)
11. Comparing (*compare, comparison*	11. Smoking and non-medical drugs (*smoke(s), cigarette(s)*)	11. Like (*like(s), adoring, fancy*)
12. Probability (*perhaps*)	12. Furniture and household fittings (*chair, table*)	12. Living creatures: animals, birds (*cat, snail*)

Note: here we use double- and single-underlining in the same way as for the preceding two tables, but we underline only the number showing a semantic tag's position in the table.

4. Conclusion

We began with two notions of style – [1] as a chosen way of using language and [2] as a chosen way of expressing meanings – and have shown two paradigms for investigating style using quantitative corpus techniques. In Section 3, however, we have concentrated mainly on the second and more general notion of style. The first paradigm (in Section 2) was a comparison of two or more matching corpora from different periods of time. The second paradigm (in Section 3) was a comparison of a focal text – the text whose style was to be analysed – and a reference corpus. In all, three different reference corpora were assembled and used.

We also used two different quantitative techniques for undertaking these analyses. The first technique was simply to count occurrences per million words in the two or more corpora being compared, so that the differences can be represented as percentages (and calculated for statistical significance) or, in our case, represented visually as line charts. The second technique is to employ the 'key feature' method of listing items in order of keyness, or distinctiveness in the focal text, as contrasted with the reference corpus, measured in terms of the significance ratio of Log Likelihood. It should be emphasised that these techniques could have been applied differently: we could have applied the 'key feature' analysis to the diachronic comparison of LOB and F-LOB or of B-LOB

and LOB, for example. Similarly, we could have applied the per-million-words analysis to the study of 'The Mark on the Wall' as compared with one of the reference corpora. The main difficulty with this was the relative shortness of the 'The Mark', which would have given undue prominence to some features occurring only a few times.

It is worthwhile, finally, noting some of the limitations as well as the future possibilities of these stylistic methods. It is only too obvious, to begin with, that this type of analysis when applied to very large quantities of electronic text would be virtually impossible without the power of the modern computer. The great advantage of the techniques illustrated here is that they can be carried out automatically and at great speed. Wmatrix also shows great adaptability to the use of a wide range of corpora. The variety of corpora capable of being used is limited only by the user's ability to assemble the corpora and load them as 'personal folders' onto the Wmatrix website.

The corresponding disadvantage is that any activity involving human scrutiny of the data is immensely slow by comparison. Although POS tagging and semantic tagging are relatively accurate, there are still plenty of 'mistakes made by the computer' that ideally need to be manually checked. Further, although at present Wmatrix can operate with grammatical tags and semantic tags, there are many other levels of analysis that at present it cannot undertake – most importantly, parsing – the systematic syntactic analysis of a text in terms of phrases, clauses and so forth. There are also some more meaning-oriented stylistic analytic tasks (e.g. identifying metaphor or irony) that cannot yet be achieved by a computer.

The present situation, then, is that certain tasks can be undertaken fast but fallibly by computer, while other tasks can be undertaken more reliably but more slowly by human beings. Wmatrix already has the advantage that it can undertake a multi-level linguistic analysis of English corpora.

In the case of per-million-words analysis, what we have presented is the outcome of both automatic and manual analysis, and is the result of a ten-year research project (see Leech et al. 2009: Chapter 2 for the techniques employed). We have mentioned some proposed examples of colloquialization, such as the increase in the progressive construction and the decrease in the passive voice, the decline of pied-piping and the increase in *that*-relativization. Ideally such investigations need more advanced annotation, i.e. parsed corpora, although existing methods (CQP searches, making use of regular-expression-type syntax) bring some of the benefits of such corpora.[29] We believe that present results are promising, and that we can look forward to a future in which more revealing analyses of style can be achieved by computer at a more abstract level.

Notes

1 See note 3 below.

2 The definition of style in terms of such comparative frequency measures has been common in linguistic thinking, and has a fairly long history – see Bloch's claim (1953: 40–44) that the style of a text is the 'message carried by the frequency distributions and transitional probabilities of its linguistic features, especially as they differ from those of the same features in the language as a whole'.

3 One departure from the ideal is that, for reasons of practical feasibility, the B-LOB corpus consists of texts from the period 1928–1934, the median year being 1931. Details of LOB and F-LOB as comparable corpora can be found from their manuals on the ICAME website (http://icame.uib.no/newcd.htm) and also from Leech et al. (2009). The third comparable corpus used in this study, B-LOB, has been compiled at Lancaster, but has not yet been released.

4 In practice, 'representativeness' like 'comparability', is an ideal to which corpora can in general only approximate (see Biber 1993, Leech 2007). However, Biber (*ibid.*, pp. 243–244), in his seminal article on representativeness, describes the Brown corpus as a well-constructed corpus with a 'good sampling frame'.

5 The reason for this is that, to avoid unfinished sentences, the compilers concluded each text sample not at the 2000th word itself, but at the sentence break *following* the 2000th word.

6 As an example of a range of semantically equivalent or similar forms between which choices are particularly complex, see Smith's (2003) investigation of recent developments in the expression of obligation and necessity by modals and semi-modals.

7 Fiction writing obviously contains a large proportion of its contractions in quoted speech passages. This is also true, to a lesser extent, of the other subcorpora. It might be supposed that the increase of contractions is due to the increasing use of quoted speech in written English in general. However, we have found that although there is such an increase, it can account for only part of the increasing use of contractions. For example, the increase of speech between LOB and F-LOB accounts for less than half of the increase of contractions of *it is* to *it's*. – see Leech et al. (2009), Section 11.3.6 and fn. 14.

8 See Hundt and Mair (1999) on 'agile' and 'up-tight' genres.

9 Historically, colloquialization has been observed further back in the diachronic development of English style – as Biber (2003) explains:
'...in the course of the nineteenth and twentieth centuries, popular written registers like letters, fiction, and essays have reversed their direction of change and evolved to become more similar to spoken registers, often becoming even more oral in the modern period than in the seventeenth century. These shifts result in a dispreference for certain stereotypically literate features, such as passive verbs, relative clause constructions and elaborated noun phrases generally'. (Biber 2003: 169)

10 This scale of formality has been documented in various ways and with varying terminology – see, for example, Biber (1988: 101–108), Nakamura (1991) and Rayson et al. (2002).

11 The decreasing use of the passive in BrE, and more particularly in AmE, is likely to have been due in part to prescriptive influences. See Leech et al. (2009, Section 7.2), and Seoane and Williams (2006: 260f.).

12 Biber et al. (1999: 926), noting the prevalence of declarative inversion structures in fiction writing, observe that 'In general, we may assume that writers of fiction make more use of the resources of the language, including options which were formerly in more frequent use'. *Upon* is one such option.

13 In spoken (conversational) corpus data, some of these features have been shown to be significantly more frequent, in the case of colloquialisms, and significantly less frequent , in the case of formal features: for example, the progressive aspect and semi-modals show a much higher frequency in speech than in writing (Leech et al. 2009, Chapters 5 and 6). Such findings strengthen the case for postulating colloquialization.

14 See Biber et al.'s Fig. 2.2 (*loc. cit.*) Lexical density is understood here as the percentage of all word tokens that belong to lexical classes, as opposed to function-word classes.

15 Sequences of proper nouns (e.g. Nicholas Winterton, Goldman Sachs) were excluded from the count of Noun + Noun sequences, on the grounds that they represented an independent stylistic phenomenon which has little to do with lexical density.

16 However, it should be pointed out that increase in the frequency of Noun + Noun sequences has been noted since the eighteenth century (Leonard 1968, Rosenbach 2006): it is not a purely twentieth-century phenomenon.

17 The notion of density or compactness clearly depends on a prior notion of the *linear extent* of a text. The obvious way to measure this is to count the number of words, which is adequate for most purposes. However, we suggest that other ways of measuring linear extent – e.g. counting syllables – may be more accurate. For example, replacing *upon* by *on* (commented on above) lowers the syllable count but not the word count. Similarly, replacing *the behavior of a patient* by *a patient's behavior* is not indisputably a saving of two words (the *'s* being arguably an extra word in the form of an enclitic postposition), whereas it is clearly a saving of two syllables.

18 This is a somewhat rough-and-ready measure of 'opportunities for using genitives', as the genitive in Present-day English is a category applicable to noun phrases, rather than to nouns (see, for example, Quirk et al. 1985: 326–328). While it is true that the genitive requires at least one noun, it is also true that the most common kind of genitive construction contains two nouns – the noun with the genitive ending and a following head noun, as in the president's daughter. There are also genitive constructions with more than two nouns: notably group genitives such as 'the president of Finland's daughter', where the genitive *'s* is attached to the second noun 'Finland', which is in an embedded position in the genitive noun phrase.

19 A selection of notable novels published in the same year as 'The Mark on the Wall' are listed at 'Literature in 1917', Wikipedia. The following three were found to be available from Project Gutenberg and other on-line resources: Florence Barclay, *The White Ladies of Worcester*; Mrs Humphrey Ward, *Missing*; Edith Wharton, *Summer*. Two of the authors are British and one (Wharton) American.

20 The one-third 1901 corpus contained one-third of each subcorpus, and each text category in proportion to their representation in the Brown-family corpus when complete. Within each text category, the texts were also matched in topic and publication with the corresponding parts of B-LOB, LOB and F-LOB.

21 In terms of Wmatrix word counts, the size of the focal text is 2,985 words, and the sizes of the reference corpora are: Three 1917 Novels: 269,842; 1901 Corpus: 342,448; B-LOB General Fiction: 56,703. Wmatrix word counts are generally slightly lower than other corpus tools because

semantically meaningful chunks, e.g. idiomatic expressions, names, places, and phrasal verbs, are counted as one item.

22 In Leech (2008: 168–176) two widely differing reference corpora were used – (a) three novels of the 1890s and (b) the General Fiction text category (K) of the B-LOB Corpus, dating from 1928–1934. In view of their disparity, it was surprising that the overall analysis was closely similar for both corpora.

23 See the list of publications on Mike Scott's webpage: http://www.lexically.net/publications/publications.htm.

24 This is a simplified version of the five-stage process presented in Rayson (2008: 521).

25 Note that these tools do not produce error-free output. The accuracy of CLAWS is in the region of 96–97%, and that of USAS is c. 91%. These accuracy rates, however, are high enough to provide a sound basis for key feature extractions, given that the most salient results show high statistical significance (see below).

26 The keyword list can include words which have 0 occurrences in List X or List Y. Negative keywords are normally less noticeable and interesting, but can be important – e.g. it is significant that 'The Mark on the Wall' makes very little use of third person pronouns such as *she* and *they*.

27 The significance measure used in Wmatrix is log likelihood, which is considered preferable to the more familiar chi-square test, and which is explained in Rayson (2008: 527–528) and at http://ucrel.lancs.ac.uk/llwizard.html.

28 It is worth mentioning that this exclamatory construction is associated with female speech, being used by more female speakers than male speakers in each age group in the conversational part of the British National Corpus.

29 Some parsed corpora are available: for example, ICE-GB and DCPSE. http://www.ucl.ac.uk/english-usage/projects/dcpse/, http://www.ucl.ac.uk/english-usage/projects/ice-gb/index.htm.

References

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

Biber, Douglas (1993), 'Representativeness in corpus design', *Literary and Linguistic Computing*, 8(4): 243–257.

Biber, Douglas (2003), 'Compressed noun-phrase structures in newspaper discourse: the competing demands of popularization vs. economy', in: Jean Aitchison and Diana M. Lewis (eds.) *New media language*. London: Routledge, 169–181.

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

Biber, Douglas and Victoria Clark (2002), 'Historical shifts in modification patterns with complex noun phrase structures', in: Teresa Fanego, María José López-Couso and Javier Pérez-Guerra (eds.) *English historical morphology: selected papers from 11 ICEHL, Santiago de Compostela, 7–11 September, 2000*. Amsterdam: Benjamins, 43–66.

Bloch, Bernard (1953), 'Linguistic structure and linguistic analysis', in: Archibald A. Hill (ed.) *Report of the Fourth Annual Round Table Meeting on Linguistics and Language Study*. Washington, D.C.: Georgetown University Press, 40–44.

Enkvist, Nils Erik (1973), *Linguistic stylistics*. The Hague: Mouton.

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

Leech, Geoffrey (2007), 'New resources, or just better old ones? The Holy Grail of representativeness', in: Marianne Hundt, Nadja Nesselhauf and Carolin Biewer (eds.) *Corpus linguistics and the Web*. Amsterdam: Rodopi, 133–149.

Leech, Geoffrey (2008), *Language in literature: style and foregrounding*. Harlow: Pearson/Longman.

Leech, Geoffrey and Mick Short (2007), *Style in fiction: an introduction to English fictional prose*. 2nd edition. Harlow: Pearson/Longman.

Leech, Geoffrey, Marianne Hundt, Christian Mair and Nicholas Smith (2009), *Change in contemporary English: a grammatical study*. Cambridge: Cambridge University Press.

Leonard, Rosemary (1968), The types and currency of Noun + Noun sequences in prose usage 1750–1950. Unpublished M.Phil. thesis, University of London.

Nakamura, Junsaku (1991), 'The relationship among genres in the LOB corpus based upon the distribution of grammatical tags', *JACET Bulletin*, 22: 44–74.

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

Rayson, Paul (2008), 'From key words to key semantic domains', *International Journal of Corpus Linguistics*, 13(4): 519–549.

Rayson, Paul, Andrew Wilson and Geoffrey Leech (2002), 'Grammatical word class variation within the British National Corpus Sampler', in: Pam Peters, Peter Collins and Adam Smith (eds.) *New frontiers of corpus research: papers from the Twenty First International Conference on English Language Research on Computerized Corpora – Sydney 2000*. Amsterdam: Rodopi, 295–306.

Rosenbach, Anette (2006), 'On the track of noun+noun constructions in Modern English', in: Christoph Houswitschka, Gabriele Knappe and Anja Müller (eds.) *Anglistentag 2005 Bamberg. Proceedings*. Trier: Wissenschaftlicher Verlag, 543–557.

Scott, Mike (2004), *WordSmith Tools version 4*. Oxford: Oxford University Press.

Seoane, Elena and Christopher Williams (2006), 'Changing the rules: a comparison of recent trends in English in academic scientific discourse and prescriptive legal discourse', in: Marina Dossena and Irma Taavitsainen (eds.) *Diachronic perspectives on domain-specific English*. Bern: Peter Lang, 255–276.

Smith, Nicholas (2003), 'Changes in the modals and semi-modals of strong obligation and epistemic necessity in recent British English', in: Roberta Facchinetti, Manfred Krug and Frank Palmer (eds.) *Modality in contemporary English*. Berlin/New York: Mouton de Gruyter, 241–266.

Wales, Katie (2001), *A dictionary of stylistics*. 2nd edition. Harlow, U.K.: Longman.

Woolf, Virginia (1917), 'The Mark on the Wall', in: Leonard Woolf and Virginia Woolf, *Two stories*. London: Hogarth Press.

This page intentionally left blank

III Focus on early English

Historical pragmatics and corpus linguistics: problems and strategies

Laurel J. Brinton

University of British Columbia

Abstract

*Corpus linguistics is "*the sine qua non *of historical linguistics" (McEnery and Wilson 2001: 123). Contemporary corpus linguistics has led to significant advances in historical linguistics, most notably in the speed and ease with which data can be retrieved. The English historical linguist has available for use a wide variety of corpora. However, none is entirely ideal. Only two corpora, the* Oxford English Dictionary *and the Helsinki Corpus, provide the full diachronic span from Old English to the present day. The OED quotation bank, though not a corpus strictly speaking, can – with caution – be fruitfully used by the historical linguist (Hoffmann 2004). At only 1.5 million words for 1000 years of language history, the Helsinki Corpus, a balanced general-purpose corpus, may prove too small for some types of searches. Apart from these sources, the historical English linguist must cobble together a variety of corpora from the individual periods of English, ranging from the* Dictionary of Old English Corpus *containing almost all extant Old English texts, to the* Middle English Dictionary *(sharing many of the weaknesses of the OED), to the rich Chadwyck-Healey corpora designed primarily for the literary scholar (and quite user-unfriendly for the linguist).*

After a review of the historical corpora available to the English linguist, this paper explores some of the problems encountered by a scholar wishing to apply corpus linguistic methodology in the field of historical pragmatics. I articulate the strategies that I have adopted in my work on pragmatic markers and, more recently, on comment clauses in the history of English (Brinton 2008). As a case study, I explore the development of the comment clause (as) you say *in the history of English. The use of a mixed qualitative/quantitative corpus-based approach allows for a detailed, empirically based description of the rise of* (as) you say; *at the same time, it permits testing of the "matrix clause hypothesis", the prevailing theory concerning the origin of comment clauses that has been extrapolated from Thompson and Mulac's synchronic work on* I think/guess. *Frequency counts of the presumed source construction (i.e.,* you say that *S) in the earlier periods cast doubt on the validity of the matrix clause hypothesis. Corpus data suggest a more nuanced view of the rise of this comment clause, namely, that a variety of structures, including relative/adverbial* as you say, *main clause* you say, *and* you say *following a fronted element all contributed to its genesis.*

1. Introduction

The use of electronic corpora has become commonplace in most traditional areas of linguistic study, and increasingly, such corpora are coming to be used in a much larger number of fields, ranging from sociolinguistics to discourse analysis, genre studies, text linguistics, and pragmatics. Historical linguistics, which is "a species of corpus linguistics" (McEnery and Wilson 2001: 123), now almost always depends upon electronic corpora. This paper addresses some of the questions and problems encountered in the use of existing historical English corpora for work focused on the sub-area of **historical pragmatics,** specifically the area that Jacobs and Jucker (1995) call "diachronic pragmatics", which "focuses on the linguistic inventory and its communicative use across different historical stages of the same language" (13). A central focus in diachronic pragmatics has been the development of one-word and phrasal pragmatic markers such as *well*, *now*, *right*, or *in fact*, as well as clausal pragmatic markers – or "comment clauses" (Quirk et al. 1985) – such as *you know* or *I mean* (Brinton 1996; 2008).

After reviewing a number of general considerations, such as the historical and diachronic corpora of English available (§2.1) and the particular problems posed by historical pragmatics for the use of corpora (§2.2), this paper presents a case study of the development of the comment clause *(as) you say* (§3). This study is intended to exemplify how a scholar working in the area of historical pragmatics, given the restrictions discussed, goes about extracting and analyzing data from a variety of corpora and the uses to which that data may be put.

2. General considerations

2.1 Corpora

The English linguist is very fortunate to have available a wide variety of historical corpora, but – unfortunately – none which is absolutely ideal. This section briefly surveys a number of these corpora; it will not attempt to provide a comprehensive accounting of these corpora, however, as several websites compile information about them.[1]

Only one corpus, the quotation bank of the *Oxford English Dictionary* (OED), provides the full span of English language history. The OED quotations were collected not for the purpose of creating a representative sample of language for the different periods but rather for the purpose of illustrating the senses (often obscure senses) of headwords. Studies such as Jucker (1994), Fischer (1997), and Hoffmann (2004) have noted a number of problems in the use of the OED quotations for corpus linguistic work, including the over-representation of certain authors (e.g., Shakespeare),[2] differing lengths of quotations, or inconsistencies in abbreviation conventions and the marking of deletions. However, Hoffmann concludes that one may safely assume that the words surrounding the headword

constitute a fair reflection of the contemporary language (2004: 20) and that the OED database of 2.4 million quotations[3] covering a time-span "unmatched by any other source of computerized data" (26) can – with caution – be fruitfully used by historical linguists. From a technical perspective, there are a number of frustrations for those using the OED quotation bank for linguistic research. Some – such as the need to eliminate multiple quotations and homographs from result lists – albeit time-consuming, are easily overcome. Others are built into the program itself and cannot be circumvented. For example, the search program of the online 3rd edition is rather limited and is inferior to that provided by the 2nd edition. With the earlier edition, searches provided the full text of each quotation, which could then be printed out in list form (see Figure 1, p. 104).

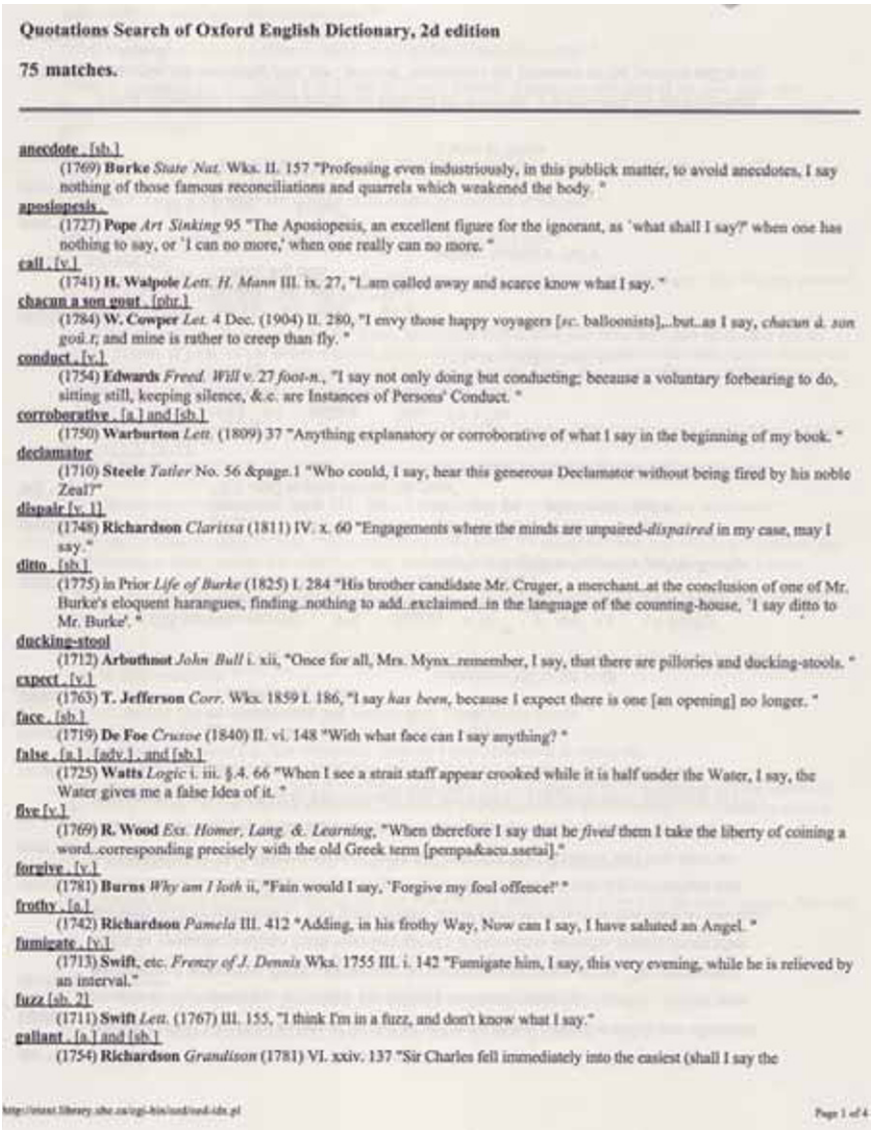
The current edition provides an initial view of only partial quotations. The user must click on each entry to get the full citation and thus has no means of printing out a complete list of citations (see Figure 2, p. 105).[4] We see here that what might seem sensible to the programmer may prove frustrating or even counter-productive to the user.

A second corpus that provides an extensive diachronic spread but which stops short at 1710 (thus omitting Late Modern and Present-day English) is the *Helsinki Corpus of English Texts* (the Helsinki Corpus/HC), a balanced corpus separated into three periods (eleven subsections). The Helsinki Corpus is an invaluable research tool for the historical linguist, but with a total of only slightly over one and a half million words covering nearly 1000 years of English language history, it may prove too small for certain types of searches, such as those in the area of historical pragmatics (where token frequency may be quite low). Nonetheless, it is an obligatory starting point for any diachronic study of English.

Apart from using these two corpora, the historical English linguist who wishes to study a linguistic feature extending from Old English to the present must use a grab-bag of different corpora for the individual periods.[5]

1) Old English:
The *Dictionary of Old English Corpus in Electronic Form* (DOEC) is an almost complete record of OE manuscripts; it includes 3047 texts, consisting of poetry, prose, interlinear glosses, glossaries, runic and Latin alphabet inscriptions. In contrast, the OE section of the Helsinki Corpus contains about one-half million words (413,250 words) and allows for smaller searches.

2) Middle English:
The quotation bank of the *Middle English Dictionary* (MED), while possessing some of the same shortcomings as the OED quotation bank, is nonetheless an invaluable source for corpus linguistic work because of its size and scope. The well-known variability of Middle English spelling presents numerous difficulties for searching, of course. Smaller Middle English corpora include the Middle English section of the Helsinki Corpus, again with slightly over one-half million words (608,570 words) and the *Corpus of Middle English Prose and Verse* of the University of Michigan, which includes 146 items in its bibliography. The *Middle*



Quotations Search of Oxford English Dictionary, 2d edition

75 matches.

anecdote . [sb.]
(1769) **Burke** *State Nat.* Wks. II. 157 "Professing even industriously, in this publick matter, to avoid anecdotes, I say nothing of those famous reconciliations and quarrels which weakened the body. "
aposiopesis .
(1727) **Pope** *Art Sinking* 95 "The Aposiopesis, an excellent figure for the ignorant, as 'what shall I say?' when one has nothing to say, or 'I can no more,' when one really can no more. "
call . [v.]
(1741) **H. Walpole** *Lett. H. Mann* III. ix. 27, "I..am called away and scarce know what I say. "
chacun a son gout . [phr.]
(1784) **W. Cowper** *Let.* 4 Dec. (1904) II. 280, "I envy those happy voyagers [*sc.* balloonists],..but..as I say, *chacun à. son goût.t*; and mine is rather to creep than fly. "
conduct . [v.]
(1754) **Edwards** *Freed. Will* v. 27 *foot-n.*, "I say not only doing but conducting; because a voluntary forbearing to do, sitting still, keeping silence, &c. are Instances of Persons' Conduct. "
corroborative . [a.] and [sb.]
(1750) **Warburton** *Lett.* (1809) 37 "Anything explanatory or corroborative of what I say in the beginning of my book. "
declamator
(1710) **Steele** *Tatler* No. 56 &page.1 "Who could, I say, hear this generous Declamator without being fired by his noble Zeal?"
dispair [v. I]
(1748) **Richardson** *Clarissa* (1811) IV. x. 60 "Engagements where the minds are unpaired-*dispaired* in my case, may I say."
ditto . [sb.]
(1775) in Prior *Life of Burke* (1825) I. 284 "His brother candidate Mr. Cruger, a merchant..at the conclusion of one of Mr. Burke's eloquent harangues, finding..nothing to add..exclaimed..in the language of the counting-house, 'I say ditto to Mr. Burke'. "
ducking-stool
(1712) **Arbuthnot** *John Bull* i. xii, "Once for all, Mrs. Mynx..remember, I say, that there are pillories and ducking-stools. "
expect . [v.]
(1763) **T. Jefferson** *Corr.* Wks. 1859 I. 186, "I say *has been*, because I expect there is one [an opening] no longer. "
face . [sb.]
(1719) **De Foe** *Crusoe* (1840) II. vi. 148 "With what face can I say anything? "
false . [a.] . [adv.] . and [sb.]
(1725) **Watts** *Logic* i. iii. §.4. 66 "When I see a strait staff appear crooked while it is half under the Water, I say, the Water gives me a false Idea of it. "
five [v.]
(1769) **R. Wood** *Ess. Homer, Lang. &. Learning*, "When therefore I say that he *fived* them I take the liberty of coining a word..corresponding precisely with the old Greek term [pempa&acu.ssetai]."
forgive . [v.]
(1781) **Burns** *Why am I loth* ii, "Fain would I say, 'Forgive my foul offence!' "
frothy . [a.]
(1742) **Richardson** *Pamela* III. 412 "Adding, in his frothy Way, Now can I say, I have saluted an Angel. "
fumigate . [v.]
(1713) **Swift**, etc. *Frenzy of J. Dennis* Wks. 1755 III. i. 142 "Fumigate him, I say, this very evening, while he is relieved by an interval."
fuzz [sb. ?]
(1711) **Swift** *Lett.* (1767) III. 155, "I think I'm in a fuzz, and don't know what I say."
gallant . [a.] and [sb.]
(1754) **Richardson** *Grandison* (1781) VI. xxiv. 137 "Sir Charles fell immediately into the easiest (shall I say the

Figure 1: Results of an OED search (2nd edition) (search performed 6/28/2002)

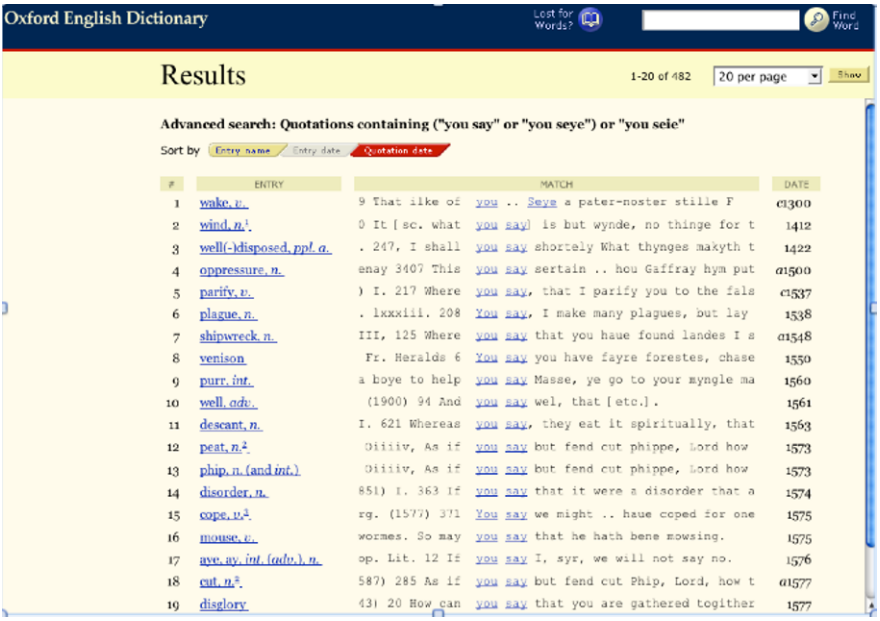


Figure 2: Results of an OED search (3rd edition) (search performed 1/31/2008)

English Collection of the University of Virginia Electronic Text Centre contains 65 texts, most of which are freely accessible online. Finally, the *Innsbruck Computer Archive of Machine-Readable English Texts* (ICAMET) contains 129 prose texts and 254 letters written between 1386 and 1688.

3) Early Modern English:

Resources for corpus work in Early Modern English are richer and more varied than those in the earlier periods. The Early Modern English section of the Helsinki Corpus covers the period 1500–1710 and again includes approximately one-half million (551,000) words. A much larger corpus is the *Lampeter Corpus of Early Modern English Tracts* (LC), which includes samples of (non-fiction, prose) texts in six domains – religion, politics, economy, science, law, and miscellaneous – from the Tract Collection, Founders' Library, University of Wales, Lampeter. It covers the period 1640–1740 and contains over a million words (1,172,102). The *Corpus of Early English Correspondence, Sampler* (CEECS) consists of two parts, each with approximately 200,000 words of correspondence, CEECS1, dating from 1418–1638, and CEECS2 dating from 1580–1680. The *Corpus of English Dialogues 1560–1760*, a 1.2 million word corpus of speech-related texts (trial proceedings, witness depositions, drama, didactic works, and prose fiction) is the newest addition to the corpora of Early Modern English. For reasons set out below, this corpus of spoken material is a very important source for historical pragmatic study.

Collections that serve primarily the needs of literary study and are hence of less usefulness for linguistic research (for reasons outlined below) are those marketed by Chadwyck-Healey and Gale. *Early English Books Online* (EEBO) contains all published books from 1475–1700; the smaller *Early English Prose Fiction* includes c. 200 works from 1500–1700. *Eighteenth Century Collections Online* (ECCO) is a digitization of "every significant English-language and foreign-language title printed in Great Britain during the eighteenth century, along with thousands of important works from the Americas".

4) Late Modern English:

The Late Modern English period provides a variety of different corpora. The corpus with the widest temporal and geographical spread is the ARCHER Corpus, a balanced corpus with British and American texts ranging from 1650–1990. Because of copyright difficulties, this remains a proprietary corpus that can only be accessed from a number of universities in the UK, USA, Germany, Sweden, and Finland.

Most recently, the *Corpus of Historical American English* (COHA), a balanced corpus of over 400 million words of American English from 1823 to the present, has become freely available over the web.[6]

Two Chadwyck-Healey corpora, the *Eighteenth-Century Fiction* and the *Nineteenth-Century Fiction* are expensive corpora again aimed in the first instance at the literary scholar. The former includes 96 complete works (epistolary novels, novels of sentiment, Gothic novels, documentary novels, allegorical and satirical texts) dating from 1700–1780, while the latter includes 250 works dating from 1786–1900.

Two smaller, freely-available corpora have been compiled by David Denison and his colleagues at the University of Manchester. The *Corpus of Late 18c Prose* (English language of the north-west in the late Modern English period) (300,000 words), contains unpublished letters written to Richard Orford between 1761 and 1790. Though not a balanced corpus, it is a good representation of business English of the period. The *Corpus of Late Modern English Prose* (1861–1919) (100,000 words) includes informal private (published) letters by British writers 100,000 words. [7]

Using texts freely available from publicly accessible archiving services such as Project Gutenberg and the Oxford Text Archive, Hendrik de Smet has compiled the *Corpus of Late Modern English Texts* (CLMET) (1728–1910), which encompasses approximately 10 million words, and the *Corpus of Late Modern English Texts Extended Version* (1728–1920), which encompasses approximately 15 million words. Though not balanced corpora, as they consist mainly of novels, letters, biographies and formal essays, their sheer size make them attractive. De Smet also has a 25 million word *Corpus of English Novels* of British and North American provenance ranging from the late 19th to the early 20th century compiled in the same fashion.[8]

5) Modern English (i.e. Late Modern to Present-day English):
The *Modern English Collection* of the University of Virginia Electronic Text Center, which describes itself as "heterogeneous collection contain[ing] fiction, non-fiction, poetry, drama, letters, newspapers, manuscripts and illustrations from 1500 to the present", is freely available online and though not a balanced corpus can supply a wealth of data.

A second corpus that is very valuable for the student of historical pragmatics is the Chadwyck-Healey *English Drama* collection of 3,900 plays in verse and prose dating from the 13th to the early 20th century. Again, although it is designed for literary study and not particularly user-friendly for the linguist, it provides a wealth of represented, colloquial speech from Middle English to the present. Figure 3 (p. 108) shows some of the difficulties in using this corpus, however. For example, very little context is shown, requiring time-consuming and laborious connections to the full text, and it is not possible to organize results chronologically or to print off results in an economical fashion.

2.2 Problems posed by historical pragmatics

Historical pragmatics presents a number of challenges for corpus studies, in the first instance because it is "historical" and in the second instance because it treats "pragmatics".

As has been pointed out, "[d]iachronic study is perhaps one of the few areas which can **only** be investigated using corpus data" (McEnery et al. 2006: 46; my emphasis). Contemporary corpus linguistics has led to significant advances in historical linguistics, most notably in the speed and ease in which data can be retrieved. However, as the foregoing review of existing historical English corpora has revealed, a truly diachronic study of English, ranging from Old English to the present day, especially if it treats a low frequency item, must resort to the cobbling together of a variety of different corpora in order to achieve the necessary diachronic spread and a sufficient number of examples. These corpora are of very different types and qualities: some are well-designed, balanced and representative corpora, others are more specialized or time-limited corpora, and others, such as the OED or MED, are not corpora at all in any strict sense. Despite such limitations, one must trust that "even if all or part of a corpus is not designed as ideally as the analyst would like, it is still possible to analyze the corpus and make generalizations based on the results that are obtained" (Meyer 2002: 121).

Corpus difficulties have led to my taking a "corpus-based" rather than "corpus-driven" approach (McEnery et al. 2006: 8, 10) in my own work in historical pragmatics. That is, I have used corpora as a starting point for collecting examples so that I might give the most complete, empirically-based description possible of the development of the linguistic forms under consideration. Moreover, I have used corpus data as a means to test and revise

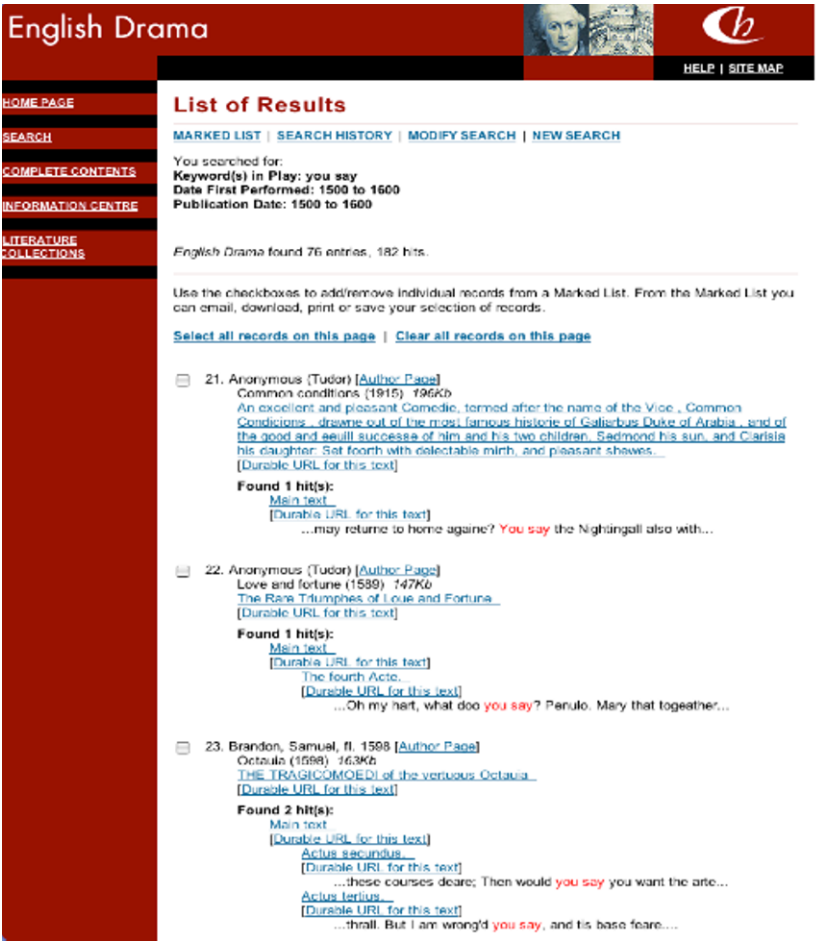


Figure 3: Results of a search of the *English Drama* corpus (search performed 2/21/2008)

hypotheses concerning the development of these forms. This could be characterized as a mixed qualitative and quantitative approach (McEnery and Wilson 2001: 76–77). Like qualitative approaches, I have aimed to give detailed descriptions, considering rare or infrequent forms and recognizing ambiguity. Quantitatively, I have restricted myself primarily to frequency counts, or what Mair has described as "home-grown statistics – rarely going beyond ad-hoc counts" (1991: 68). Such practice is not uncommon for historical linguists (see Meyer 2002: 120). For those studying pragmatics especially, calculations of significance would often not be meaningful.[9] While such a mixed qualitative/quantitative approach may appear rather unsystematic or even methodologically "impure", Mair (1991: 68) advocates it as the best approach for low frequency

phenomena, for grammatical irregularities, blends, or hybrids, and for "categories defined by a mix of structural, semantic and functional criteria" – a description that perfectly captures pragmatic markers (see Mair 1991: 68; also Schmied 1993). Such an approach has even been recommended more generally, as when McEnery and Wilson (2001: 77) observe that "[t]here has recently been a move in social science research towards multi-method approaches which largely reject the narrow analytical paradigms in favour of the breadth of information which the use of more than one method can provide".

The study of pragmatics, on the other hand, raises a set of distinctive problems for historical corpus study.

First, the study of pragmatics is typically associated with oral discourse, and the earlier periods obviously present a dearth of such data. But there is both authentic speech (e.g., court records) and represented or constructed speech (e.g., drama, prose fiction) dating back to the Middle English period. Furthermore, pragmatic markers, though typically associated with oral discourse, occur in written discourse as well, albeit with different forms and functions.[10] And finally, written discourse itself is now increasingly recognized as a subject of pragmatic study.

Second, as McEnery et al. (2006: 108) point out, "meanings dependent upon pragmatics cannot easily be detected automatically". In the case of pragmatic markers, one is dealing with forms that are multifunctional. Delicate or nuanced decisions about meaning are necessary to distinguish pragmatic from non-pragmatic uses; these uses are not normally distinguished in any strictly formal way. Stringent decisions about indeterminate or ambiguous forms are often difficult to make with absolute confidence; categorization may remain fuzzy. McEnery et al. suggest that automatic extraction of pragmatic forms could happen if the corpus were annotated manually.[11] However, when working with pre-existing corpora, the opposite approach is necessary, namely, that of collecting the data automatically and then sorting it manually. While it is often possible to narrow the search down in some mechanical way, this often proves unreliable for a number of reasons. Many clausal pragmatic markers consist of high frequency verbs such as *say* or *see*; the problems for searching for these verbs in their pragmatic uses is compounded, for example, by the extensive overlap between forms of these *see* and *say* in Middle English.[12] Although pragmatic markers are often sentence-initial or sentence-final, they are not restricted to such positions. Pragmatic markers may be set off by commas, but even if punctuation were a reliable criterion, which especially in older texts it is not, search programs do not typically recognize punctuation, and those that do do not return completely reliable results.[13] Figure 4 (p. 110) shows an example of a MED search for *as it were*.[14] Manual checking of the examples reveals that only the three circled examples can be analyzed as metalinguistic pragmatic markers (what have been termed "indirect conditionals"), whereas the others are real conditionals with the meaning 'as if it were' and hence propositional and not pragmatic in function.

A third problem is that a pragmatic approach often requires access to the larger context in which a form occurs in order to achieve a complete semantic/pragmatic interpretation. While this is possible with many corpora, it is not easily possible in the case of the MED and OED quotations. Here, it would be necessary to consult the original printed text, a time-consuming and laborious task.

Finally, pragmatic and semantic meanings are by definition subjective. As a consequence, the interpretations of particular examples may be debatable and results may not be entirely replicable. It is for this reason that it is often necessary to eschew strict statistical analysis in favor of simple frequency counts, which themselves must be considered rough figures in some cases.


a1450 *Dc.291 Lapid.*(Dc 291) 34: Aspites is reed & shinynge..She doth awey the briddes fro be land bat is sowe..when a man puttith hit in be sonne beem, hit yeueth bryghtnes **as hit were** fire.

a1450 *Mandev.(3)* (BodeMus 116) 23/5: Ther fley out the graue **as it wer** the hed of a forschapyn beste, foul and hodous [?read: hedous].

a1450 *Methodius(2)* (Add 37049) 105/31: And be bestes bai sal bynde at be grafes of sayntes, **as it wer** to a mawnger.

a1450 *MS Sln.2463* in *EETS 102* (Sln 2463) 252 n.4: A ffystule in be wyke of be eye, hit comyth of concours of rennyng of humores to be corner of be eye besyde be nose..bey..abyden there & maken an emynence, **as hit were** a lupyne; & therfore of sume hit is cleped lupinus.

a1450(?c1430) Lydg. *DM(1)* (Hnt EL 26.A.13) 18/153: Sire archebisshop, whi do 3e 3ow with-drawe So frowardli **as hit were** bi disdeyne?

a1450(1408) **Vegetius(1)* (Dc 291) 106b: Some makeb **as hit were** a grene of roopes wib a ridyng knotte, and in be comynge of be strook bei caccheb be hede of be Ram in bilke snare and pulleth a litel a side, and so letteb him of his strook.

a1450(1408) *Vegetius(1)* (Dc 291) 66a: On be whiche bank..moot be made..enbatayling of defence, **as hit were** on a toun ober a castelle walle.

a1456(a1402) * Trev. *Nicod.*(Add 16165) 104b: I sawe [vr. seye] Ihesus **as hit were** bright blasing of light.

a1456(a1402) * Trev. *Nicod.*(Add 16165) 109b: Panne came a gret voyce, **as hit were** a thondre.

a1456(a1402) * Trev. *Nicod.*(Add 16165) 113b: Ye..shal speke with no man, but ye shal beo, **as hit were**, dombe.

a1475 *Liber Cocorum* (Sln 1986) p.39: 3olke of egge ben shalt bou take, That harde is sobun; lay in to bo top **As hit were** a gyldene knop.

a1475 *Liber Cocorum* (Sln 1986) p.5: Make bo flesshe to seme, iwys, **As hit were** raw, and 3yt hit nys.

a1475 *Rev.St.Bridget* (Gar 145) 53/17: Thy wysdom is **as it war** the see, that for gretenes may nott be drawe owt

Figure 4: Sample search results for *as it were* in the MED (search performed 2/15/2008)

3. Case study: *(as) you say*

In the space remaining, I will present a case study in historical pragmatics and thereby hope to illustrate the advantages of – as well as some of the problems inherent in – the corpus linguistics approach in this field. This is a small subsection of a larger study (Brinton 2008) of comment clauses in the history of

English.[15] This work focuses on the history of comment clauses formed with common verbs of perception and cognition in a variety of syntactic forms, including present-tense verbs with first- and second-person subjects (*I mean, you see*), imperative verbs (*look, see, say*), adverbial/relative clauses (*as it were, if you will*), and nominal relative clauses (*what's more, what else*). Apart from exploring the development of individual comment clauses, one goal of this study has been to critically examine what I call the "matrix clause hypothesis" (Thompson and Mulac 1991) concerning the rise and development of comment clauses.

3.1 Comment clauses

A "comment clause", according to Quirk et al. (1985: 1112–1118) is a parenthetical disjunct, an adverbial element conveying either the speaker's comment on the form of what is being said (a "style" disjunct) or the speaker's observations on the content of the utterance (a "content" disjunct). Quirk et al. describe comment clauses as functioning as hedges expressing tentativeness over truth value, as expressions of the speaker's certainty, as expressions of the speaker's emotional attitude toward the content of the matrix clause or as claims to the hearer's attention (1985: 1114–1115). A comment clause thus belongs to the larger class of pragmatic (or discourse) marker. Although definitions abound (see, e.g., Brinton 1996: 29–40; Schourup 1999), pragmatic markers are typically seen as elements that are syntactically independent, do not affect the propositional content of the utterance, and serve pragmatic (textual, subjective, intersubjective, metalinguistic) functions. Comment clauses differ from prototypical pragmatic markers such as *well* or *so*, however, in being clausal in origin.

Quirk et al. (1985: 1112–1120) distinguish between finite comment clauses such as *I know, I guess, I say, you see, you know, as you say, as I remember, what's more surprising* and non-finite comment clauses such as *to be honest, broadly speaking, roughly speaking, put in another way*). They identify three types of finite clauses:

a) those such as *I believe* which resemble (syntactically defective) matrix clauses with a transitive verb or adjective otherwise requiring a *that*-clause complement;
b) those such as *as you know* which resemble finite adverbial or relative clauses; and
c) those such as *what is more important* which resemble nominal relative clauses.

Comment clauses in medial or final position are parenthetical disjuncts, generally form a separate tone unit, and are marked by increased speed and lowered pitch and volume (Peltola 1982/1983: 102; Quirk et al. 1985: 1112, 1113). In initial position, however, their syntactic status may be indeterminate between main clause and parenthetical (Biber et al. 1999: 197, 1076–1077).

Comment clauses have been fairly extensively studied in Present-day English, but diachronic studies are quite limited.[16] The prevailing theory of the origin and development of comment clauses has been extrapolated from Thompson and Mulac's (1991) synchronic study of the matrix clause-type comment clauses *I think* and *I guess* in Present-day English. They propose a synchronic sequence, as shown in the (1a–c):

(1) a. **I think** that we're definitely moving towards being more technological.
b. **I think** Ø exercise is really beneficial, to anybody.
c. It's just your point of view you know what you like to do in your spare time **I think**.

Thompson and Mulac argue that *I think* followed by *that* in (1a) is a matrix clause, *I think* without *that* in (1b) is indeterminate between a matrix clause and a parenthetical disjunct, and *I think* in (1c) is clearly parenthetical, as it is no longer restricted to sentence-initial position. In this position it serves as a unitary particle expressing epistemicity. There is reversal of the matrix clause/complement clause structure, the original complement clause being reanalyzed as the matrix clause and the original matrix clause now serving as a parenthetical disjunct. A crucial condition in Thompson and Mulac's theory is that reanalysis depends upon the greatest frequency of indeterminate structures: "those subjects and verbs occurring most frequently without *that* are precisely those which occur most frequently as [parentheticals]" (1991: 317). This "matrix clause hypothesis" recalls earlier theories of "slifting" or "sentence lifting" (Ross 1973) or "complement preposing" (Hooper 1975), which in contrast to Thompson and Mulac, suggest that it is the *that*-clause rather than the original matrix clause which is moved.

3.2 *(As) you say* in Present-day English

Quirk et al. (1985: 1116) list *as you say* as a comment clause belonging to their second type; specifically, it is a relative comment clause (equivalent in meaning to 'which you say').[17]

Fitzmaurice (2004) argues that *you say* is a pragmatic marker with a focusing function; it may also be interactive since by using it, the speaker is drawing attention to a proposition for his or her own communicative ends while attempting to engage the addressee and keep the interaction going (442–443). Nevertheless, she finds *you say* to be "largely quotative and … descriptive in meaning" in the ARCHER Corpus (442). Its use as a comment clause never exceeds a frequency of 0.1/1000 words (442). But she admits that even in its quotative function, *you say* is intersubjective because it expresses "the speaker's interpretation of what the interlocutor has said as well as recapitulating the actual utterance of the interlocutor" (443).

The pragmatic functions of *you say* extend beyond the purely evidential functions recognized for the related form *they say*. In the case of *you say*, the content of the speech is presumably obvious to both interlocutors, as may not be the case with *they say*. Thus, the speaker must have a secondary (non-evidential) reason in uttering *you say*, such as to remind the hearer of what he or she has said on a previous occasion or to confirm understanding or interpretation. Present-day English corpus evidence suggests that speakers use *you say* in two ways:

a) to query what the interlocutor has said, in which case it is generally an interrogative sentence tag (2a–b), or
b) to highlight information expressed by the interlocutor in order to take issue with this information (2c–f).

(2) a. And since then there's just been the two of us. You're an actor, **you say**? (1991 Brett, *Corporate Bodies: A Charles Paris Mystery* [FLOB]).
b. Upriver, **you say?** In the jungle? Well, good luck to you, boy (1986 "Harrison's Heart of Darkness the Mosquito Coast", *Time Magazine* 1 Dec. [TIME]).
c. Simple, **you say**, yet how many people force down meals on a diet that they would not dream of choosing if they were not on that diet? (1989 Ashcroft, *Get Slim and Stay Slim: The Psychology of Weight Control* [BNC]).
d. There's a Frank Sinatra song that ends: "Here's to the winners all of us can be". So tell that to the country's 650,000 unemployed **you say**? (1986 Robbins, "The One That Got Away", *Sydney Morning Herald,* 7 July [ACE]).
e. Easy, **you say**, apply back cyclic. Well, yes, it can be that easy if you are only moving fairly slowly … (1990 Day, *Learning to Fly Radio Controlled Helicopters* [BNC]).
f. And third, after it is over, **you say**, wasn't it a wonderful experience (1986 "Liberty's Ringmaster of Ceremonies", *Time Magazine* Feb. 7 [TIME]).

The first usage is primarily "descriptive" or referential. The second is more obviously non-referential as it may accompany information not actually uttered by the interlocutor but implicitly assumed by his or her argument. More importantly, this usage points to the epistemic nature of *you say*, or the speaker's (relatively) low level of commitment to the truth value of the accompanying proposition, since it is often used as a means to introduce either an explicit or implicit disagreement.

In contrast, speakers use the adverbial *as you say* to express agreement with the interlocutor's ideas. Often *as you say* has a metalinguistic function in that it accompanies a figure of speech used by the interlocutor, or his or her quoted words repeated (approvingly) by the speaker (3c–e). This contrast between *you say* and *as you say* is consistent with the difference in function noted

between *as*-comment clauses and their corresponding *as*-less variants (see, e.g., Potts 2002; Blakemore 2006), namely, that the *as*-clause asserts the truth of the matrix clause while the *as*-less form does not. For this reason, the *as*-less variant is used for disagreement or interrogation, while the *as*-variant is used for approbation.

(3) a. But, **as you say**, a fixed identity, a shell, is also a trap, is no solution (1990 *Reynolds, Blissed out: The Raptures of Rock* [BNC]).
b. But, **as you say**, rumours don't have to be true, and the blind assassin has got hold of the wrong rumour (2000 Atwood, *The Blind Assassin* [CanE]).
c. Or maybe … you are planning … one of those jaunts to Oxford or Woodstock, to get a breath of old stone **as you say** (1991 Scruton, "A Mistake", *A Dove Descending and Other Stories* [FLOB]).
d. "Yea, such would give me, **as you say**, a foot in both camps. What of the lass herself?" (1990 Wiat, *The Child Bride* [BNC]).
e. Rather, the duty of the high court is to uphold the laws of this country and, as you say, "develop a higher loyalty" than mere politics (1984 *Time Magazine* Nov. 29 [TIME]).

Because of the use of *say* as a general verb of communication, the overall frequency of *(as) you say* as a comment clause in Modern English is difficult to determine. Excluding cases in which *you say* introduces direct speech, I have found that parenthetical *you say* represents 7% of the total uses of *you say* in the selection of Present-day English corpora I surveyed (see Table 1). If instances of indeterminate *you say* (where *you say* is clause initial and *that* does not precede the following clause) are included, the percentage rises to 20%, and if *as you say* is included,[18] parenthetical uses of *(as) you say* account for over one-quarter of the uses of *you say*. The frequency with which *you say* is followed by a full complement clause (with *that*), however, is fairly low (10%).

Table 1: Frequency of *(as) you say* types in Present-day English corpora

	you say (parenthetical)	*you say* (initial)	*you say that*	*as you say*	*so you say*	Total instances of *you say*
ACE	1	1	0	0	0	20
BNC[a]	6	9	11	8	1	100
FLOB	4	3	1	5	0	32
FROWN	5	1	0	0	0	33
Strathy[a]	2	24	13	3	2	100
WC	2	1	0	0	0	26
Total	20 (6%)	39 (13%)	25 (8%)	16 (5%)	3 (1%)	311

[a] Random sampling of 100 instances of *you say*

3.3 The development of *(as) you say*

Given the existence of both the adverbial type *as you say* and the matrix type *you say* in Present-day English, the question of the source construction thus arises. Fitzmaurice (2004: 445) suggests that parenthetical *you say* originates as a main clause expressing epistemic stance followed by a complement nominal clause. More importantly, she sees *as you say* as a comment clause that develops **from** *you say*. For Fitzmaurice, *you say* is intersubjective and *as you say* is interactive. Does the historical evidence bear out such a progression?

The DOEC provides 45 examples of *ge secgað/þu secge*. These are roughly divided among main clause constructions ('you say that S', 19 examples), adverbial/relative constructions (*þe/þæt ge secgað* 'as/which you say', 10 examples), and other constructions (16 examples). There are no examples of parenthetical 'you say'.

Figure 5 (p. 116) presents the frequencies of the different *you say* constructions in Middle and Early Modern English in the corpora examined.

There are also no examples of parenthetical *you say* in Middle English. Instances of initial *you say* followed by a *that*-less nominal clauses (4a) – structures which are indeterminately main clause or parenthetical – outnumber initial *you say* followed by a *that*-clause (4b). When a fronted element precedes *you say* + *that*-less complement, the resulting structure is also indeterminate: *you say* may be interpreted either as a clause-medial parenthetical or as a main clause (4c). Example (4d) could arguably be seen as an early, rare parenthetical:

(4) a. **ȝe say** þan þe angell made hir with child, Nay, sum lyke an angell has hyr begiled (c1400 *Life of Saint Anne* (1) (Min-U Z.822.N.81) 767 [MED]).
'you say then the angel made her with child, nay, something like an angel has beguiled her'

b. **Thou saist that** we prechen but fallace and fables, and leve the gospel (1402 *Friar Daw's Reply* (Dgb 41) 89 [MED]).
'you say that we preach only falsehoods and fables and leave the gospel'

c. A blysful lyf **þou says** I lede; Þou woldez knaw þerof þe stage (c1400 ?c1380) Pearl (Nero A.10) 410 [MED]).
'a blissful life you say I lead: you would know the condition thereof'

d. Couetise, **ȝe say**, es godd of þe lyuer … he hase in his hande a byrnand fyrebrande whare-wit he styrres þe luste of lechery (c1440 *The Prose Alexander* (Thrn) 83/21 [MED]).
'strong desire you say is the god of the liver … he has in his hand a burning firebrand wherewith he stirs the lust of lechery'

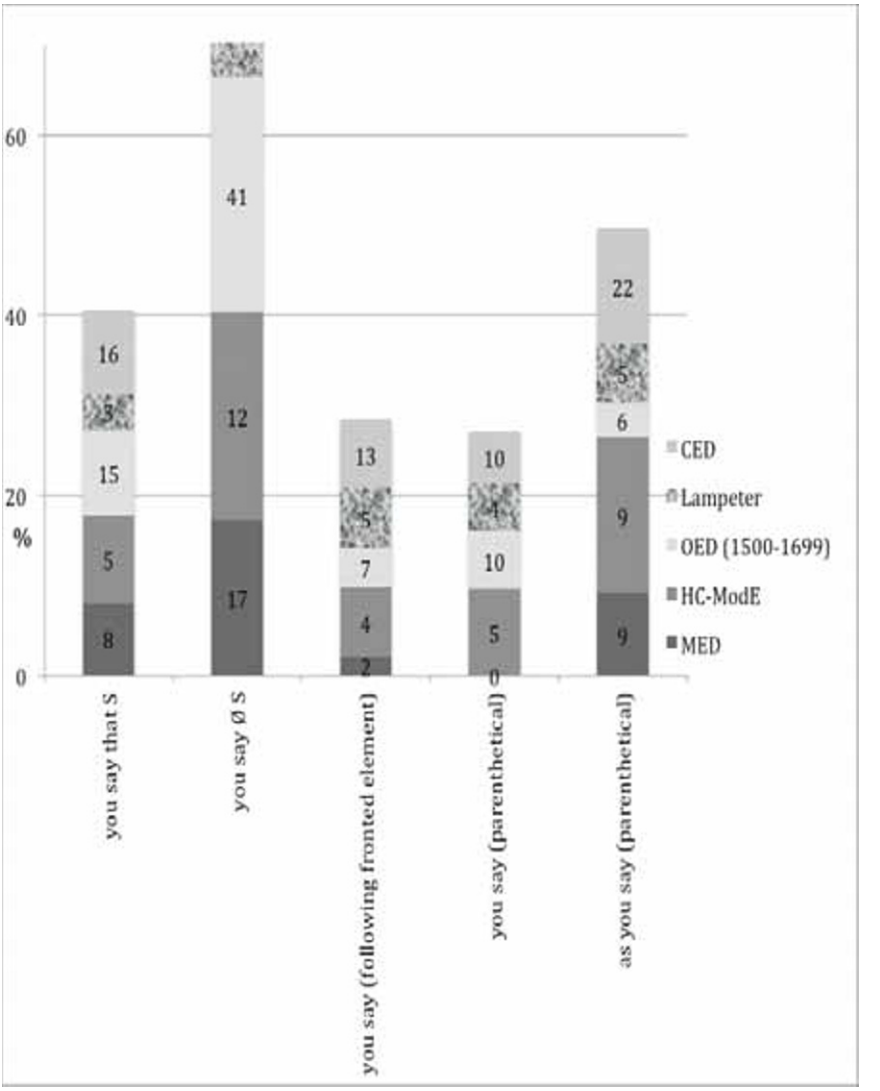


Figure 5: *(As) you say* in Middle and Early Modern English

As Fitzmaurice (2004) has noted, speakers use *you say* in order to draw attention to a proposition for their own communicative ends. In (4a) and (4c), for example, *you say* accompanies a clause that the speaker then explicitly refutes, while in (4b) there is an implicit refutation of the charge of preaching falsehood and fallacy.

In contrast to *you say,* parenthetical *as you say* constitutes roughly 9% of the examples of *you say* in Middle English (5a–b). Its frequency rises to roughly 18% in the Early Modern English section of the Helsinki Corpus (5c–g):

(5) a. We er noght drunkin **als ȝe say**, It ne es bot vnþren tide [Vsp: undrin] of þe day (a1400 *Cursor Mundi* [Göt Theol 107] 18972 [MED]).
'we are not drunk, as you say, it is not but the third hour of the day [9:00]'
b. Þou … sittis, **as þou sais**, in sege as ane Aungell (c1450 (?a1400) *Wars of Alexander* (Ashm. 44) 1872 [MED]).
'you … sit, as you say, in the seat like an angel'
c. Your realme to the which you be bothe (**as you saye**) inheritoure, and by your people accercited and vocated vnto (a1548 E. Hall, *Chronicle* (*The Union of the two Noble and Illustre Famelies of Lancestre and Yorke*) 40 [OED]).
d. If I speake this rashlie and foolishlie, **as you say**, and your self learned as you boast, and I vnlearned, I shall be the more easily ouerthrowne (1593 Gifford, *A Dialogue concerning Witches* B3R [HC]).
e. Faith, **as you say**, there's small choice in rotten apples (1593–4 Shakespeare, *The Taming of the Shrew* I. i. 134–35 [Evans]).
f. If it be **as you say** a trifle, the more to blame you (1601 *Concerning Churching of Women* [CED]).
g. if you stay here the law **as you say** will very speedily pursue you (1662 Dauncey, *The English Lovers* [CED])

As in Present-day English, we can see here that *as you say* is used to express approbation or agreement with the interlocutor's ideas. It may also be used in a metalinguistic sense, as in (5b).

Unambiguous examples, in which *you say*[19] is a parenthetical in medial or final position, appear in the late 16th century:

(6) a. The text itself, **you say**, is sufficient to convince this absurdity (1583 Fulke, *A Defense of the Sincere and True Translations of the Holie Scriptures into the English Tong* x. 391 [OED]).
b. Well, on Mistress Ford, **you say**, – (1597 Shakespeare*, The Merry Wives of Windsor* II, ii, 47 [Evans]).[20]

Syntactically indeterminate instances of *you say* following fronted constructions continue to occur in Early Modern English:

(7) a. knowing that which **you say** proceedeth from a deere care that you haue (1605 Erondell, *The French Garden* [CED])
b. O that / I knew this husband, which, **you say**, must charge his / horns with garlands! (1606–7 Shakespeare, *Antony and Cleopatra* I, ii, 3–5 [Evans]).

c. Is this youre sonne, who **ye say** was borne blind? (1611 *King James Bible* [HC]).
d. the Bread, which after consecration, **you say**[,] is turned into the Body of Christ (1641 *Dialogues Betwixt Three Travellers* [CED]).

However, the most common construction in Early Modern English as in Middle English is clause-initial *you say* followed by a *that*-less nominal clause (8).[21] The construction is more frequent than clausal complements with an overt *that* complementizer (see Figure 5).

(8) a. I cannot tell what you will doe, for **you say** my horse hath broken into your corn (1594 *A Knacke to Knowe a Knaue* [CED]).
b. **You say** you will deale rationally in those ways (1649 *Triall of Lieut. Collonell John Lilburne* [CED]).
c. **You say**, *We preach another Gospel*: You do but *Say* it, and I thank God, *you can Do no more* (1674 Penn, *A just Rebuke to one & twenty Learned and Reverend Divines* [LC]).
d. **You say** you saw him the 29th at *Tixhall* Bowling green (1685 *The Trial of Titus Oates* IV, 85.C2 [HC]).

Example (8c) makes clear the interactive function of *you say*, as the speaker explicitly comments on the interlocutor's restriction to speech rather than action.

Representative examples of the different forms of *(as) you say* in Late Modern English are given in (9):

(9) a. **You say that** your time is very well employed (1746–71 Chesterfield, *Letters to his Son* [CLMET]).
b. **You say**, you should like to see your young hounds run a trail-scent (1781 Beckford, *Thoughts on Hunting* (1802) 85 [OED]).
c. As for the boasted Cleopatra, which **you say** was drawn from your own wife, I believe the copy (1751 Smollett, *The Adventures of Peregrine Pickle* [CLMET]).
d. You make your own pens, **you say**. Nib them a little broader (1752 E. Synge, *Letters* 14 Aug. (1996) 455 [OED]).
e. A Train of Thinking which sometimes I get into … ; I hope, only symptomatically, **as you say** (1742 Richardson, *Pamela* (1785) III. xli. 391 [OED]).
f. A Natatile Beet, **do you say**? … Who ever heard of, or ever read the Name of a swimming Beet? (1725 Bailey tr. *Erasmus' Colloquies* 443 [OED]).

Figure 6 (p. 119) shows the percentage of *(as) you say* types in Late Modern English in the corpora that I have examined. From the mid-18th century to the present, Fitzmaurice finds a rise in use of *you say* in drama, but a decrease

in its use in letters from a high point in the early 18th century (2004: 441) in the ARCHER Corpus.

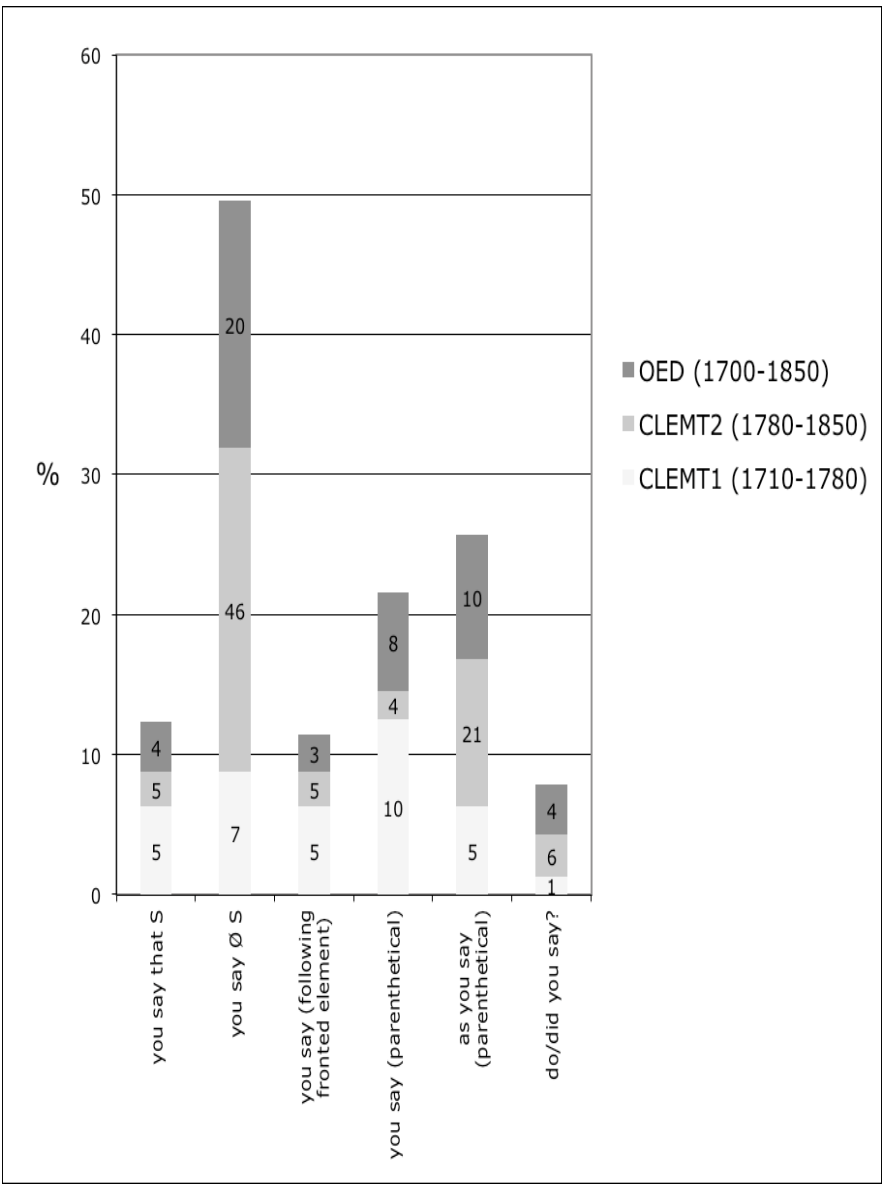


Figure 6: *(As) you say* in Late Modern English

3.4 Accounting for the development of *(as) you say*

Average percentages of the different *you say* structures in the historical corpora are shown over time, from Middle English through Late Modern English, in Figure 7.

There does not appear to be historical evidence that *as you say* develops from *you say*, as suggested by Fitzmaurice (2004, see above). In fact, parenthetical *as you say* predates parenthetical *you say*: parenthetical *as you say* can be found in the early 15th century, while parenthetical *you say* does not appear (in my corpora) until the late 16th century. *As you say* occurs with much the same frequency from Middle English through Late Modern English, though there is a slight decline in the modern period (with the possible exception of British English; see the relative frequency of *as you see* in the FLOB and BNC data given in Table 1).

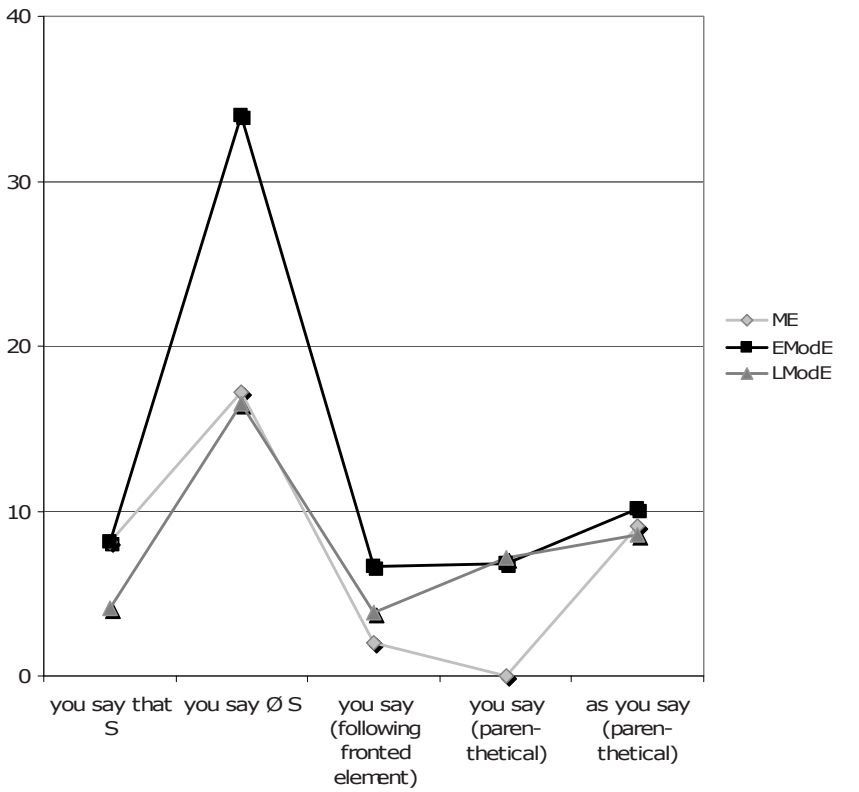


Figure 7: The frequency of different *(as) you say* structures over time

The "matrix clause hypothesis" – implicit in Fitzmaurice's argument – encounters several problems in the case of *as you say*. First, the hypothesis predicts that the source construction, namely *you say that* S, should be the most frequent construction in the time period immediately prior to the development of the indeterminate construction. Yet *you say* with an explicit *that* complementizer is of low frequency during the Middle English period. Its frequency never exceeds 8%. In fact, the adverbial construction *as you say* is marginally more common than *I say that* S in Middle English (9% as compared with 8%). Second, although indeterminate *you say* Ø S structures are the most frequently occurring forms during all periods,[22] they still constitute only one-third of the total instances of *you say* at their height in Early Modern English. Third, it is not clear that these structures should necessarily be understood as resulting from *that*-deletion. Historically, *that* deletion is a complex process, with no simple path from the *that* to the zero form. Rissanen (1991) concludes that "omission" is perhaps not an optimal term because "[a]t the level of spoken expression zero may well have been the unmarked object link throughout the history of English" (287). He finds zero to be scantily attested in Old English and early Middle English, but to have gained ground in late Middle English and to have reached its height in the late 16th/early 17th century. Following this large rise, the tide of zero forms is reversed in the "norm-loving eighteenth century" (288). In confirmation, Finegan and Biber (1995) see a consistent rise in the use of *that* from 1650–1990 in the ARCHER Corpus. Moreover, individual verbs have behaved differently with respect to *that* over time: the zero complementizer is less common with *say, tell,* and *see* than with *know* and *think*, according to Rissanen (1991). Finegan and Biber (1995) find that *think* is the only verb that occurs consistently with a majority of zero forms in their data. Aijmer (1997: 8–10) concludes as well that with *I think* the zero complementizer may have been the unmarked link in speech throughout Old English and Middle English.[23] Thompson and Mulac's (1991) hypothesis, based as it is on the behavior of *think* and *guess*, may therefore be unduly influenced by the fact that *think* stands out as more often taking a zero-complementizer than other verbs.

These facts call into doubt the aptness of the matrix clause hypothesis alone for explaining the rise of parenthetical *you say*. Given the early appearance of parenthetical *as you say* (including the existence of the relative/adverbial *þe/ þæt ge secgað* 'which/as you say' in Old English), it might serve as a possible source of parenthetical *you see*. That is, *you say* would evolve from *as you say* via deletion of the adverbial/relative complementizer *as*. Note that such a source does not involve the reversal in syntactic hierarchy implicit in the matrix clause hypothesis.

The predominance of sentence-initial, indeterminate matrix *you say* S in Middle and Early Modern English suggests strongly that this structure contributes to the rise of parenthetical *you say*. However, both the rarity of *you say that* S structures in the early period and the complexities of *that*-deletion in the history of English (indeed the claim by some scholars that *that*-less forms may even be

the norm until the eighteenth century), the initial step in the matrix clause hypothesis is called into question.

An additional source for parenthetical *you say* may be the construction in which a relative or interrogative object pronoun or stressed NP from the subordinate clause is fronted, resulting in the indeterminate structures shown in (4c), (7a–d), and (9c). Here *you say* may be reanalyzed as a medial parenthetical. Thus, the contribution of this construction to the rise of parenthetical *you see* cannot be discounted. Therefore, it is perhaps most correct to think of the blending of three constructions, the relative/adverbial *as you say* (with deletion of *as*), the sentence-initial *you say* and *that*-less nominal complement (with syntactic reversal of main and subordinate clause), and indeterminate *you say* following a fronted element, as the source of parenthetical *you say*.

What seems clear from the data, however, is that, at least from Middle English onwards, the complement with explicit *that* following *you say* has not played an important role, even in written documents.

4. Conclusion

Many of the historical corpora and text collections of English discussed in this chapter, especially the Chadwyck-Healey and Gale corpora, offer rich resources for the literary scholar and social historian. For example, by performing keyword searches of *interesting* on the ECCO (*Eighteenth Century Collections Online*) corpus, at varying levels of subject specificity, Law (2011) is able to track the semantic and grammatical shifts of the word *interesting* through the 18th century; she concludes that the shift in use of *interesting* from verb to adjective, and its shift in attachment from event-based to landscape and person-based nouns, signals the subject's turn towards less agential, more affective, and more ethical subject-object relations – a turn produced by the experience of rapid change resulting from the rise capitalism and colonialism. More generally, Franco Moretti's *Graphs, Maps, Trees* is a call for literary scholars to practice a methodology he terms "distant reading" – a way studying of literary history that (1) takes large amounts of "data" (in this case, novels) (2) renders them into abstract models (graphs, maps, and evolutionary trees) and (3) argues for the significance of these patterns and trends in literary-historical terms.[24]

Provided the scholar is patient and persistent, these primarily literary corpora can also be combined with historical corpora more specifically designed for linguistic study, such as the *Corpus of English Dialogues*, as well as the quotation banks of the *Oxford English Dictionary* and the *Middle English Dictionary*, to provide an aggregate corpus with a comprehensive chronological spread, a wide distribution of genres – including speech or speech-like genres – and, most importantly, a sufficient size to undertake studies in historical pragmatics. While some of these corpora/text collections place limitations on the scholar, for example, by providing insufficient context – especially important for pragmatic study – or by not allowing easy viewing and printing of the data, the case study in

this paper was intended to show that the use of such an eclectic set of corpora, with qualitative and quantitative corpus linguistic methodology, can succeed. It can lead to advances, both descriptive (the semantic and syntactic development of *(as) you see* comment clauses in English) and theoretical (the complex and multiple origins of comment clauses more generally), in our understanding of language change on the discourse level. As observed by Christian Mair, "[t]he role of the corpus, after all, is not only to provide a limited and representative data-base for statistical analysis, but also to provide authentic and realistic data, the close reading of which will allow the linguist to approach grammar from a functional and discourse perspective" (1991: 77).

Notes

1 See the lists compiled by David Denison, Richard Xiao and David Lee (http://www.llc.manchester.ac.uk/intranet/ug/useful-links/computing-resources/#corpora, http://www.lancs.ac.uk/postgrad/xiaoz/papers/corpus%20survey.htm and http://www.uow.edu.au/~dlee/CBLLinks.htm, respectively). See also Xiao (2008) as well as Rissanen's now somewhat dated review article (2000).

2 By Willinsky's (1994: 211) count, there are over twice as many quotations from Shakespeare as his nearest competitor (Walter Scott) in the first edition of the OED. He also points to some obvious omissions from the OED, including the Romantic poets, Chancery English, working-class presses of the nineteenth century, and "the entire body of women writers" (177).

3 These are being added to continually as the 3rd edition progresses. For information on *The Oxford English Corpus*, see http://oxforddictionaries.com/page/oec.

4 An advantage of the 3rd edition, however, is that it is possible to reorganize the quotations in a number of different ways, such as chronologically, whereas in the older edition the quotations appear invariably under the headword.

Note that a new interface for the OED was launched in 2010, but despite a number of additional features (such as direct links to the *Middle English Dictionary* and *Historical Thesaurus*), the difficulties of performing corpus searches remain.

5 I am restricting my discussion to more generalized, multi-purpose corpora and am not including specialized corpora restricted by genre or domain such as the *Corpus of Early English Medical Writing* (late 14th century to

1750) or *Lexicons of Early Modern English*, which consists of word entries from monolingual dictionaries dating from 1480–1702, nor newspaper corpora, such as the *Zurich Corpus of English Newspapers* (1671–1791) or the *ProQuest Historical Newspapers*, many beginning in the mid-19th century. On the distinction between generalized and specialized corpora, see McEnery et al. (2006: 15).

I have also not included the Penn-York-Helsinki parsed historical corpora (Old English prose, Old English poetry, Middle English, Early Modern English, Early English). The parsed corpora are rather expensive and not as readily available as the other corpora discussed here, although they are obviously important resources.

6 This corpus was not available when this paper was originally written (February 2008) and was therefore not used for the case study presented in Section 3.

7 The former is available through the Oxford Text Archive and the latter directly from David Denison.

8 The *Corpus of Nineteenth-Century English* (CONCE), covering the period 1800–1900 and consisting of one-million words is being compiled by Merja Kytö, Juhani Rudanko and Erik Smitterberg but is not yet publicly available. Nor is the older *Century of Prose Corpus* (COPC) covering the period 1680–1780 and consisting of one-half million words.

9 Nonetheless, I have tried to be more quantitatively rigorous than McEnery and Wilson (2001: 123) suggest when they claim that historical linguistics "has tended to take a more selective approach to empirical data, simply looking for evidence of particular phenomena and making, at most, rather rough estimates of frequency".

10 Because of the greater (and sometimes exclusive) reliance on written documents in the earlier period, I chose to use primarily the written corpora for Present-day English as my starting point in identifying the functions of the pragmatic markers I have studied.

11 See Stenström (1984) on the difficulties of tagging discourse markers in a corpus.

12 One presumed Middle English example of "I say" that I collected required the combined expertise of two medieval colleagues with knowledge of the

story-line of the original text to decipher it as actually an example of "I see". (My thanks to Sîan Echard and Robert Rouse for their help.)

13 For example, the VIEW search program (designed by Mark Davies for the *British National Corpus,* the *Time Archive*, and the *Corpus of American English*) allows for searches incorporating punctuation, but I have found that those involving quotation marks are unreliable.

14 I put the raw results of the MED search into a Word file and then format and sort them. Additional problems with MED searches – apart from problems posed by variable spelling – include the duplication of citations, which is much more extensive than in the OED, and the difficulty of organizing the examples chronologically when dates may be preceded by "?" or "a" or both.

15

16 See Akimoto (2000) on *pray*, (2002) on *I'm afraid*; Brinton (1996) on *I think/guess*, etc.; Fischer (2007a, 2007b) on *I think/guess*, etc.; Fitzmaurice (2004) on *you see, you say, you know*; Nevanlinna (1974) on *as who say/saith*; López-Couso (1996) on *methinks*; Molina (2002) on *I'm sorry*; Palander-Collin (1999) on *I think/methinks*; Traugott (1995) on *let us, let alone, I think*; Traugott and Dasher (2002) on *let's, I pray, I promise*; Wischer (2000) on *methinks*.

17 *You say* would belong to Quirk et al.'s (1985: 1114–1115) matrix-clause type comment clause, though they do not list it. They do note the use of the tag *wouldn't you say?* (1115).

18 *As you say*, which is typically parenthetical, is overwhelmingly more frequent in spoken English than in other genres. It has a frequency of 12/million in the spoken subsection of the *Corpus of American English* and 24 times/million in the spoken subsection of the *British National Corpus*.

19 *Thou say'st* survives into the 17th century but is not common.

20 Shakespeare is cited from the printed *Riverside Shakespeare* (Evans 1997).

21 The large number of *you say* forms in the *Corpus of English Dialogues* is a result of the genre of texts in this corpus, most significantly trials and

witness depositions, which are in large part concerned with what people say or have said on occasion. The oral nature of both types of texts would contribute to a preponderance of *that*-less complements (see below, §3.4 on the status of "*that*-deletion").

22 *That*-less complements are the majority form in both Middle English and Early Modern English as well as in Present-day English in the corpora I examined.

23 Cf. Tagliamonte and Smith's (2005) study of modern English dialects, which find that over 90% of the instances of *I think/guess/mean* and *you know* have a zero complementizer, with *think* most frequently lacking the complementizer (also see Wulff 2008, who found that *think, say, mean*, and *know* occurred most often with a zero complementizer).

24 My thanks to Anita Law for bringing this work to my attention.

References

Electronic and online sources

ARCHER Corpus. *A Representative Corpus of Historical English Registers 1650–1990*. http://llc.stage.manchester.ac.uk/research/projects/archer/.

British National Corpus [BNC]. http://www.natcorp.ox.ac.uk/. Searched using VIEW (Variation in English Words and Phrases) developed by Mark Davies, Brigham Young University. http://corpus.byu.edu/bnc/.

A Corpus of English Dialogues 1560–1760. 2006. Compiled by Merja Kytö and Jonathan Culpeper in collaboration with Terry Walker and Dawn Archer. http://www.helsinki.fi/varieng/CoRD/corpora/CED/index.html.

The Corpus of Historical American English 400+ million words, 1810–2009 [COHA]. Searched using VIEW (Variation in English Words and Phrases) developed by Mark Davies, Brigham Young University. http://corpus.byu.edu/coha/.

A Corpus of Late 18c Prose (The English language of the north-west in the late Modern English period). Compiled by David Denison, Linda van Bergen, and Joanna Soliva. http://www.llc.manchester.ac.uk/subjects/lel/staff/david-denison/corpus-late-18th-century-prose/.

A Corpus of Late Modern English Prose. 1994. Compiled by David Denison, Linda van Bergen and Graeme Trousdale. http://www.llc.manchester.ac.uk/subjects/lel/staff/david-denison/lmode-prose/.

The Corpus of Late Modern English Texts [CLEMT] and *Corpus of Late Modern English Texts Extended Version*. Compiled by Hendrik de Smet. http://perswww.kuleuven.be/~u0044428/.

Corpus of Middle English Prose and Verse. 2006. Available through the *Middle English Compendium*. http://quod.lib.umich.edu/c/cme/.

Dictionary of Old English Corpus, Web Corpus [DOEC]. 2004. Ed. by Antonette di Paolo Healey with John Price Wilkin and Xin Xiang. University of Toronto. http://tir.doe.utoronto.ca/index.html.

Early English Books Online [EEBO]. Chadwyck-Healey Ltd. http://eebo.chadwyck.com/home.

Early English Prose Fiction. Ed. by Holger Klein et al. Chadwyck-Healey Ltd. http://collections.chadwyck.com/home/home_eepf.jsp.

Eighteenth Century Collections Online [ECCO]. Gale Digital Collections. http://mlr.com/DigitalCollections/products/ecco/.

Eighteenth-Century Fiction. Ed. by Judith Hawley et al. Chadwyck-Healey Ltd. http://collections.chadwyck.com/home/home_c18f.jsp.

English Drama. Ed. by John Barnard et al. Chadwyck-Healey Ltd. http://collections.chadwyck.com/home/home_ed.jsp.

ICAME (International Computer Archive of Modern and Medieval English). *Collection of English Language Corpora*. 1999. Compiled by Knut Hofland, Anne Lindebjerg, and Jørn Thunestvedt. 2nd ed. CD-ROM. Bergen: Norwegian Computing Centre for the Humanities. (See http://khnt.hit.uib.no/ icame/manuals/).
Australian Corpus of English [ACE]
Corpus of Early English Correspondence, sampler [CEECS]
Freiburg-Brown Corpus of American English [FROWN]
Freiburg-LOB Corpus of British English [FLOB]
Helsinki Corpus of English Texts, diachronic part [HC, Helsinki Corpus]
Innsbruck Computer-Archive of Machine-Readable English Texts [ICAMET]
Lampeter Corpus of Early Modern English Tracts [LC]
Wellington Corpus of Written New Zealand English [WC]

Middle English Collection. University of Virginia Electronic Text Center. http://etext.lib.virginia.edu/collections/languages/english/mideng.browse.html.

Middle English Dictionary [MED]. 2001. Electronic version available through the *Middle English Compendium*. http://quod.lib.umich.edu/m/med/.

The Modern English Collection, University of Virginia Electronic Text Center, http://etext.lib.virginia.edu/modeng/modeng0.browse.html.

Nineteenth-Century Fiction. Chadwyck-Healey Ltd. http://collections.chadwyck.com/marketing/home_c19f.jsp.

Oxford English Dictionary [OED]. 2000–. 3rd ed. online (in progress). Oxford: Oxford University Press. http://www.oed.com/.

Strathy Corpus of Canadian English [CanE]. Strathy Unit, Queen's University. http://www.queensu.ca/strathy/Projects.html.

The Time Magazine Corpus (100 million words, 1920s–2000s) [TIME]. Searched using VIEW (Variation in English Words and Phrases) developed by Mark Davies, Brigham Young University. http://corpus.byu.edu/time/.

Secondary sources

Aijmer, Karin (1997), '*I think* – an English modal particle', in: Toril Swan and Olaf Jansen Westvik (eds.) *Modality in Germanic languages: historical and comparative perspectives* (Trends in Linguistics, Studies and Monographs 99). Berlin and New York: Mouton de Gruyter, 1–47.

Akimoto, Minoji (2000), 'The grammaticalization of the verb *pray*', in: Fischer, Rosenbach and Stein (eds.), 67–84.

Akimoto, Minoji (2002), 'On the grammaticalization of the parenthetical "I'm afraid"', in: Jacek Fisiak (ed.) *Studies in English historical linguistics and philology: a Festschrift for Akio Oizumi*. Frankfurt am Main: Peter Lang, 1–9.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan (1999), *Longman grammar of spoken and written English*. Harlow, Essex: Pearson Education.

Blakemore, Diane (2006), 'Divisions of labour: the analysis of parentheticals', *Lingua*, 116: 1670–1687.

Brinton, Laurel J. (1996), *Pragmatic markers in English: grammaticalization and discourse functions* (Topics in English Linguistics 19). Berlin and New York: Mouton de Gruyter.

Brinton, Laurel J. (2008), *The comment clause in English: syntactic origins and pragmatic development* (Studies in the English Language). Cambridge: Cambridge University Press.

Evans, G. Blakemore (ed.) (1997), *The Riverside Shakespeare*. 2nd ed. Boston and New York: Houghton Mifflin.

Finegan, Edward and Douglas Biber (1995), '*That* and zero complementisers in Late Modern English: exploring ARCHER from 1650–1990', in: Bas Aarts and Charles F. Meyer (eds.) *The verb in contemporary English*. Cambridge: Cambridge University Press, 241–257.

Fischer, Andreas (1997), 'The *Oxford English Dictionary* on CD-ROM as a historical corpus: *to wed* and *to marry* revisited', in: Udo Fries, Viviane Muller and Peter Schneider (eds.) *From Ælfric to* The New York Times: s*tudies in English corpus linguistics* (Language and Computers: Studies in Practical Linguistics 19). Amsterdam and Atlanta: Rodopi, 161–172.

Fischer, Olga (2007a), 'The development of English parentheticals: a case of grammaticalization?' in: Ute Smit, Stefan Dollinger, Julia Hüttner, Gunther Kaltenböck and Ursula Lutzky (eds.) *Tracing English through time: explorations in language variation. A Festschrift for Herbert Schendl on the occasion of his 65th birthday* (Austrian Studies in English 95). Vienna: Braumüller, 103–118.

Fischer, Olga (2007b), *Morphosyntactic change: functional and formal perspectives* (Oxford Surveys in Syntax and Morphology). Oxford: Oxford University Press.

Fischer, Olga, Anette Rosenbach and Dieter Stein (eds.) (2000), *Pathways of change: grammaticalization in English*. Amsterdam and Philadelphia: John Benjamins.

Fitzmaurice, Susan (2004), 'Subjectivity, intersubjectivity and the historical construction of interlocutor stance: from stance markers to discourse markers', *Discourse Studies,* 6 (4): 427–448.

Hoffmann, Sebastian (2004), 'Using the *OED* quotations database as a corpus – a linguistic appraisal', *ICAME Journal*, 28: 17–30.

Hooper, Joan B. (1975), 'On assertive predicates', in: John B. Kimball (ed.) *Syntax and semantics*. Vol. 4. New York: Academic Press, 91–124.

Jacobs, Andreas and Andreas H. Jucker (1995), 'The historical perspective in pragmatics', in: Andreas H. Jucker (ed.) *Historical pragmatics: pragmatic developments in the history of English* (Pragmatics & Beyond, New Series 35). Amsterdam and Philadelphia: John Benjamins, 3–33.

Jucker, Andreas H. (1994), 'New dimensions in vocabulary studies: review article of the *Oxford English Dictionary (2nd edition)* on CD-ROM', *Literary and Linguistic Computing*, 9 (2): 149–154.

Law, Anita (2011), '*Interesting*, 1700–1800'. MS, University of British Columbia.

López-Couso, María José (1996), 'On the history of *methinks*: from impersonal construction to fossilized expression', *Folia Linguistica Historica*, 17: 153–169.

Mair, Christian (1991), 'Quantitative or qualitative corpus analysis? Infinitival complement clauses in the Survey of English Usage corpus', in: Stig Johansson and Anna-Brita Stenström (eds.) *English computer corpora. Selected papers and research guide*. Berlin and New York: Mouton de Gruyter, 67–80.

McEnery, Tony and Andrew Wilson (2001), *Corpus linguistics: an introduction*. 2nd ed. (Edinburgh Textbooks in Linguistics). Edinburgh: Edinburgh University Press.

McEnery, Tony, Richard Xiao and Yukio Tono (2006), *Corpus-based language studies: an advanced research book* (Routledge Applied Linguistics). London and New York: Routledge.

Meyer, Charles (2002), *English corpus linguistics: an introduction*. Cambridge: Cambridge University Press.

Molina, Clara (2002), 'On the role of meaning in the historical development of discourse markers'. Paper presented at the 12th International Conference on English Historical Linguistics, Glasgow, UK, August 2002.

Moretti, Franco. 2005. *Graphs, maps, trees*. New York: Verso.

Nevanlinna, Saara (1974), 'Background and history of *as who say/saith* in Old and Middle English literature', *Neuphilologische Mitteilungen*, 75: 568–601.

Palander-Collin, Minna (1999), *Grammaticalization and social embedding: I THINK and METHINKS in Middle and Early Modern English* (Mémoires de la Société Néophilologique de Helsinki 55). Helsinki: Société Néophilologique.

Peltola, Niilo (1982/1983), 'Comment clauses in Present-day English', in: Inna Koskenniemi, Esko Pennanen and Hilkka Aaltonen (eds.) *Studies in classical and modern philology.* Helsinki: Suomalainen Tiedeakatemia, 101–113.

Potts, Christopher (2002), 'The syntax and semantics of *as*-parentheticals', *Natural Language & Linguistic Theory*, 20: 623–689.

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

Rissanen, Matti (1991), 'On the history of *that*/zero as object clause links in English', in: Karin Aijmer and Bengt Altenberg (eds.) *English corpus linguistics: studies in honour of Jan Svartvik*. London and New York: Longman, 272–289.

Rissanen, Matti (2000), 'The world of English historical corpora: from Cædmon to the computer age', *Journal of English Linguistics*, 28: 7–20.

Ross, John Robert (1973), 'Slifting', in: Maurice Gross, Morris Halle and Marcel-Paul Schützenberger (eds.) *The formal analysis of natural languages: proceedings of the First International Conference.* The Hague and Paris: Mouton, 133–169.

Schmied, Josef (1993), 'Qualitative and quantitative research approaches to English relative constructions', in: Clive Souter and Eric Atwell (eds.) *Corpus-based computational linguistics.* Amsterdam and Atlanta, GA: Rodopi, 85–96.

Schourup, Lawrence (1999), 'Discourse markers', *Lingua*, 107: 227–265.

Stenström, Anna-Brita (1984), 'Discourse tags', in: Jan Aarts and Willem Meijs (eds.) *Corpus linguistics: recent developments in the use of computer corpora in English language research* (Costerus, New Series, 45). Amsterdam: Rodopi, 65–81.

Tagliamonte, Sali and Jennifer Smith (2005), '*No momentary fancy!* The *zero* "complementizer" in English dialects', *English Language and Linguistics*, 9: 289–309.

Thompson, Sarah and Anthony Mulac (1991), 'A quantitative perspective on the grammaticalization of epistemic parentheticals in English', in: Elizabeth Closs Traugott and Bernd Heine (eds.) *Approaches to grammaticalization*. Vol. 2. (Typological Studies in Language 19). Amsterdam and Philadelphia: John Benjamins, 313–329.

Traugott, Elizabeth Closs (1995), 'Subjectification in grammaticalisation', in: Dieter Stein and Susan Wright (eds.) *Subjectivity and subjectivisation: linguistic perspectives.* Cambridge: Cambridge University Press, 31–54.

Traugott, Elizabeth Closs and Richard B. Dasher (2002), *Regularity in semantic change* (Cambridge Studies in Linguistics 96). Cambridge: Cambridge University Press.

Willinsky, John (1994), *Empire of words: the reign of the OED*. Princeton: Princeton University Press.

Wischer, Ilse (2000), 'Grammaticalization versus lexicalization: "*methinks"* there is some confusion', in: Fischer, Rosenbach and Stein (eds.), 355–370.

Wulff, Stefanie (2008), 'A multifactorial approach to *that*-deletion in English complement constructions'. Paper presented at the American Association of Corpus Linguistics Conference, Provo, Utah, March 2008.

Xiao, Richard (2008), 'Well-known and influential corpora', in: Anke Lüdeling and Merja Kytö (eds.) *Corpus linguistics: an international handbook* (Handbücher zur Sprach- und Kommunikationswissenschaft/ Handbooks of Linguistics and Communication Science 29.1–2). Berlin and New York: Mouton de Gruyter, 383–456.

This page intentionally left blank

‘Upon these *Heads* I shall discourse’: lexicographical and corpus evidence for senses and phrases

Claudia Claridge

University of Duisburg-Essen

Abstract

This paper uses the set of body part terms to investigate the use and the treatment of transferred senses in Early Modern English. The data basis is provided by four eighteenth-century dictionaries (Kersey 1715, Bailey 1730, Martin 1749, Johnson 1755) and by EModE corpora such as the CED, CEEC, and ZEN. Points to be investigated for the dictionaries are the kinds, the number and the ordering of transferred senses listed, the presence of accompanying usage notes, as well as the differences and similarities between the dictionaries. The corpus analysis will provide evidence about the kinds and relative frequencies of senses in real language use, as well as about their typical contexts and collocations. Finally, the degree of overlap between the dictionaries and the corpora will be discussed.

1. Introduction

Corpora have come to have a logical connection to dictionaries, in the sense that in much modern lexicography electronic corpora form the data backbone, supplying and justifying the words, phrases, senses and particular usages to be included in dictionaries. The COBUILD project and the *British National Corpus* supported by dictionary publishers Longman and Chambers are well-known instances of this practice. In reversal of this approach, any existing dictionary, whether originally based on a corpus or not, can of course be evaluated (and potentially improved) by counterchecking its entries against a(nother) corpus. In spite of the remarkable flourishing of historical corpus linguistics, it has remained largely disconnected from the study of English historical dictionaries from Cawdrey (1604) onwards. An exception is Barnbrook’s (2005) comparison of Johnson’s prescriptive usage notes with evidence from the Helsinki Corpus and Alexander Pope’s works. Combining the two fields of research could certainly provide insights that are not possible with a purely *intra*disciplinary approach. The present contribution intends to make a start in this direction. It aims at counterchecking entries in four eighteenth-century dictionaries in the light of authentic Early Modern English usage as attested by three corpora. The question is which particular senses and usages are recorded in dictionaries and how does this fit to the ones attested by corpora. Needless to say, a comprehensive assessment of this aspect would be a major undertaking, so that a small pilot study is to illustrate the potential usefulness and also the problems of such an

approach. This pilot is based on a small selection of body part words, namely *head, face, eye, foot,* and *leg*, which are included in all the dictionaries and are frequent enough in corpora to make generalisations possible. Another reason for choosing terms of this kind is that they are frequently found in non-literal and idiomatic usages[1] (following from the anthropocentric and 'embodied' nature of language), which is a special focus of the present study.

2. Sources of data: lexicography meets corpus linguistics

Both dictionaries and corpora hold up a mirror to the language of a given period, but they do so in different ways. Dictionaries tell us something about a period's approach to language, its degree and type of (meta)linguistic awareness, its attitudes to language and its usage as well as of course about its stage of lexicographical development in the purely technical sense. Corpora tell us something about a period's linguistic behaviour in various contexts in all its richness, at least if the corpora are well constructed and well chosen. Language is a tool/instrument for transporting content and producing sense in corpora and it is an 'object' abstracted away from individual uses in dictionaries. However, neither corpora nor dictionaries (or rather: their respective makers) are perfect. Not even the best corpus is truly representative[2] and comprehensive, and even the best lexicographer will miss, skew or misrepresent some linguistic 'facts' – and this point is the more relevant the more we go back in time. But the two sources coexisting for a period and forming the basis of investigation may not be liable to omissions, biases and faults to the same extent. Thus, they can not only be used to corroborate (or falsify) each other's descriptions, but also to complement each other, each providing information that the other may lack.

After these general remarks let me now introduce the data to be made use of here, starting with the lexicographical sources. The first half of the eighteenth century witnessed the birth of general lexicography in the modern sense, with John Kersey's *Dictionarium Anglo-Britannicum* from 1708 being the first dictionary to conclusively break with the so-called hard-word tradition (Starnes and Noyes 1991). It is thus one of the dictionaries to be included here, along with Nathan Bailey's *Dictionarium Britannicum* (1730), Benjamin Martin's *Lingua britannica reformata* (1749) and Samuel Johnson's *A Dictionary of the English Language* (1755[3]). All four include the common vocabulary of English, as well as each of them a certain amount of more specialised words. The number of entries thus varies between them, with Martin being the smallest (c. 24,500 words), followed by Kersey (c. 35,000), Johnson (42,773) and, with the highest number, Bailey (c. 48,000). As concerns the more specialised lexical components, the following information is found in the prefaces. Kersey claims to include terms relating to arts and sciences from many fields, legal phrases, proper names, hard words found in noted authors and difficult words from other languages. Bailey says he lists hard and technical words from many fields, legal terminology, terms relating to classical culture/ancient history and proper names. Martin also enters

words from the sciences and from various technical fields, as well as, according to his preface, phraseological units. Johnson dispenses with many specialized, technical words and pays considerably more attention to the common words of English. Both Martin and Johnson claim to exclude obsolete words and both of them explicitly say that they include both 'original' / 'primitive' and 'figurative' / 'metaphorical' meanings of words. These latter two are the first dictionary makers to evolve and to expound on a more theoretical and systematic lexicographical approach. As to the sources for the entries, all lexicographers of the time based themselves at least partly on other dictionaries, both monolingual and bilingual ones, whether they acknowledged it or not. Martin, for instance, mentions other dictionaries as guidelines for sense differentiation, such as the bilingual *Latin Dictionary* by Ainsworth and the *Royal French Dictionary.* Notably, he also mentions 'popular speech' as a source of input (p. viii). Authentic speech, or rather writing, is the major source for Johnson, who used the literary output from the time of Sidney to the Restoration, which he called the 'wells of English undefiled' (preface), as his 'corpus'. All of them also used specialized works, of a lexicographic or general nature, to a smaller or larger extent, such as John Harris' *Lexicon technicum* (1704) used by Kersey.

I have chosen three corpora to compare to these dictionaries, namely the *Corpus of English Dialogues* (CED), a sub-section of the *Corpus of Early English Correspondence* (CEEC, parsed version),[4] and the *Zurich English Newspaper Corpus* (ZEN). This selection is intentionally removed from the known or likely sources of the dictionaries, covering the potentially more private (CEEC), more colloquial, spoken-like (CED) and less high-brow, more utilitarian (ZEN) registers, on the whole more everyday types of linguistic output. The corpus sources thus contrast with the comparatively more literary and technical / specialised sources used by the dictionary compilers and are therefore a good way to evaluate how representative the dictionaries nevertheless are with regard to the common core of the lexicon. This may then show how much the sources used have biased the dictionaries and how far they have removed them from everyday language use. Basic information about these corpora is summarized in Table 1 (additionally, the Helsinki Corpus is made use of for some aspects).

Table 1: Corpus sources

	Time	***Words***	***Texts***	***Types of texts***
CED	1560–1760	1,157,720	168	trials, witness depositions, comedy, handbooks, prose fiction
CEEC sub-corpus	1650–1710	396,864	16	letters
ZEN	1661–1791	1,627,162	349	newspapers
Total	1560–1791	3,181,746	533	

In the case of Johnson, there is another interesting avenue to pursue. As he has produced a great variety of texts besides his dictionary, there is the possibility to compare his own linguistic usage to his dictionary entries. I have chosen five (parts of) texts for this purpose, which together cover a fairly wide range of types of writing and come to 409,611 words. These texts are: (1) *Rasselas, Prince of Abyssinia* (literary narrative), (2) *A Journey to the Western Islands of Scotland* (travel writing), (3) the *Rambler*, (4) the *Idler* (both essays) and (5) *Lives of Poets,* here Addison, Savage and Swift (biography and literary criticism).[5]

The corpora and Johnson's text have both been used in a typical corpus-linguistic way, with KWIC-concordances being produced and analysed for the search terms chosen.[6] The dictionary data is available in less easily analysable forms. The four dictionaries are included in PDF-facsimiles of the original/old editions in the resource *Eighteenth-Century Collections Online* (Thomson-Gale).[7] It is possible to enter search terms (e.g. the word *head*) after an individual dictionary has been chosen, but there are certain drawbacks to the output: (i) it is not in concordance form nor in any otherwise collected form, but one needs to click on the individual hits in sequence to work one's way through the dictionary, and (ii) it is not completely reliable as it is based on the old typeface; it finds irrelevant things, while it also misses occurrences. It is, however, possible to download and print out parts of the works selected. Therefore, the basis of the research presented here are simply the entries of the respective keywords to be investigated. It will not be possible to say anything about the overall range of uses (senses and phrases) of, e.g., *eye* in the dictionaries, and thus to evaluate what the compilers had at their disposal/in their knowledge, but nevertheless may not have chosen to include in the entries proper.

3. Words, senses and uses – and their treatment

Polysemy can be seen as the norm in language, as the great majority of words will have more than one established sense. Most existing polysemy can probably be covered by assuming the following routes of meaning extension from a given sense: generalisation, specialisation, metaphor and metonymy (e.g. Cruse 2004). Particularly in the latter cases, we are originally dealing with deliteralization processes, thus raising the question of distinguishing literal from non-literal meanings. With respect to the body part items used here (*head, face, eye, leg, foot*), this tricky problem can be dealt with in a very pragmatic commonsensical way, namely by taking as the really literal meaning that one which refers to the physical and visible anatomical part of humans and vertebrate animals.[8] All other senses are here taken to be transferred in the sense that they must originally have derived from this sense, be it metonymically (e.g. *we numbered twenty heads*, i.e. persons), metaphorically (e.g. *the eye of the storm*) or in some other way. If the words in question occur in specific collocations, (semi-)fixed phrases and full-blown idioms, they also often tend to have a non-literal meaning, e.g. *foot the bill, face to face, an eye for an eye and a tooth for a tooth*.

The established, i.e. conventional, senses of polysemous items are as a rule listed in modern dictionaries, in contrast to, e.g., creative metaphors and perhaps common, but nevertheless ad-hoc extensions (such as metonymies of the kind *the hamburger wants to pay*). As sense differentiation is not uncontroversial, there will be differences as to the number of senses distinguished by particular dictionaries and also as to the order in which they are presented, depending on the criteria used (e.g. frequency, chronology). There is furthermore the question of classification of senses, that is whether or not they are labelled as a (certain type of) figurative meaning – which also partly depends on which meaning has been given as the basic one (cf. Drosdowski 1989: 798). Osselton (1995a: 16f., 18, 21) remarked on the fact that modern dictionaries are in themselves fairly inconsistent in applying the label 'figurative' (for example marking *locust* 'person with destructive propensities' as figurative, but not *wolf* 'rapacious, greedy person'), do not agree with each other in applying the label, and furthermore use a variety of labels for non-literal uses (besides fig. also slang, colloquial, informal etc.) without clearly explaining their criteria.[9] The lexicographical treatment of phrasal units is even more problematic and often not systematically done by modern dictionaries, with learner dictionaries scoring somewhat better in this respect. The question is first of all how many and which phraseological units are listed at all. If some are included, the following aspects would need to be considered: where a phrasal unit is to be listed (under which main entry, where within the entry), what is the lemma form of the unit, what meta-information is to be given and how it is to be labelled (cf. Burger 1989). All of these aspects are treated differently in extant dictionaries.

In historical lexicography, the provision of and differentiation into different senses evolves gradually, the larger and less hard-word oriented the dictionaries become. In this process, non-literal senses will also increasingly have been listed, even though not often explicitly recognized as such. Drosdowski (1989: 799f.) found on the whole a very unsystematic and haphazard treatment of metaphorical senses in early German lexicography before Adelung's dictionary (i.e. before the 1790s). Often metaphorical senses are not listed, and if they are, they are as a rule not labelled. He singles out Stiegler's dictionary of 1691 as one early example where intermittent, inconsistent marking occurs.[10] The situation is undoubtedly similar in early English lexicography, although it still awaits more investigation. As to the lexicographic treatment of phraseology, Knappe (2004: 450) concludes that this was as a rule largely ignored, partly because English-Latin lexicography still exerted a strong influence on the selection of lemmas to be defined. Exceptions to this general situation are Bailey, who covered specialized phrases and proverbs in many of his works, Johnson, and especially Wilkins and Lloyd's *An alphabetical dictionary* from 1668 (Knappe *ibid*.).

The practice of the four dictionaries under scrutiny in this study will be treated in Section 4, so that I will concentrate here on their meta-comments as far as available. Neither Kersey nor Bailey say anything about polysemy and their treatment of it in the front matter of their works, nor about the phraseology of everyday English; they do mention legal phrases as being covered, however.

Furthermore, they mention a large number of specialized fields, whose vocabularies will be included, which also implies the listing of specialized senses. Neither of them talks about the lexicographical decisions related to these aspects. The matter is different and drastically improved with Martin and Johnson. Martin devotes Section v of his preface to this question and its lexicographical treatment, identifying the 'accurate enumeration and distinction of the several significations of each respective word' (p. vii) as a chief desideratum, which had so far been neglected by dictionaries. His own work was to remedy this by listing the different senses in the following order: (i) etymological or original sense, (ii) general, popular significations, (iii) figurative, metaphorical uses, (iv) humorous, poetical and burlesque uses, and (v) scientific meanings – to which are added at the end of the entry compounds and phraseological units involving the head word in question, both of which may of course be based on any of the preceding sense types. He does not elaborate on the above sense categories by giving definitions. Martin is the first to number the senses listed in an entry and thus to clearly distinguish them. Johnson remarked in the preface to his dictionary that 'the tropes of poetry will make hourly encroachments, and the metaphorical will become the current sense', thus acknowledging the role of transferred uses in semantic change. Like Martin, Johnson numbered the senses provided, and also attempted to proceed from the 'primitive [sense of the word] to its remote and accidental signification'. In the *Plan* (1747) he was fairly specific as to the ordering of senses, but seems to have arrived at a less strict attitude by the time the dictionary appeared, triggered probably by the practical problems of distinguishing senses and of ordering them. Particle verbs are highlighted by Johnson in particular, as they lead to new, transferred meanings and constitute an important instance of English phraseology. Otherwise, he does not comment on phrasal units. Neither of these lexicographers comments explicitly on a labelling practice with regard to non-literal and phrasal uses.

4. The lexicographers' view

The aspects to be considered here are: (i) how many entries and senses are attributed to the words in question?, (ii) how are they ordered?, (iii) how many and which non-literal senses and phrasal uses are provided?, (iv) how are they treated (e.g. sequencing, labelling)? and finally (v) how do the dictionaries compare with regard to the preceding points?

Let's begin by an overall numerical presentation of entries, senses and phrases provided for the chosen lemmas, as given in Table 2. As can be seen, the two earlier dictionaries exhibit the tendency to give a greater number of entries to these lemmas than the two later ones, though less pronounced in Kersey. One could read into this that they treat the senses covered by these entries as more independent than those combined in one entry and thus take something of a stance of homonymy in these cases (though certainly not in precisely these terms). As many of the senses listed by Kersey and Bailey are fairly specialized

ones prevalent in or restricted to certain fields of knowledge and registers, this is not an illogical procedure from their point of view. The connection between *head* 'the uppermost or chief part of the body' and *head* 'shank or longest part of [an anchor]' (Bailey) is after all not that easy to make. With regard to entries, it is also apparent that nominal uses dominate. Johnson is the most consistent in including verbal instances, with the exception of *leg*. All the other lexicographers are less inclusive. As verbal uses for these particular words will have a tendency to be (somewhat) figurative and/or phrasal, this is an instance of interesting senses being potentially overlooked. While it is easy to count the entries, the matter is more complicated with regard to senses for Kersey and Bailey, who in contrast to Martin and Johnson do not use numbering. A new entry of course indicates a new sense, but within an entry the punctuation used must serve as a guideline. Unfortunately, there is neither explicit comment by the authors on this nor is there consistency.

Table 2: The treatment of five lexemes in four dictionaries

		Kersey 1708	**Bailey 1730**	**Martin 1749**	**Johnson 1755**
	entries	n: 5	n: 11	n: 1	n: 1
head					v: 1
	senses	5	11	5	35 (31 + 4)
	phrases	-	-	-	(2)
	entries	n: 3	n: 8	n: 5	n: 1
face		v: 2	v: 2	v: 1	v: 2
	senses	10 (8 + 2)	13 (9 + 4)	14 (11 + 3)	15 (9 + 6)
	phrases	(1)	-	-	1 (2 senses)
	entries	n: 1	n: 16	n: 1	n: 1
eye					v: 2
	senses	3	16	5	17 (15 + 2)
	phrases	-	-	-	-
	entries	n: 1	n: 3	n: 2	n: 1
leg	senses	3	3	2	4
	phrases	-	-	-	-
	entries	n: 2	n: 4	n: 1	n: 1
foot			v: 1		v: 2
	senses	5	6 (5 +1)	8	21 (16 + 5)
	phrases	2	3	-	1

(Notes: n = noun, v = verb; () under senses subdivides senses into first nominal and secondly verbal ones; () under phrases indicates 'implicit' entry, i.e. in the examples)

Commas, full stops, colons and semi-colons co-occur within entries, of which I have accepted everything except for commas as sense dividers. Kersey, for example, uses a full stop in the entry for *eye* and a colon in *face* to distinguish

different senses. It has to be admitted that excluding commas as dividers may lead to awkward results at times, as illustrated by a comparison by one of Bailey's entries for *face* and Martin's:

(1) a. **Bailey**: FACE, visage, countenance, presence, appearance, shew**;** state of affairs, condition, *etc.*
b. **Martin**: FACE 1 countenance, visage, looks 2 condition of affairs 3 presence or sight 4 exterior part of a building 5 appearance, outside 6 confidence, assurance 7 grimace, wry-face

Bailey's five terms preceding the semi-colon correspond to Martin's senses 1, 3, 5 and potentially also 6, but count as one sense for Bailey in the present study. Note also the 'etc.' at the end of his entry, which may or may not point to specific further senses. Given this counting, one can see that often in the early dictionaries, especially Bailey, the number of entries and senses coincide. In other instances the difference might be only slight, e.g. *foot, face*. It is only in Martin and in particular Johnson that a higher number of senses based on few entries are listed. The difference with regard to senses is especially striking with regard to *head* in Johnson.

It is possible to make a distinction of senses into more general ones and more specialised ones, in order to investigate the distribution in the dictionaries. General senses are common, everyday uses of a word; they are often characterised by a fairly wide or vague semantic range and by a fairly unrestricted referential application. Specialised senses, on the other hand, have a narrower semantic range and are restricted to specific registers. While both groups can contain literal and transferred senses, the latter may be more common in the second group. This group of senses can also receive lexicographical marking, whether explicitly by a diatechnical label or implicitly by a specific phrasing of the definition. The following senses for *leg* can illustrate the distinction:

(2)

a. general senses:	**b. specialised senses:**
1 The limb by which we walk, particularly the part between the knee and the foot.	LEGS [in *Trigonometry*] the two Sides of a right angled Triangle, when the third is taken for the Base.
2 An act of obeisance.	LEGS [in a *Ship*] small ropes of the Martnets that go thro' the bolt Ropes of the Main and Fore Sail. (Bailey)
3 To stand on his own legs; to support himself.	
4 That by which anything is supported on the ground: as the *leg* of a table. (Johnson)	

In (2b) we see Bailey's explicit labelling by means of indicating in square brackets the technical/scientific etc. field or walk of life that the term belongs to; in Kersey the same purpose is achieved but the field is given in italics without brackets. In some cases, the marking is more implicit, as in 'HEAD *of an Anchor*,

is the Shank or longest Part of it' (Bailey), where *anchor* refers to the nautical register. There are instances where the dictionaries do not agree in their (non-)labelling of senses:

(3) a. [in Mechanick Arts] the upper parts of inanimate and artificial Bodies, as the Head of a Nail (Bailey, s.v. *head*)
b. The top of any thing bigger than the rest. (Johnson, s.v. *head*)

Bailey marks this meaning as 'technical' whereas Johnson does not (his examples show that he intends the same meaning as Bailey). While Johnson's classification as a common sense seems more logical here than Bailey's classification, he also leaves (4) unmarked, which is less plausible as the sense is reminiscent of hunting jargon.

(4) State of a deer's horns, by which his age is known. (Johnson, s.v. *head*)

A numerical comparison of such general and specialised senses (as understood here) yields the following picture:

Table 3: Distribution of senses

	Kersey	**Bailey**	**Martin**	**Johnson**[11]
general	11 (39%)	15 (29%)	26 (76%)	90 (96%)
specialised	17 (61%)	37 (71%)	8 (24%)	4 (4%)

This result confirms Osselton's (1995b: 10) statement that the early dictionaries were strong on 'Terms of Art', i.e. technical terms. This also means that they carried on the hard-word tradition in a very specific way, namely by turning partly into encyclopedias of knowledge (cf. Hayashi 1978: 88). The technical tendency is especially strong in Bailey, where it goes hand in hand with a pronounced encyclopedic approach of the dictionary. In contrast, general senses dominate in Martin and Johnson. Table 4 presents a listing of the fields the specialised senses fall into, together with the occurrence in which dictionaries:

Table 4: Fields for specialised senses given in the dictionaries

	Kersey	Bailey	Martin	Johnson
agriculture[12]	1	1		
anatomy		1		
architecture	3	5	2	1
astrology	1	1		
astronomy		1		
botany	1	1	1	1
fine arts & literature	1	2		1
hieroglyphically		3		
horses		1		
hunting				1
jewelry		2		
mechanic arts		1		
medicine		3		
military	7	8	3	
printing		1		
sea-faring	2	4		
tailoring		1	1	
trigonometry	1	1	1	

The most common fields here are architecture, e.g. *eye* as 'the middle of the scroll of the Ionic capital', and military affairs, e.g. *face* of a bastion (fortification), both found in Bailey and Kersey. For both of the examples given here, one might argue that they are not necessary for a general dictionary. What these and the other examples given above (2b)–(4) also show is that the specialised senses are usually transferred ones, i.e. the more specialised senses a dictionary provides the greater will be the amount of non-literal senses given.

In contrast to the labels for specialised registers, not a single label referring to the figurative use of an item has been found within the entries investigated. This may partly be due to the lack of an appropriate metalanguage. However, terms like metaphor, metonymy, personification etc. were familiar from rhetoric. Metaphor could have been used as a label for Martin's *face* 4 (1b) and Johnson's *leg* 4 (2a), for instance, while *face* 7 (Martin) and *leg* 2 (Johnson) are metonymies. What is noticeably also lacking are usage labels of the prescriptive kind: there are no indications in these entries of good or bad usage.

The next aspect to be treated concerns the ordering of senses. The first – and probably unsurprising – thing to notice is that the sense 'physical body part' is always the first one to be given in these entries. Furthermore, unlabelled senses usually precede diatechnically marked senses, as in (5), where Kersey starts with three general meanings (separated by semi-colons), followed by two architectural ones and finally an astrological sense (separated by colons):

(5) FACE, Visage, Looks; State or Condition of Affairs; Appearance : In *Architecture*, a flat Member that has a great Breadth and small Projecture or Jutting out : Also the Front or outward Part of a great Building : In *Astrology*, the third part of every Sign. (Kersey)

While in (5), general and specialised senses are combined in one entry, the latter usually get separate entries in Bailey, sometimes also in Martin. In Kersey, two further specialised (military) senses of *face* follow (5) as entries in their own right, leaving it unclear why the senses get different treatment. In order to maintain the sequence general-specialised Bailey even separates different verbal uses. He first lists the nominal general senses, followed by the entry with the verb in its general meanings (e.g. *to look toward*), followed in turn by six specialised nominal senses, and then produces again a verbal entry, marked as [in *Military Affairs*] and defined 'turn the face and whole body according to the word of command'. This second verbal use does not immediately follow an entry with a military sense, but one labelled as [in *Astrology*], and precedes a specialised nominal sense from masonry. This leads to the fact that very often there seems to be no logical system of ordering the sequence of specialised senses/entries, cf. the following extract from Bailey's section on *eye*:

(6) EYE (literal sense)
EYE [with *Architects*] the middle of the scroll of the *Ionic* capital, cut in the form of a little rose; also any round window made in a pediment, an *Attic*, the reius of a vault, &c.
(… 10 entries …)
Bullock's EYE [*Architect.*] a little sky-light in the covering or roof, intended to illuminate a granary or the like.
(… 4 entries …)
EYE *of a Volute* [*Architect.*] the centre of the volute, or that point where the Helix or spiral, of which it is formed, commences, … (Bailey)

The first literal or general entry is followed by a meaning from architecture, a field which recurs several times for *eye*, but is spread out across the page and interrupted by the listing of various other specialised senses.

For Martin and Johnson it is also interesting to look at the ordering of general senses, as they provide a sufficient number of them. The entry for the noun *face* is to serve as an example here, for which Martin provides seven, and Johnson nine general meanings, presented in (7).

(7)

a. Martin:	**b. Johnson:**
1. the countenance, visage, looks.	1. The visage.
2. the condition of affairs.	2. Countenance; cast of the features; look; air of the face.
3. presence, or sight.	3. The surface of any thing.
4. the exterior part of a building.	4. The front or forepart of any thing.
5. appearance, outside.	5. State of affairs.
6. confidence, or assurance.	6. Appearance; resemblance.
7. grimace, or a wry face.	7. Presence; sight.
(+ four specialised entries)	8. Confidence; boldness.
	9. Distortion of the face.

There is some agreement between them. Both start with the same (two) senses, as Martin's 1 corresponds to Johnson's 1 and 2, and both end with the same two senses (6=8, 7=9); senses 4 are also identical, as front of *house* appears in Johnson's illustrations. Disagreement in the sequencing is found with regard to Martin's 2 and 3 versus Johnson's 5 and 7. If we take Martin at his word, as stated in his preface, the original/etymological meaning (here 1) should be followed first by general and popular senses, then by figurative/metaphorical ones. Martin might thus have seen 2 and 3 as more general, common meanings than those following. 2 presents a figurative, abstract meaning (his third rank), while 3 is more metonymical in nature, thus from his perspective probably less figurative. It is noticeable in both dictionaries that more metaphorical, more abstract senses often precede metonymic senses (e.g. Martin's 2 and 4 before 7), thus perhaps reacting to different frequencies and thus commonness of those senses.

As to phrases, it is clearly visible in Table 2 above that they are neglected in these dictionaries. We find seven phrases explicitly listed, one for *face* (Johnson) and six for *foot* (Kersey, Bailey; Johnson); compare footnote 1 for the higher instances found in a modern dictionary. Interestingly, Martin, who mentions phraseology explicitly in his front matter, does not include a single phrasal unit in the entries examined here. Explicit listing means that the phrase is listed as a headword with its own entry, or as a numbered sub-entry as in the following examples:

(8) a. To GAIN or LOSE GROUND FOOT BY FOOT, is to do it regularly and resolutely, defending every Post to the utmost, or forcing it by dint of Art and Industry. (Kersey)

b. *To be on the same* FOOT *with another*, is to be under the same circumstances. (Bailey)

c. 6. *On* FOOT. Walking; without carriage.
Israel journeyed about six hundred thousand *on foot*. Ex. xii. (Johnson)

In Kersey bold-face and use of a special font mark entry status (8a contains an approximation), while in Bailey and Johnson this is indicated by smallcaps. The rest of the phrase is italicized, a procedure Bailey also uses for collocations and compounds, e.g. A FOOT *bank*. These cases represent what Pinnavaia (2006: 156) calls a conscious treatment of phraseology, while the following is rather unconscious, as it shows no explicit awareness on the part of the lexicographer that he is dealing with a phrasal unit. Those phrasal occurrences indicated in brackets in Table 2 are thus not presented as phrasal (sub-)headwords, but dominate the relevant part of the entry, as in (9), where all the examples given by Johnson contain the phrase *make head against*.

(9) Resistance; hostile opposition.
Then **made** he **head against** his enemies … Spenser.
… Bolingbroke **made head against** my power. Shakespeare.
Two valiant gentlemen first **making head against** them … Raleigh's Apology.
… by which he can **make head against** it. South. (Johnson, s.v. *head*, sense 11, examples shortened)

This phrase is actually the only instance where *head* has the meaning 'resistance' and thus Johnson actually defined the phrasal meaning, but without saying so. There are quite a few of such hidden phrase entries in Johnson, which means that Johnson provides more senses than the word taken in isolation warrants. The senses 7, 9, 11 and 14 of *foot*, for instance, represent in fact phrases and their meanings; the same goes for senses 7 and 25 of *head*. Additionally, more phrases are found prominently as illustrations under various senses. Some examples, together with the sense under which they are listed are found in (10):

(10) a. we **are** not **upon the same foot** with our fellow subjects in England 'state, character, condition' (s.v. *foot*)
b. if such a tradition were at any time **set on foot** 'state of incipient existence' (s.v. *foot*)
c. while other jests are something rank **on foot** 'motion, action' (s.v. *foot*)
d. let it **lie on my head** 'person as exposed to any danger or penalty' (s.v. *head*)
e. with the duke of Marlborough **at the head** of them 'place of command' (s.v. *head*)
f. the indisposition .. is at last **grown to** such **a head** 'crisis, pitch' (s.v. *head*) (Johnson)

Note that both Kersey and Bailey gave the phrase found in (10a) lemma status and provided a definition. Interestingly, Johnson does not list any phrasal verbs for these items, although *face* is found with such uses at this time. Pinnavaia's (2006) investigation of idioms related to food and drink in Johnson confirms the low incidence of phrasal items. Her search, which was based on 225 different

items, yielded only thirty-four occurrences of idiomatic expressions. While she tentatively assumes that there may simply have been fewer idiomatic expressions in use and thus available for inclusion (*ibid*. 159), it is much more likely that the lexicographers were not so aware of their presence, especially when they were not strictly speaking word-like, such as phrasal verbs, or were too inconspicuous to be noticed (e.g. *from head to foot*), and/or did not see the dictionary as the right place for them.

5. The corpus evidence in comparison to the dictionaries

The corpora were searched for all possible word forms and spellings of the lexemes under consideration (the latter especially important in case of the CEEC), so that, for example, plural and verbal past tense uses were also found. Occurrences of the items within proper names, notably *head* in pub names (frequent in ZEN), were deleted from the results. It turned out that, as assumed, the frequency of the items was high enough to make a semantic classification feasible. An important first step consisted in dividing the strictly literal sense, i.e. the physical body part meaning, from all other uses. The results are presented in Table 5. In the case of *head, face, eye* and *leg* the basic body part sense is indeed the most frequently used sense overall and also in the individual corpora taken separately. The only exception to this pattern is *head* in the CEEC. Surprisingly, the overall situation is reversed in the case of *foot*, where the transferred senses dominate in all three corpora, most strikingly in ZEN. As noted above, the body part sense is listed first in all of the four dictionaries, which is borne out not only by its basicness, but also, in four of the five cases, by the frequency of this sense in actual usage.

Table 5: Literal *vs*. transferred uses

		CED		CEEC		ZEN		**Total**	
			%		%		%		%
head	literal	393	65.5	43	42	648	71	1,084	67
	transf.	207	34.5	60	58	264	29	531	33
face	literal	276	75	24	71	150	59	450	69
	transf.	91	25	10	29	105	41	206	31
eye	literal	273	77	45	58	131	69	449	72
	transf.	80	23	33	42	59	31	172	28
leg	literal	126	95	17	81	135	100	278	96.5
	transf.	6	5	4	19	-	-	10	3.5
foot	literal	90	46	17	22	92	14.5	199	22
	transf.	106	54	59	78	541	85.5	706	78

Verbal occurrences here are by definition transferred, as they can only metonymically relate to the body part sense. As we have seen, they are not

consistently listed by the dictionaries. *Leg*, which is given only as noun by all four dictionaries, has no verbal occurrences in the corpora. Neither has *eye*, which is listed as a verb solely by Johnson. According to the *OED* both items existed as verbs at the time in question, but were perhaps of too low frequency to come to the attention of the lexicographers. Verbal *foot*, which is listed by Bailey and Johnson, is found twice in the corpora. *Head* with 24 verbal occurrences is again only provided by Johnson, while *face*, which has 69 corpus instances, was apparently frequent enough to be noticed by all four dictionary makers.

The next step in the analysis is the sense differentiation of the non-literal occurrences. This was carried out independently from the dictionary data, by attempting to give as precise meaning paraphrases as possible for each occurrence and then by sorting them into semantic groups, yielding the corpus 'senses' to be used below. These transferred usages combine meanings of the item as such, of the item in particular collocations and of the item in fixed phrases and idioms. The degree of fixity and/or idiomaticity will in some cases reach such an extent that it will only make sense to list the whole phrase as an item with a particular meaning. In the following I will go through the individual items, describing their corpus use and comparing it to the picture presented by the four dictionaries.

5.1 *Head*

Head as a verb occurs in the senses 'lead' (20 instances) and 'behead' (4), both of which are listed by Johnson (in the same order), who adds two more senses not found in the corpora used here. Nine nominal senses occur more than once, while the category 'other' in the following diagram covers a variety of rare, and sometimes not quite clear, figurative uses. The occurrence of non-basic senses of nominal *head* in the three corpora is illustrated in Figure 1 (p. 148).

The senses are not equally distributed across the corpora. The most prominent sense in the CED and the CEEC is clearly the 'mind' reading (e.g. *he had bad designs in his head*), while ZEN favours the 'topic', 'in command' (e.g. *at the head of the army*), 'leader' (e.g. *supreme head of the church*) and 'main institution' (e.g. *head house*) readings. None of these meanings are found in Kersey, and only one, the upper part 'of object' meaning is listed by Bailey, who treats it as a specialised sense. Otherwise none of the specialised senses provided by Kersey and Bailey are attested in the present corpus evidence; as many of their senses are of a military nature the ZEN, which covers such topics, could in principle have provided evidence. The fact that it does not, and that the Helsinki Corpus also yields none of these specialised meanings, may be taken as an indication of the rareness and register-restrictedness of such senses. Martin's senses are somewhat difficult to compare, as his paraphrases tend to be very brief (cf. *face* in (7) above) and helpful illustrative examples are lacking. His sense 3 'the front or forepart, as of an army etc' corresponds to the 'in command' sense, while his sense 4 'chief or principal' can correspond to both the 'leader' and 'main institution' senses listed here. The 'of object' sense is defined by him

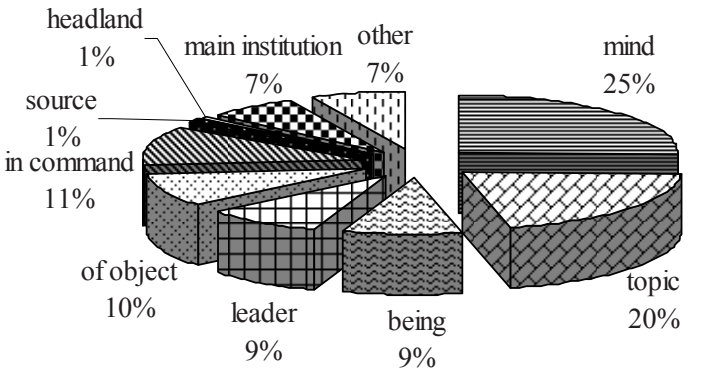


Figure 1: Non-basic senses of nominal *head* in the three corpora

(sense 2), as is the 'source' sense (5, e.g. *head of the river*). The two most frequent corpus senses are curiously not mentioned by him at all. Unsurprisingly, Johnson is more thorough, covering everything except for the 'headland' sense. The 'mind' sense as classified here corresponds to both his sense 9 (understanding, faculties of the mind) and 21 (the brain), while the 'topic' sense is unequivocally his 23 (principal topics of discourse). The sense 'being', i.e. *head* metonymically standing for the whole human or animal being can potentially be identified with several of Johnson's senses, namely 2 (person as exposed to danger), 4 (counting of animals), 8 (presence of beings), and 14 (counting of people). One also finds defined senses which cover instances from the 'other' group, such as the 'intention' sense. Johnson thus can be said to define everything of any note and frequency, Martin to a lesser extent, while Bailey and Kersey are on the whole insufficient in their coverage of senses. Bailey and Kersey of course offer specialised senses, which may not have been very common, but Johnson also provides additional general senses not found in the corpus material, namely his senses 6, 13, 18, 19, 26, 27, 28 and 29. Some of his senses may have been obsolete already at this time, such as perhaps 28 (power, armed force), which he attested only by Shakespearean examples. For the others a more comprehensive corpus search is needed – though of course the number of available corpora is limited.[13]

In addition to the meanings above, some which are already collocationally fairly fixed, there are also phrases which are even more fixed and idiomatic in nature. Most of them occur more than once. The highest number is found in the CED (11 types / 22 tokens), followed by CEEC (7 / 7) and by the ZEN (4 / 6). In order of frequency the phrases found are: *from head to foot* (5), *be upon sb's head*

(4), *bring* (etc.) *to a head, hold* (etc.) *head against sb, hit the nail on the head, over head and ears, bring* (etc.) *the house over our heads* (3), *bring* (etc.) *sth over/upon one's head, by the head and shoulders, head and heels, hand over head* (2), and *beat into sb's head, draw* (etc.) *upon one's own head, head or tail* (1). Of those we find only two listed explicitly in the dictionaries, both by Johnson: *head and ears* (sense 3), *head and shoulders* (sense 31). Johnson furthermore defined a verb followed by *head against sb* but without identifying it as a phrase, cf. (9) above. He also provided indirect evidence for other idiomatic uses, namely *let it lie on my head* (under sense 2, corresponding to *be upon sb's head* above?) and *grow to such a head* (under sense 25). None of the other three lexicographers provided any of the above or other phrases.

5.2 *Face*

Moving on to *face*, one finds seven more or less different verbal uses, namely the (i) spatial situation of objects vis-à-vis each other, (ii) people turning and looking in a direction, (iii) people involved in confrontation, (iv) covering the façade of a building, and (v) the three phrasal verbs *face out*, *down* and *about*. Of those Kersey, Bailey and Martin have (ii), who additionally offer 'stare somebody in the face' (all three) and 'cover sleeves of a garment' (Bailey, Martin). Johnson gives senses (i)–(iv), thus covering with (i) the most frequent corpus meaning (particularly in ZEN), and also offers some more of his own. None of the dictionaries lists the three phrasal verbs – which is especially surprising for Johnson, as he is fairly comprehensive for other verbs in this respect.

Figure 2 presents the nominal transferred senses found in the corpora; it is followed by an illustration of the more common uses.

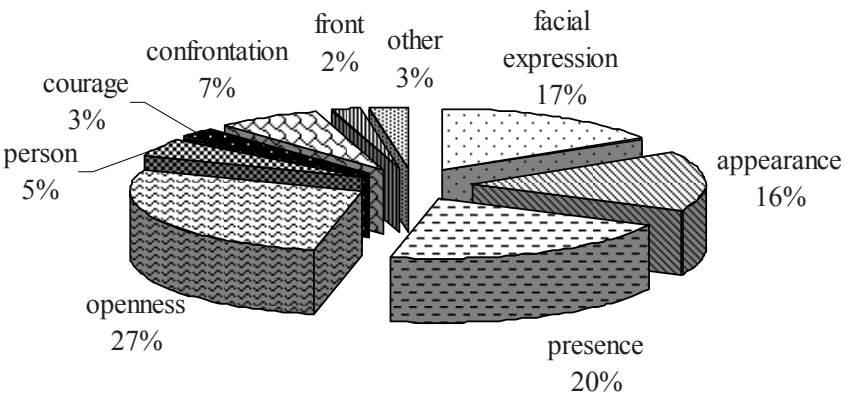


Figure 2: Non-basic senses of nominal *face* in the three corpora

(11) a. facial expression: I must make an angry **face** outwardly, though I smile inwardly. (CED d1cchapm)
b. appearance, condition etc.: The French, at Brusels, seem'd at first exceedingly joyful upon the Death of the King of England, thinking that might change the **face** of Affairs, … (ZEN, *Postboy*, 1701)
c. presence: if he would stand in the Kings **Face** (CED d4hosam)
d. openness, frank communication: he call'd me Cuckold to my **face** (CED d3cdryde)
e. confrontation: she laugh'd violently … and drew the Curtain in my **face** (CED d5cgarri)

Most of these senses are actually present in the dictionaries, cf. Bailey's and Martin's general definitions quoted in (1) above. Kersey is the least detailed on these general senses, missing such common meanings as 'facial expression' and 'presence'. Johnson and Martin both cover the senses 'facial expression', 'appearance', 'presence', 'courage', 'confrontation', and Johnson additionally 'front'. The relative frequencies found here do not bear out the sequencing of senses as provided by Martin and Johnson. The sense I have termed 'openness' here might be covered by the 'presence' meaning in dictionaries, though it is better kept apart from it. It remains unclear whether Johnson and Martin had this nuance in mind without spelling it out explicitly. It is a usage that is particularly common in the CED. The metonymic sense 'person' (as in *I do not see so many Faces as are mentioned in that Act.* – CED d3tsling) is also not found in the dictionaries. It is not overly frequent and quite a number of its representatives are insults (*Do filthy Face, do if thou darst*. – CED d4cshadw), i.e. vulgar language, which may account for the exclusion. Two further dictionary senses, the 'exterior part of a building' (Martin) and 'surface of any thing' (Johnson) do not correspond perfectly with the corpus evidence either, as there it is always the front part (not simply any outer part) that is indicated. The specialised, technical senses provided by Kersey, Bailey and, to a lesser extent, Martin are again not found in the corpora. Of course, this is largely due to the nature of the corpora used here, which do not encourage the occurrence of such uses. However, a search of the Helsinki Corpus (EModE section), which has a wider range of texts including more formal, 'academic' ones, also yields none of specialised senses among its 122 instances of *face*.

Fixed expressions and idioms including *face* are found in the CED (4 types / 13 tokens) and ZEN (4 / 12), but interestingly not in the CEEC. The most common phrase with fourteen instances, *face to face*, is also listed explicitly in Johnson (main entry), but in none of the others. The other corpus phrases found are *the face of the earth* (6), *fly in the face of* (3), *set one's face against* (1) and *Janus face* (1). None of them is found explicitly or implicitly in the dictionaries. Bailey, who is the one to pay most attention to the facts of classical civilization, explains the two faces of Janus under this entry, but without giving any indication that one can use this phrase in English as an established formula.

5.3 *Eye*

The agreement between the dictionaries and the corpora with regard to *eye* is on the whole rather small. Verbal *eye*, listed only by Johnson, is found only twice (both CED), which may account for the oversight of the other lexicographers. None of the specialised nominal senses given by Kersey, Bailey and Martin (from the fields of architecture, botany etc.) are attested in the corpora; that is also true for the Helsinki Corpus. What is even more striking is that there is also no match with the more general senses provided by Martin, either. These are 'loop, or small hole' and, in the plural, 'spectacles', both also found for the relevant period in the *OED*. The senses found here are shown in Figure 3.

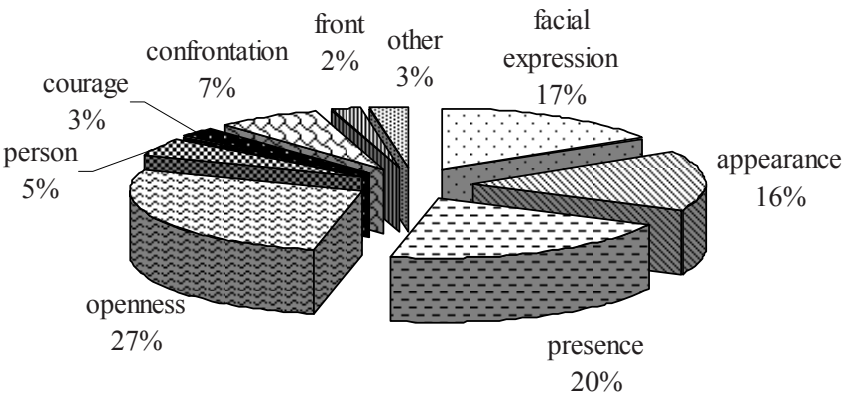


Figure 3: Non-basic senses of nominal *eye* in the three corpora

Johnson's fifteen senses overlap with the corpus evidence only in three cases (beyond the literal meaning). His sense 2 'sight, ocular knowledge' is also the most frequently attested corpus sense. His senses 7 'notice, attention, observation' and 8 'opinion formed by observation' correspond to attention and opinion in Figure 3, which occur with some frequency. He does not give the sense where *eye* metonymically stands for person (with seeing highlighted, of course), as in *attract the admiration of many eyes*, which is more frequent in the corpora than the two senses previously mentioned. There is also none of his senses that is a good match for the 'understanding'-reading, as in *I will open your eyes on the unhappy business*. Eleven of Johnson's senses are not found in the corpora, some of which might simply be rare (e.g. 14 'shade of colour') while some might have been obsolete or obsolescent, e.g. senses 3, 4, 5, and 9, all of which are illustrated by only one example from Shakespeare (3, 4, 9) or Dryden (5).

Many of the senses of *eye* are collocationally fairly fixed, such as *have an eye on/to* (observe, intent etc.) and *in the eyes of* (opinion). Most phrasal uses are clearly linked to the basic senses of eye, such as *the naked eye, eagle-eyed* or *believe one's eyes*, which all work with the 'sight' sense. In one case we find an almost proverbial case, based on a scriptural passage:

(12) If men in general would take *the beam out of their own eyes*, they might then perfectly see the *moat in others*. (ZEN, *The Middlesex Journal*, 1771)

The dictionaries again list none of the phrasal uses explicitly. Johnson's illustrations of course contain *have an eye* and *in the eyes of* under the relevant senses, but no further phrases are hidden in the remainder of the entry. Nor are any found in the other dictionaries. Again, the omission of (12) in Bailey (not found under *beam* or *moat* either) may seem surprising given his cultural leanings, but the *Dictionarium Britannicum* in contrast to his other dictionaries is notable for its exclusion of proverbs (as which he might have classified the item).

5.4 *Foot*

While *foot* is the most frequently attested item of those treated here, it is semantically less diversified than the preceding lexemes. Furthermore, the overlap between lexicographical and corpus evidence here is greater than in all other cases, pointing to the possibility that these senses were somewhat more salient for contemporary observers. It can certainly be argued that the degree to which the senses are lexicalized is greater than in most other cases treated here, e.g. in the case of *foot* as a measurement unit, as a metrical term in literature and as a term for infantry. These three and two more senses, 'foot of an object' and the 'end' of something (e.g. a page), are given together with their proportion in Figure 4 (p. 153).

All occurring senses are found in the dictionaries. Johnson has all of them; Martin has three, lacking 'end' and 'poetry'; Kersey and Bailey have 'measure', 'poetry' and 'of object', treating them partly as specialised entries (with a label). It is interesting that the latter two lexicographers, who so often list military senses, do not give the 'infantry' sense, which is the one most frequently found. Neither of them lists *footmen* or *foot soldier* separately, so that this meaning is completely lacking in their dictionaries. The frequency of this sense is certainly due to the prominence of military reporting in ZEN; with a different selection of sources the 'measurement' sense might come in first, which is here found in second place. 'Infantry' is given as sense 4 and 8 in Martin and Johnson, respectively, while 'measure' is surprisingly the second to last of all of Johnson's meanings (sense 15). Johnson again has a number of senses not found in the corpora, many of which seem to be fairly idiomatic in nature. The verbal use, found twice in CED, is given by Johnson and by Bailey.

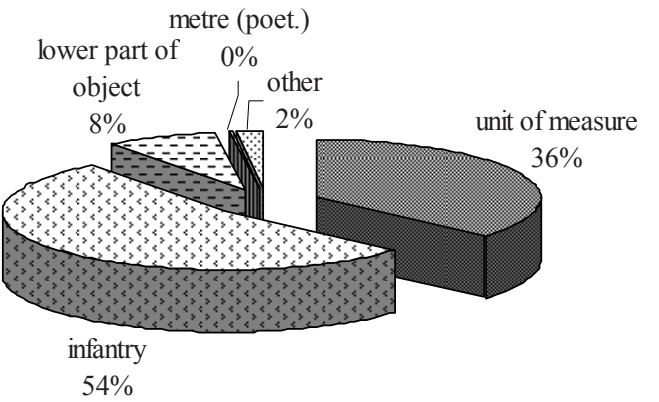


Figure 4: Non-basic senses of nominal *foot* in the three corpora[14]

In contrast to the few basic meanings, phrases are fairly common with *foot*. 164 tokens were found, distributed among fourteen types. The greatest number of types occurs with twelve in CED (62 tokens), while ZEN yields most tokens (91, nine types). The phrases in order of frequency are: *on foot* / *a foot* 'in progress' (42 tokens), *on foot* 'walking, not driving' (32), *on a* [adjective] *foot* 'manner' (30), *on foot* 'ready' (15), *at sb's feet* (14), *set a/one's foot* (10), *from head to foot* (5), *hand or foot, 'sfoot* (4), *on the foot of* 'reason' (3), *tread* etc. *under foot* (2), *by the foot* 'closely', *foot by foot*, and *foot in the grave* (1). Three of those are explicitly in the dictionaries: *on foot* 'walking' by Johnson (sense 6), *foot by foot* and *on the same foot* (cf. *on a* [adj] *foot*) by Bailey and Kersey, though by the latter only with a clearly military meaning. In Johnson's entry one can furthermore find hidden away *on the same foot* (under sense 9), *set on foot* (sense 11), and *on foot* 'progress' (sense 14). Martin does not offer any phrasal uses.

5.5 *Leg*

Leg is the least diverse item of the five investigated. Neither in the dictionaries nor in the corpora are there any verbal uses. In the overwhelming majority of cases the noun occurs in its most literal sense, in ZEN exclusively so. Only ten occurrences are transferred, representing four distinct usages. One is the unsurprising but here rare 'leg of an object', such as of a chair, while the other three are more phrasal / idiomatic in nature. The most frequent meaning is a time- and culture-bound one, namely that of 'obeisance', as found in (13 a/b).

(13) a. The present came, which Iack seeing, **made legs to** the gentlewoman, (CED, d2farmin)
b. and parted in great Anger with the Usuall Ceremony of a **Leg** and a Courtesy, that you would have dyed w=th= Laughing to have seen us. (CEEC, osborne)
c. It was well he could undertake such a journey from that place where they affect to have men **preach** themselves **off of their very leggs**, yet they themselves stand where they did. (CEEC, duppa)
d. Upon my Soul, I pity the poor Creature! -- She is now **upon her last Legs**. (CED, d5cgarri)

Johnson is the only lexicographer to list the 'leg of object' and 'obeisance' meanings. (13 c/d) are phrasal uses found, both of which have a rather colloquial flavour, which may account for their absence in dictionaries. The specialised meanings from the fields of trigonometry and sea-faring (Kersey, Bailey, Martin) are in turn not found in the corpora, nor are they present in the Helsinki Corpus.[15]

5.6 *Johnson* vs. *Johnson*

As stated above, Johnson's dictionary can also be compared to his own linguistic usage, in order to see whether there was any noticeable influence on, say, the ordering of senses. This is not only of 'personal' interest, but also relevant from the point of view that his own writing was certainly much closer to his stylistic ideal (as voiced in the preface to the dictionary) than the corpus texts used above. If we look at the distribution of body-part sense versus transferred meanings, we find a difference to the situation as presented in Table 5 above. With Johnson, only one item, *leg*, shows a preponderance of the literal meaning (with in fact 12 to 0 occurrences). In all other cases, he used the words more in their transferred senses, especially clearly so with *eye* (203 transferred *vs*. 55 body part meaning). Nevertheless, he always lists the body-part sense first in the dictionary.

Let us now look at the senses he uses and compare them with his dictionary senses.

Table 6: *Head* as used and defined by Johnson

Sense	Text occurrences	Dictionary – no. of sense
brain, mind, understanding	15	9, 21
person / being	8	2, 4, 14
head / front of object	5	15, 17, 20
top person / leader	4	5
be at head/front of sth./in command	4	7
parts of writing / topics	1	23

In spite of the fact that the most common corpus use of *head* is also his own most frequent sense, Johnson did not give it a more prominent listing. One might therefore be led to say that there was no biasing influence of his own usage. However, the second most frequent corpus sense ('topic') being his least used sense and being relegated to sense 23 in his dictionary entry might speak for the opposite conclusion. What evidence do the other items provide?

Eye is the term most frequently used by him of those investigated. The metonymic 'sight' reading is his most common transferred use, in agreement with the corpora, and is listed in second place in the dictionary. The difference to the treatment of *head* could be that in the case of *eye* we find the simpler metonymy with the common pattern 'instrument > action, process', while 'mind, understanding' for *head* implies 'container > contained = instrument > process'. Thus, not so much frequency of use but semantic/cognitive 'closeness' of a sense to the original body part meaning may have influenced the sequencing of senses in this instance.

Table 7: *Eye* as used and defined by Johnson

Senses	Text occurrences	Dictionary – no. of sense
sight, seeing, looking	93	2
person	33	-
observation, watching	28	7
opinion, view	14	8
understanding	3	15 (?)

The meaning 'person' (e.g. *he thinks each eye surveys him with contempt,* Rambler), also metonymic though of a different kind (part > whole), is Johnson's next frequent meaning in use, which he did not list in the dictionary at all – once more in contrast to the entry for *head*. And again, one can argue that ease and salience of semantic connection is stronger for the *head* than for the *eye* metonymy: head is the bigger and more prominent body part of the two and the metonymy works more directly (whereas *eye* needs the prominent presence of 'perception': instrument > seeing > actor/experiencer = person). *Foot*, which also marginally occurs in Johnson's writings for person (in a non-military sense) and is also not listed as such in his dictionary, is in its degree of salience clearly similar to *eye*, not to *head*. *Foot* 'person' also makes a fairly ad-hoc impression, akin to the *hamburger* = 'customer in a restaurant' mentioned in Section 2, a meaning modern dictionaries would not list either. The precise type of semantic process involved in a given sense may thus play an important role for the degree of metalinguistic awareness as seen in lexicography. *Face* 'person' proves somewhat of an exception to this conclusion. It occurs in the corpora, was used by Johnson himself and is certainly closer to *head* than to *foot* and *eye* in this respect – but nevertheless it is not a defined sense in the dictionary.

As to the other senses of *face* Johnson used in his writings, these are 'appearance' (sense 6 in the dictionary), 'facial expression' (sense 9) and

'presence' (sense 7), in their order of frequency. *Foot* is mostly used by Johnson in its 'measurement' meaning, which is listed as late as sense 15, however. His less frequently used senses 'lower part of an object' and 'walking' are listed earlier, being represented by senses 2, 3 and 6, respectively. As indicated above, *leg* is actually only used in its basic sense in the Johnsonian texts. However, there are two occurrences of *leg of pork* (i.e. a particular cut of meat), which are of interest, as they do not occur in the corpora but can be assumed to represent a fairly common usage. This meaning is not listed separately by Johnson in the dictionary, although it is in some modern dictionaries.

On the whole, I think that one can conclude from these five entries that Johnson was not particularly biased by his own usage of the words in the writing of the dictionary. On the one hand, he did not include some senses he himself used and he did apparently not privilege his own frequent senses. On the other hand, he included a considerable number of senses which do not find a correlate in the works of his investigated here.

6. Discussion and conclusion

Let me summarize and discuss the findings. On the whole there is a not inconsiderable overlap between the dictionaries and the corpora with regard to different senses of words. However, this applies much more to Martin and Johnson than to Kersey and Bailey. Johnson is generally in better agreement with the corpus data than Martin, which can certainly be accounted for by his use of a corpus of attestations (whereas Martin's precise manner of working on the dictionary is not known). In some cases, his sources might have led Johnson to posit more sub-senses than necessary, cf. the comparison above of *head* in the dictionary *vs*. the corpora. The degree of overlap can vary greatly with items studied, as has been shown for *eye* (bad) vs. *foot* (good). This means that for a general assessment of the quality of the dictionaries, more entries need to be included. It also raises the question why the lexicographers reacted so differently to these two items: why is it 'easier' to describe some items than others. Salience of sense (differentiation) has been used above as an explanation, but this point would need more research.

While both Kersey's and Bailey's works have been described, and partly rightly so, as great improvements in the course of English lexicography, on the evidence given here they cannot truly be called *general* dictionaries of the English language. While they do include the general, common words of the language, they do not necessarily list their common meanings. Kersey and Bailey lack, for example, verbal *head*, the most common meanings of nominal *head*, basic meanings of *face*, as well as fairly common meanings of *eye* and of *leg*. Neither of those two can thus be called representative of English usage as such. What is surprising in this respect is the great economic success Bailey's dictionaries enjoyed during the eighteenth century. The attitude of the buying public seems to have favoured the fact that, in Hayashi's (1978: 86) words, in this

work 'a dictionary of words cooperates with a dictionary of encyclopaedic articles', with a certain emphasis on the latter. Both Bailey's and Kersey's dictionary reflect and cater for an attitude to language that puts greater value on the specialized word and sense than on the common one.

On the other hand, none of the corpora used here, including the Helsinki Corpus, provide any information on these specialised word uses. Also some general senses of Johnson and Martin have not been corpus-attested here, for example *eye* 'spectacles' provided by Martin but not by Johnson. Did the latter not include it because he did not find it attested?[16] It is of course well known that lexical and semantic phenomena need to be studied with the help of larger and diversified corpora than frequently occurring morphosyntactic features. As historical corpora are – and will be for the foreseeable future – fairly small and as the register/genre/text type coverage in them is not fully comprehensive,[17] contemporary dictionaries could be seen as useful in filling the gaps in our lexico-semantic knowledge of Early Modern English. Even if one used all possibly available corpora together it is questionable whether one would find instances of, say, the architectural and military senses listed by Bailey and Kersey, as texts of this nature are usually not included. This leaves us with the question of how far we can trust the dictionary evidence – 'ghost words' are known to have been included in early dictionaries, which usually means seventeenth-century ones, but Kersey and Bailey need not be immune to this procedure. In the absence of other sources, we can countercheck their entries against the *OED*'s. Bailey's specialised senses of *face*, for example, are mostly found there as well: 'façade' and *face of stone* cf. *OED* s.v. *face* sense 12b, in astronomy cf. *OED* 11c, *face* in fortification cf. *OED* 17a. His sense *face of a gun* also appears (*OED* 19), but interestingly attested only by Bailey's own entry and by one other source named 'Symth Sailor's Word-bk (1867)' – perhaps a case of a really doubtful sense? A further point to mention is that even if a dictionary reliably proves the existence of a word and a meaning, we still only know part of the story – the missing elements are the word's use in context and its frequency. Larger corpora would also come in handy in checking on whether and when senses have become obsolete or obsolescent even though still listed in dictionaries (due to the sources used by them), for examples some of the senses of *eye* listed by Johnson.

Despite their fairly small size, the corpora used here have produced a greater range of collocations, fixed expressions and idioms than given by the dictionaries. This proves that even small corpora can be useful in studying phraseology, an area that is still very much underresearched in historical linguistics. This result also confirms Knappe's (2004) assessment that early dictionaries are weak in this area. The question is why this is so. One reason might have to do with linguistic attitudes: the embracing of a certain stylistic ideal (visible for example in Johnson's selection of sources) and a certain degree of prescriptivism. Many idiomatic expressions tend to be fairly colloquial, informal and can potentially be perceived as clichés. Their exclusion could thus be seen as a conscious stylistic decision on the part of the lexicographers. Not much is known about Kersey's and Bailey's attitude in this area, except for that

the latter did include many proverbs in his other dictionaries; he seems not to have been put off by everyday language and clichés in general. Given what Martin and Johnson say in their prefaces, a deliberate demotion of phrasal evidence also seems unlikely. Other reasons are the lack of awareness of phrasal units and the lack of sufficient data, which are partly connected. As mentioned above, Johnson sometimes treated phrasal uses as special senses of the headword. The attention of early lexicographers seems to have focused very much on the individual word and its individual sense, neglecting the word's environment. Johnson's treatment of phrasal verbs is an exception, if one looks at his entries for phrasal verbs with *take* or *put*, for example. However, as stated above he lists none of the corpus-attested phrasal verbs with *face* – the difference may have to do with the fact that the senses of *take/put* (almost) completely disappear in the phrasal uses, making the output really opaque. A use like *face down* is less striking in its semantic change. Of course *face* plus particle may also have been less frequent than, e.g., *take up*, which leads over to the data question. The more lexicographers worked with other dictionaries, which was a common procedure, the fewer phraseological units they will have come across to include in their own dictionary. Johnson, who worked with a corpus of attestations, would have needed to note down a phrasal expression several times, and to sort these occurrences together, in order to recognize something as an established phrase. In many cases he might not have done that, also because many phrases seem so 'normal' and he himself remarked in the preface how often common, basic sense of words (substitute phrases here) were left unattested in his database.

A last point to be mentioned, namely the question of labelling practice, again relates to the type and degree of metalinguistic awareness present. There is a considerable amount of labelling according to field or register, both implicit and explicit; sometimes there is even 'overzealous' marking as, e.g., when Bailey marks the *head of a nail* with the label '[mechanical arts]'. One can argue that this represents awareness of a pre-linguistic nature, as professions, occupations, and topics are extralinguistic facts of life. The treatment of senses as separate main entries in Kersey, Bailey and sometimes in Martin, may simply reflect extralinguistic classifications, probably not really a linguistic classification as different lexemes. Meta*linguistic* awareness, on the other hand, is present in Martin's and Johnson's numbering of senses within one entry. Neither of them shows any overt awareness of the type of senses listed in these five entries by giving such labels as 'metaphorical' or 'figurative', however. Johnson was well aware of such processes (cf. preface) and he did use labelling elsewhere, as the entry for *to abase* shows: 'To cast down, to depress, to bring low, almost always **in a figurative and personal sense**'. *Figurative* is found one hundred times in the dictionary,[18] which does not seem very frequent, but of course other terms can fulfil the same or similar purpose. Questions are which terms he uses, which items and senses he marks and why. The way Johnson, and as far as possible also Martin, treat figurative meanings, I think, merits further research.

Notes

1 Cf. for example Fernando (1996: 124). The *Longman Dictionary of the English Language* (1984), for instance, lists twenty phrasal units in the main entry for *eye*, fifteen for *foot*, twelve for *head*, ten for *face*, and five for *leg*.

2 Cf. Biber (1993) on representativeness in corpus linguistics.

3 Only the first edition of Johnson's dictionary will be investigated here, in contrast to the fairly common approach of combining the first and fourth (1773) editions in Johnsonian research.

4 I have chosen a sub-corpus of letters written after 1650, which is a fairly arbitrary cut-off point. The corpus files used are listed in the references section.

5 All Johnsonian texts were downloaded from Project Gutenberg.

6 The WordSmith programme was used for this purpose.

7 URL: http://galenet.galegroup.com. Johnson's *Dictionary* is also available from Cambridge University Press (McDermott 1996), which allows full text searching, but not extracting and downloading the search results. It is in principle possible to export the complete text from the programme and then run searches, but as this is not possible with the other dictionaries, the procedure was dispensed with for the sake of comparability.

8 For treatments of the question of literal meaning, cf. Ariel (2002) and Giora (1997).

9 Because of this situation and other problems, Osselton (1995a) is in favour of completely abandoning the label 'figurative' (or similar labels).

10 Drosdowski (1989: 800) quotes from the entry for *Fuß* 'foot': 'sed per metaphoram multis aliis rebus tribuitur, ut […] die Füße an Stülen / Bänken / Tischen / Betten […], Fuß des Berges / radix montis. Fuß einer Seule / spirula, basis.' – i.e. foot of a chair etc, foot of a mountain, foot of a column.

11 Johnson does not have any diatechnical marking in the relevant entries; the four specialised senses are thus my interpretation and include cases like (4).

12 Fields given in italics are those which have not been explicitly labelled, but decided on by me.

13 It would be possible to furthermore use the Lampeter Corpus, the *Century of Prose Corpus* and parts of the ARCHER corpus. A representative corpus of the eighteenth century, which would certainly be useful here, does not exist.

14 Zero percent for poetic metre in the diagram is due to rounding down as done by MS Excel.

15 The Helsinki Corpus, however, yields a specialised, apparently nautical, sense not found in the dictionaries: *we kept our cowrse due sowth stil and passed before the wynd with our mayn yerd a crosse al the way, abowt 30 legs comonly or more in 24 howrs* (HC cediar2a).

16 The *OED* gives this use, s.v. *eye* sense 26b.

17 CED, CEEC and ZEN are by definition special purpose corpora with restricted coverage, while the Helsinki Corpus is a general purpose corpus, but nevertheless cannot be called fully comprehensive. These sources could of course be supplemented by the following: Lampeter Corpus (pamhlets), *Century of Prose Corpus* (literature), *Corpus of Early English Medical Writing*, and *A Representative Corpus of Historical English Registers* (ARCHER). ARCHER at present includes newspaper reportage, journals/diaries, letters, fiction prose, legal opinion, medical writing, (other) science writing, advertisements, drama, fictional conversation, and sermons/homilies.

18 Full-text search carried out with McDermott's CD-Rom edition. Not all hits necessarily represent a semantic label.

References

Primary sources

Dictionaries

Bailey, Nathan (1730), *Dictionarium Britannicum.*
Johnson, Samuel (1755), *A Dictionary of the English Language.*
Kersey, John (1708/15), *Dictionarium Anglo-Britannicum.*
Martin, Benjamin (1749), *Lingua Britannica reformata.*
(all from: *Eighteenth-Century Collections Online*, URL http://galenet.galegroup.com)
McDermott, Anne (ed.) (1996), *Samuel Johnson. A Dictionary of the English Language.* On CD-Rom. Cambridge: Cambridge University Press.

Corpora

CED = *A Corpus of English Dialogues 1560–1760* (2006). Compiled under the supervision of Merja Kytö (Uppsala University) and Jonathan Culpeper (Lancaster University).
CEEC = *Parsed Corpus of Early English Correspondence, text version* (2006). Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Jukka Keränen, Minna Nevala, Arja Nurmi and Minna Palander-Collin, with additional annotation by Ann Taylor. Helsinki: University of Helsinki and York: University of York. Distributed through the Oxford Text Archive. Files used for the sub-corpus: basire, browne, conway, corie, duppa, essex, fleming, haddock, jones, marvell, minette, osborne, pepys, pretty, prideau, tixall.
Helsinki Corpus = *The Helsinki Corpus of English Texts* (1991). Department of English, University of Helsinki. Compiled by Matti Rissanen (Project leader), Merja Kytö (Project secretary); Leena Kahlas-Tarkka, Matti Kilpiö (Old English); Saara Nevanlinna, Irma Taavitsainen (Middle English); Terttu Nevalainen, Helena Raumolin-Brunberg (Early Modern English). See http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/index.html.
ZEN = *Zurich English Newspaper Corpus*. Version 1.0 (2004). Compiled by Udo Fries et al. Englisches Seminar, Universität Zürich.

Samuel Johnson's works

Idler
A Journey to the Western Islands of Scotland
Lives of the Poets (Addison, Savage and Swift)
Rambler
Rasselas, Prince of Abyssinia
(all from Project Gutenberg, URL: http://www.gutenberg.org/wiki/Main_Page)

Secondary sources

Ariel, Mira (2002), 'The demise of a unique concept of literal meaning', *Journal of Pragmatics*, 34: 361–402.

Barnbrook, Geoff (2005), 'Johnson the prescriptivist? The case for the prosecution', in: Jack Lynch and Anne McDermott (eds.) *Anniversary essays on Johnson's* Dictionary. Cambridge: Cambridge University Press, 92–112.

Biber, Douglas (1993), 'Representativeness in corpus design', *Literary and Linguistic Computing*, 8 (4): 243–257.

Burger, Harald (1989), 'Phraseologismen im allgemeinen einsprachigen Wörterbuch', in: Franz J. Hausmann, Oskar Reichmann and Herbert E. Wiegand (eds.) *Wörterbücher – dictionaries – dictionnaires. Ein Internationales Handbuch zur Lexikographie*. Vol. I. Berlin: De Gruyter, 593–599.

Cawdrey, Robert (1604), *A Table Alphabeticall.* London: Printed by I. R(oberts).

Cruse, Alan (2004), *Meaning in language.* 2nd ed. Oxford: Oxford University Press.

Drosdowski, Günther (1989), 'Die Beschreibung von Metaphern im allgemeinen einsprachigen Wörterbuch', in: Franz J. Hausmann, Oskar Reichmann and Herbert E. Wiegand (eds.) *Wörterbücher – dictionaries – dictionnaires. Ein Internationales Handbuch zur Lexikographie*. Vol. I. Berlin: De Gruyter, 797–805.

Fernando, Chitra (1996), *Idioms and idiomaticity.* Oxford: Oxford University Press.

Giora, Rachel (1997), 'Understanding figurative and literal language: the graded salience hypothesis', *Cognitive Linguistics*, 8 (3): 183–206.

Hayashi, Tetsuro (1978), *The theory of English lexicography, 1530–1791.* Amsterdam: Benjamins.

Knappe, Gabriele (2004), *Idioms and fixed expressions in English language study before 1800.* Frankfurt/Main: Lang.

OED = Oxford English dictionary. Second edition, version 3.1 (2002). On CD-Rom. Oxford: Oxford University Press.

Osselton, N.E. (1995a [1986]), 'Figurative words: modern practice and the origins of a labelling tradition', in: N.E. Osselton, *Chosen words. Past and present problems for dictionary makers*. Exeter: University of Exeter Press, 16–24.

Osselton, N.E. (1995b [1990]), 'The character of the earliest English dictionaries', in: N.E. Osselton, *Chosen words. Past and present problems for dictionary makers*. Exeter: University of Exeter Press, 1–15.

Pinnavaia, Laura (2006), 'Idiomatic expressions regarding food and drink in Johnson's *Dictionary of the English Language* (1755 and 1773)', *Textus*, 19: 151–166.

Starnes, DeWitt Talmage and Gertrude Elizabeth Noyes (1991 [1946]), *The English dictionary from Cawdrey to Johnson: 1604–1755*. With an introductory article and a bibliography by Gabriele Stein. Amsterdam: Benjamins.

This page intentionally left blank

Prayers in the history of English: a corpus-based study

Thomas Kohnen

University of Cologne

Abstract

Although vernacular English prayers form a fairly important genre, especially during the Late Medieval and Early Modern periods, there are hardly any linguistic studies on them. This article investigates typical (text-) linguistic and discourse-functional properties of prayers (personal pronouns, performative formulas and pattern of address). The results suggest that prayers are – despite their unidirectional character – an interactive and performative genre, manifesting a partly idiosyncratic use of language, which does not seem to have changed very much across the centuries. The study also reveals interesting links to other genres and spheres of discourse (e.g. conversational interaction, administrative writing and personal letters).

1. Introduction

In the history of the English language, prayers constitute a somewhat neglected genre. There are hardly any linguistic descriptions, let alone corpus-based studies (but see Crystal and Davy 1969). This is in contrast to the real importance prayers seem to have had. Prayer books were among the most popular texts in Late Medieval and Early Modern England. They accompanied the daily lives of many people and they provided accessible models for expressing their most intimate hopes and fears. In his recent book on the English people and their prayers the historian Eamon Duffy says that "the history of prayer ... is as difficult to write as the history of sex". Both prayer and sex are "intensely personal" and "not readily accessible to objective analysis" (Duffy 2006: ix). Given this exceptionally intimate and seemingly unapproachable, but also extremely popular nature of prayers, one could say that prayers are a very attractive, even a sexy object for a corpus-linguistic investigation.

This paper falls into three parts. I will first give a very short overview of the history of (private) English prayers, with special emphasis on the Early Modern period, and of the prayer corpus used in this study. I will then deal with typical (text-) linguistic and discourse-functional properties of prayers. These are personal pronouns, performative formulas and patterns of address. In the conclusions section I will shortly discuss the results of this investigation against the background of other genres in the history of English.

2. English prayers: background and data

In the context of Christian religion, the general communicative setting of prayers seems to have been quite stable across the past centuries. Prayers are unidirectional and involve a transcendental addressee (God or a saint). Prayers may be public or private, they may be performed together or alone, they are in many cases fairly short and often come as published collections of prefabricated and rather fixed items.

The basic manifestation of prayer, "prayer proper", takes the form of a first person (*I*) who addresses God (or a saint) using the second person (*thee*; see example (1)).

(1) O Holy Lord God Almighty, … behold here **I** prostrate my self before **thee**, … (Anne D. Morton, *The Countess of Morton's Daily Exercise*, 1666, A4)

There are also some minor manifestations of prayer where this setting is altered. For example, the person praying may talk *about* God, especially in sections devoted to adoration and praise. In this kind of prayer the person praying provides adoring and reverential terms for God ("Glory be to the father ...") or invites other people to join in worshipping him ("Prayse ye the lorde."; see example (2)).

(2) **Glory be to the father, to y^{e} sonne, and to the holy ghoste**. As it was in the begynnynge: as it is nowe / and euer shalbe. Amen. **Prayse ye the lorde.** (*Prymer in Englyshe and in Laten*, 1536, 22^{v})

In a similar way, a prayer may be introduced by a request (mostly in the first person, *let us pray*) directed at other people to join in prayer, with the request and the following supplications actually constituting this prayer (see example (3)).

(3) **Let us pray.** WE beseech thee, O Lord, defend us from all perils of mind and body ... (*The Primer, or, Office of the Blessed Virgin Mary,* 1658, 309)

In addition, prayers may contain other texts, mostly extracts from the Bible. These are usually sections from the psalms or narrative passages taken from the gospels (see example (4)).

(4) The .xciiij. psalme. COme & let vs ioyfully gyue thankes vnto the lorde: let vs reioyse in god our sauyoure / let vs approche in to his presens with prayse and thankes geuynge / and synge we vnto hym in Psalmes. (*Prymer in Englyshe and in Laten*, 1536, 22^{v})

It goes without saying that in such sections the communicative setting, with the person praying addressing God, is not necessarily reflected in the linguistic

structure of the text (for example, in (4) God is referred to in an impersonal way). In this investigation all the different "varieties" of prayer mentioned above are included in the data.

When we look at prayers in the history of English, we find only few vernacular prayer collections till about 1400. Some more vernacular prayers appear in the course of the 15th century, but the main share of prayer collections belongs to the Early Modern English period and the following centuries. The type of prayer collection which proved to be most important for the development of vernacular prayer is the primer or so-called *Book of Hours* (on English primers see Duffy 1992: 209–265; Duffy 2006; Butterworth 1953; see also Littlehales 1892). Primers were prayer-books or devotional manuals for the use of the laity. They contained a copy and later the English translation of different parts of the Breviary and Manual, with various added vernacular prayers. These collections of vernacular prayers were later expanded, revised and developed further into special collections (for example, for women).

One particular feature of the vernacular prayer collections was that they contained a common pool of devotions which were copied from book to book and even survived with small alterations the religious upheavals of the 16th century (Duffy 2006: 80, 139, 164–168). The prayers were mainly concerned with safety and salvation, protection and pardon from sin and similar issues.

The users of primers were mostly lay people, chiefly from the rising middle classes (among them many women). Later the users were spreading further down the social scale. Depending on the class of the user, primers could come as precious luxury editions or as cheap mass products. But it seems that the contents of these Early Modern prayer collections were rather similar across social class and even across denomination. So prayer books seem to be a rather stable genre.

Primers or Books of Hours have been called the most popular book in the late Middle Ages. The primer was also the chief product of print technology both in terms of numbers and editions (Duffy 2006: 4, 28). Thus one can say that in terms of reception, popularity and circulation prayers were among the most important genres in Late Medieval and Early Modern England.

This study is based on a preliminary collection of prayers that will be part of the *Corpus of English Religious Prose* (which is presently compiled at the University of Cologne; see Kohnen 2007). Due to the development and availability of vernacular prayers, the focus was on the 16th and 17th centuries. The focus was also on prayers in the context of private devotion. Extracts were selected from primers as well as later collections (see the Appendix for a detailed list).

The prayer corpus was divided up into three sub-corpora, each covering a period of fifty years (see Table 1).[1] In all, the number of words in the corpus adds up to ca 257,000 words.

Table 1: Sub-corpora of the prayer corpus

Subcorpus 1	1525–1574	90,913 words
Subcorpus 2	1575–1624	98,434 words
Subcorpus 3	1625–1674	67,640 words

3. Linguistic analysis

In my linguistic analysis I will deal with typical (text-)linguistic and discourse-functional properties of prayers. These are personal pronouns, explicit performatives and patterns of address. The features were selected because they seem to reflect the basic functional profile of the genre and thus can be used as a basis for studying stability or change in the genre and for a comparison with other genres.

3.1 Personal pronouns

One prominent and predominant feature of prayers is the high frequency of first-person and second-person pronouns (see examples (5) and (6)). This is a reflection of the basic communicative situation of prayer, with the person praying addressing a transcendental authority.

(5) **I** thanke **thee** for blessing **me** this day past, and **I** intreat **thee** good God, so to continue thy blessings, and to increase them more and more toward **me**, that **I** may feele and find, that **thou**, O Lord, art my euerlasting shield, and succour. (Michael Sparke, *The Crums of Comfort with Godly Prayers*, 1628, D12)

(6) O Gloryous iesu ... / **I** praye **the** that **I** may haue true confessyon / contricyon and satisfaccyon or ["before"] **I** dye / and that **I** may se and receyue thy holy body ... without synne. And that **thou** wyt ["will"] my lorde god forgyue **me** all my synnes… / and that **I** may ende my lyfe in the true fayth of holy chyrche / (*Prymer of Salysbury Vse*, 1527, ciii^{v})

The data reveal a remarkable stability in the high frequency of first-person and second-person pronouns across the sub-corpora, first-person pronouns ranging between 40 and 45, second-person pronouns around 25 per 1,000 words (Figure 1, p. 169).

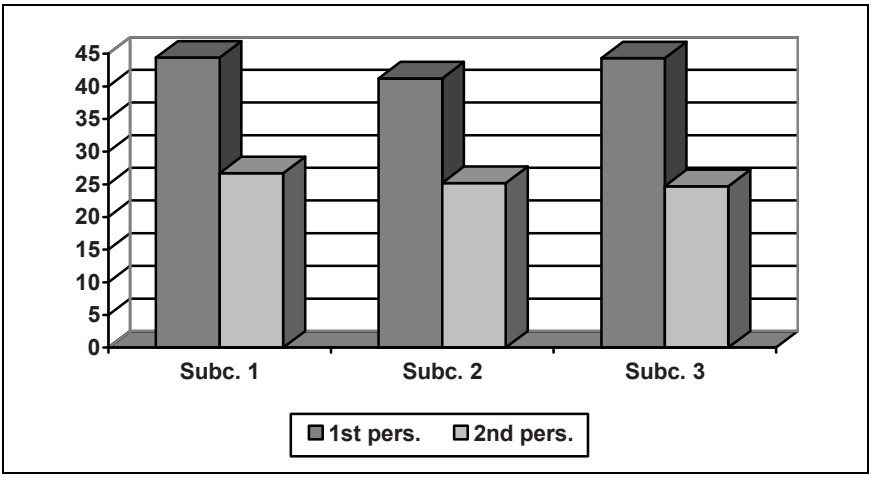


Figure 1: First-person and second-person pronouns in the sub-corpora of the prayer corpus (frequency per 1,000 words)

When we split up the first-person pronouns into singular and plural pronouns (see Figure 2, p. 170), we notice an interesting development, a decrease in singular pronouns and an increase in plural pronouns. The reason for this may be a more communal orientation of Protestant prayer, a feature which has sometimes been called typical of the Reformation.[2]

The general high frequencies of first-person and second-person pronouns clearly reflect the prominent position of addressor and addressee in prayers: persons praying often refer to themselves as well as to the addressee (God, a saint). In this prayers are similar to texts which are associated with spoken language and interaction, for example, conversation, drama, trials, and letters. Such genres are often called interactive. Are prayers an interactive genre?

With prayers the term "interactive" may sound inappropriate due to the unidirectional nature of the communication. On the other hand, the term may be appropriate because it reflects the high involvement of the addressor with the addressee and the fact that the addressor performs speech acts which directly aim at the addressee. This is, for example, also typical of a letter writer who is writing a letter in anticipation of an answer (which in fact may never arrive).

In deciding whether prayers may be called interactive we may also compare the frequencies found in prayers with data in other typically interactive genres. Figure 3 (p. 170) below contains some prominent examples.

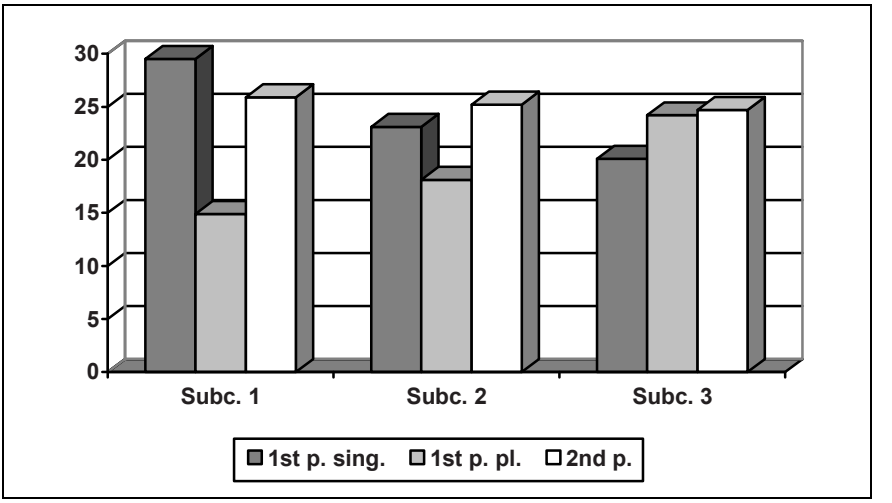


Figure 2: First-person (singular / plular) and second-person pronouns in the sub-corpora of the prayer corpus (frequency per 1,000 words)

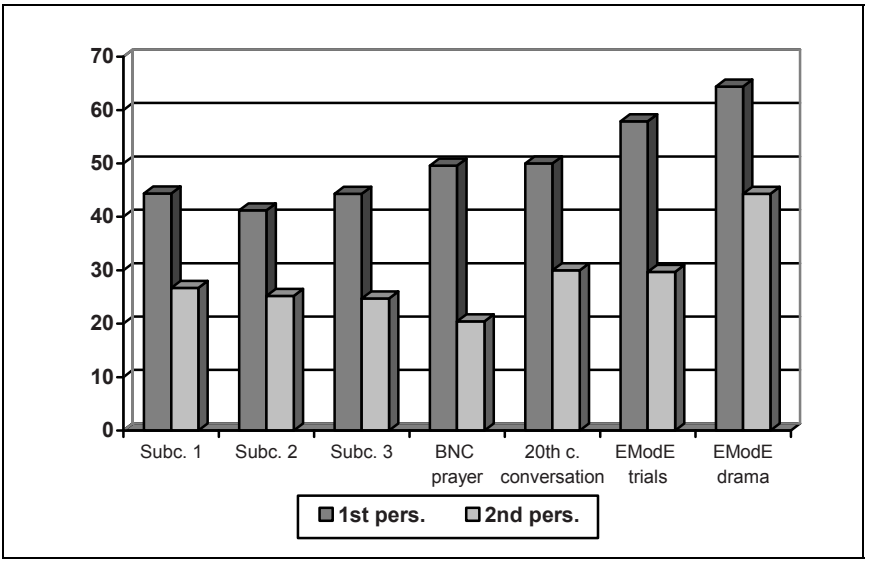


Figure 3: First-person and second-person pronouns in prayer sub-corpora and other genres (frequency per 1,000 words)

I looked at the collection of prayers contained in the *British National Corpus*. The file is rather small (7,243 words) but may be quite typical of (public)

contemporary prayer.[3] Here we find an even higher frequency of first-person pronouns (49.6) and a slightly lower frequency of second-person pronouns (20.4). But the basic general proportion of first- and second-person pronouns seems to have prevailed throughout the centuries. Biber et al., in their *Grammar of Spoken and Written English*, say that the highest frequencies for first- and second person pronouns are found in conversation (1999: 334). The frequencies given are quite similar to those in the prayers, although slightly higher (50 for first-person and 30 for second-person pronouns). It is quite striking how the proportion of first- and second-person pronouns shown in the prayers basically resembles conversation. It is also remarkable that none of the other registers mentioned by Biber et al. reach the frequencies noted for prayers. Here fiction comes closest, with 25 for first-person and 11 for second-person pronouns.

I also included the frequencies reported by Culpeper and Kytö (2000: 184–185) for first- and second-person pronouns in their corpus of Early Modern English trial proceedings and dramas (comedies). The frequencies of first-person pronouns in trials (57.9) and drama (64.4) are even higher than in prayers and 20th-century conversation. This applies to second-person pronouns in drama as well (44.3), while trials (29.7) are similar here to the Early Modern English prayers and 20th-century conversation.

The striking similarities of the proportions shown in the data suggest that prayers clearly rank among the typically interactive genres, which show more or less close connections to spoken interaction.

3.2 Performative formulas

Another quite conspicuous feature of prayers is the high frequency of explicit performative formulas, that is, expressions which make explicit the speech act which the addressor performs. In most of the texts of the present corpus the constitutive speech acts of prayers are made explicit by means of an explicit performative formula. These are directive speech acts, that is, asking the addressee to perform an act (for example, *pray, beseech and entreat*; see examples (7)–(9)), acts of thanking God (examples (9) and (10)), acts of confessing / professing (one's sins or God's true divine nature; example (11)) and acts of praising / worshipping God (example (10)).

(7) Oh kill **I beseech thee** sweet Iesu, and vtterlie extinguish in me all inordinate lusts; pull out, and plucke vp by the rootes what vice soeuer is in me, and take quite awaie whatsoeuer displeaseth thee in me. (Thomas Bentley, *The Fift Lampe of Virginitie*, 1582, 10)

(8) Wherefore **we pray and besech** thy maiestye, that at no tyme thou suffer vs to be vnthankefull vnto these exceding great benefites, nor yet vnworthy of thy greate merytes, … (Cuthbert Tunstall, *Certaine Godly and Deuout Prayers*, 1558, 14)

(9) **I thanke thee** for blessing me this day past, and **I intreat thee** good God, so to continue thy blessings, … (Michael Sparke, *The Crums of Comfort with Godly Prayers*, 1628, D12)

(10) O Lorde Iesu Cryste / **I worshipe prayse & tha~ke the** which wast taken / bownde & wykedly entreted of thy enmys. Make me fre fro~ all vyces & to be neglecte lytel to be set by and suffer gladly bothe rebukis and iniury. (*The Mystik Sweet Rosary of the Faythful Soule*, 1533, 29)

(11) I Cover not, I do not dissemble, I do not extenuate and lessen my Sins: **I freely confess them**, I have done exceedingly amiss, I call it often to mind, and I condemn my self for it. (Anne D. Morton, *The Countess of Morton's Daily Exercise*, 1666, D5)

The frequencies of the major explicit performatives in the prayer sub-corpora are given in Figure 4.

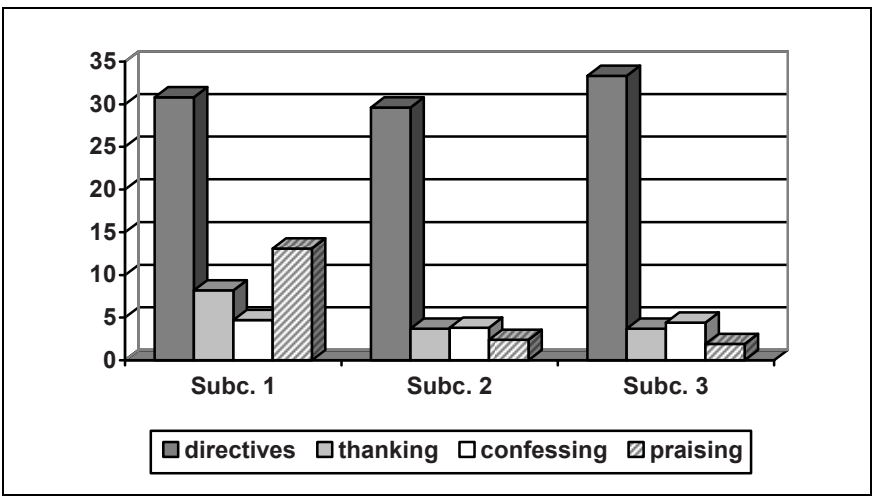


Figure 4: Major explicit performatives in the sub-corpora of the prayer corpus (frequency per 10,000 words)

The most prominent performatives are directives. Frequencies remain here at a high level (around 30) in all sub-corpora. The most frequent verb used is *beseech* (536 items), followed by *pray* (108 items), *entreat* (29 items), *require* (11 items) and some others (*ask*, *appeal*, *demand* etc.) which show a rather low incidence (1–5). On the whole, the number of different directive speech-act verbs seems to be quite stable across the periods (7–8), although some verbs are not found either

in the earlier period (*beg, entreat, plead*) or in the later period (*require, ask, demand*).

The frequencies of the other performatives are not quite as spectacular but still fairly high. With *thank* they range between 4 and 8, with *confess* around 4 and with verbs of praising between 13 and 2. The decrease in thanking and praising is quite remarkable. It seems difficult to speculate about the reasons.

The four classes of performatives are also found in the collection of contemporary prayers stemming from the *British National Corpus*. In fact, the frequency of directives is extremely high here (233 in 10,000 words), which may be due to the compact character of the prayers, with many petitions and hardly any reflective and contemplative passages. Among the other performatives, verbs of thanking have 55, praising 4.1 and confessing 1.4 items in 10,000 words. So it seems that performatives are typical of prayer. Why is that so? In order to give a first answer to this question, we will look at some more data in other genres and focus on directive performatives. Figure 5 shows the frequencies of directive performatives in several corpora and genres.

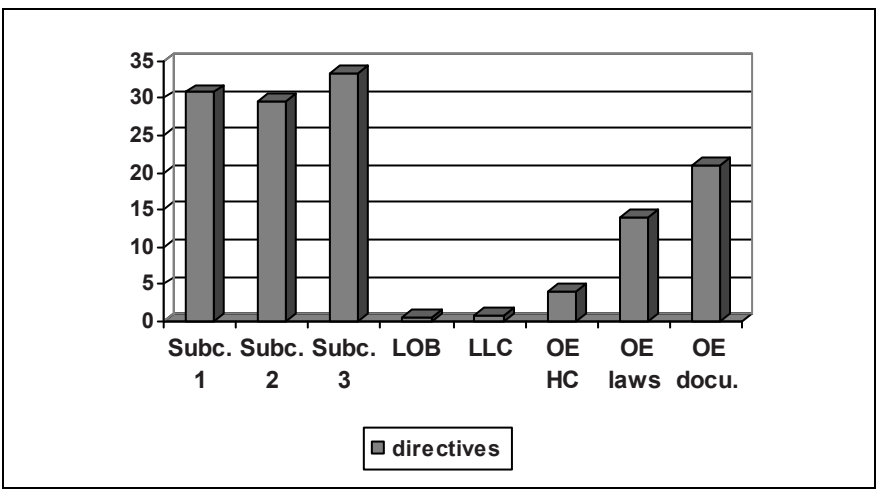


Figure 5: Directive performatives in several corpora and genres (frequency per 10,000 words)

In an investigation of directive performatives (Kohnen 2000a, b) I found that corpora of modern written and spoken English showed an extremely low frequency of performatives (0.6 in LOB and 0.9 in LLC), whereas the Old English section of the Helsinki Corpus had 4 items in 10,000 words. Among the Old English genres which showed the highest frequencies were laws (with 14) and documents (with 21). How does this help to explain the high frequency of directive performatives in prayers?

This comparison shows first of all that directive performatives in prayers seem to be much more frequent than anywhere else in contemporary English. The only genres showing relatively high frequencies are Old English genres stemming from a special field. Laws and documents are genres which have been shown to have an oral background, where stating what you are doing is most important because it ensures the proper performance and validity of the speech acts contained in the genre (enacting a law, making a will, making a lease etc.; Kohnen 2000a; on the oral nature of Old English laws see also Danet and Bogoch 1994). My suggestion is that prayers show this oral feature in a similar way and that the explicit performative ensures the proper performance and validity of the speech acts of the prayer. This is another feature of prayers which locates them in the field of orality and interactive language use.

3.3 Patterns of address

The third discourse-functional feature concerns patterns of address. Given the interactive nature of prayers, it is hardly surprising that we find a large quantity of addresses and a high frequency of selected address terms. These are, of course, mostly designations of God. By far the most common is the address term *Lord*. I have looked at the instances in the three sub-corpora where *Lord* is used as an address term, that is, used to appeal directly to the addressee of the prayer. Here I distinguished the ordinary cases from the patterns where the address term *Lord* is followed by an apposition or a relative clause or both apposition plus relative clause. This construction seems to be quite peculiar to prayers and religious language (see examples (12) and (13)).

(12) O My souerayne **lorde Ihesu the veray sone of almyghty god and of the moost clene & glorious vyrgin Mary / that suffred the bytter deth for my sake and all mankynde vpon good fryday & rose agayne the thyrde daye**. I beseche the lorde haue mercy vpon me that am a wreched synner but yet thy creature. **And for thy precyous passion saue me** & kepe me from all perylles bothe bodyly & goostly / (*Prymer of Salysbury Vse*, 1527, ccvii^{r})

(13) O Most sweet **Lord Jesus Christ**, **the true God**, **who from the bosome of the highest almighty father was sent into the world to release sins, to redeem the afflicted**, … : vouchsafe, O Lord Jesus Christ **to absolve, and deliver me** thy servant **out of the affliction and tribulation in which I am put.** (*The Primer, or, Office of the Blessed Virgin Mary*, 1658, 245)

Figure 6 (p. 175) shows the frequency of *Lord* used as an address term and the combination with apposition and/or relative clause. The frequency of the address term *Lord* is fairly high and fairly stable (ranging between 5 and 6.5). This corroborates the basic interactive nature of the genre. The frequency of the

address term *Lord* plus apposition and/or relative clause is, given the peculiarity of the construction, fairly high as well, although decreasing (1.6–0.7). A comparison with the BNC prayers shows that a similar situation can still be found in the late 20th century (9.8 for the address term *Lord*), although the construction with apposition and/or relative clause has a very low frequency (0.3).

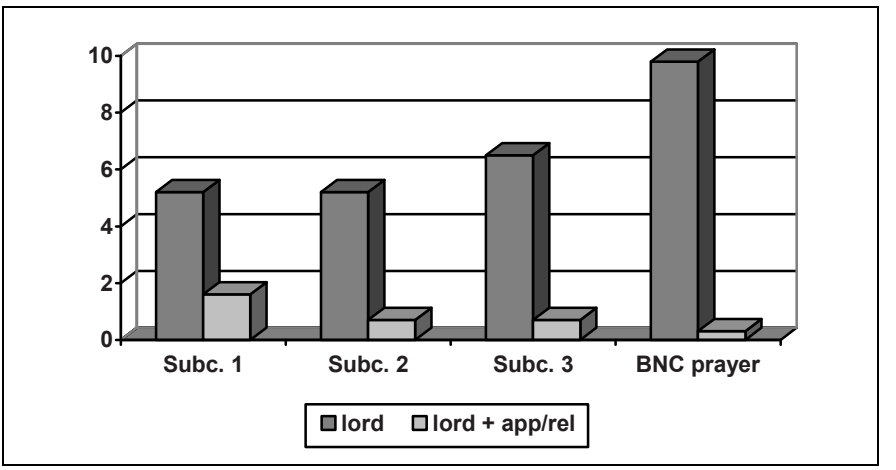


Figure 6: The address term *Lord* in prayers (frequency per 1,000 words)

What is the function of the appositions and relative clauses following *Lord*? As can be seen from the examples, the apposition and the relative clause contain additional information which supports the subsequent petition. In (12) the relative clause contains among other things a reference to Christ's passion ("that suffred the bytter deth"), a point which is resumed in the subsequent petition ("for thy precyous passion saue me"). In (13) there is a similar relationship between relative clause and subsequent petition ("who ... was sent into the world to release sins, to redeem the afflicted, … : vouchsafe, ... to absolve, and deliver me thy servant out of the affliction and tribulation in which I am put").

As Meyer (1991: 179) has shown, appositions are particularly frequent in genres which typically have less shared information, for example, formal written texts. Appositions are used here to provide the information which cannot be supposed to be shared. The same may, of course, be claimed for non-restrictive relative clauses. What about the prayers? Quite obviously, here the information given in the apposition and relative clause must be taken to be known to God. Why, then, are there so many appositions and relative clauses? One reason for the inclusion of the information may be to recall to the person praying the most essential and most relevant facts of their religious faith. This might be so because the prayers are usually readymade products, adapted to potential users with limited theological background knowledge (on this see also Crystal and Davy

1969: 170). If this analysis is correct, it might very well betray the artificial nature of prayers. The information given in the appositions and relative clauses is in real fact for the addressor, not for the addressee.

But there may be another explanation which might again reveal the proximity of prayers to oral genres. Most of the information given in the appositions and relative clauses may be what Ellen Prince (1981) has called "unused" information. "Unused" information can be assumed to be known to both addressor and addressee as background knowledge but has not been mentioned in the communicative activity so far. A typical situation where many "unused" topics are continually raised from a large inventory of common background knowledge is a conversational situation where two people have known each other for long. The items are all shared knowledge but they are mentioned in order to explain the relevance of the point which is being discussed. A similar situation may be assumed for prayer. Here the person sometimes is even negotiating with God, mentioning several points which are in fact common knowledge but which may underline the point of the prayer. Example (14) may illustrate this aspect.

(14) Heare thou my God, for I am despised; turne their shame vpon their owne heads: for they are puft vp with pride, as the stomach that is choaked with fat. **O Lord of hosts**, **thou righteous searcher**, **which knowest the reines and the verie heart**, let me see them punished, if it be thy will: for vnto thee doo I commit my cause. (Anne Wheathill, *A Handfull of Holesome (though Homelie) Hearbs*, 1584, 60)

The request presented in (14) is not actually a very Christian supplication ("turne their shame vpon their owne heads ... let me see them punished"). But the person praying can actually offer a plausible justification in the form of arguments presented as "unused" information contained in the apposition and relative clause following the address term *Lord* ("thou righteous searcher, which knowest the reines and the verie heart"). If God knows "reines and heart" of the enemies, he must know that they are wicked; if he is a righteous judge, he must punish them. Example (14) nicely illustrates that patterns of address typical of prayers serve mostly interactional functions.

4. Conclusions

This was a short corpus-based investigation of some text-linguistic and discourse-functional features of prayers, which revealed important elements of the genre profile and interesting, maybe unexpected, points of comparison to other genres.

Although marked by a unidirectional character, prayers may be called an interactive genre, at least with regard to personal pronouns and address terms. Seen in this perspective, prayers must be located in close proximity to conversation and written manifestations of spoken interaction. Here prayers

should be compared to more data from everyday conversation, drama, trials and personal letters.

In addition, prayer is a performative genre. In prayers the constitutive speech acts are typically realised by performative formulas. This suggests links to orality and formulaic language use. Here prayer could be compared to (formerly) oral genres (for example, wills, laws, charms etc.).

On the other hand, prayers show a partly idiosyncratic language use (see, for example, the patterns of address), and they form a conservative genre which does not seem to have changed a lot during the centuries. Thus, although prayers seem to have been very popular, they reflect a special language use, typical of the register of religious language. Here a comparison to genres or domains which do not seem to have changed very much across the centuries (for example, administrative writing) seems to be rewarding.

But there is another aspect which sets prayer apart from many other genres. Prayer texts only form the script for a complex speech event. Although we can read the words which were supposed to be said, we do not know which words were actually said. In this prayers are similar to texts of plays, which usually need to be somehow performed.

These various aspects of prayers show that they form a highly attractive genre, revealing often quite unexpected links and similarities to other genres and spheres of discourse.

Notes

1 The slight difference in the number of words in each period is due to the availability of texts (in the process of compilation) and the period boundaries. At a later stage in the final compilation of the *Corpus of English Religious Prose* the period boundaries will be altered, forming a common grid for all genres of the corpus.

2 Heal, for example, in her monograph on the Reformation in Britain and Ireland, states in the context of prayer that Hooker's "affirmation of the power of collective worship could stand as exemplary of the ambition of all Protestant reformers" (2003: 428).

3 In the *British National Corpus* this is Text GXO.

References

Corpora

BNC = *British National Corpus*. See http://www.natcorp.ox.ac.uk/.

Corpus of English Religious Prose. Under compilation by Thomas Kohnen, Tanja Rütten, Ingvilt Marcoe, Kirsten Gather and Dorothee Groeger, University of Cologne. See http://www.helsinki.fi/varieng/CoRD/corpora/COERP/index.html.

Helsinki Corpus = *The Helsinki Corpus of English Texts* (1991). Department of English, University of Helsinki. Compiled by Matti Rissanen (Project leader), Merja Kytö (Project secretary); Leena Kahlas-Tarkka, Matti Kilpiö (Old English); Saara Nevanlinna, Irma Taavitsainen (Middle English); Terttu Nevalainen, Helena Raumolin-Brunberg (Early Modern English). See http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/index.html.

LLC = *The London-Lund Corpus of Spoken English*. Compiled by Jan Svartvik, Lund University. See http://www.helsinki.fi/varieng/CoRD/corpora/LLC/index.html.

LOB = *The Lancaster-Oslo/Bergen Corpus*. Original version (1970–1978) compiled by Geoffrey Leech, Lancaster University, Stig Johansson, University of Oslo (project leaders), and Knut Hofland, University of Bergen (head of computing); POS-tagged version (1981–1986), compiled by Geoffrey Leech, Lancaster University, Stig Johansson, University of Oslo (project leaders), Roger Garside, Lancaster University, and Knut Hofland, University of Bergen (heads of computing). See http://www.helsinki.fi/varieng/CoRD/corpora/LOB/index.html.

Secondary sources

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan (1999), *Longman grammar of spoken and written English*. Harlow: Pearson Education Limited.

Butterworth, Charles C. (1953), *The English primers, 1529–1545: their publication and connection with the English Bible and the Reformation in England.* Philadelphia: University of Philadelphia Press

Crystal, David and Derek Davy (1969), *Investigating English style*. Harlow: Longman.

Culpeper, Jonathan and Merja Kytö (2000), 'Data in historical pragmatics: Spoken discourse (re)cast as writing', *Journal of Historical Pragmatics*, 1 (2): 175–199.

Danet, Brenda, and Bryna Bogoch (1994), 'Orality, literacy, and performativity in Anglo-Saxon wills', in: John Gibbons (ed.) *Language and the law*. London: Longman, 100–135.
Duffy, Eamon (1992), *The stripping of the altars. Traditional religion in England 1400–1580*. New Haven: Yale University Press.
Duffy, Eamon (2006), *Marking the hours. English people and their prayers 1240–1570*. New Haven: Yale University Press.
Heal, Felicity (2003), *Reformation in Britain and Ireland*. Oxford: OUP.
Kohnen, Thomas (2000a), 'Corpora and speech acts: the study of performatives', in: Christian Mair and Marianne Hundt (eds.) *Corpus linguistics and linguistic theory. Proceedings of the 20th ICAME Conference, Freiburg im Breisgau 1999*. Amsterdam: Rodopi, 177–186.
Kohnen, Thomas (2000b), 'Explicit performatives in Old English: a corpus-based study of directives', *Journal of Historical Pragmatics*, 1 (2): 301–321.
Kohnen, Thomas (2007), 'From Helsinki through the centuries: the design and development of English diachronic corpora', in: Päivi Pahta, Irma Taavitsainen, Terttu Nevalainen and Jukka Tyrkkö (eds.) *Towards multimedia in corpus studies* (Studies in Language Variation, Contacts and Change in English 2). Helsinki: Research Unit for Variation, Contacts and Change in English. Available at http://www.helsinki.fi/varieng/journal/index.html.
Littlehales, Henry (1892), *The prymer or prayerbook of the lay people in the Middle Ages. Part II. Collation of MSS.* London: Longmans, Green & Co.
Meyer, Charles F. (1991), *Apposition in contemporary English.* Cambridge: Cambridge University Press.
Prince, Ellen F. (1981), 'Towards a taxonomy of given-new information', in: P. Cole (ed.) *Radical pragmatics.* New York: Academic Press.

Appendix: The texts of the prayer corpus

Subcorpus 1 (1525–1574) 90,913 words

Prymer of Salysbury Vse, 1527
Mystik Sweet Rosary of the Faythful Soule, 1533
Deuoute Prayers in Englysshe, 1535
The Pater Noster Spoken of y^e^ Sinner, 1535
Prymer in Englyshe and in Laten, 1536
The Rosary with the Articles of the Lyfe & Deth of Iesu Chryst, 1537
Catharine Parr. *Prayers or Meditacions*, 1545
A Boke of Prayers Called y^e^ Ordynary Faschyon of Good Lyuynge, 1546
Deuout Meditacions, Psalmes and Praiers, 1548
Cuthbert Tunstall, *Certaine Godly and Deuout Prayers*, 1558
Thomas Becon, *The Pomavnder of Prayer*, 1561
A Good and a Godly Prayer, 1563
A Fourme of Prayer to be Vsed in Priuate Houses, 1570

Subcorpus 2 (1575–1624) 98,434 words

Thomas Twyne, *The Garlande of Godlie Flowers*, 1580
A Prymmer or Boke of Private Prayer, 1580
Richard Day, *A Booke of Christian Prayers*, 1581
Thomas Bentley, *The Fift Lampe of Virginitie*, 1582
Anne Wheathill, *A Handfull of Holesome (though Homelie) Hearbs*, 1584
Edward M. Dering, *Godly Private Prayers*, 1597
Thomas Sorocold, *Supplications of Saints*, 1612

Subcorpus 3 (1625–1674) 67,640 words

Michael Sparke, *The Crums of Comfort with Godly Prayers*, 1628
Jeremy Taylor, *The Golden Grove*, 1654
The Primer, or, Office of the Blessed Virgin Mary, 1658
Ann Douglas Morton, *The Countess of Morton's Daily Exercise*, 1666

Semantic drift in Shakespeare, and Early Modern English full-text corpora

Ian Lancashire

University of Toronto

Abstract

How can we detect Early Modern English semantic deviation, which Manfred Görlach describes as "a difficult and largely unsolved problem for the history of the English lexicon"? Lexical and phrasal neologisms stand out from the simplest word-lists, but readers must understand the meaning of words in context before recognizing semantic drift. New diachronic corpora such as Early English Books Online (EEBO) *and* Lexicons of Early Modern English (LEME)*, two web-based corpora, could help, were their vocabulary to be lemmatized and associated with specific senses in the* Oxford English Dictionary. *Only* LEME *tries to do so, however, and its more than half a million word-entries are not uniformly analyzed yet. Three instances of semantic deviation in Shakespeare's plays serve to illustrate the challenge: the "pricking" of a witch's thumb in Macbeth (a means of torture), the villain's name "Aron" in Titus Andronicus (a new starch obtained from a weed of that name), and the term "acting" (for "enacting") as used by Brutus in describing his dream of a conspiracy to assassinate Caesar. Because these innovative senses are undocumented in this period, because no monolingual English dictionaries survive from it, and because these instances of drift disappeared soon afterwards, they are hard to find. Manfred Görlach's problem will remain in force for some time to come if we have to rely on literary text analysis to locate semantic deviation.*

Shakespeare's couplet in *Macbeth*, "By the pricking of my Thumbes, / Something wicked this way comes" (IV.i.44–45), prompted the titles of Ray Bradbury's novel *Something Wicked this Way Comes* (1962) and of Agatha Christie's novel *By the Pricking of My Thumbs* (1968), but what the weird chatter of the second of the bearded witches meant has been anyone's guess.[1] Why does one of Shakespeare's odd "sisters" complain that tingling in her thumbs portends the imminent entrance of a king of the Scots, Macbeth? Did she, as some have thought, suffer from compulsive ergotism? But what does a disease acquired through the eating of contaminated grain have to do with the approach of a man who, depending on one's perspective, is either "braue *Macbeth* (well hee deserues that Name)" or "this dead Butcher"?

Scholars date *Macbeth* about 1606, just three years after James I succeeded Elizabeth to the English throne. Shakespeare's play, which showed a

pageant-like prophetic line of kings down to James I, would have appealed to this Scots King. Its witches alone would have appealed to him because he had authored *Daemonologie* (1597), a book that gave witches, and their detection, some publicity. It describes a telltale but secret witch's mark to which the devil gifted insensibility to pain no matter how she was "nipped or pricked" there (1924: 33), but only as long as the witch obeyed him. This passage specifies pricking as a type of torture used to extract confessions from suspected witches by Scottish prosecutors. Its ancient instrument was first called the pilliwinkis (*DSL* 1590–91–; *OED* 1397–), and later the thumbikins (*DSL* and *OED* 1684–) and thumbscrews (*OED* 1794–). A word-entry in *Lexicons of Early Modern English* (*LEME*), from the Spanish-English dictionary by Richard Perceval and John Minsheu (1599), documents a comparable practice: "*Tráto de cuérda*, a kinde of torment by tying the thumbs to make confesse."[2]

The authoritative *OED* explanation, that "*pricking of* (also *in*) *one's thumbs"* meant "an intuitive feeling or hunch; a premonition or foreboding" when used to allude to *Macbeth* ("pricking, n.," 1b), while generally correct, misses the point. The second witch would have known that, when a King of the Scots greets a woman who might be demonically possessed – and Macbeth asks, "How now you secret, black, & midnight Hags? / What is't you do?", two lines later – he might bring with him some pilliwinkis, pricking her thumbs to elicit an answer.

Although Shakespeare's phrase "By the pricking of my thumbs" is not neologistic, the phrase nonetheless is as innovative as most "hard words" of the time. 400 years later, the *OED* still does not associate this participial noun, or the phrase Shakespeare coined from it, with witch torture. Shakespeare derives a novel sense of an old participial noun, "pricking," from a new sense of its old verb, "prick," that is first observed in James' *Daemonologie* in 1597. When Shakespeare invests a word with a part of speech it has never had before, or when he borrows a construction from Latin, linguists credit him with contributing new words to English by derivation. His other form of lexical creativity – taking a sense from one word and adding it to another related word by *semantic derivation* (from an English term) – is harder to prove, mainly because evidence for it readily escapes notice: word-searches locate new strings much more easily than they do novel senses of those strings. Researchers cannot rely on data mining but must do close reading. Manfred Görlach says that the increase in the "semantic ranges" of a term is "a difficult and largely unsolved problem for the history of the English lexicon" (1991: 199–200). Terttu Nevalainen analyzes the types of semantic derivation, either precipitated by "language-external factors within the same conceptual field, or … intentionally extended to new items in another field" (433–434), but what activates these mechanisms and leads speakers to transfer a sense of one word to another word, often a related form of that word, is unclear. Equally obscure are the grounds for the long-term acceptance, or rejection, of semantic changes. Not only do we need very large databases of contemporary idiolects to help us detect semantic drift, we have to understand semantic appropriation. Because English had no monolingual dictionary, no guide to the meaning of words, Shakespeare and his contemporaries enjoyed great freedom.

Playwrights and acting companies valued language for the entertainment it afforded audiences and disseminated lexical innovations quickly. What radio, TV, and film do today, the Rose, the Theatre, and the Globe did in the 1590s.

Intentionality is fraught with problems. Do we today dare claim that we can know what Shakespeare was thinking? Understanding his text, independently of that, is hard enough, given the four centuries that have elapsed since he wrote. Happenstance, not motive, might play a part: when devising the second witch's witty allusion, he could have appropriated gossip about the Scots' witch-trials or other publications rather than read James' own prose. Early English Books Online / Text Creation Partnership (EEBO/TCP), a collection of nearly 15,000 digitized books in the *Short-Title Catalog* and Wing, has 78 occurrences of "pricking" in 45 texts published from 1600 to 1606. At least one concerns demonic possession. John Darrel narrates the case of William Somers of Nottingham, of whom trial was made, "by pricking of pyns, whereat he neuer styrred though a pyn being somewhat greate and crooked was thrust vp to the head" (p. 11). Because none of the 84 occurrences of the word "thumbs" in 51 texts from 1590 to 1610 concerns witches, no match to Shakespeare's phrase can be found.

EEBO/TCP, a collection of nearly 33,000 digitized books, which enables readers to trace semantic changes, is a true corpus. Like the Helsinki Corpus, EEBO/TCP selects texts from an entire period so as to be representative, and enriches their texts by encoding them, but unlike Helsinki does not excerpt equal-length samples from these texts. Helsinki offers, for each text, extralinguistic information such as date of composition and of publication, text type, author, etc., and Anthony Kroch and others have made syntactically-tagged altered versions of the ME and EModE sections of the Helsinki Corpus (PPCME2, PPCEME). Each EEBO/TCP text has the standard bibliographical fields, but researchers are now beginning to tag all the words in EEBO/TCP texts by their lemmata, their dictionary headwords. It is hard to find all variant forms of any vocabulary item in EEBO/TCP now, except by trawling its complete word-index. By linking each word-form to its *OED* headword, however, researchers will be able to find all uses of that word, and hence all its semantic values. In turn, that index enables readers to compare the semantic preferences of Shakespeare with those of his contemporaries.

Another corpus tool for the study of semantic derivation is my *Lexicons of Early Modern English* (*LEME*), published online by the University of Toronto Library and the University of Toronto Press in 2006.[3] More than 1,200 manuscripts and printed books from the 230 years from 1470 to 1700 include lexical entries, so that *LEME*'s current 588,000 word-entries, drawn from 176 bilingual, polyglot, and monolingual dictionaries and glossaries, and treating eight varieties of English and 36 other languages, amount to one-eighth of an estimated total of four million word-entries that survive from the period. *LEME* searches – being freely searchable online, its site giving a bibliography of all known texts – now supplement searches of the *Oxford English Dictionary* and Shakespeare editions by checking what his contemporaries had to say about the

words he (and anyone else in the period) used. Unlike EEBO/TCP, *LEME* explicitly describes the changing semantics of headwords.

LEME grew out of the *Early Modern English Dictionaries Database* (*EMEDD*), some sixteen dictionaries that I and my student Mark Catt made available from 1996 to 1999. I started transcribing dictionaries about 1988 as founding director of the Centre for Computing in the Humanities. My first paper on them was at the ACH/ALLC conference in Tempe, Arizona, in March 1991 in a session on the Renaissance Knowledge Base. David Richardson at Cleveland, Roy Flannagan at Ohio, and I then proposed to create a large digital corpus of English literature, to be served by the *EMEDD*. The subsequent NEH grant application was turned down, however, maybe because industry was on the verge of releasing some huge full-text corpora. In 1996 the Modern Language Association of America published the manual for the *TACT* text-analysis concordancer, which contained my small digital library of English literature on CD-ROM; and in 1997–98 I put online Renaissance Electronic Texts, a small Web series with three works, the 1623 Elizabethan homilies, Edmund Coote's *The English School-maister*, and Shakespeare's sonnets, co-edited with Hardy Cook. However, Chadwyck-Healey's English poetry database, which would become *Literature Online* (*LION*), came online in 1996. It decided me to focus on the dictionaries. They posed transcription and encoding problems that even *LION* and largescale projects like EEBO/TCP (begun in 2000) could not resolve easily. *EMEDD* had grown popular with researchers; and so I persevered. About five years ago, the Canada Foundation for Innovation (CFI) funded TAPoR (Text Analysis Portal for Research), directed by Geoffrey Rockwell. They and the Social Sciences and Humanities Research Council of Canada have generously supported my *LEME* work. After two years' programming by Dr. Marc Plamondon, the help of a half-dozen research assistants, and the great Information Technology team at the University of Toronto Libraries, *LEME* replaced *EMEDD*, at double its size.

LEME resembles a historical period dictionary, a text archive, and a diachronic corpus like EEBO/TCP when purposed for research in historical lexicography. Unlike any period dictionary, *LEME* relies only on dead lexicographers.[4] Like an archive, it enriches texts with editorial apparatus, encoded information about functional segmentation, language, and bibliography. *LEME* produces, in response to search requests, historical word profiles that often supplement information in the *OED*. These profiles tell us to which vocabulary a lexicon contributes (the mother tongue, or the new "hard" sub-languages of England's professions and guilds). Profiles also can be used to locate hitherto undocumented words, to give revised chronological limits, to identify etymology, and to document senses in the language of the times. *LEME* profiles may also be analyzed in groups, lexicon by lexicon, to uncover the contribution of an individual lexicographer to the English language, or decade by decade, to estimate the respective sizes of the "two tongues" of English (mother and hard).

LEME word profiles help us to understand the minutiae of Early Modern English, but only if dictionary headwords, whose spelling is unpredictable, are

manually mapped to the *OED* headword form. *LEME*'s lemmatization process has two stages. To any word-entry's form or explanation segment, I add a lexeme attribute that gives, for any explained word, its *OED*-headword spelling. Lexeme attributes are thus part of the base encoded lexicon text that we upload for conversion into the database. *LEME* uses these attributes to make a modern-spelling index to the important English words in its half-million word-entries. The second step in lemmatization uses different software. We have begun semi-automatically lemmatizing all English words whatsoever, lexicon by lexicon, chronologically from the beginning. In this way, *LEME* builds up a supplementary database of old-spelling-to-lemma equivalents. The more lexicons we process, the easier it is to lemmatize others. We use this lemmatization database to create a chronological index of English words. With this tool, we will be able to know when lexicographers used a word, and dropped it.

LEME word-searches, in this way, do not retrieve modern definitions but, instead, unruly descriptions. The *OED* elegantly defines the word "owl" in its various forms, *howle*, *howlet*, *oule*, *owl*, and *owle*, as "Any bird of prey of the order Strigiformes (which comprises the families Strigidae and Tytonidae), typically nocturnal and characterized by a large rounded head, raptorial beak, soft plumage, upright posture, and large eyes directed forwards and surrounded by a shallow cone of radiating feathers." This definition uses just enough information, general and then increasingly specific, to identify uniquely this bird in its avian world. Earlier lexicographers held themselves accountable to no such definitional standard. They describe a thing more than define a word. Thomas Thomas observes the owl's cat- and lion-like eyes (hence the adjective "owl-eyed") and listens to its howl (1587), a sound that Randle Cotgrave terms a "skreeke, or cry" (1611), and Thomas Blount a "whoop" (1656, antedating *OED* 1658). Being "hoodded, muffled about the head" (Florio 1598), it "sits in the day time" in a "solitarie place, or corner" (Cotgrave 1611), and – being a "night-bird" (Kersey 1702) – goes hunting "Mise and Rats" (Cotgrave). Captured, it is "tide to a stocke to catch other birdes with" (Florio 1598), a humiliation well described by Cotgrave: "a Fowler hid in a thicke bush, or tree, stucke full of lime-twigs, and hauing an Owle fast pearched neere to him, cries like a bird, and pinching a liue one, makes her crie; which others hearing, flie thither to rescue her from th'Owle, and so become intangled." The freedom of *LEME* glossographers from definition opens their entries up to capture new senses, and to refine their dating.[5]

Because *OED* does not cite about 95.7 percent of *LEME* word-entries, word-profiles like that on "owl" supplement the *OED* with antedatings, similes, anecdotes, and neglected senses. Each lexical text added to *LEME* brings a raft of new information. The 2,500 lemmas in John Stanbridge's *Vocabula* in 1510, for example, antedate the present *OED* on 75 occasions for terms like angling rod (1552), barber's shop (1579), the quinch (1571), scumming (1530), scythe stone (1688), strangullion (1547), and unweave (1542).[6] Some phrases such as "she dove" and "honey season," are not found in the *OED*.[7] I have not been able to identify about twenty words. Some seem *bona fide*, like "fusor" (for bell founder), "ulcerary" (another form of "ulcerative", perhaps), "the in ryne" (for

Latin "liber", perhaps a pellicle; cf. "rind," n. 1, 5b), and "in barke" (for Latin "codex"; cf. "mesophloem," n.).[8] Then half a dozen already documented words appear to bear an unexpected sense: "the lothe" (a disease), "a turne or a keruer", "the byrlynge yrons" (for Latin "fullonia forceps"), "bouty chese" (for Latin "pinguis"), and "a tacket / or a forest" (for Latin "saltus").

Researchers in lexicography want to attack bigger problems than the semantics of specific words: how big was the mother tongue, and how big the larger vocabulary that individuals recognized but seldom used? Early Modern English experienced "the fastest vocabulary growth in the history of English in proportion to the vocabulary size of the time", according to Terttu Nevalainen (1999) and others. Was there truly a vocabulary explosion in the sixteenth century? *Ordered Profusion* (Finkenstaedt, Wolff, Neuhaus and Herget, 1973), based on *OED* first-occurrence dates, identifies 1560–1660 as the peak period of expansion. To help answer those questions, *LEME* must supplement its dictionaries corpus by lemmatizing a large number of representative non-lexical texts, especially in the incunabular and early Tudor periods. The first text in this supplementary corpus consists of the sixteen Paston letters dated 1473, my *terminus a quo*, the year in which Caxton began printing. It would be reasonable to begin with the entire *Middle English Dictionary* headword list, except that it is not lemmatized to the *OED*. Once integrated with early lexicons like the Pepys manuscript of the *Medulla Grammatice*, *Promptorium Parvulorum* (1499), legal lexicons by John Rastell, and herbals by Peter Treveris and Richard Banckes, the combined lexical and supplementary text corpora give *LEME* a reasonably generous snapshot of the English language at the beginning of the Early Modern period.

By far the larger user-community for *LEME*, however, is not linguistic but literary and historical researchers. For all these people, even for New World immigrants like myself, *LEME* has something. One of the least expected *LEME* words is "Canada," an English word on loan from Inukitut "kanata" (meaning 'settlement'). Herbalist Thomas Johnston (1633) and glossographer Elisha Coles (1676) tell us that the Indian sun and the marigold of Peru, that is, the common sunflower, whose seeds are tasty if boiled with butter, vinegar, and pepper, had another English name, "Batfafas de Canada," which they Englished as "Potatoes of Canada." Two other *LEME* lexicographers, Richard Perceval and John Minsheu (1599), explain the quite different Spanish term "cañada" to mean "a cragge or cliffe, a rocke, a caue: a way to driue sheepe, a sheepe or goates walke": a sense that the *OED* already records in a late nineteenth-century glossary of mining terms. Although none of these terms instances semantic drift – only our habit of calling our omnipresent honking and defecating Canada geese "Canadas" shows semantic transfer – if we encode words in corpora like EEBO/TCP and *LEME* for their lemmas, researchers can trace elusive semantic derivations that can substantially change the way we read a text.

Ten years ago I used the *EMEDD* to show that the name of Shakespeare's first villain, Aron in *Titus Andronicus*, is not Aaron. Shakespeare did not take this name from the prose pamphlet that is the source for Titus' tragedy, but named

Aron after a very common weed otherwise called wake robin, the burning herb, or dragon's mace, cuckoo pintle, calf's foot, ramp, starchwort, priest's hood, and priest's pintle. Shakespeare believed that men resembled plants in having powers, virtues, and vices, and the audience that saw this Moor create havoc on the Rose Theatre stage in the early 1590s understood that Aron was the namesake of a common, noxious black-spotted weed in part because of his dramatic fate (which is unique in the period drama). The Romans bury him in the ground and leave him to starve, "breast-deep in earth" and "fast'ned in the earth" (5.3.179–83), a punishment that is true to his namesake.

To name a character after an English weed, turning a noun into a proper name, is semantic as well as part-of-speech derivation because the remorseless enemy of Titus Andronicus and his family, the ancestor of Iago in Shakespeare's *Othello*, assumes the features, powers, and uses of this weed as documented in sixteenth-century lexicons and herbals. Semantic derivation often manifests itself in metaphors like this. Richard Banckes' herbal (1525) notes that aron is "bytter and pryckynge vpon the tonge" and functions as a laxative and as a powder "to frete awaye the superfluyte of flesshe." Sir Thomas Elyot (1538) compares its leaves "to Dragons, but broder, and hauynge blacke spottes" and both he and William Turner (1548) affirm that it "groweth moche about hedges" as well as "in euery hedge almost in Englande about townes in the spryng of the yere." John Maplet (1567) adds that it "groweth only in shadowie places, and such as be hedged, so kept away fro*m* the Sunnes heate." Robert Dodoens' herbal (1578), dedicated to Elizabeth, has most to say about the black-spotted weed whose stalk, cod, or hose reminded the average European of an erect pintle or penis:

> Cockowpynt hath great, large, smoth, shining, sharpe poynted leaues, much larger than Iuy leaues, & spotted with Blackish markes of blacke and blew: amongst them riseth a stalke of a spanne long, spotted here & there with certaine purple speckles, and it carieth a certayne long codde, huske, or hose: open by one syde like the proportion of a haares eare, in the middle of the sayd huske, there groweth vp a certayne thing lyke to a pestel or clapper, of a darke murry, or wanne purple colour: the whiche after the opening of the velme or huske doth appeare, whan this is gone, the bunche or cluster of beries also or grapes, doth at length appeere, whiche are greene at the first, and afterwarde of a cleare or shining yellowish red colour, lyke Corall, and full of iuyce in eache of the sayde berries, is a small harde seede or twaine. The roote is swelling rounde lyke to a great Olife, or smal bulbus Onion, white and full of Pith or substaunce, and it is not without certayne hearie stringes by it: with much increase of small yong rootes or heades.

John Gerarde's *Herball*, published in 1597 after *Titus Andronicus*, offers a briefer description: aron is a small member of the family of dragons (682) with "spots of diuers colors like those of the adder" (681), is found in England, Africa, Egypt,

"generally in all places hot and drie, at least in the first degree", and ... "groweth in woods neere vnto ditches vnder hedges, euerie where in shadowie places" (p. 685). New to Gerarde's account is aron's importance in the production of starch. "The most pure and white starch is made of the rootes of Cuckowpint," he writes, "but most hurtfull for the hands of the laundresse that hath the handling of it, for it choppeth, blistereth, and maketh the hands rough and rugged, and withall smarting" (p. 685). Gerarde's illustrations reveal its erect, pintle-like stalk (see Figure 1).

Figure 1: *Arum maius*. Great Cockow pint. *Arum minus.* Little Cockow pint.

These illustrations and descriptions show that Shakespeare's Aron is a personified weed. Dragon-like, he is named "the devil" (5.1.145) and compared to an adder (2.3.35). He possesses the plant's black-"spotted, detested, and abominable" body (2.3.74) and its "bitter tongue" (5.1.150). His garb and sword visually highlight his namesake's pintle. Demetrius orders Aron to keep "your lath glued within your sheath/ Till you know better how to handle it" (41–42), alluding to the weed's erect sheath or calyx, its whorl of sepals that envelops what he later calls his "deadly-standing eye" (2.3.33) and what Lucius terms "wall-ey'd slave" (5.1.44).[9] Because Aron takes Tamora as his mistress, fathers her bastard, and assists in the rape of Lavinia, he associates well with the weed's

bawdy name "pintle." Last, Aron proves hard on the hands, as Gerarde says of the starchy plant: Titus loses one, and Lavinia two hands, to the Moor's treachery.

What prompted Shakespeare to name his first villain, Aron, after a weed? Herbals were a neglected field of science in the early 1590s, as John Gerarde, a gardener to William Cecil (Elizabeth's great statesman, the man who kept England together from the start of her reign, and patron to Gerarde's great book), says in 1597. Yet aron had a special importance for the Crown at this time: its role in the English starch industry (centuries later, its starchy core was called arrowroot). Phillip Stubbes in 1583 attacked the "liquide matter which they call Starch" as a devil's tool to make the same kind of "great ruffes" that we see Shakespeare wearing in the woodcut portrait to the 1623 folio.[10] Acting companies, like anyone who pretended to dress well, must have starched their clothes often. The Crown accordingly awarded Richard Young in 1588 "the exclusive right to import, make and sell starch" in England for seven years, one of a series of monopolies that Cecil promoted because they increased the Queen's revenues (Peckham 20). This monopoly was then transferred to Sir John Pakington on 6 July 1594 for another eight years (Hulme 1900: 49). When aron emerged as a potentially cheaper substitute for starch, the starch monopolists complained, the following year, about infringements against their rights, an unpopular lobbying that generated "violent attacks" (Peckham). Aron's association with a controversial vanity employed by acting companies, and disliked by puritans like Stubbes, probably explains why the Moor in *Titus Andronicus* was its namesake. Shakespeare attributed to his chief villain something featured in stage clothing. Semantic drift took place in the one area of his business life that occupied Shakespeare most.

Another example of semantic derivation occurs in Brutus' soliloquy in Act II of Shakespeare's *Julius Caesar*, a speech that must have sent shivers up the spine of many who gathered for its first performance at the Globe on 21 September 1599. The previous year's death of William Cecil, fears of a second Spanish Armada, troubles in Ireland, and their old queen's lack of a bodily successor put Londoners on edge. Brutus confesses his own fear at what he, Cassio and others had been mooting, a public assassination attempt on Caesar. He compares his state of mind to a dread-paralyzed council in a "little Kingdome":

Betweene the acting of a dreadfull thing,
And the first motion, all the Interim is
Like a Phantasma, or a hideous Dreame:
The Genius, and the mortall Instruments
Are then in councell; and the state of a man,
Like to a little Kingdome, suffers then
The nature of an Insurrection.[11]

These deceptively simple verses, much later, haunted T. S. Eliot when he penned the following lines in his apocalyptic 1925 poem, "The Hollow Men":

Between the idea
And the reality
Between the motion
And the act
Falls the Shadow

Like many editors of Shakespeare's *Julius Caesar*, however, Eliot inverted the meaning of Brutus' words by saying that a "hideous Dreame" occurred between the "first motion" and the "acting." Shakespeare said the opposite: the "acting" precedes the "motion." The *Oxford English Dictionary* defines the noun "acting," quoting Brutus' words as an illustration, as "The process of carrying out into action; performance, execution." But how can there be an "interim" between, as it were, an "action" (in this sense) and its "first motion"?

The simplest of words, a participial noun "acting" does not mean what either Eliot, or Shakespeare's editors, or the *Oxford English Dictionary* believe it does, although they all properly credit Shakespeare with lexical inventiveness elsewhere. When Brutus uttered "acting," he used as novel a word in English as two other words in his speech would shortly become – "interim" and "genius," both of which are plain Latin words, as their italicization in the First Folio signals, that anyone might expect a Roman like Brutus to speak. Contemporary Early Modern dictionaries in 1604 and 1607 are first to register these Latinisms as imported English words, as "hard words," which they still are to some of us. Robert Cawdrey (1604) explains "genius" as "the angell that waits on man, be it a good or euill angell." Dr. John Cowell, a legal lexicographer, is first to use "interim" as an English word in 1607. A third italicized, obviously Latinate word ("Phantasma") was translated by schoolmaster John Baret in 1574 as "A vaine vision, a false imagination: a vision of that which is not"; Shakespeare's fellow poet John Marston made it English in 1598. Was the noun "acting" as problematic for the Globe audience in 1599 as those other Latinate words would have been?

Here diachronic lexicography is instructive. The noun "act," with several meanings, goes back to the fourteenth century, but the verb "act" first turns up in 1594, coined by Shakespeare's fellow dramatists and poets Robert Greene and Michael Drayton from the old noun (by what is now termed zero derivation) to mean "to perform a command" and "to perform something on the stage." The present participle "acting" (as in the sentence "he is acting the role of Brutus") arrived automatically with the verb "to act," but the participial adjective and the participial noun forms for "acting" – two more zero derivations – soon followed. Poet Samuel Daniel first used "acting" as an adjective in 1597 ("The acting spirits"), and Shakespeare's Friar Lawrence used "acting" as a noun when he bent Juliet to his tragic, simulated-death plot in Romeo and Juliet (iv. 1.120; 1595). John Florio also used "acting" as a noun in his Italian-English dictionary of 1598: "Pre, vsed much in composition, and set before other wordes, as a going, or acting before." This lexical innovation in the vocabulary of Shakespeare's own craft took place chaotically. Writers in 1599 did not articulate multiple meanings

precisely. Words acquired and lost multiple senses, not by the prescription of the language industries and professions, but by the literary imaginations of individuals.

What seems to have happened is that Shakespeare transferred to his new noun "acting" a specific sense of the old noun "act," that is, "Something transacted in council, or in a deliberative assembly; hence, a decree passed by a legislative body, a court of justice, etc. (L. actum, pl. acta.)" (*OED*). The central metaphor of Brutus' soliloquy makes this meaning, "enacting," highly probable. He tells us that, when his guardian angel and the "mortall Instruments" of a state or kingdom are "in councell," a revolt may ensue. Brutus means by the noun "acting" something "transacted" by this inner "councell" of his genius, his good or bad angel, in advance of when its mortal instruments, whether they be his knife-wielding hand or a signed paper committing the conspirators to a murder, actually put this "decree" into effect. Enacting can, indeed, precede the "first motion."

It has taken four centuries to detect this meaning because even educated readers today interpret Shakespeare as if he were writing modern English, and the *OED* entry for the verbal noun "acting" does not offer the sense, "enacting." Yet, his verb "act," adopted from the identical noun in 1594, was as novel to him then as the verb "to text-message" is today: that phrase first occurred in 1994, a zero derivation from the noun, which appeared in 1978.

Why did Brutus' speech, as Shakespeare meant it, lead a Globe audience to shiver? I think that the explanation inheres in the powerful multivalent word "acting." The central metaphor of Brutus' soliloquy explicates this novel word as being about the enacting of decrees, but we do not need an expert in unpacking word-meaning to see that the man who performed the part of Brutus was also play-acting. Who, then, was responsible for plotting an assassination, a character or an actor? Plays in Elizabethan London were censored because they put dangerous ideas into their spectators' minds. And the worst idea was to kill a reigning monarch. Modern horror films like *Se7en* (1995) and *Zodiac* (2007) frighten with stories of serial killers who strike randomly at ordinary persons, not at prime ministers or kings. Shakespeare also struck fear into theatre-goers, but with stories of assassins like Sir Piers Exton *(Richard II*) and Sir James Tyrrell (*Richard III*) who, by targeting monarchs, threatened England with political and social disorders like the Wars of the Roses. Like them, Brutus strikes down social order itself by killing its guarantor.

Shakespeare contributed to semantic variation by investing a word that his fellow professionals coined, "acting," with his own novel sense, one that associated actors with political traitors. He would shortly learn, the hard way, the thin line that divided plays and treason. Friends of the earl of Essex would revive Shakespeare's old *Richard II* (1595), which depicts the murder of an English king, on 7 February 1601, as a prelude to their failed rebellion against Queen Elizabeth (Hammer 2008). Afterwards, Shakespeare's company must have had some serious explaining to do in the trial that led to the beheading of Essex.

Most discussion of semantic derivation generalizes from well-documented, successful changes in a word's meaning over many decades, such as the gradual shift of "silly" to denote foolish rather than innocent behaviour. Semantic drift would have been especially frequent in a century without published language standards. My three examples all failed to alter a word's meaning lastingly, to judge from the inability of editors and readers, after four centuries, to grasp what Shakespeare meant. Although the changes he rings on semantics occur in his mother tongue, where they might have had more impact than so-called "hard words," they emerged from and affected a narrow professional area, acting. One change reflected the interests of Shakespeare's new royal patron (his company was "the King's men"), and two others related to his job as an actor (insofar as he wrote and possibly played the parts of traitors, and insofar as players wore starched ruffs). The full-text and lexicographic corpora that brought to light these failed innovations thus illustrate the conditions in which short-lived semantic variation takes place. To that extent, I hope, they shed light on what Manfred Görlach refers to as "a difficult and largely unsolved problem."

Notes

1 All quotations from Shakespeare are taken from the *Riverside Shakespeare* (1997).

2 See *LEME* URL http://leme.library.utoronto.ca/lexicon/entry.cfm?ent=237-25637.

3 That *LEME*, a reasonable start on a period historical dictionary to follow Toronto's *Dictionary of Old English* and Ann Arbor's *Middle English Dictionary*, got published we owe to the generous support of the Canada Foundation for Innovation, the Social Sciences and Humanities Research Council, and a six-university research network, Text-analysis Portal for Research (TAPoR).

4 It now has 27 large lexicons. Eight are bilingual English and Latin works: the EETS edited text of the missing Monson manuscript of the Catholicon Anglicum, the Pepys MS of Medulla Grammatice (ca. 1480), *Promptorium Parvulorum* (1499), Sir Thomas Elyot's *Dictionary* (1538), dictionaries by Huloet (1552), John Withals (1556), John Baret (1574), Thomas Cooper (1584), and Thomas Thomas (1587). Another eight map English to a Romance language: French works by John Palsgrave (1530) and Randle Cotgrave (1611), Spanish by del Corro (1590), Stepney (1591), and John Minsheu (1599), and Italian by William Thomas (1550) and John Florio (1598, 1611). One, by William Salesbury (1547), is

Welsh-English. Ten are English monolingual dictionaries: Edmund Coote (1596), Robert Cawdrey (1604), John Cowell (1607), John Bullokar (1616), Henry Cockeram (1623), Thomas Blount (1656), Edward Phillips (1658), Wilkins (1668), Coles (1676), and Kersey (1702).

5 For example, *LEME* explanations indicate that owls come in several types. John Stanbridge's *Vocabula* (1510) antedates the first *OED* citation of a screech owl by more than eighty years (Shakespeare's in *2 Henry VI*, 1593). Thomas Thomas describes it as "an vnluckie kinde of bird (as they of olde time said) which sucked out the blood of infants lying" (1587), and John Higgins (1585) calls it a lich-owl for its interest in dead bodies. John Florio depicts the horn-coot "with feathers on each side of his head like eares" (1598, antedating *OED* 1650; cf. Elyot 1538). The owl also gives its name to the "sea-owle", a fish called a lump, paddle, or puddle, having "great eyes and teeth like a saw, it keepes euer by the shore" (Florio 1598; antedating *OED* 1601).

6 The others are "bedde borde" (1530), "bemynge knyfe" (1530), "blawbole" (1530), "boke seller" (1527), "botum" (1524; "bottom," n. 7), "brede bakynge / or bakers crafte" (1757; "bread," n. 1, 9), "chese racke" (1530), "coppys" (1538), "cowe house" (1530), "deynty mouthed" (1530), "dog flee" (1841), "donge pyke" (1530), "filipendulo" (?1540), "flat nosyd" (1530), "furryer" (1576), "fyre cryket" (1530), "fysshe catcher" (1530), "gad be" (1530), "gardyn mynt" (1530), "gome tothe" (1535), "grosser" (1545), "guttre stone" (1530), "handfastynge" (1530), "harowe pynne" (1530), "hedge sparowe" (1530), "herbe ryall or lurcke" (1530), "heruest man" (1530), "hey mower" (1530), "hey tyme" (1530), "hogges trough" (1530), "hony man" (1552), "hore heded" (1561), "horse myll" (1530), "horse tamer" (1530), "kell" (1530), "latter math" (1530), "nyght gnat" (1530), "pryket / or a tegge" ("tegge" 1537), puke" (1530), "ray fysshe" (1611), "redd hede" (1664), "renger" (1530), "ruen chese" (1539), "saw dust" (1530), "scryche oule" (1593), "seruys tree" (1530), "shere flockes" (1585), "slycer (1530), "slypper maker" (1889), "snayle shell" (1530), "swemynge in the hede" (1530), "swynes pockes" (1530), "sythe stone" (1688), "tarfyche" (1530), "the foreman of the shoppe" (1574), "the plough ere" (1530), "to berk" (1545), "to cloke" (1514), "to sclate" (1530), "to sesterne" (1587), "to syt a sonnynge / or to sonne" (sunning, to sun; 1519), "to twyfalowe" (1557), "to warble" (1530), "tymbre worme" (1530), "wake robyn" (1530), "water pompe" (1530), "waterysshe" (1530), "well stomacked" (1540), and "wyddynge chambre" (1552).

7 The others are "brede of the hande," "fery barge" (for Latin "hyppago"), "he doue," "hony faule" (for Latin "melligo"), "salt mete or sauce," and "to crye warre."

8 The others are "a befe" (for Latin "frons"), "a blanket," "a lodix," "a creuys" (for Latin "Licusta"), "a lana" (wool), "a nedle / or the b*uns*" (for Latin "acus"), "a pasuet" (for Latin "rapa"), "a wynde" (for Latin "galerita"), "a wynde" (for Latin "picus"), "a wynderlynge" (for Latin "nauum"), "aswole" (for Latin "mamella"), "glutter" (for Latin "gluttino" or glue), "ransyn" (for Latin "ramisis"), "the cornys" (for Latin "Arbutum"), "the pastures" (for Latin "suffragines"), "the vynbred of the salet" (for Latin "Hec buccula"), "to teue / or pounde" ("teue" … for Latin "tundo"), and "to wyder" (for Latin "perasco").

9 See "eye, n.", *OED*, 10.b–c (the remains of the calyx …. the centre of a flower).

10 *The Anatomie of Abuses* (1583): d8.

11 TLN 684–690. I quote from the First Folio edition (1623) at the Internet Shakespeare Editions Web site (http://ise.uvic.ca/Annex/Texts/JC/F1/Work).

References

Electronic and online sources

DSL. 2004–. *Dictionary of the Scots Language*. Edinburgh: Scottish Language Dictionaries. URL: http://www.dsl.ac.uk/dsl/index.html.

EEBO-TCP. 2000–. *Early English Books Online / Text Creation Partnership*. Ann Arbor: University of Michigan. URL http://www.lib.umich.edu/tcp/eebo/.

EMEDD. 1996–99. *Early Modern English Dictionaries Database*. Toronto: CHASS, University of Toronto.

Helsinki Corpus. 1991. *The Helsinki Corpus of English Texts*. Department of English, University of Helsinki. Compiled by Matti Rissanen (Project leader), Merja Kytö (Project secretary); Leena Kahlas-Tarkka, Matti Kilpiö (Old English); Saara Nevanlinna, Irma Taavitsainen (Middle English); Terttu Nevalainen, Helena Raumolin-Brunberg (Early Modern English). See http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/index.html.

LEME. 2006–. *Lexicons of Early Modern English*. Toronto: University of Toronto Press, and University of Toronto Libraries. URL: http://leme.library.utoronto.ca.

OED. Oxford English Dictionary Online. 2011. Oxford: Oxford University Press. http://www.oed.com/.

PPCEME. Kroch, Anthony, Beatrice Santorini, and Lauren Delfs. 2004. *The Penn-Helsinki Parsed Corpus of Early Modern English*. http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-2/index.html.

PPCME2. Kroch, Anthony, and Ann Taylor. 2000. *The Penn-Helsinki Parsed Corpus of Middle English*. 2nd edn. http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-3/index.html.

RET. 1997–98. *Renaissance Electronic Texts*. Toronto: CHASS, University of Toronto. URL: http://www.library.utoronto.ca/utel/ret/ret.html.

TAPoR. 2003–. *Text Analysis Portal for Research*. Hamilton, ON: McMaster University. URL: http://portal.tapor.ca/portal/portal.

Other

Banckes, Richard (1525), *Here Begynnyth a Newe Mater, the whiche Sheweth and Treateth of ye Vertues [and] Proprytes of Herbes, the whiche is Called an Herball*. London: Rycharde Banckes. STC 13175.1.

Darrel, John (1600). *True Narration of the Strange and Greuous Vexation by the Devil, of 7. Persons in Lancashire, and William Somers of Nottingham*. STC 6288.

Dodoens, Rembert (1578), *A Nievve Herball, or Historie of Plantes*. London: Gerard Dewes. STC 6984.

Eliot, T. S. (1925), *Poems, 1909–1925*. London: Faber and Gwyer.

Elyot, Sir Thomas (1538), *The Dictionary of Syr Thomas Eliot*. London: T. Bertheleti.

Finkenstaedt, Thomas and Dieter Wolff, with contributions by H. Joachim Neuhaus and Winfried Herget (1973), *Ordered profusion: studies in dictionaries and the English lexicon*. Heidelberg: Carl Winter.

Finkenstaedt, Thomas, Ernst Leisi and Dieter Wolff (1970), *A chronological English dictionary listing 80 000 words in order of their earliest known occurrence.* Heidelburg: Carl Winter.

Florio, John (1598), *A Worlde of Wordes, or, Most Copious, and Exact Dictionarie in Italian and English.* London: Arnold Hatfield for Edward Blount.

Gerarde, John (1597), *The Herball or Generall Historie of Plantes.* London: Edm. Bollifant for Bonham Norton and John Norton. STC 11750.

Görlach, Manfred (1991), *Introduction to Early Modern English*. Cambridge: Cambridge University Press.

Grenander, M. E. (1977), '*Macbeth* IV.i.44–45 and convulsive ergotism,' *English Language Notes,* 15: 102–103.

Hammer, Paul E. J. (2008), 'Shakespeare's *Richard II*, the play of 7 February 1601, and the Essex rising,' *Shakespeare Quarterly* 59: 1–35.

Hulme, E. Wyndham (1900), 'The history of the patent system under the prerogative and at Common Law,' *Law Quarterly Review* 16: 44–56.

James I (1924), *Daemonologie (1597) Newes from Scotland declaring the Damnable Life and death of Doctor Fian: a notable Sorcerer who was burned at Edenbrough in Ianuary last (1591)*. London: J. Lane.

Jones, Richard Foster (1953), *The triumph of the English language: a survey of opinions concerning the vernacular from the introduction of printing to the Restoration*. Stanford: Stanford University Press.

Kytö, Merja (comp.) (1996), *Manual to the diachronic part of the Helsinki Corpus of English Texts. Coding conventions and lists of source texts*. Helsinki: Department of English, University of Helsinki. URL http:// http://khnt.hit.uib.no/icame/manuals/HC/INDEX.HTM.

Lancashire, Ian (1997), 'Understanding Shakespeare's *Titus Andronicus* and the EMEDD,' in: Ian Lancashire and Michael Best (eds.) *New scholarship from old Renaissance dictionaries: applications of the Early Modern English Dictionaries Database*. A special issue of *Early Modern Literary Studies*. URL: purl.oclc.org/emls/emlshome.html.

Lancashire, Ian (2002), '"Dumb Significants" and Early Modern English definition,' in: Jens Brockmeier, Min Wang and David R. Olson (eds.) *Literacy, narrative and culture*. Richmond, Surrey: Curzon, 131–154.

Lancashire, Ian, John Bradley, Willard McCarty, Michael Stairs and Russ Wooldridge (1996), *Using* TACT *with electronic texts: a guide to text-analysis computing tools, version 2.1 for MS-DOS and PC DOS*. New York: Modern Language Association of America.

Maplet, John (1567), *A Greene Forest, or A Naturall Historie*. London: Henry Denham. STC 17296.

Minsheu, John (1599), *A Dictionarie in Spanish and English, first published into the English tongue by Ric. Perciuale*. London: Edm. Bollifant.

Nevalainen, Terttu (1999), 'Early Modern English lexis and semantics,' in: Roger Lass (ed.) *The Cambridge history of the English language*. Volume 3: *1476–1776*. Cambridge: Cambridge University Press, 332–458.

Peckham, Brian W. (1986), 'Technological change in the British and French starch industries, 1750–1850,' *Technology and Culture* 27 (1): 18–39.

Shakespeare, William (1997), *The Riverside Shakespeare*. Ed. G. Blakemore Evans. 2nd edn. Boston: Houghton Mifflin.

Stanbridge, John (1510), *Vocabula Magistri Stanbrigi*. London: Wynkyn de Worde. STC 23178.

Turner, William (1548), *The Names of Herbes in Greke, Latin, Englishe, Duche and Frenche with the Commune Names that Herbaries and Apothecaries Vse*. London: S. Mierdman for John Day and William Seres. *STC* 24359.

Corpora and the study of the history of English

Matti Rissanen

University of Helsinki

Abstract

In the last four decades, the use of corpora has become the standard method of analysing and explaining the diachronic development and synchronic stages of the English language. Corpora have reduced the time spent on finding evidence for linguistic phenomena of past centuries and exercised a significant influence in bringing theoretical and use-based analyses closer together. They have also improved the potential for pragmatic and discourse-based analysis of the history of English, focusing on the negotiation of meaning between the speaker/writer and hearer/reader.

The present paper offers a brief survey of the history of historical corpora over the last few decades. Chronological, regional and genre coverage of existing corpora is described and attention called to the areas in which new corpora are needed. A corpus-based analysis of the development of the adverbial sub-ordinator provided (that) *from Middle to Present-day English illustrates the kind of information that can be found in English historical corpora.*

1. Introduction

The last four decades have witnessed remarkable new initiatives in the study of the history of English, thanks to the emergence and establishment of corpus linguistics. The use of corpus evidence has become the standard method for analysing and explaining the diachronic development and synchronic stages of the language of the past, from the beginnings of English to our own time.[1]

The advances brought about by corpus linguistics are closely connected with the new developments in the evidence-based variationist approach to the study of the history of English. The basis of this approach is the analysis of various linguistic ways of expressing roughly the same meaning, the lexico-grammatical potential of language, to use Halliday's terminology (1973), attention being focused on the language-external factors affecting the choice of the variant expression. Change in language is largely explained through changes in the quality and strength of these factors while keeping the language-internal trends of change in mind at the same time.

The most obvious language-external factors causing variation and change can be grouped in the following way:

- Sociolinguistic variability and the influence of social change and mobility on the development of language, including gender, participant relationship, level of formality and social networks.
- Textual variability, i.e., the interrelation between the genre or topic of text and linguistic expression, including discourse situation and medium.
- Regional variability and the differentiation and amalgamation of dialects and regional varieties, including contact phenomena.

Based on these language-external factors, the historical linguist will particularly observe the following types of large-scale change within a speech community:

- Changes in the structure of society (e.g., weakening of class society; change from agrarian to industrial; urbanisation).
- Changes in culture and learning (e.g., religious innovation; increasing literacy; new developments in science and learning).
- Changes in foreign policy (e.g., invasions; immigration; other contacts with speakers of foreign languages).

Corpora have radically reduced the time and diminished the toil and trouble in finding evidence for linguistic phenomena of past centuries. As a result, corpora have exercised a significant influence on bringing theoretical and use-based analyses closer together. There is now little excuse for offering research findings on the history of language based on dictionary evidence only (unless the dictionary with its apparatus of examples is also available as a massive corpus, as is the case with the *Middle English Dictionary* and the *Oxford English Dictionary*). Conversely, simply collecting examples of a linguistic phenomenon and presenting this material without any generalizations or theoretical considerations is much less acceptable now than it was in the early years of historical language studies when years of reading was the only way to collect evidence.

The new opportunities offered by extensive and wide-ranging corpus evidence have also meant a considerable impetus for pragmatic and discourse-based analysis of the history of language, focusing on the negotiation of meaning between the speaker/writer and hearer/reader (see, e.g., Traugott and Dasher 2002). Collecting sufficient and reliable evidence for this kind of analysis, with due attention to the external factors causing variation and change, would be practically impossible without the support of corpora.

We could even go as far as to say that without the support and new impetus provided by corpora, evidence-based historical linguistics would have been close to the end of its life-span in these days of rapid-changing life and research, increasing competition on the academic career track and the methodological attractions offered to young scholars.

There are also other, less obvious and less direct ways in which corpora have provided new initiative for research. Both multi-genre, multi-purpose corpora and those focusing on a particular genre or regional variety encourage –

indeed almost compel – the historical linguist to pay attention to interdisciplinary aspects of his or her research. Corpora may also be valuable resources for scholars of other disciplines. The relationship between corpus-based linguistic research and cultural, social and even political history are the most obvious. The intricacies of class society can be approached through corpora, including private and official correspondence; regional corpora supply information on immigration and contacts and relationships with other nationalities; trends in legislation can be approached through corpora concentrating on statutes and official documents; the findings in natural science and medicine in centuries past appear in medical text corpora, and so on.

In the present paper I will first offer a brief survey of the history of historical corpora over the last few decades. I will then discuss the chronological and regional or genre coverage of existing corpora and call attention to the areas in which new corpora are most urgently needed. Finally, I will give an example of the use of various corpora in the study of the development of the adverbial subordinator *provided (that)* from Middle to Present-day English.

2. On English historical corpora

2.1 A short history of English historical corpora

It is certainly not a mere coincidence that the idea of creating English historical corpora and databases appeared soon after the first systematic publications describing change through variation had come out. To mention only some of the pioneers of the variationist approach, Weinreich, Labov and Herzog published their powerful formalization of variation-based historical linguistics in 1968, while Samuels (1972) applied this approach to the development of English. Romaine (1982) developed the theory of variation in her discussion of a more focused syntactic topic, relative clauses, and James and Lesley Milroy successfully combined aspects of change with the analysis of present-day spoken varieties of English in a number of monographs and articles (e.g., 1978; Milroy, J. 1985; 1992; Milroy, L. 1987). Another source of inspiration for the compilers of English historical corpora was the *Survey of English Usage* and the epoch-making introductory chapter in the *Grammar of Contemporary English* (Quirk et al. 1972: 2–32), published again in revised form in the *Comprehensive Grammar of the English Language* (Quirk et al. 1985: 3–34). Present-day English corpora, the *Brown Corpus* of American English, issued in the 1960s, and the *Lancaster-Oslo/Bergen Corpus* of British English, published in the following decade, established the basic pattern and encouraged the planners of historical corpora.[2]

The earliest electronic resource containing substantial information on the history of English was the *Dictionary of Old English Database*, created in the late 1970s and arising from the Toronto *Dictionary* project. The *Augustan Prose Sample* and the *Century of Prose Corpus*, which consist of late 17th and 18th century texts and completed in the early 1980s, were early attempts for more

focused corpora. The *Helsinki Corpus of English Texts* project was started in the 1980s, and the diachronic corpus was completed and ready for general distribution in 1991.

The Helsinki Corpus, which covers the period from the eighth to the early 18th century, was the first attempt to create a structured long-diachrony corpus of English texts. The remaining centuries up to our own times were soon covered by *A Representative Corpus of Historical English Registers* (ARCHER).[3]

The 1990s was essentially a period of more focused historical corpus projects, targeted mainly at late Middle and Early Modern English. The corpora which resulted from these projects provide valuable detailed information on the influence of the various external factors listed above: sociolinguistic, genre-based and regional (see below, under 2.2.2 and 2.2.3). These focused corpora were supplemented by corpora concentrating on one author only. The most notable among the large multi-purpose corpus projects was the *Middle English Compendium*, which includes the *Corpus of Middle English Prose and Verse*, and the *Middle English Dictionary Database.* The electronic version of the *Oxford English Dictionary* was also issued in the 1990s.

The compilation of versions equipped with grammatical and syntactic tagging was another important development in the field of historical corpora. The first version of ARCHER was pioneering, and the *Penn-Helsinki Parsed Corpus of Middle English Texts* and *Penn-Helsinki Parsed Corpus of Early Modern English Texts*, completed in the 1990s, include enlarged versions of the corresponding parts of the Helsinki Corpus with detailed morphological and syntactic tagging. The *York-Toronto-Helsinki Parsed Corpus of Old English Prose* was completed a little later. The tagged and parsed version of the *Corpus of English Correspondence*, prepared at the Research Unit for Variation, Contacts and Change in English (VARIENG), in Helsinki, in cooperation with York University, was completed in 2006.

WebCorp, which uses the internet as data source, began a new era in the creation of corpora, although its usefulness for illustrating the earlier history of English is understandably restricted. Another development, supported by the internet and inspired by gigantic present-day corpora, is the compilation of huge historical corpora: the hundred-million-word *Time Magazine Corpus*, consisting of ten million words from each decade, covers the period from 1923 to 2006, and the *Corpus of Historical American English* (some 400 million words) comprises 19th- and 20th-century texts. The commercial *Literature Online* (LION) and *Early English Books Online* (EEBO) are excellent, although costly, resources.

2.2 Coverage of available corpora and the need for new ones

At the moment, there are some thirty major English historical corpora available, multi-purpose or focused, covering the main groups of extralinguistic factors causing variation and change in English fairly satisfactorily. By a rough count, they amount to more than 40 million words, excluding 20th century corpora. The

available corpora give a fair picture of the development of English vocabulary and grammar from the earliest times to our own days. The chronological coverage of the corpora is uneven, however, and does not give us a sufficient amount of information on all genres or regional varieties, or the language use of different social groups. More corpora are needed and their use should be made easier and more efficient by new software developments, as regards both search engines and annotation.

In the following, the extant corpus resources within each major period of English are briefly surveyed.

2.2.1 Old English

The situation in Old English is the most satisfactory. Practically all extant texts, over three million words, can be found in the *Toronto Dictionary of Old English Corpus*, which includes more than one manuscript of many texts. The most desirable developments for the Toronto Corpus are parameter coding and a retrieval program that would make it easier to sort out the texts by date, dialect, and genre, and to create partial corpora according to these parameters. The Old English part of the Helsinki Corpus, although its size is only one-seventh of the Toronto Corpus, is helpful here. The *York-Toronto-Helsinki Parsed Corpus of Old English Prose* (c. 1.5 million words) and the *York-Helsinki Parsed Corpus of Old English Poetry* (c. 70,000 words) are valuable in grammatical studies of Old English.

2.2.2 Middle English

The available Middle English corpora also offer a good basis for wide-ranging research, although the degree of coverage of all extant texts achieved in Old English would of course be impossible. The structured Middle English part of the Helsinki Corpus (608,000 words) offers a feasible starting-point, particularly when supplemented by the *Penn-Helsinki Parsed Corpus of Middle English*, which is about double the size of the Helsinki Corpus, and the 3-million-word sampler corpus of the *Innsbruck Computer Archive of Middle English Texts*. The two last-mentioned corpora only include prose texts, which to some extent diminishes their usefulness, particularly in vocabulary studies. The half-a-million word *Corpus of Middle English Medical Texts* extends from mid-14th to the end of the 15th century and gives us valuable information on the early development of scientific writing and vernacularisation. The earliest letters included in the *Corpus of Early English Correspondence*, whose Middle English part amounts to c. 400,000 words, date from the 1420s. *A Corpus of English Religious Prose*, being prepared in Cologne, will include a large number of Middle English religious texts. Of the corpora representing regional varieties, the *Helsinki Corpus of Older Scots* includes texts dating from the 15th century, a period crucially important from the point of view of the earliest development of the Standard or

standards. The earliest texts of *A Corpus of Irish English* date from the 13th century.

The cornerstone of Middle English corpus studies is, however, *The Middle English Compendium*, a marvellous product of the half-century-long *Middle English Dictionary* project. The *Compendium* is divided into three parts: an electronic version of the *Middle English Dictionary*, a *HyperBibliography of Middle English Prose and Verse*, based on the Dictionary bibliographies, and a *Corpus of Middle English Prose and Verse*. The electronic *Dictionary* allows searches both by the head-word and by the words and phrases included in the huge collection of examples. The corpus part includes almost 150 complete texts and, by a rough estimate, 17 or 18 million words, both prose and verse. The main problem of this corpus is its search program, which is less user-friendly: the user must check the occurrences text by text, as they are not given as one file. The texts of the corpus have not been systematically selected but its coverage is excellent: statutory and other administrative texts, Chaucer's major writings and other important literary works, the *Paston Letters*, etc. We can regard the Middle English corpus situation as highly satisfactory even at present largely thanks to this corpus.

Other invaluable new sources are the *Linguistic Atlas of Early Middle English* (LAEME) and the *Linguistic Atlas of Older Scots* (LAOS). These are up-to-date, corpus-based 'daughter atlases' of the *Linguistic Atlas of Late Mediaeval English*, which was compiled a few decades ago. The first versions of both LAEME and LAOS are now available, although both projects are still on-going.

As to future developments, focused Middle English corpora including all or most manuscripts of some more popular texts would be extremely valuable. The LAEME and LAOS projects offer a good start in this area. LAEME includes samples of more than one manuscript of *Ancrene Riwle* and of various other Early Middle English texts. A structured and coded corpus of statutes and documents from the end of the 14th century on would also be useful. We have a good start with the corpus consisting of the Parliament Rolls (1275–1504).

2.2.3 Modern English

The student of Early Modern English can start his or her research with the half-a-million-word Early Modern English part of the Helsinki Corpus and the 1.7-million-word *Penn-Helsinki Parsed Corpus of Early Modern English* and continue with more focused second-generation corpora which provide information on regional variation (Older Scots, Irish), sociolinguistic variability (*Parsed Corpus of Early English Correspondence*), genre variation (*Lampeter Corpus of Early Modern English Tracts*, the *Zurich English Newspaper Corpus*, *Early Modern English Medical Texts*, and the *Lexicons of Early Modern English* (derived from the *Early Modern English Dictionaries Database*, see Lancashire in this volume). Speech-based expressions in 17th and 18th century texts can be approached through *A Corpus of English Dialogues 1560–1760.*

Three multi-purpose corpora, the *Penn Parsed Corpus of Modern British English*, ARCHER (see note 3), and the *Corpus of Late Modern English Texts* cover the later Modern English period from the mid-17th to the 20th century. The size of CLMET, which is based on the texts included in the Gutenberg Project and the Oxford Text Archive, is impressive, roughly ten million words, as against the c. one million words of PPCMBE and less than two million of ARCHER. While its accuracy in reproducing the original texts may not be ideal, it is certainly adequate for lexical and syntactic studies. American English texts from the beginning of the nineteenth century on can be found in the huge 400 million-word *Corpus of Historical American English.*

It is also possible to study linguistic developments in the 20th century by comparing the *Lancaster-Oslo/Bergen Corpus* (British English) and the *Brown Corpus* (American English), which represent the language of the 1960s, with the *Freiburg-Lancaster-Oslo/Bergen Corpus* and the *Freiburg-Brown Corpus*, which include texts taken from corresponding genres but written in the 1990s. American English from 1923 until the 21st century can also be approached through the 100 million-word *Time Magazine Corpus.*

At the time of the updating of the present survey (August, 2011), the *Corpus of Late Modern English Medical Texts* and the 18th-century continuation of the *Corpus of English Correspondence*, both compiled at VARIENG in Helsinki, are nearing completion. There are also a number of highly focused corpora in preparation, including witchcraft pamphlets, tobacco pamphlets, bluestocking letters, pauper letters, etc. New LOB versions, which will cover the earlier decades of the 20th century, are in preparation at Lancaster (see Leech, Smith and Rayson in this volume).

A Modern English corpus of legal and documentary texts, a continuation of the corresponding Middle English corpus mentioned above, would be useful.

A series of historical multi-purpose corpora of regional varieties of English would be an important future project. At present, we have historical corpora of Older Scots, Irish English, Canadian English and Australian English. A corpus of Early American English and one of historical New Zealand English are in preparation. Particularly interesting additions to this list might include historical corpora of Indian English and South-African English. With corpora of this kind, we would have a historical counterpart to the *International Corpus of English*, which would certainly open new vistas in the study of immigrant varieties, and the influence of contact and new environments on the English language.

There are also other developments to be considered in the world of English historical corpora. One is the possibility of creating corpora which would give more multiplex information using the internet. Corpora of this kind could include the original manuscript of the text, the edited transcript and comments, explanations and background information either already available on the internet or prepared and edited for the purpose. This information would of course be available via appropriate links. The focused genre corpora in preparation will in all probability be developed in this way.

2.3 Corpora and interdisciplinary research

As was mentioned at the beginning of this paper, existing corpora and the corpus projects in progress can promote and facilitate research even in other disciplines. This interdisciplinarity is natural in view of the present trend in linguistic research which emphasizes the influence of extralinguistic factors on variation and change in language.[4]

Old English corpora include a large number of texts which offer excellent material for the study of the introduction and establishment of Christianity in Anglo-Saxon England, the earliest English society and Anglo-Saxon concepts of the surrounding world. For instance, the short text of the voyages of Ohthere (Ottar) and Wulfstan on the Baltic Sea and around the North Cape and Kola Peninsula to White Sea is unique and of immense value as the earliest detailed geographical description of the Northern world.[5]

Middle English corpora, particularly the *Corpus of Middle English Prose and Verse*, the *Innsbruck Archive* and the *Penn-Helsinki Parsed Corpus*, which consist of full texts, offer useful material not only for the study of early literature but also for socio-historical research. In the Late Middle English and Early Modern periods, letter corpora are invaluable for this purpose, and medical corpora provide information on the early stages of natural sciences. Newspapers from the 17th and 18th centuries can be studied in the *Zurich English Newspaper Corpus.*

2.4 *Corpus Resource Database*

In 2007, a database incorporating relevant basic information on existing English corpora, both historical and present-day, was initiated at VARIENG in Helsinki. At the time of writing this survey, this *Corpus Resource Database* (CoRD) contains information on close to 50 corpora, a number which is steadily increasing. The editors of CoRD hope that the compilers of new English corpora will send relevant basic information on their corpora for inclusion in this database. The address is www.helsinki.fi/varieng/CoRD.

3. The rise and development of the connective *provided (that)*

In the remaining part of this paper I will give an example of the use of major historical corpora in analysing the emergence and development of a syntactic-lexical detail in the English language, the adverbial connective *provided (that).* The story of the grammaticalisation of this item is fairly simple but, at the same time, typical of English adverbial subordinators.

3.1 Middle English

The adverbial subordinator *provided (that)* appears in late Middle English. It is worth noting that the verb *provide* was not borrowed from French but directly from the Latin *providere* (*OED*, s.v. *provide* v.; *MED* s.v. *providen* v.). The first question worth asking is whether the grammaticalised subordinator *provided (that)* developed from the verb in Middle English or whether it goes back to the Latin connective *proviso quod.* Information from dictionaries and corpora implies that the second alternative is more likely: both the verb and the connective appear at roughly the same time in Middle English, the first half of the 15th century (examples (1) and (2)).[6]

(1) Sant gregor gaf ansuer honest,
And o þat man þat was in were
þe soth he sceud him al clere,
And **prouide** him, wit quik resun,
þat at þis resurrectioun,
Wit all his limes hale and fere
Sal cum befor þe demstere.
(a1400 Cursor Mundi Cotton Vesp. A 3 22914–22920 CMEPV)[7]

(2) The remayndre of hem to the seid Erle and to heirs for ever, **provyded ever, that**, ȝif it better plese … (1430 Doc in Flasdiek Origurk. 94 *MED*)

Example (1) is, however, the only pre-1410 instance of the verb *provide* out of some 180 occurrences of the *provide* stem in CMEPV. Since the other Middle English corpora (see Section 2.2.2 above) and the dictionaries confirm the late introduction of the verb, it is hardly likely that there would have been sufficient time for the grammaticalisation of the connective from the verbal stem in the Middle English period. The influence of the verbal use on the later grammaticalisation of *provided (that)* is discussed below.

If the foreign origin of the subordinator use is accepted, *provided (that)* belongs to the same group of Middle English adverbial connectives as *except* (from Latin) or *save* (from Old French). With these two connectives, too, the verb appears in Middle English at roughly the same time as the connective (see Rissanen 2002; 2009).

The distribution of occurrences in Middle English corpora indicates that the subordinator *provided (that)* was introduced into the language through official writing, i.e., documents and statutes, Chancery or Signet Office texts (see Table 1).

Table 1: *Provided (that)* in the Middle English part of the Helsinki Corpus, the *Corpus of Middle English Prose and Verse*, and the *Innsbruck Computer Archive of Middle English Texts* (absolute figures)

	HC (ME4) 1420–1500	**CMEPV**	**ICAMET**
Statutes and documents	10	29	18
Other	–	2	2

All the ten instances in the Helsinki Corpus occur in statutes or official documents (examples (3) and (4)). The prevalence of the officialese is also obvious in the larger CMEPV and ICAMET. *Provided also* and *provided always* were, in fact, set idioms in the statutes, see examples (4)–(5).

(3) be it aggreed and accorded by the same auctorite, that oure sov~eygne Lord the kyng … have full power … to graunt to ev~y of such p~sones p~teccion … **Provided that** this acte be not available to eny p~sone for eny entre sen the first day of this p~sent p~liament. (1488–1491 Statutes 2,529 HC)

(4) the l~res patentes … stand and be goode and effectuell to the same Thomas after the teno=r= and effecte of the same l~res patentes, the seid Acte not withstondyng. **Provided also that** this acte extend not ne be p~judiciall to Henry Erle of Northumberlond (1488–1491 Statutes 2,533 HC)

As with most adverbial connectives, the loss of the particle *that* marking the subordinate clause can be regarded as a sign of the continuation of the grammaticalisation process.[8] The earliest instances can be found in late Middle English texts (example (5)).[9]

(5) **Provided** alwey, this Acte..extende not..unto oure Oratrice and true Bedewoman, Petronille Mounferant. (1464 RParl. 5.542b *MED*)

One instance can be found of *provided* used with a noun phrase in Middle English corpora (example (6)).

(6) and I shal fight ayenst them thre without fawte, **prouyded** alwayes the noble and juste jugement of your Court / one after another (c. 1500 Melusine 78 MECPV)

Neither *MED* or *OED* recognize this usage, which was never established in English. In this respect *provided* differs from most other adverbial connectives which can be used both as subordinators and prepositions.

3.2 Modern English

For brief descriptions of Modern English corpora, see Section 2.2.3, above. In the Early Modern English period, the subordinator *provided (that)* gradually extends its domain beyond officialese. In the Helsinki Corpus and the *Penn-Helsinki Corpus of Early Modern English*, there are occurrences in historical, philosophical, religious and scientific texts (Fabyan, translation of Boethius, Boyle, Clowes), handbooks (Markham, Langford), law court trials (Throckmorton) and even fiction (John Taylor, Aphra Behn, *Penny Merriments*).[10] As can be seen from the figures in Table 2, there seems to be some increase in the use of *provided (that)* in the group "Other" in the second half of the 17th century; the occurrences are, however, few in comparison to those in statutes and other official texts, in which the idiomatic use continues.

Table 2: *Provided (that)* in the Early Modern English parts of the Helsinki Corpus and in the *Penn-Helsinki Parsed Corpus of Early Modern English*. Absolute figures. (Figures per 100,000 words in brackets.)

HC	EModE1 1500–1570	EModE2 1570–1640	EModE3 1640–1710
Statutes and official corresp.	18 (99.5)	13 (74.5)	24 (126.0)
Other	1 (0.6)	2 (1.2)	7 (4.6)
PPCEME			
Statutes[11]	56 (222.2)	42 (165.6)	58 (204.7)
Other	2 (0.4)	6 (1.0)	9 (1.7)

The use of the subordinator in a colloquial context, as in example (7), is remarkable:

(7) The Rival cruelly vext; got a red hot iron, and comes again, tell her he had brought her a Ring, **provided** she would give him another kiss; (1684–85 Penny Merriments 159 HC)

Also in Shakespeare's comedies, see (8):[12]

(8) HORTENSIO I promised we would be contributors,
And bear his charge of wooing, whatsoe'er.
GREMIO And so we will, **provided that** he win her.
GRUMIO I would I were as sure of a good dinner.
(Taming of the Shrew, 1.1. 214–217 Oxford Shakespeare)

The *Corpus of English Dialogues* confirms that *provided (that)* can also be used in less formal dialogue contexts, including comedy and fiction; see Table 3:

Table 3: The connective *provided (that)* in the *Corpus of English Dialogues.* Absolute figures. (Figures per 100,000 words in brackets.)

Corpus of English Dialogues	
1560–1599	0
1600–1639	4 (2.0)
1640–1679	5 (2.5)
1680–1719	7 (2.4)
1720–1760	5 (2.2)

The figures are low and there is no increase in frequency in the dialogue genres from the beginning of the 17th to the mid-18th century. The restricted use of this subordinator is also implied by its rarity in private correspondence, only five isolated instances occurring in the 2.1-million-word *Parsed Corpus of Early English Correspondence* (1410–1681). This suggests that *provided (that)* remained a somewhat literary expression and probably never became a natural element in real-life informal spoken language.

The figures given by the *Lampeter Corpus* (1640–1740, c. 1.2 million words of pamphlet texts) indicate that *provided (that)* did not become popular even in more formal contexts outside legal or documentary language in Early Modern English. This corpus includes 41 instances of *provided (that)*, or 3.5 per 100,000 words.

This survey of corpus evidence suggests that the establishment of *provided (that)* in Early Modern English was decisively supported by its frequent, partly formulaic use in statutory language. However, it never became a high-frequency connective, despite also being accepted in less formal kinds of writing and even in colloquial written dialogue. Its use seems to have reached more or less the status it has in present-day English about 1700.

Information on later Modern English usage can be derived from the *Representative Corpus of Historical English Registers* (ARCHER) and the *Corpus of Late Modern English Texts* (CLMET), among others. The size of ARCHER is c. 1.7 million words, both British and American English, covering the period from the mid-17th century to the 1990s. The figures for *provided (that)* can be seen in Table 4.

The frequencies are low and fairly uniform except for the somewhat unexpected peak in the first half of the 19th century. The absolute figures are, however, so low that no conclusions can be based on them. Further research might inquire whether the printed prose of the early 19th century shows systematic stylistic differences from the use of language in the preceding or following decades.

Table 4: The connective *provided* (*that*) in ARCHER. Absolute figures. (Figures per 100,000 words in brackets.)

ARCHER	
1650–1699	6 (3.3)
1700–1749	2 (1.1)
1750–1799	5 (1.1)
1800–1849	13 (7.2)
1850–1899	6 (1.7)
1900–1949	5 (2.8)
1950–1997	8 (2.2)
Total	45 (2.5)

Genrewise, CLMET consists of a less varied selection of texts than ARCHER, but its larger size, c. 10 million words, makes the results more reliable. Table 5 confirms the early stabilization of the low frequency of *provided (that)*:

Table 5: The connective *provided* (*that*) in CLMET. Absolute figures. (Figures per 100,000 words in brackets.)

CLMET	
1710–1780	97 (4.6)
1780–1850	79 (2.1)
1850–1920	61 (1.5)
Total	237 (2.4)

The higher frequency of the connective in the 18th century part of CLMET is of some interest. As this corpus is based on long samples of text, it is easy to trace the authors who particularly favour *provided.* Table 6 lists those who use the subordinator *provided (that)* more than five times per 100,000 words.

The total number of authors in CLMET is 72 (15, 29, and 28 in the three sub-periods respectively). Some authors in the 1710–1780 period seem to be more fond than others of using *provided (that)*. The high frequency in the two texts by Sterne is striking. It is possible that the increase in the popularity of *provided (that)* can be explained by stylistic trends and/or authorial preferences but, as mentioned above, more research is needed on the possible influence of stylistic factors on the choice of less frequent and formally-sounding adverbial connectives.

Our excellent set of late 20th-century corpora, LOB and F-LOB, and Brown and Frown, representing British English and American English from the1960s and 1990s provide information on recent change within one generation; see Table 7.

Table 6: Texts and authors favouring the connective *provided (that)* in CLMET. Absolute figures. (Figures per 100,000 words in brackets.)

CLMET	
1710–1780	
Sterne, *Tristram Shandy, Sentimental journey*	21 (13.3)
Chesterfield, *Letters to his son*	17 (8.5)
Smith, *Wealth of nations*	16 (8.0)
Hume, three philosophical texts	11 (5.6)
1780–1850	
Hogg, *Memoires of a sinner*	8 (9.5)
De Quincey, *Confessions opium-eater*	3 (7.7)
Anne Bronte *Agnes Grey, Wildfell Hall*	13 (6.5)
1850–1920	
Butler, three texts, varying genres	12 (5.9)

Table 7: The subordinator *provided* (*that*) in late 20th-century corpora. Figures for miscellaneous documentary texts and scientific texts (H–J) and fiction (K–R) given separately. Absolute figures. (Figures per 100,000 words in brackets.)

	A–R	H–J	K–R
LOB (BrE 1960s)	48 (4.8)	26 (11.8)	6 (2.7)
Brown (AmE 1960s)	22 (2.2)	11 (5.0)	2 (0.8)
F-LOB (BrE 1990s)	19 (1.9)	9 (4.1)	1 (0.5)
Frown (AmE 1990s)	17 (1.7)	8 (3.6)	0

Not surprisingly, even today this subordinator seems to be favoured in the formal register of administrative or scientific texts, categories H and J in these four corpora, and the occurrences in fiction (categories K–R) are most infrequent, see examples (9)–(11).

(9) Today Mr James said: "Despite appalling difficulties during the last year the group has survived. **Provided** it can now achieve the essential level of new capital support to maintain and develop its market position, I believe it will again become a valuable investment." (A38:67–68 F-LOB)

(10) In certain cases students may be awarded support for pre-thesis studies on campus, **provided that** they intend to carry out their thesis research at Argonne. (H28:13–14 Frown)

(11) I puzzled over and analysed the wording … By around four thirty a.m. I had come to the conclusion that it probably was, but then stewed over how the hell Sexton thought I could find where Stover had stashed his savings (**provided** there were any, of course, and that Stover had not blown them on some extravagance or other we hadn't yet caught up with (L11:66 F-LOB)

The most significant detail in Table 7 is the relatively high frequency of *provided* in LOB. This may imply that British English non-fiction written prose of the 1960s was more formal than non-fiction American English prose. The decrease in the frequency of this connective in the second half of the 20th century, both in British and American English, is considerable.

3.3 Grammaticalisation of *provided (that)*

The development of *provided (that)* is also interesting from the point of view of the stage of grammaticalisation this connective reaches in the course of its development. Two syntactic features can be observed here: the use or absence of the subordination marker *that*, and the character of the elements, such as *always* and *also*, that may be placed between *provided* and the following subordinate clause.

As we saw under 3.1, the earliest instance of the absence of *that* can be found in the 15th century (example (5)). In almost all Middle English instances, however, *provided* is followed by *that.* The increase in the frequency of the instances without *that* coincides with the slight popularisation of *provided* in genres other than statutory texts in the second half of the 17th century (see Table 2). While all the occurrences in the Helsinki Corpus sub-sections EModE1 and EModE2 (1500–1640), and the 25 statutory text instances in EModE3 (1640–1710) have *that*, all seven instances in other text genres display *provided* without *that.* The figures from PPCME (Table 2) confirm this, since eight of the nine instances in non-statutory genres in sub-period 3 have *provided* without *that*. The *Corpus of English Dialogues* (Table 3) shows a similar trend, only one of the 17 post-1640 instances occurring with *that.* Thus both syntactic simplification and the increased frequency shown by the spread of its use to genres outside the officialese suggest further grammaticalisation of *provided (that).*

Against this background, it is somewhat surprising that the complete grammaticalisation of *provided* never took place. The main factor seems to be that *provided* never completely lost its verbal character. The following example from the *Corpus of English Dialogues*, in which *provided* is obviously wavering between a verb and a connective, is illustrative:

(12) but were I the Physitian that could cure his malady, and had so good judgment of his affects as of mine owne, charity would I should minister unto his Disease, what effect soever the potion would worke? **provided**

> this, that he disclosed his griefe in time: mistake me not *Marianus*, and pardon me if I conceale what I would utter my thoughts are mine own. (1641 Marianus 177 CED)

Furthermore, in idiomatic expressions such as *provided always/also*, the connective is separated from the subordinate clause by elements not forming part of the grammaticalized item. See examples (13) and (14):

(13) … the seid dep~tyng of such Souldeours and also theire reteignours if it be trav~sed be tried in the same Shire where they be for such cause arrested and arrayned. **Provided all wey** that no Capteyn be charged by this acte for lak of his nombre reteigned (1509–1543 Statutes 3,27 HC)

(14) the said Assessment shall by order of the said Justices be levied by the Overseers of the High-ways by Distresse and Sale of the Good~ of Persons so assessed …
Provided neverthelesse and be it enacted That no such Assessment or Assessment~ made in any One Yeare for enlargeing of High-ways shall exceed the Rate of Six Pence in the Pound (1695–1699 Statutes 7,211 HC)

We could even claim, based on the figures derived from CLMET, that the development towards complete grammaticalisation stopped and was perhaps even reversed from the second half of the 19th century on. Of the 61 post-1850 occurrences of *provided (that)*, no less than twelve, i.e., c. 20 per cent, have *that*. The earliest of these dates from 1865 and the latest from 1903. It is also worth noting that in seven of these twelve instances *provided* and *that* are separated by other elements (*only, of course, always*), examples (15), (16). In the pre-1850 parts of CLMET, the figures are different, only 14, or c. 8 per cent of the 176 instances occurring with *that.*

(15) let the pupils read during that time just whatever they like, **provided only that** they keep silence and read. (Wells, Mankind in the Making CLMET3)

(16) The muscular fibre … does so with the greater energy the more often it is stimulated, **provided, of course, that** reasonable times are allowed for repose. (Butler, Unconscious Memory CLMET3)

The return of *provided that* is quite dramatic in the corpora representing the second half of the 20th century. In the LOB Corpus, *that* follows *provided* in 21 out of the 48 instances (c. 44 p.c.) and in Brown, in 8 out of the 22 instances (c. 36 p.c.).

Thus, *provided* obviously belongs to that group of adverbial connectives with which the grammaticalisation was never complete.[13] This may be due both to the tendency to separate this connective from the subordinate clause by other elements and to the high frequency of the verbal use of *provide* in comparison to

the connective use. In LOB, there are 259 instances of the verbal use of *provide*, as against the 48 instances of the connective use. In F-LOB the corresponding figures are 385 as against 19. Even in ARCHER and CLMET, the verbal use prevails. Similar failure to reach complete grammaticalisation can be traced, particularly with connectives associated with verbs, such as *save* (cf. Rissanen 2009).

4. Conclusion

This survey of the present stock of English historical corpora and of the information these corpora can provide for the origin, development and present-day uses of an adverbial connective is necessarily superficial and does not cover all historical corpora now in use. Furthermore, important new corpora, either multi-genre or focused on particular genres, social groups or regional varieties, are being compiled in various parts of the world. I hope, however, that even this brief overview has illustrated the new initiatives and vistas for multi-faceted use-based research into English, past and present, excellently described and discussed by Elizabeth Closs Traugott (2008). Corpora have changed and will change the dimensions of linguistic analysis in many ways, offering an opportunity to deal with large quantities of textual evidence, tackle problems which would previously have been regarded as too time-consuming, and reveal connections between linguistic phenomena which would previously have remained hidden under masses of unstructured evidence. They have also given us a new sense of responsibility for basing our conclusions on solid evidence, more easily verifiable or falsifiable than before. We also have a better opportunity to combine our results on language use in various periods with more theoretical considerations of language and its development.

At the same time, we have to keep in mind that even the best corpus represents only a slice of linguistic reality, not its entirety (Rissanen 1989). Furthermore, corpora can be successfully and sensibly used only if the user – student or scholar – has sufficient mastery of the language of the period or periods he or she is studying, including the political and cultural background of the period. Finally, corpora give us excellent quantifiable evidence of language use, but collecting and calculating this evidence is not enough, since the core of our research is the analysis and interpretation of the material collected in this way – and in this the computer will never replace the human brain.[14]

Notes

1 It is a question raised and discussed in various contexts (e.g., at the ICAME Conference at the University of Michigan, Ann Arbor in 2005) whether corpus linguistics, i.e., corpus compilation, software developments and corpus-based studies, is a more or less independent

branch of linguistics or just a methodology designed to support and intensify in-depth analysis and explanation of present and past stages of the language. It is worth noting that there are professorial chairs of corpus linguistics at various major universities. We should not forget, however, that systematic study of language, based on careful analysis of extensive textual evidence, has been carried out for centuries. First and foremost, corpus linguistics is a handy shorthand term covering various corpus-related research activities, both theoretical and practical.

2 The main compilers of the corpora are listed in the Appendix.

3 This project is still in progress and is being hampered by copyright problems, although the corpus can be used in more than half-a-dozen centres in Europe and the United States.

4 Reference can be made to Elizabeth Closs Traugott's (2008) plenary lecture at the 20th IAUPE Conference in Lund (see especially pp. 200–201).

5 The latest and very thorough study of these voyages is Valtonen (2008).

6 According to *MED* (s.v. *providen* v. 6), the connective use of the present participle form *providing* appears in the early 15th century. This form remains marginal, however; no examples can be found in HC, PPCME or PPCEME.

7 The abbreviated title of the dictionary or corpus from which the example is taken is indicated at the end of the reference line. For a list of corpora, see Appendix.

8 In this paper, the Middle and Modern English developments of *provided (that)* are treated in terms of (further) grammaticalisation, although the connective was originally borrowed from Latin.

9 The *OED* gives the earliest instances of *provided* without *that* as late as the early 17th century.

10 For the list of texts included in the Helsinki Corpus, see Kytö (1996).

11 To simplify genre-based word counting, official letters have been included in the "other" group in the PPCEME figures. Their role is, however, negligible: no instances of the connective in the first and third sub-period and one in the second sub-period.

12 There are altogether nine occurrences of *provided (that)* in the *Oxford Shakespeare*.

13 It is not surprising that Quirk et al. (1985: 1002–1003) defines *provided (that)* as a marginal subordinator. It is questionable, however, whether we can regard the Late Modern English development of *provided (that)* as an example of degrammaticalisation (cf. the discussion, e.g., in Hopper and Traugott (2003: 130–139)).

14 The research reported in this paper was supported in part by the Academy of Finland Centre of Excellence funding for the Research Unit for Variation, Contacts and Change in English at the Department of English, University of Helsinki.

I wish to express my sincerest thanks to the anonymous referee for excellent comments on my paper.

References

Biber, Douglas, Edward Finegan and Dwight Atkinson (1994), 'ARCHER and its challenges: compiling and exploring a Representative Corpus of Historical English Registers', in: Udo Fries, Gunnel Tottie and Peter Schneider (eds.) *Creating and using English language corpora. Papers from the Fourteenth International Conference on English Language Research on Computerized Corpora, Zürich 1993*. Amsterdam and Atlanta, GA: Rodopi, 1–13.

Corpus Resource Database (CoRD) (2008–), http://www.helsinki.fi/varieng/CoRD/index.html

Culpeper, Jonathan and Merja Kytö (1997), 'Towards a Corpus of Dialogues, 1550–1750', in: Heinrich Ramisch and Kenneth Wynne (eds.) *Language in time and space. Studies in honour of Wolfgang Viereck on the occasion of his 60th birthday*. Stuttgart: Franz Steiner Verlag, 60–73.

De Smet, Hendrik (2005), 'A Corpus of Late Modern English Texts', *ICAME Journal*, 29: 69–82.

Fries, Udo (2001), 'Foreign place names in the ZEN Corpus. Language contact in the history of English', in: Dieter Kastovsky and Arthur Mettinger (eds.) *Language contact in the history of English.* Frankfurt am Main: Peter Lang, 117–129.

Halliday, M. A. K. (1973), *Explorations in the functions of language*. London: Edward Arnold.

Healey, Antonette di Paolo (1999), 'The Dictionary of Old English Corpus on the World-Wide Web', *Medieval English Studies Newsletter*, 40: 2–10.

Hickey, Ray (2003), *Corpus Presenter: software for language analysis. With a manual and* A Corpus of Irish English *as sample data.* Amsterdam: John Benjamins.

Hopper, Paul J. and Elizabeth Closs Traugott (2003), *Grammaticalization*. 2nd edition. Cambridge: Cambridge University Press.

Kytö, Merja (comp.) (1996 [1991]), *Manual to the diachronic part of the Helsinki Corpus of English Texts: coding conventions and lists of source texts*. (3rd ed.). Helsinki: Department of English, University of Helsinki.

Kytö, Merja and Terry Walker (2006), *Guide to* A Corpus of English Dialogues 1560–1760 (Studia Anglistica Upsaliensia 130). Uppsala: Acta Universitatis Upsaliensis.

Lancashire, Ian (1994a), 'Early Modern English Renaissance Dictionaries Corpus: an update', in: Merja Kytö, Matti Rissanen and Susan Wright (eds.) *Corpora across the centuries. Proceedings of the First International Colloquium on English Diachronic Corpora, St Catharine's College Cambridge, 25–27 March, 1993.* Amsterdam and Atlanta, GA: Rodopi, 143–149.

Lancashire, Ian (1994b), 'Toronto English Renaissance Dictionaries Database', *Medieval English Studies Newsletter*, 30: 6–8.

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

MED = Middle English Dictionary. See http://quod.lib.umich.edu/m/med/.

Milroy, James (1985), 'Linguistic change, social network and speaker innovation', *Journal of Linguistics*, 21: 339–384.

Milroy, James (1992), *Linguistic variation and change.* Oxford and Cambridge Mass.: Blackwell.

Milroy, James and Lesley Milroy (1978), 'Belfast: change and variation in an urban vernacular', in: Peter Trudgill (ed.) *Sociolinguistic patterns in British English*. London: Edward Arnold, 19–36.

Milroy, Lesley (1987), *Language and social networks.* Oxford: Basil Blackwell.

Nevalainen, Terttu and Helena Raumolin-Brunberg (2003), *Historical sociolinguistics: language change in Tudor and Stuart England*. London: Longman.

Nurmi, Arja (1999), 'The Corpus of Early English Correspondence Sampler (CEECS)', *ICAME Journal*, 23: 53–64.

OED = Oxford English Dictionary. See http://www.oed.com/.

Pintzuk, Susan and Ann Taylor (1997), 'Annotating the Helsinki Corpus: the *Brooklyn-Geneva-Amsterdam-Helsinki Parsed Corpus of Old English* and the *Penn-Helsinki Parsed Corpus of Middle English*', in: Raymond Hickey, Merja Kytö, Ian Lancashire and Matti Rissanen (eds.) *Tracing the trail of time: proceedings of the Diachronic Corpora Workshop, Toronto (Canada) May 1995*. Amsterdam and Atlanta, GA: Rodopi, 91–104.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik (1972), *A grammar of contemporary English.* London and New York: Longman.

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

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

Rissanen, Matti (2000), 'The world of English historical corpora: from Cædmon to computer age', *Journal of English Linguistics*, 28: 7–20.

Rissanen, Matti (2002), '"Without except(ing) unless...": on the grammaticalisation of expressions indicating exception in English', in: Katja Lenz and Ruth Möhlig (eds.) *Of dyuersite & chaunge of langage: essays presented to Manfred Görlach on the occasion of his 65th birthday*. Heidelberg: C. Winter Universitätsverlag, 77–87.

Rissanen, Matti (2009), Grammaticalisation, contact and adverbial connectives: the rise and decline of *save*, in: Shinichiro Watanabe and Yukiteru Hosoya (eds.) *English philology and corpus studies: a festschrift in honour of Mitsunori Imai to celebrate his seventieth birthday.* Tokyo: Shohakusha, 135–152.

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

Romaine, Suzanne (1982), *Sociohistorical linguistics: its status and methodology.* Cambridge: Cambridge University Press.

Samuels, M. L. (1972), *Linguistic evolution, with special reference to English.* Cambridge: Cambridge University Press.

Schmied, Josef and Claudia Claridge (1997), 'Classifying text- or genre-variation in the *Lampeter Corpus of Early Modern English Texts*', in: Raymond Hickey, Merja Kytö, Ian Lancashire and Matti Rissanen (eds.) *Tracing the trail of time: proceedings of the Diachronic Corpora Workshop, Toronto (Canada) May 1995*. Amsterdam and Atlanta, GA: Rodopi, 119–135.

Taavitsainen, Irma and Päivi Pahta (eds.) (2004), *Medical and scientific writing in late Medieval English.* Cambridge: Cambridge University Press.

Taavitsainen**,** Irma, Päivi Pahta and Martti Mäkinen (2005), *Middle English medical texts*. Amsterdam and Philadelphia: John Benjamins.

Traugott, Elizabeth Closs (2008), 'The state of English language studies: a linguistic perspective', in: Marianne Thormählen (ed.) *English now: selected papers from the 20th IAUPE Conference in Lund 2007* (Lund Studies in English 112). Lund: Centre for Languages and Literature, Lund University, 199–225.

Traugott, Elizabeth Closs and Richard B. Dasher (2002), *Regularity in semantic change.* Cambridge: Cambridge University Press.

Valtonen, Irmeli (2008), *The North in the Old English Orosius. A geographical narrative in context* (Mémoires de la Société Néophilologique de Helsinki 73). Helsinki: Société Néophilologique.

Weinreich, Uriel, William Labov and Marvin Y. Herzog (1968), 'Empirical foundations for a theory of language change', in: Winfred P. Lehmann and Yakov Malkiel (eds.), *Directions for historical linguistics: a symposium.* Austin, Texas: University of Texas Press, 95–195.

Appendix

Selection of English historical corpora

Names of compilers or corpus project leaders, and/or contact persons are given in brackets. More information on many of the corpora listed below can be found in the *Corpus Resource Database* http://www.helsinki.fi/varieng/CoRD/index.html. See also the list of references, above.

Old English (c. 750–1150)

1. *The Dictionary of Old English Corpus* (DOEC), c. 3.5 million words (project leader Antonette diPaolo Healey).
2a. *The Helsinki Corpus of English Texts* (HC), Old English part, c. 500,000 words (project leader Matti Rissanen).
3. *The York-Toronto-Helsinki Parsed Corpus of Old English Prose* (YCOE), c. 1.5 million words (project leader Susan Pintzuk).

Middle English (c. 1150–1500)

2b. *The Helsinki Corpus of English Texts* (HC), Middle English part, c. 500,000 words (see 2a above).
4. *The Penn-Helsinki Parsed Corpus of Middle English* (PPCME), c. 1.1 million words (project leader Anthony Kroch).
5. *Innsbruck Computer Archive of Middle English Texts* (ICAMET), c. 5 million words. Sampler corpus c. 3 million words (project leader Manfred Markus).
6. *The Corpus of Middle English Medical Texts* (1375–1500) (MEMT), Helsinki, c. 500,000 words (project leader Irma Taavitsainen).
7. *A Linguistic Atlas of Early Middle English 1.1* (1150–1325) (LAEME), Edinburgh, c. 650,000 words (Margaret Laing).
8. *A Linguistic Atlas of Older Scots, Phase 1* (1380–1500) (LAOS), Edinburgh, (Keith Williamson).
9. *The Middle English Compendium* (MEC), Ann Arbor, Michigan (project leader Frances McSparran).
9a. *The Corpus of Middle English Prose and Verse* (CMEPV), c. 18 million words (see 9 above).

9b. *The Middle English Dictionary* (MED) (see 9 above).
10. *Parliament Rolls of Medieval England* (1272–1509).

Early Modern English (c. 1500–1700)

2c. *The Helsinki Corpus of English Texts* (HC), Early Modern English part, c. 500,000 words (see 2a above).
4a. *The Penn-Helsinki Parsed Corpus of Early Modern English* (PPCEME), c. 1.8 million words (see 4 above).
6a. *Early Modern English Medical Texts (EMEMT) Corpus* (1500–1700), Helsinki, c. 2 million words (project leader Irma Taavitsainen) (see 6 above).
11. *The Helsinki Corpus of Older Scots* (1450–1700) (HCO), Helsinki, c. 850,000 words (Anneli Meurman-Solin).
12. *The Corpus of Irish English* (14th–20th c.), Essen, c. 550,000 words (Raymond Hickey).
13. *The Corpus of Early English Correspondence Sampler* (1417–1681) (CEECS), Helsinki, c. 500,000 words (project leader Terttu Nevalainen).
13a. *The Parsed Corpus of Early English Correspondence* (1410?–1681) (PCEEC), 2.2 million words (see 13 above).
14. *The Lampeter Corpus of Early Modern English Tracts* (1640–1740), Chemnitz/Zwickau, c. 1.1 million words (project leader Josef Schmied).
15. *A Corpus of English Dialogues 1560–1760* (CED), Uppsala and Lancaster, c. 1.2 million words (project leaders Merja Kytö and Jonathan Culpeper).
16. *Early Modern English Dictionaries Database* (EMEDD) Toronto, c. 500,000 words (Ian Lancashire).

Late Modern English (c. 1700–1990)

4b. *Penn Parsed Corpus of Modern British English* (PPCMBE), c. 1,000,000 words (see 4 and 4a above).
17. *A Representative Corpus of Historical English Registers* (1650–1990) (ARCHER), University of Northern Arizona and University of Southern California, c. 1.7 million words (Douglas Biber and Edward Finegan; later versions: international team).
18. *Corpus of Historical American English* (1810–2009) (COHA), c. 400 million words (Mark Davies).
19. *A Corpus of Late Modern English Texts* (1710–1920) (CLMET), Leuven, c. 10 million words (Hendrik De Smet).
20. *Zurich English Newspaper Corpus* (1661–1791) (ZEN), c. 1.6 million words (project leader Udo Fries).
21. *The Century of English Prose Corpus* (1680–1780) (COPC), Cleveland, Ohio, c. 550,000 words (Louis Milić).

22. *Corpus of Early Ontario English, pre-Confederation section* (1776–1850) (CONTE-pC), British Columbia, Vancouver, c. 225,000 words (Stefan Dollinger).
23. *A Corpus of Late Modern English Prose* (1861–1919), Manchester, c. 100,000 words (David Denison).
24. *The Brown Corpus* (1960s) (Brown) 1,000,000 words (Henry Kučera and W. Nelson Francis).
25. *The Lancaster-Oslo/Bergen Corpus* (1960s) (LOB), 1,000,000 words (project leaders Geoffrey Leech and Stig Johansson).
26. *The Freiburg-Brown Corpus* (1990s) (Frown), 1,000,000 words (project leader Christian Mair).
27. *The Freiburg-Lancaster-Oslo/Bergen Corpus* (1990s) (F-LOB), 1,000,000 words (see 25 above).
28. *Time Magazine Corpus* (1923–2000s), 100 million words (Mark Davies).
29. *The Corpus of Contemporary American English* (1990–2011) (COCA), c. 425 million words (Mark Davies).
30. *A Corpus of Oz Early English* (1788–1900) (COOEE), c. 2 million words (Clemens Fritz).
31. *WebCorp*. Birmingham City University (UK) (project leader Antoinette Renouf).

The status of onset contexts in analysis of micro-changes

Elizabeth Closs Traugott

Stanford University

Abstract

König and Vezzosi (2004: 239) commented about work on grammaticalization that: 'Very rarely … do we find detailed discussions of the onset contexts that set such a process into motion and the conditions that such contexts must meet'. However, understanding local context-derived inference is now a high priority in a number of areas of linguistics, ranging from computational semantics to discourse analysis of conversation to work on micro-changes in historical linguistics. In this paper I discuss the hypothesis that a subset of linguistic contexts, specifically 'bridging contexts', are a key factor in morphosyntactic change. I begin by outlining the history of the hypothesis, which originated in work on semantic change, and shifted emphasis from implicatures and invited inferences associated with a changing expression (e.g. Traugott and König 1991) to the linguistic contexts in which two (or more) meanings are possible, but one is 'only contextually implicated' (Evans and Wilkins 2000: 549). Subsequently Heine (2002) and Diewald (2002) hypothesized that the development of "bridging" or "critical" contexts is a necessary and distinct stage in grammaticalization, using synchronic variation as data I then go on to discuss the extent to which we can find evidence in diachronic corpora of English for these hypotheses. I also evaluate Hansen's (2008) claim that in bridging contexts new meanings are backgrounded against Heine's that they are foregrounded. Two case studies are presented: the development in late Middle English and Early Modern English of be going to *with temporal meaning (e.g. Danchev and Kytö 1994), and of pseudo-clefts with ALL and WHAT (Traugott 2008a, 2010).*

1. Introduction[1]

In 1994 Bybee, Perkins, and Pagliuca said: 'Facts about grammaticization suggest that grammatical meaning is constituted from a set of diachronically related uses with meanings that are contextually determined to a large degree' (Bybee, Perkins, and Pagliuca 1994: 281). However, ten years later König and Vezzosi commented about work on grammaticalization that 'Very rarely … do we find detailed discussions of the onset contexts that set such a process into motion and the conditions that such contexts must meet' (König and Vezzosi 2004: 239). Understanding local context-derived inference is a now a high priority in work on micro-changes in historical linguistics (see e.g. Krug 2002; Eckardt 2006; Kempson and Cooper 2008). Understanding such inferences is also a high

priority in a number of other areas of linguistics ranging from computational semantics (e.g. Bobrow et al. 2007) to discourse analysis of conversation (e.g. Ewing 2005). It is therefore a good time to take stock of various hypotheses about onset contexts. One hypothesis that has been put forward is that 'bridging contexts' are a key factor in morphosyntactic change (Heine 2002; Eckardt 2006; Hansen 2008). Evans and Wilkins originally conceptualized 'bridging context' as a contribution to work on semantic change, and defined it as a context in which two (or more) meanings are possible, but one is 'only contextually implicated' (Evans and Wilkins 2000: 549). Heine, Eckardt, and Hansen likewise interpret 'bridging contexts' as pragmatic and semantic.[2] Another hypothesis is that onset contexts for grammaticalization involve not only semantic but also structural opacity. Diewald (2002) has called these 'critical contexts'. My focus will be on comparing the concepts of 'bridging' and 'critical' contexts and what evidence there is from empirical historical data that they can be identified as a necessary and distinct stage in grammaticalization. I define grammaticalization as 'the process by which grammar is created' (Croft 2006: 366),[3] and assume that this process involves micro-steps (also known as 'gradual' change) (Traugott and Trousdale 2010). My case studies are two constructions: one the well-known case of *be going to*, the other much less well known: a subset of pseudo-clefts with ALL and WHAT, specifically those with *do* (e.g. *What/All I did was (to) voice support for her*).

Any set of terms is fraught with problems, and those in semantics and pragmatics are no exception (see Kempson 1996, Kearns 2006, and references in them). There is as yet no full agreement on whether there is a useful distinction to be made between semantics and pragmatics, and if so, where to draw the boundaries between them (e.g. Bach 2004; Recanati 2004; Hansen 2008). Here I assume it is useful to distinguish between them and that the distinction is essentially between a theory of coded, conventionalized meaning on the one hand, and a theory of defeasible meaning on the other, although it may not always be possible to uniquely determine which is which. I further assume not only an information-structure-oriented approach to pragmatics such as Grice (1989 [1975])) espoused, but also a more interactional one, as espoused by Keller (1994 [1990]).

I start by outlining some types of language-internal context[4] that have been considered operative in the onset of grammaticalization (Section 2). I then move on to the empirical data (Section 3), and suggest how 'linguistic context' might be modified in future work (Section 4).

2. Approaches to language-internal context

In this section I distinguish four different conceptualizations of language-internal onset context that have been used in work on grammaticalization. The first is primarily structural and distributional, the second adds pragmatic implicatures or 'invited inferences' that arise out of these structural and distributional contexts.

The third appears to conceptualize onset context as pragmatic only; this is a 'bridging' context, narrowly defined. The fourth is a 'critical' context involving pragmatic, semantic, and structural opacity. In the first two conceptualizations, it is typically the item undergoing grammaticalization that is conceived as having the potential to change; in the latter two the potential for change is thought also to lie in the context itself.

2.1 Structural contexts

Much work on grammaticalization and morphosyntactic change is couched in terms of an item X being modified in morphosyntactic and semantic contexts. A representative statement is:

> Strictly speaking, it is never just the grammaticizing element that undergoes grammaticization. Instead, it is the grammaticizing element *in its syntagmatic context* which is grammaticized. (Himmelmann 2004: 31, italics original)

For example König and Vezzosi (2004: 239) identify 'sentences with other-directed transitive verbs and third-person singular subjects' as the onset contexts in which complex reflexive anaphors arise out of intensifiers. More particularly:

> In object positions and in contexts of binding these compounds become more and more frequent and ultimately obligatory as part of a paradigm of inflexional forms… Moreover, the original emphatic forms lose their focal stress in most contexts (phonological attrition). (König and Vezzosi 2004: 229)

Here the semantic category of both the verb and the subject as well as syntactic position are relevant to changes in the *–self* compound. In other work 'context' can be understood in terms of more strictly morphosyntactic distribution:

> At the methodological level, secondary grammatical categories are delimited against each other by their distribution, where prototypical members of the primary grammatical categories may be used in the specification of contexts. Primary grammatical categories are delimited against each other by their distribution, where secondary grammatical categories may be used in the specification of contexts.[5] (Lehmann 1993: 6)

Recent work on diachronic collostructional analysis (Hilpert 2008) operationalizes analysis of structural, especially lexical context, over time.

2.2 Invited inferencing contexts

While syntactic and semantic factors are without question crucial in morphosyntactic development, they alone among 'language-internal factors' would appear not to account adequately for onset of grammaticalization.[6] Pragmatic factors have been argued to play a large role too. One proposal regarding such factors is the invited inferencing theory of semantic change, or IITSC (e.g. Traugott and König 1991; Traugott and Dasher 2002). The concept of invited inferencing draws on the distinction between conversational and generalized implicatures as developed in Grice (1989 [1957]), Levinson (1995, 2000), and Geis and Zwicky (1971), but modifies these proposals. The term 'invited inference' is used to highlight the interactive nature of the process. The proposal is that speakers exploit conversational implicatures and invite addressees to infer meaning. These innovative utterance-token meanings may become conventionalized as generalized implicatures and may eventually be semanticized into an item X. Invited inferencing is a process that may be involved in semantic change alone, and is therefore independent of grammaticalization.

Invited inferencing is usually considered to be a necessary, but not sufficient, condition for the onset of grammaticalization. It is necessary because negotiation of meaning by interlocutors and parsing by addressees by hypothesis open the door to novel uses.[7] It is not sufficient because change does not have to occur, and if it does, the changes that lead to grammaticalization involve increased abstraction and schematicity, in contrast to changes in lexicalization that lead to increased idiosyncracy (Brinton and Traugott 2005). Furthermore, factors such as routinization (Bybee 2006; Detges 2006) as certain discourse purposes are adopted (Waltereit and Detges 2007) are of course also relevant.

The concept of invited inferencing depends in part on that of pragmatic ambiguity (see Horn 1985 on metalinguistic negation and Sweetser 1990 on conditionals and conjunctions). Pragmatic ambiguity as developed by Sweetser accounts for the observation that in some instances 'it is possible for a linguistic form to have only one semantic value, but multiple functions nevertheless' (Sweetser 1990: 10). An example of how pragmatic ambiguity and invited inferencing can be seen to work together is provided by Krug's (2002) discussion of the transition from *want* 'lack' > 'need' > 'desire', and from transitive verb to auxiliary. Krug draws on ambiguity (presumably pragmatic, although he does not specify that), indeterminacy, and invited inferences to account for the changes. In the case of 'lack' > 'need', there is the implicature 'If somebody lacks something, he or she will usually also need it (otherwise stating the lack would be unnecessary or odd)' (p. 140). He says 'a full third of the verbal uses in ARCHER drama and fiction from the period 1650–1699 are ambiguous' (*ibid*.), as in (1):

(1) your Daughter (do ye conceive me) wants a Husband; and I want a Wife (do ye conceive me;) Now what are we born for in this world, but to supply one another's wants? (1671 ARCHER, Caryll.D1 [Krug 2002: 140])

Krug comments: 'While the semantics of WANT in WIFE AND HUSBAND WANTED leans perhaps towards "need", it certainly possesses traces of "lack" and "desire"' (*ibid*.) and goes on to say:

> With so many relatively frequent coexisting senses, it comes as no surprise that sometimes the meaning of WANT is indeterminate, in particular between the two senses 'lack' and 'need', which are doubly motivated by entailment, and, conversely, by invited inferences.[8] (p. 141)

Note that here the semantics of the lexical item *want* is in focus and spoken of metalinguistically as an actor: it 'leans' toward 'need'; ambiguities are 'motivated by' pragmatic factors. The potential for change is in the lexical item and its pragmatics in use.

Although I have distinguished structural and invited inferencing approaches to context for reasons of clarifying differences between methodologies, practitioners often combine the two approaches. For example, Kytö and Romaine (2005) combine them in their discussion of the development of *be/have like to + V* 'imminently likely to V' into an 'avertive' or modal auxiliary marking 'action narrowly averted' (Kuteva 2001). They refer to the structural morphosyntactic and semantic contexts for the onset of the construction in later Middle English as past tense, conditional *if* or *but*, and infinitive verbs with semantically negative prosody,[9] as in (2):

(2) on of the mynysteris of the said Cathedrall Churche was sette afire, and began to brenne, and yf hit hadde had his course *lyke to have sette* a fyre and brende the cheif and grete parte of the citee.

'one of the minsters of the said Cathedral church was set afire and began to burn, and if it had had its course, would have come close to setting a fire and burning the chief and great part of the city'.
(1447 Shillingford, *Letters* [ICAME: Helsinki; Kytö and Romaine 2005: 3])

Kytö and Romaine also discuss invited inferencing, specifically the development of the implicatures of counterfactuality and narrowly averted eventhood, in these contexts (p. 5). In addition they draw attention to 'the need to contextualize semantic change within larger discourse structures, typically across sentence-boundaries' (*ibid*.), that is, within co-text. They conceptualize the actual grammaticalization of the construction, i.e. its appearance independent of conditional marking, as an instance of the absorption of meanings from that co-text into the semantics of the construction.

A profile-shift from locating the potential for change in the expression to locating it in its changing context was suggested by Heine, Claudi, and Hünnemeyer when they referred to changes arising out of invited inferences,

perspectivization, schematization, and prototype extension as 'context-induced reinterpretations'. The metalinguistic discourse is not about changes being passively 'motivated by' contexts, but about contexts metaphorically conceived as active drivers: '[o]nce one of the arrays of conversational implicatures is conventionalized, then context-induced reinterpretation may be said to come in' (Heine, Claudi, and Hünnemeyer 1991: 101). In later work Heine (2002) privileged non-conventionalized implicatures as a necessary, but not sufficient, step in grammaticalization, and it is to this approach to context that I now turn.

2.3 Bridging contexts as pragmatic and semantic contexts

The term 'bridging' is widely used in historical linguistics for any kind of ambiguity, whether pragmatic, semantic, or syntactic. However, use with reference to pragmatic implicatures appears to originate in the 1970's with several papers by Clark and Haviland (e.g. Clark 1975, Clark and Haviland 1977). It is part of early work promoting serious study of language use and processing in addition to, or instead of, competence. Bridging is conceived as a synchronic activity that hearers engage in. The assumption is made that speakers are obliged by a 'given-new contract' to make a syntactic distinction between given and new, as in (3a). However, speakers sometimes fail to do this, as in (3b):

(3) a. John fell. What he did was trip on a rock. (Clark 1975: 172)
b. Horace got some picnic supplies out of the car. The beer was warm. (Clark and Haviland 1977: 21)

Regarding (3a), clefts and pseudo-clefts are considered to be syntactically designed to cue bridging to givenness; the implicatures include something like (i) *John fell because he did something* and, crucially for Clark, (ii) that 'something' is the antecedent for *what he did.* Regarding (3b), Clark and Haviland assume a given-new contract and Grice's (1989 [1975]) Cooperative Principle and Maxim of Relation. They hypothesize that hearers build an inferential bridge of the type: (i) the picnic supplies contain a quantity of beer, and (ii) 'that quantity is being referred to by the given information of the target sentence' *The beer was warm* (Clark and Haviland 1977: 21).

Early proposals like this gave way to arguments about how speakers and hearers negotiate common ground in general, not only given and new (e.g. Clark 1996). The linguist engaged in historical pragmatics must in essence do what Clark, Haviland, and many others since them have said language-users do: build inferential links between clauses or parts of clauses to determine what speakers/writers might have meant. The task is more difficult, however, for the historical linguist, because general cultural knowledge has to be inferred too, and so reliance is almost exclusively on the text.

The notion of 'bridging context' developed by Evans and Wilkins (2000) is rather different from the concept of 'bridging' outlined above, although it too concerns implicatures. The main differences in approach are:

a) 'Bridging' is a type of linguistic context, not an activity that speakers engage in.
b) It is not restricted to constructing givenness.
c) The prime data are lexical, not syntactic.
d) It is part of a larger project testing claims that patterns of polysemy and semantic extension are cross-linguistically replicated, for example, in the domain of perception verbs like SEE > KNOW (Viberg 1984, Sweetser 1990).

Evans and Wilkins use largely synchronic 'text and context' to 'describe (or reconstruct) bridging contexts, the places where extended meanings commonly have their genesis' (p. 550), and point out that in order to fully understand such contexts one needs to understand the cultural scripts that underlie them. Their prime objective, therefore is to account for semantics and semantic change.[10]

Evans and Wilkins argue that when a later meaning B arises in addition to an earlier meaning A, this often happens

> because a regularly occurring context supports an inference-driven contextual enrichment of A to B. In these contexts, which we term BRIDGING CONTEXTS, speech participants do not detect any problem of different assignments of meaning to the form because both speaker and addressee interpretations of the utterance in context are functionally equivalent, even if the relative contributions of lexical content and pragmatic enrichment differ. (Evans and Wilkins 2000: 550; caps original)

Enfield paraphrases: a bridging context is:

> [A] speech context in which something inferrable as utterance-meaning from an input sentence-meaning happens also to be true, and thus not defeasible in that context. (Enfield 2005: 318)

Evans and Wilkins and Enfield draw on work by Sweetser on pragmatic as well as semantic ambiguity, and by Traugott and others on invited inferencing, but shift attention from changes in form-meaning pairs to the context in which such pairings occur. From this perspective it is not the semantics of WANT that 'leans' in any direction (Krug 2002: 140 cited above), but rather the context that provides the cue to a preferred reading.

As mentioned at the end of Section 2.2, Heine, Claudi, and Hünnemeyer (1991) identify 'context-induced reinterpretation' as following the conventionalization of implicature diachronically. Using a 1998 version of the Evans and Wilkins (2000) paper, Heine (2002) put forward the stronger hypothesis that grammaticalization involves four stages, abstracted from the continuum of change (see also Heine and Kuteva 2007):

(4) I. [T]here is an expression with a 'normal' or source meaning occurring in an array of different contexts.
II. … [T]here is a bridging context giving rise to an inference to the effect that, rather than the source meaning, there is another meaning, the target meaning, offering a more plausible interpretation of the utterance concerned.
III. … [T]here is a new type of context, the switch context, that no longer allows for an interpretation in terms of the source meaning. Switch contexts may be viewed as a filtering device that rules out the source meaning.
IV. Finally, no longer being associated with the source meaning, the target meaning is now open to further manipulation: It is freed from the contextual constraints that gave rise to it. … I will refer to this situation as the conventionalization stage. (Heine 2002: 86)

As I understand this proposal, at Stage I conversational implicatures abound. Some of these begin to be generalized and invite attention to the implicature. These are bridging contexts, which are characterized further as in (5):

(5) a. While the target meaning is the one most likely to be inferred, it is still cancellable (see Grice 1967),[11] that is, an interpretation in terms of the source meaning cannot be ruled out.
b. A given linguistic form may be associated with a number of different bridging contexts.
c. Bridging contexts may, but need not, give rise to conventional grammatical meaning. (Heine 2002: 84–85)

Bridging contexts alone will not lead to change. The onset of grammaticalization is identified with the emergence of switch contexts. Part of Heine's purpose in the article is to distinguish (initial) switch from conventionalized stages of grammaticalization. His hypotheses are based on variation in synchronic data (but see Heine and Miyashita 2008 for a historical study of the development of raising, functional *drohen* 'threaten').

Although Stages I–IV in (4) appear to be discrete, the distinctions between them are meant to be on a continuum (p. 86). What is important about the proposal for our purposes is that it is these bridging contexts at Stage II that are hypothesized *in addition to* the initial implicatures (Stage I) to be necessary (but not sufficient) for grammaticalization to take place. This is stronger than any prior hypotheses about the role of linguistic pragmatics and semantics in grammaticalization.

Heine nowhere says that structural change is excluded at Stage II, and indeed he includes structural contexts such as passives in his examples. However, the bridging of reflexive |*'é* 'self' with passive in North Khoisan, of *taka* 'want' with 'about to' in Standard Swahili, and of *dabei* 'on that occasion' with 'still' in

German is treated as semantic, and as occurring in a subset of pre-existing environments. What is by hypothesis new is the foregrounding of the 'target meaning' in the restricted context (p. 86). Although critical of Heine's examples, Eckardt likewise appears to assume that bridging examples come into being as a result of a shift in pragmatic and semantic status alone since she refers to writers and readers understanding that a certain kind of sentence used in a certain kind of context not only gives rise to certain implicatures but can be used in 'a conventional way to express a certain kind of proposition' (Eckardt 2006: 118–119).

Presumably 'foregrounding of the target meaning' means the meaning that will be associated with the grammaticalized item is salient or preferred over the original, source meaning in the bridging context. By contrast, Hansen (2008: 63) claims that in bridging contexts the 'target interpretation … is still backgrounded with respect to the source meaning, and only moves into the foreground when we reach Stage III'. Her reason is that the 'actual reinterpretations' have not yet occurred, i.e. the inferences have not yet been (and may never be) semanticized, therefore they cannot be exploited.

2.4 Critical contexts

Working with historical evidence from the development of modals in German, Diewald (2002) proposes an alternative model in the same volume as Heine (2002), and has elaborated on the model since (e.g. Diewald 2006, Diewald and Ferraresi 2008). While both Heine (2002) and Diewald (2002) agree that there is considerable similarity between their perspectives, there are significant differences as well. Diewald sees the pre-condition for grammaticalization as Stage I in (6). 'Untypical context' implies a Stage 0, with typical contexts. Grammaticalization can be triggered only if a period of multiple opacity occurs, i.e. when several meanings may be implicated. This is Stage II, the emergence of 'critical contexts'. These may remain opaque in larger contexts, or one reading may be preferred; in particular further context may lead to defeasing of implicatures (Diewald 2002: 111). Grammaticalization itself does not occur until Stage III, at which time 'isolating contexts' separate out and crystallize the new meaning and function from among competing options:

(6) I: 'Untypical contexts': the development of implicatures in contexts 'which show clusters of contextual features that had not been customary before' (Diewald 2002: 109).
II: 'Critical contexts' with 'multiple structural and semantic ambiguities' or opacities that invite 'several alternative interpretations, among them the new grammatical meaning' (p. 103). At this stage semantic and structural possibilities that were distributed over different contexts 'accumulate in one specific critical

context' (p. 109). Stage II is hypothesized to disappear in later development (Diewald 2006: 3).

III: 'Isolating contexts': 'specific linguistic contexts that favor one reading to the exclusion of the other' (Diewald 2002: 103).

Diewald (2002) ends with emphasis on the fact that in her view 'critical' contexts are not pragmatic or semantic only, but are 'defined by semantic *and* structural ambiguity' (p. 117; italics original). They may arise as the effect of independent changes elsewhere in the system (p. 117), e.g. changes in the Middle High German verbal morphological paradigm allowed for the development of a critical context in which *hân* 'have' + past participle could be grammaticalized as a deontic modal (Diewald 2006: ft. 11, p. 20). They drive change by 'creating' ambiguity and 'inviting' reinterpretation (Diewald and Ferraresi 2008: 101). Finally, Diewald notes that Heine's (2002) Stages I and II 'would have to be subsumed partly under the untypical contexts and partly under the critical contexts' (Diewald 2002: 117).

Figure 1 outlines the two models. Note that in both, grammaticalization (Gzn) occurs at Stage III.

<table>
<tr><th>Heine</th><th>Diewald</th></tr>
<tr><td rowspan="2">Stage I: 'normal' use</td><td>Stage 0: 'normal' use</td></tr>
<tr><td rowspan="2">Stage I: 'untypical' context</td></tr>
<tr><td rowspan="2">Stage II: 'bridging' context (pragmatic, semantic)</td></tr>
<tr><td>Stage II: 'critical' context (multiple opacity: pragmatic, semantic, structural)</td></tr>
<tr><td>Stage III: 'switch' context (Gzn)</td><td rowspan="2">Stage III: 'isolating' context (Gzn; reorganization and differentiation)</td></tr>
<tr><td>Stage IV: conventionalization</td></tr>
</table>

Figure 1: Heine's and Diewald's models compared (based on Heine 2002 and Diewald 2006).

The importance of discourse context, and of the pragmatic inferencing that occurs in it has recently been questioned in Fischer (2007); for example, she

attributes discourse marker meanings of expressions like *anyway, I think*, 'to their own lexical source-concept(s), to general semantic principles of change (notably metaphor)' (p. 312), and to analogical matching with other formulae. This approach partly falls out from her interest in promoting analogical thinking as a motivation for morphosyntactic change, and in part to a view of change as primarily internal to the language rather than the result of speaker-hearer negotiation of meaning. Since I regard change as the outcome of just such speaker-hearer negotiation in language use, I regard linguistic discourse context as essential to grammaticalization (it may, however, be irrelevant in certain cases of lexical semantic change, such as preemption of meaning for metatextual purposes (e.g. of 'construction' in the sense of 'syntactic string/phrase' to refer to a form-meaning pair in construction grammar)).

2.5 Questions to be addressed

Diewald (2006) provides a brief account of the development of German modal auxiliaries and the evidence they provide for the three Stages in her model, including the appearance of a large number of highly opaque examples in one specific 'critical' context prior to the appearance unambiguous ones. Diewald and Ferraresi (2008) provide a highly detailed account of the development of the German modal particle *eben* (cognate with English 'even'), and the evidence that it too provides for the three Stages. In both cases, the surge of opaque examples at Stage II is shown to abate when Stage III is reached (see also Krug's (2002) observation cited in Section 2.2 that 'a full third' of the examples of *want* in the second half of the seventeenth century are ambiguous).

In the next section I seek to answer the following questions with respect to Heine's and Diewald's models, testing them against two other data sets:

a) Is there evidence that the onset of grammaticalization is enabled by bridging contexts, understood as primarily pragmatic/semantic contexts (Heine), or rather by one specifiable critical context, understood to have pragmatic, semantic, and morphosyntactic properties (Diewald)?
b) Is there evidence for the hypothesis that Stage II, whether construed as bridging or critical, is necessary for grammaticalization (Diewald and Heine)?
c) Is there evidence that Stage II is short-lived (Diewald)?
d) Is there evidence that at Stage II the new meaning is foregrounded as Heine (2002) suggests, or backgrounded as Hansen (2008) suggests? The issue is not addressed explicitly in Diewald (2002); rather, the potential undecidability of many examples in a critical context is highlighted, as is the hypothesis that over time one reading came to be favored (p. 112).

At the end of Section 4 I will ask a further question:

e) Assuming that some version of the hypothesis that contexts drive change is plausible, is there some limit on the number of preceding and following clauses that should be imposed by the analyst?

3. Two examples

The development of *be going to* in the late Middle and Early Modern English periods has been discussed at considerable length by many researchers. Among reasons this development is interesting is that a biclausal structure is reduced to a monoclausal one that eventually participates in the auxiliary system.[12] Since it takes place within the 'envelope' of an originally complex clause construction, it does not require the analyst to look outside of this envelope when structural morphosyntactic, semantic, or invited inferencing contexts are considered. However, as has often been shown, one may need to investigate the larger discourse context/co-text in order to determine whether implicatures in a particular example are defeasible.

The development of ALL- and WH-clefts has received little attention to date. The pseudo-clefts are also clearly not traditional instances of grammaticalization. Among reasons their development is worth considering here is that, since there is no lexical element involved, pseudo-clefts are instances of the grammaticalization of non-lexical material, a topic of considerable interest recently (see e.g. Diessel 1999 on the grammaticalization of demonstratives; Molencki 1997, Sorva 2007 on the development of concessive *albeit*; and Dufter 2008, Lehmann 2008 on the development of *c'est*-clefts in French). Furthermore, the cleft expressions occur clause-initially, and therefore there is often no prior text within the sentential envelope to provide context. The analyst must therefore look to prior and following discourse to establish what kinds of pragmatic contexts are relevant.

3.1 *Be going to*

Although the development of verbs of motion, most especially *go*, but also *come*, has been a staple example in work on grammaticalization (see e.g. Bybee, Perkins, and Pagliuca 1994, Hopper and Traugott 2003 [1993]), remarkably little work has been done citing the linguistic contexts in which *be going to* first developed in English. Notable exceptions are Danchev and Kytö (1994) and Eckardt (2006). Drawing on the Helsinki Corpus, the Shakespeare Corpus, and the Elizabethan and Jacobean Drama Corpus, Danchev and Kytö explore the semantics associated with *be going to* in early examples, specifically movement/spatial displacement, temporality/futurity/modality/intention, aspectuality, and expressivity. While they sometimes cite extensive prior and subsequent context, they are more concerned with showing that multiple meanings are available, and exploring the 'hierarchy of semantic features' involved than with teasing apart semantic and pragmatic factors. For example, of (7), a very early example of participial *going* without the *be*-verb, they say that 'the semantic features of MOVEMENT, INTENTION and NEAR FUTURE seem to co-exist in a hierarchy that is difficult to determine' (p. 61):

(7) And thane come Englissh folk to the seid Merchauntz of the Maryknyght and bad theym beware whom they had lefte yn their Ship saying that yt was likely be taken And there vppon the said persones of the ship of Hull *goyng to* do the said wrong yaf to oon henry wales Gentilman duellyng abowte the coste of Develyn x marcz to lette and arreste the seid Maister and Merchauntz wan they come downe toward their Ship cleped Maryknyght

'And then English people came to the said merchants of the Maryknight and warned them about those they had left in their ship, saying that it was likely to be seized. And then the said people of the ship of Hull, going to do the said wrong, gave to one Henry Wales, gentleman, who lived on the coast of Develyn, ten marks to stop and arrest the said master and merchants when they came down toward the ship called Maryknight'.
(1483 *Chancery English,* 174 [ICAME: Helsinki; Danchev and Kytö 1994: 61; free translation added])

From the perspective of invited inferences, one might say the following: *going to* is semantically an expression of motion for a purpose; in the context of *do* there is a strong inference of intention and futurity (or at least prospective temporality) derived from purposive *to*; use in a non-finite modifying clause demotes activity and promotes intention. In (7) nothing prevents the motion verb meaning from being understood, but absence of any mention of going to a place in the immediately prior context (And there vppon the said persones of the ship of Hull) backgrounds it.

If we consider the larger context, however, we find that the intention/futurity meaning is only very locally foregrounded, since the sentence preceding (7) is (8):

(8) And w(h)ile the seid Maister and Merchauntes of the seid ship called Maryknyght were at diner the said persones of the Ship of Hull hyred theym hors priuely and rode downe to the seid shippes And there the same sonday they toke the said ship cleped Maryknyght lade thanne with Stokfyssh oyle and lynnencloth and other Merchaundises to the value of xvC li.

'And while the said master and merchants of the said ship called Maryknight were at dinner the said people of the ship of Hull secretly hired themselves horses and rode down to the said ships. And the same Sunday they seized the said ship called Maryknight, which was laden at the time with stockfish oil and linen and other merchandise to the value of 1551 marks'.

Motion (*rode downe*) has therefore been predicated of the persons of the ship of Hull, and so is salient in the larger context. From this perspective we cannot

really tell whether the 'new meaning' of futurity was preferred or not in *goyng to do the said wrong* in (7). Considering that the English folk and the master and merchants *come*, and everyone is therefore in motion, we might conclude that it is backgrounded. The analysis depends, then, on how much prior context we consider.

Danchev and Kytö go on to cite what has often been regarded as the first example of future-oriented *be going to* (with the *be*-verb), example (9):

(9) Therefore while thys onhappy sowle by the vyctoryse pompys of her enmyes *was goyng to* be broughte into helle for the synne and onleful lustys of her body.

'Therefore, while this unhappy soul by the victorious procession of her enemies was going to be brought into hell for the sin and unlawful lusts of her body'. (1482 *Revelation to the Monk of Evesham,* 43 [ICAME: Helsinki; Danchev and Kytö 1994: 61; free translation added)

They draw attention to the linguistic context: passive, past tense, and cite Hopper and Traugott's observation (1993: 83)[13] that motion is demoted in this context because there is no human agent, and the destination of the journey (hell) is an adjunct of *be broughte* not of *was goyng to*. Hopper and Traugott use this example to illustrate invited inferencing, and argue that because this inferencing is associated with the morphosyntactic context it is a kind of conceptual metonymy. They point out that in this particular case, extralinguistic knowledge of the belief that after death the soul goes on a journey with the purpose of being rewarded or punished for action in life is also important for understanding the passage. But in fact, we do not need extra-linguistic knowledge, if we allow our search in linguistic co-text to be far-reaching enough, since the revelation starts a page earlier with a scene in which the dreamer hears a great noise and sees a company of wicked spirits 'leading' her soul, which is a physical entity that they toss *as a tenyse balle* (Arber 1869: 42). Presumably the audience would have this in mind when hearing or reading (9).

In the corpus of *Early English Books Online* I have found a similar, slightly earlier example of *be going to*, also in the passive, and also in the past tense:

(10) Also ther passed a theef byfore alexandre that *was goyng to* be hanged whiche saide O worthy king saue my lyf for I repente me sore of my mysdedes.

'Also a thief who was going to be hanged passed before Alexander, and said "O worthy king, save my life, for a repent me deeply of my sins"'. (1477 Mubashshir ibn Fatik, Abu al-Wafa', 11th C; *Dictes or sayengis of the philosophhres* [LION: EEBO])

Striking here is the absence of any *by*-phrase, and if one were to extract only the head and the relative ('theef … that *was goyng to* be hanged'), as is often done in citations, demotion of motion and promotion of futurity would appear salient. However, *passed* at the beginning of the citation suggests that motion may have been on the writer's (or reader's) mind. This appears to be true of other early examples in *Early English Books Online*, many of them from *Froissart's Cronycles of Englande* (1523 and 1525). Typically there is a verb of motion in the larger co-text, reinforcing the motion meaning, although the immediate local context might convey intention or futurity. For example, in (11) *went* appears three times in the sentence preceding the one with *am goynge to*, and all refer to its subject, Sir Garses:

(11) Than this sir Garses went to delyuer them and as he wente sir Olyuer Clesquyn mette him & demaunded wheder he went and fro whens he came. I come fro my lorde the duke of Aniou and *am goynge to* delyuer the hostages. To delyuer them quod sir Olyuer abyde a lytell and retourne agayne with me to the duke.

'Then this Sir Garses went to deliver them, and as he went, Sir Oliver Clesquyn met him and demanded whither he went and from whence he came. "I come from my lord the Duke of Anjou and am going to deliver the hostages". "Wait a little to deliver the hostages", said Sir Oliver, "and return with me to the Duke"'. (1525 Froissart, 3rd 4th *Book of Cronycles of Englande* [LION: EEBO])

In sum, in the examples discussed so far, there is a possibility that a non-motion reading can be inferred, but it is not strong, and may be an artifact of the hindsight of knowing that the *be going to* future developed later. A review of the examples with the *be*-verb in the Helsinki Corpus, Early Modern English period 1500–1710, reveals that prior to 1700 most are probably opaque examples in which futurity is at best a weak inference. (12) is particularly interesting, since the first line of Allwit's second speech, which starts with *I am going to bid*, parallels the first line of his first speech, where *I'le goe bid* appears. 'I'll go' expresses future/ intention + motion. Danchev and Kytö (1994: 66), citing only the first two lines and the last line of (12), suggest the two expressions are 'mutually reinforcing synonyms'. This presumably means they interpret both *I'le goe bid* and *I am going to bid* as expressing intention/future. But when we look at the intervening lines, we find that Allwit discourses about walking forth, and running to and fro. We may therefore conclude that *am going to bid* is an ambiguous example, where both motion and intention/futurity are available. The two strings are therefore strictly speaking not 'synonyms' (furthermore, a synonym should require *I am going to go bid*):

(12) Enter Allwit

All. I'le goe bid ['summon'] Gossips presently my selfe,
That's all the worke I'le doe, nor need I stirre,
But that it is my pleasure to walke forth
And ayre my selfe a little: I am ty'd to nothing
In this business; what I doe is meerely recreation,
Not constraint.
Here's running to and fro, Nurse vpon Nurse…
(3 lines omitted)
Enter Sir Walter Whorehound.

S. Walt. How now (I aske) ?

All. I *am going to* bid Gossips for your Worship's child Sir,
A goodly Girle I faith, giue you ioy on her.

(1630 Middleton, *Chaste Maid* II.ii.1–13, Staged 1611–1613 [ICAME: Helsinki; see Danchev and Kytö 1994: 66])

In (13) we find *be going to* in the context of temporal expressions: the time the letter is written and the time that Cousin Dalison will have arrived at Dean, but there is no reason to think that this is not an opaque example, since the cousin's travel plans are under discussion:

(13) It is now about 12 of the clock, Mooneday noone and my Cozin Dalison *is going to* take water for Gravesend. Shee will bee at Deane Tuesday night. (1662 Oxindon, *Letters* [ICAME: Helsinki])

Likewise, in (14), where *this weeke* occurs after *is,* showing the string is not fully grammaticalized here, the topic is plans to travel, in this case to emigrate:

(14) Worthy Mr Ennis, who being turned out of his living here for not swearing and therfore not capacitated to exercise his ecclesiastick function in his own country, Scotland, *is* this weeke *going to* try whither he cannot more quietly live among ye heathens in America. (1692 Hatton, *Letters* [*ibid*.])

The first examples in the Helsinki Corpus that look as if they are in 'switch' or 'isolating' contexts where motion is virtually ruled out are two involving be married. It is true that one often goes to a place for the purpose of being married, but in both cases the linguistic context is the writer's evaluation of the situation, not travel, as in (15):

(15) There is one Mr Colson I am shure my Lady has seen at diner with my Unckle *is going to* be married, which one would wonder at, there being nothing to be liked in him but his fin diamond ring. (1699 Hatton, *Letters* [*ibid*.])

However, as Danchev and Kytö point out, there is a relatively early grammar in which *be going to* is defined as equivalent to *be about to*:

(16) About to, or *going to*, is the signe of the Participle of the future ….: as, my father when he was about [to] die, gave me this counsell. I *am [about]* or *going [to] read*. (1646 Poole, *Accidence* 26 [Danchev and Kytö 1994: 67; square brackets original])

This example is notable not only for the metalinguistic statement about future semantics, and the association of *be going to* with be *about to* (see Garrett 2012 for the importance of this association),[14] but also for the collocation of *be going to* with *read*, which is an unlikely (but not impossible) collocate of motion with a purpose. We may conclude, then, that *be going to* had grammaticalized for at least some speakers by the end of the first half of the seventeenth century, but that most instances were either motion or opaque examples, not temporal. In other words, during most of the seventeenth century individual speakers differed in whether they did or did not have a grammaticalized *be going to*. The new meaning was not conventionalized until the end of the century. By this time most speakers can be concluded to have the auxiliary as well as the motion meaning; a change has occurred (assuming change is not innovation in the individual, but involves spread to a community, see e.g. Weinreich, Labov, and Herzog 1968, Milroy 1992).

Eckardt cites Mossé (1938) as saying that toward the end of the sixteenth century 'ambiguous uses of *going to* get more frequent' (Eckardt 2006: 93) and as saying that 'around 1650, the construction can be seen as a firmly established part of English' (p. 94). In support of this she cites the example from Poole and other examples from the period without context. I have suggested that when we look at larger contexts, more examples should be considered to be ambiguous than is commonly assumed.

With respect to the questions posed in Section 2.5, it appears that the development of the *be going to* future depends at least in part on the availability of defeasible pragmatic meanings arising from the surrounding linguistic context, and that a Stage II can legitimately be posited. Since the increase in the availability of this reading is also a function of new morphosyntactic and lexical contexts, such as the passive and especially verbs that are not strongly compatible with motion,[15] this stage is best associated with Diewald's 'critical' context. There does, however, not appear to be one specific context in which semantic and structural possibilities that were distributed over different contexts 'accumulate in one specific critical context', so there is no 'critical context' in the narrow sense. It remains an empirical question whether the notion of critical context should be expanded from a single context to a limited set of contexts. The hypothesis that Stage II is a necessary precursor for grammaticalization is supported by the example of *be going to*. While ambiguous examples continue to be used (Diewald would suggest these are a continuation of Stage I), it is noticeable that there is a

surge of opaque examples in the seventeenth century; this levels out around 1700. Therefore, there is some evidence that Stage II is relatively short-lived.

It appears furthermore that in order to understand the micro-changes that *be going to* underwent, we need to look at larger co-texts, at a minimum the sentences preceding and following the one in which it occurs, and preferably more. When we do this, the prominence/salience/foregroundedness of the new reading becomes less easy to determine.

3.2 The development of pseudo-clefts

My second example is the development of a subset of ALL- and WH-pseudo-clefts. These involve reanalysis of the information structure and syntax of complex strings of the type ALL/WHAT NP DO BE X, where X is a non-finite clause (for studies of the development of a larger range of sub-types of ALL- and WH-pseudo-clefts, see Traugott 2008a, 2010).

Before going into the history of these pseudo-clefts, it is useful to consider aspects of their structure in contemporary English. The literature on WH-pseudo-clefts is vast, going back to the early 1970s with groundbreaking analyses by Higgins (1979),[16] using constructed examples, and Prince (1978), using a mix of data from newspapers and other written texts as well as constructed examples. One of the current debates about pseudo-clefts concerns the function of WH-clefts, most especially whether it is primarily given-new information-packaging (e.g. Prince 1978), or, at least in conversation, management of interaction (e.g. Hopper and Thompson 2008). ALL-clefts have, by contrast, received almost no attention other than Bonelli (1992). Examples that are especially relevant to the discussion that follows are:

(17) a. Nikki Caine, 19, doesn't want to be a movie star. ***What she hopes to do is be a star on the horse-show circuit***. (10/10/1976 *Today*, p. 44 [Prince 1978: 887])
b. … classical music was just 'music', and therefore ***all one had to do was to listen to it***. (COBUILD 10 Million Corpus [Bonelli 1992: 36])

Key points are that in pseudo-clefts such as these:

a) There are two clauses, one of which is a (reduced) relative, one of which involves a copula.
b) Some part of the construction (typically the relative) must be given or at least recoverable.
c) The focus constituent (X following the copula) is construed as an exhaustive, exclusive listing (Nikki only wants to be a star on the horse-show circuit, not any other kind of star; the speaker had only to listen, not do anything else such as dance).
d) In ALL- and WH-clefts DO refers to the same event as V in X.
e) In ALL- and WH-clefts temporality matches across the two clauses.[17]

f) In ALL-clefts *all* does not mean 'everything' and is replaceable by *only* (it is 'downward inferential', see Horn 1996: 18); ALL-clefts typically signal that the speaker/writer regards the focus as less than adequate (Kay 2002).

Early examples of ALL-clefts that I have found appear around 1600, and of WH-clefts around 1660.[18] Both types are found at this period almost exclusively with DO + non-finite clause or with SAY + finite clause. Here I consider only the first type, while recognizing that the two types may have influenced each other. The questions I seek to answer are i) what were the historical antecedents of the DO-type?, and ii) what evidence is there for either bridging or critical contexts?

Prior to the first examples that appear to be pseudo-clefts we find purposive clauses. They are prospective, i.e. future-oriented, because of the purposive, and *all* as in (18a) means 'everything':

(18) a. I loue thee dearer then I doe my life,
And ***all*** *I did, was* to aduance thy state,
To sunne bright beames of shining happinesse.
(1601 Yarrington, *Two Lamentable Tragedies* [LION: EEBO])

b. Shal. Will you, upon good dowry, marry her?
Slen. I will do a greater thing than that, upon your request, cousin, in any reason.
Shal. Nay, conceive ['understand'] me, conceive me, sweet coz. ***What I do is to pleasure*** ['please'] **you,** coz. Can you love the maid?
(?1597 Shakespeare, *Merry Wives of Windsor* I.i.250 [LION: Shakespeare])

c. Mistake mee not faire Knight, (said shee) for by my past thoughts I protest he is the God of my desiers, ***what I did, was to deceiue the Pagans,*** who are waking Dragons that neuer sleepe about mee. (1612 Markham, *Meruine* [LION: EEBO])

Criteria a) and b) for pseudo-clefts are fulfilled: there are two clauses, one of which is a (reduced) relative, and a givenness relationship can be inferred. However, the other criteria are not fulfilled. What was done was an action separate from and earlier than X, but meant to contribute to X (advancing the addressee's state, giving pleasure, deceiving the pagans).

Early Modern English examples of the ALL-cleft that do fulfill all six criteria include:

(19) a. But, if Mr. Husband give over before you, gett an inhibition ('legal stay'): (…) there is no possibilitie of overthrowing the new election which shalbe made when the place is voyd, and if it be so allready, or shalbe so, ***all you can doe is to do some good for the tyme to come,***

which if you can doe conveniently, and without much trouble, it wilbe woorth your labour (1624 Oliver Naylor to John Cousin [CEEC])

b. There was no possibility of my leaving the Army to fetch her out of that Convent … I never could obtain Leave to be absent, but remain'd most part of the Winter there; ***all I could do was to order some Soldiers, that went for France, to call at Charleville,*** but I never heard from them since. … I conjur'd him not to carry things to that Extremity; but he was inexorable, and ***all I could do was to acquaint Isabella with it***… a Brute, who kept no Measures with any one, I thought fit to put off my Courtship to another Time. ***All I did was to tell my eldest Brother of it***. (1697 Saint-Evremond, *Female Falsehood* (trans.) [LION: EEBO])

c. What need'st thou woman such a whyning keepe?
Thy sonn's as well as anie man ith' lande,
Why ***all he did, was bidd a man but stande***,
And told him coyne he lackt; there's those doe worse,
Then bidd an honest man deliver's purse. (1616 Goddard, *A Mastiff VVhelp* [*ibid*.])

In these examples *all* is understood to be downward inferential, the focus X is an exhaustive listing; it is not purposive, and its temporality is the same as that of the first clause. In each example *do* is highly bleached and could be omitted (*You can only do some good; I could only order some soldiers to call; I could only acquaint Isabella with it; I only told my brother; he bade a man stand*). In other words, *do* is a kind of pro-verb for the verb in the infinitive clause. The loss of the purposive meaning of *do* presumably allowed for the loss of *to* in (19c); here the *but* 'only' overtly expresses the exhaustiveness associated with the construction suggesting that exhaustiveness has not yet been fully semanticized in the ALL NP DO BE to V construction.

Examples of the WH-cleft that fulfill the first five criteria for pseudo-clefts are:

(20) a. there were no reason that my presence should bring any constraint to your actions, neither was it the occasion that drew me to put you out of those fancies, ***but what I did was onely to shew you these other pictures, and to ask your opinion of them;*** tell me then, I pray you, what you think of the Princesse which is next to *Armazia?* (1640 Duverdier, *Love and Arms*, trans. from French by Early of Pembroke [LION: EEBO])

b. although he [Mr. Baxter] sometimes pretends only to Preach to some of many thousands, that cannot come into the Temples, many of which never heard a Sermon of many years; and to this purpose he put so many Quaere's to me, concerning the largeness of Parishes, and the necessity of more Assistants, thereby to insinuate, ***That what he did, was only to Preach to such, as could not come to our Churches;*** yet,

> when he is pinch'd with the point of *Separation,* then he declares … (1661 Stillingfleet, *Unreasonableness of Separation* [*ibid*.])

It is striking that both of these early examples include *only* in X. Like *but* in (19c), *only*, being a focus marker, carries the meaning of exhaustiveness. This suggests that the pseudo-cleft construction had not crystallized fully yet. However, *do* and the verb in X refer to the same action and share the same temporality.

I turn now to the question whether the development of this particular subset of pseudo-clefts shows evidence of a stage where bridging or critical contexts predominate. In the case of ALL-clefts studied here, the contexts are typically not opaque, but rather general metalinguistic discussion of how one person's *all* may be considered inadequate:

(21) My heart is confident and bold within,
Since ***all I did was but to punish sinne***:
If in some circumstances, faile I shall,
To be accuser, witnesse, Iudge and all,
My witnesse-bearing thus I iustify…
(1622 Aylett, *Susanna* [LION: EEBO])

In sum, the chief local linguistic contexts that appear to have contributed to the development of the pseudo-clefts are:

a) X-clauses restricted by *but* 'only', *only, merely*,
b) Negative, contesting contexts.

These are the contexts in which unambiguous examples throve in the seventeenth century, even after the new meaning was unambiguously available. Over time, association of the construction with exhaustiveness led to the disappearance of focus particles like *only* in X-clauses, but the pseudo-clefts continue to be favored in contesting contexts, as the contemporary examples in (17) show.

I have not found convincing ambiguous early examples in the sense that there is one meaning and a second that is implied and not defeasible. There are, however, a couple of examples which appear to be genuine ambiguous examples because two perspectives are represented, one of which is pragmatically inferrable. Consider first (22) with *all*; here the verb *drive* is roughly equivalent to *do* in its sense of intentional activity:

(22) By all which your Honours may perceive, how he hath falsly traduced the Commissioners of the Navie, the Masters, Wardens, and Assistants of the Trinitie-house; the principall men of the Corporation of the Ship-wrights; ***and all he drives at, is by his unjust aspersions to bring the Parliament and them at ods, that so he might accomplish his own ends.*** (1646 mscb.sgm [ICAME: Lampeter])

From the perspective of the person referred to (*he*), actions performed are for a purpose (causing disagreement), but from the perspective of the writer, the result of his actions can be inferred to be an exhaustive listing (*he only brings the Parliament and the commissioners at odds*). This inference derives from the iterative aspect and from the explicit purposive that follows (*so that he might accomplish*) and demotes the potential purpose reading of *to* in the prior clause. The writer's negative perspective clearly dominates (cf. *falsly, unjust*), and colors *all*, so that the latter can be inferred to be downward inferential (but probably not as strongly so as in a genuine ALL-cleft, where a 'less than adequate' reading would be expected in addition to the 'only' reading). From the hind-sight of four centuries in which ALL-clefts have been used, (22) appears to allow both meanings. In this instance the new meaning is preferred, but the older one is not defeased.

A similar kind of opaque example with a WH-cleft is illustrated by (23):

(23) And that there had lately appeared to him a Vision, which bad him, arise and Dig and plow the Earth, and receive the Fruits thereof; that their intent is to restore the Creation to its former condition. That as God had promised to make the barren Land fruitful, so now ***what they did, was to renew the ancient Community of injoying the fruits of the Earth, and to distribute the benefit thereof to the poor and needy, and to feed the hungry and cloath the naked***. (1682 Whitlocke, *Memorials* [LION: EEBO])

Without the prior sentence and *their intent is to restore the Creation to its former condition*, the infinitives might be understood to corefer with *do* (*they renewed, distributed, and fed*), in which case the reading would be pseudo-cleft. But in the context of *intent*, the older meaning (*did for the purpose of renewing…*) is also inferrable.

The pseudo-clefts are therefore unlike *be going to* in that little evidence is provided of significant amounts of ambiguity, whether primarily pragmatic/semantic, or pragmatic/semantic and structural. There is no evidence of a critical context Stage II, in the narrow sense, that is, of a specific construction which creates ambiguity and invites reinterpretation. Rather, the relevant contexts for the change are contesting discourses, including those in which a narrow interpretation of X as exhaustive is triggered by uses with *only/but/merely*. One approach that could save the notion that an ambiguous Stage II is necessary might be to argue that the development of the pseudo-clefts is not a case of grammaticalization. However, Traugott (2008a) shows why the development of the class of pseudo-clefts in general (i.e. where V is *say*, *do*, or any of the other verbs that later came to be used, such as *mean, think*) is an instance of grammaticalization without lexical source. Suffice it here to mention that the reorganization of the purposive clause is typical of grammaticalization: *do* and the purposive are bleached (so much so that *to* is lost),[19] and the whole construction comes to serve an information-structure function, at least in writing.

A more comprehensive study of the pseudo-clefts is likely to show that a variety of expressions contributed to their development, among them more lexical expressions such as:

(24) The only good act that ever my brother did, was to bring us acquainted, and is indeed all that he has to live on. (1658 Brome, *The Weeding of the Covent-Garden* II.ii [LION: Prose Drama])

This is consistent with Patten's (2010) hypothesis that clefts in general should be considered part of a larger network of specificational constructions. The point here is that neither bridging nor critical contexts narrowly construed appear to have been a necessary stage in the development of the pseudo-clefts, at least as far as the corpora investigated can provide empirical evidence for hypotheses. Stage II in both models is therefore optional, and not required.

4. Conclusion

From this brief study we may conclude that:

1) The hypotheses of 'bridging' and 'critical' contexts usefully shift attention away from a concept of changes originating in the pragmatics internal to structures to a concept of the changes originating in the immediate and potentially the larger co-texts.
2) Inferential pragmatics are key to enabling grammaticalization, but bridging contexts understood in terms of pragmatics and semantics alone are too restrictive to trigger grammaticalization. Indeed, we would in fact expect that changes in the manipulation of inferential pragmatics should be accompanied by other developments, such as the structural shifts that Diewald identifies with critical contexts, given that grammaticalization necessarily involves morphosyntactic and 'host-class', as well as semantic-pragmatic expansion (Himmelmann 2004).
3) There is no evidence that the available options are accumulated in just one critical context prior to grammaticalization.
4) There may not always be a 'Stage II' in which bridging or critical contexts necessarily precede the development of a new grammatical use.
5) Therefore there may be no Stage II that is short-lived.
6) It is more important to determine the range of 'contexts of origin' in a particular case of grammaticalization than to focus on just one type (see also Eckardt 2006: 94).
7) As mentioned in Section 2.5, Hansen (2008) regards bridging examples as backgrounding the new meaning, whereas Heine (2002) regards them as foregrounding it. If the issue is one of whether one reading is preferred over or more prominent than another, then the textual evidence discussed here suggests that whether the original meaning or the new pragmatic

> meaning is back- or fore-grounded is a function of the co-text. If it is one of availability above the level of consciousness, as Hansen implies when she refers to availability for exploitation, this too depends on co-text, and how much intentionality we attribute to speakers/writers. For example, how intentional is the ambiguity that we can read into (22)?

In sum, a purely structural approach to onset/triggering contexts is not sufficient. Pragmatic factors need to be taken into account. 'Critical' contexts that combine pragmatic with structural factors are more appropriate than 'bridging' contexts in accounting for the onset of grammaticalization. However, a modification of Diewald's hypotheses is called for. The fundamental difference is that Stage II is optional, and critical context is to be understood in an extended sense. It does not imply restriction to one specific context. It should be understood only as pragmatic, semantic, and structural context.

Readers will have noticed that I have cited ambiguous examples that occur after some clear cases of the new grammaticalized constructions came into being. This is contrary to Eckardt's (2006: 97) methodology, which is not to count such examples of *be going to* after 1646, because of example (16) from Poole. Her reason is presumably that the 'landscape' that speakers draw on and addressees interpret is by hypothesis different once the new grammatical meaning has been isolated.[20] Specifically, we are dealing not with pragmatic, but with semantic ambiguity. My reason is that we cannot know from one writer's use whether an innovation has spread to the community. More importantly, opaque examples appear throughout the life of polysemous constructions,[21] not only as points of origin, but by hypothesis as points of maintenance of the link between the originally related constructions. Moore (2007) even suggests that they may form a 'feedback loop' leading to avoidance of ambiguity in certain genres, and hence changes in different genres. It has long been noted that old meanings and uses may persist and by hypothesis constrain the newer polysemies (e.g. König and Traugott 1982, Bybee and Pagliuca 1987, Hopper 1991). Bybee, Perkins, and Pagliuca (1994: 16) also point to the retention of 'certain more specific semantic nuances of the source construction'. The extent to which opaque contexts contribute to persistence in both 'source' and 'target' polysemies remains to be investigated. So does whether the effects of bridging contexts is different in different registers. It seems likely that there might be significant differences in speech and writing, since stress, intonation, and phonology are all often cues to grammaticalization in speech but may be considerably less so in writing. Indeed, it is at least conceivable that naïve contemporary speakers might find *be going to* potentially ambiguous in writing, but extremely unlikely that they would do so in speech.

Two issues deserve further consideration. One concerns whether ambiguity is a prerequisite for grammaticalization, and if so, whether certain domains are more likely to evidence ambiguity than others. It has long been thought that structural ambiguity is a prerequisite for reanalysis (see e.g. Timberlake 1977, Haspelmath 1998). Intuitively, this is a plausible assumption,

given than structural reanalysis involves parsing, and differentiation of older and newer interpretations of a string. However, this assumption has been challenged by Harris and Campbell, who say: 'reanalysis does not depend upon opacity or upon lack of evidence supporting the old analysis' (1995: 72), a position shared by Detges and Waltereit (2002). Detges and Waltereit suggest that ambiguity is a 'natural result of' but not a prerequisite to reanalysis (p. 170). While Haspelmath (1998) regards ambiguity as a prerequisite to reanalysis, he says: 'there is no such requirement for grammaticalization change' (p. 326). Heine's and Diewald's move was to focus on pragmatic and semantic ambiguity in grammaticalization. Here I have shown that on the one hand we have evidence that, in certain domains and in certain languages, there is a clearly attested stage in which pragmatic, semantic, and even structural ambiguity abounds before a grammaticalization change crystallizes. In English *be going to* and *want* are good examples, in German, the modals and modal particles. But other domains show very little evidence of such ambiguity prior to the crystallization of a new grammatical construction. These include the pseudo-clefts discussed here and the quantifiers and degree modifiers derived from partitives like *a lot/bit/shred of* discussed in Traugott (2008b). In the latter cases some ambiguous examples occur, but they are infrequent, and may post-date clear examples of grammaticalized constructions. This issue is likely to come to be of particular importance as the role of analogy or constructional attraction in grammaticalization is further explored (see e.g. Detges and Waltereit 2002, Fischer 2007, Traugott 2008b, Trousdale 2010, Kiparsky 2012), since it is not immediately obvious in what ways ambiguity might be a major factor in analogical or constructional change, which privilege matching to or extension of existing patterns, rather than differentiation.

Since ambiguity, most especially pragmatic ambiguity, is largely a function of context, the second issue that needs considerable attention is how much context the researcher should take into consideration in order to determine whether ambiguity is an issue, and if so, what kind it is. The conclusions drawn here suggest that if we are to understand micro-changes and seek to do statistical analyses on them, we need to take into consideration not only the immediate structural contexts of expressions we are investigating, but also the co-text. As a practical matter, it seems that for the kinds of data investigated here, ten clauses of prior text and three of following text may be sufficient to provide a proper analysis. However, for some other types of investigation more following context may be necessary. For example Traugott and Pintzuk (2008) advocate the use of ten clauses prior and following for study of information structure in Old English. It would be useful in future work to establish a norm that can be used across all historical analyses.

Notes

1 Many thanks to Gabriele Diewald, Bernd Heine, Merja Kytö, and Graeme Trousdale for comments on an earlier draft. I am of course responsible for any errors of fact or interpretation.

2 The term is, however, sometimes used more loosely to refer to any ambiguous context, syntactic as well as semantic. I will use it in the narrow sense given here.

3 Lehmann (2004: 155) objects that this characterization renders the concept too 'wide and heterogeneous'.

4 Non-linguistic contexts are also of great importance, e.g. contact, genre and register, and changing ideologies. For reasons of space, I will limit myself to linguistic ones.

5 'Primary' is understood as a 'grammatical category such that its class comprises words with lexical meaning or higher-level grammatical units' (parts of speech and 'all syntactic categories'); 'secondary' is understood as a category 'such that its class comprises grammatical signs', or morphology (e.g. inflections) (Lehmann 1993: 4).

6 I ignore such phonological and phonetic factors as stress and intonation here.

7 For challenges to the hypothesis that pragmatic factors necessarily precede structural shifts, see e.g. Fischer (2007).

8 The 'entailment' and 'inference' of 'lack' are need and desire, respectively (Krug 2002: 140). However, one can lack a car, a vase full of roses, or a copy of a book without necessarily needing one. Krug appears to use the two terms 'conversational implicature' and 'invited inference' as alternates. However, they imply rather different foci of attention on the analyst's part, the first on the speaker alone, the second on the speaker as member of an interactional dyad.

9 For the concept of semantic prosody (also referred to as 'harmony') see e.g. Stubbs (1995), who points out that certain items tend to collocate with members of a negatively or positively-oriented set, e.g. *cause* with negatively oriented NPs (*accident, cancer, crisis*).

10 'Bridging contexts' appear to have much in common with pragmatic interpretations of 'vagueness' (for some discussion of the problems attendant on distinguishing polysemy and vagueness, see Geeraerts 1993, Tuggy 1993).

11 Grice (1967) is the same as Grice (1989 [1975]) in the references.

12 Harris and Campbell (1995), among many others, regard reduction of biclausal structure as typical of auxiliation. By contrast, Fischer (2007: Chapter 5) questions the generality of biclausal sources of auxiliaries, especially modal auxiliaries in English.

13 This is Hopper and Traugott (2003: 89).

14 However, the presence of the very early passive examples in (10) and (11) suggests that the development of passives with *be going to* that Garrett (2012) identifies from 1630 on may not be as supportive as he suggests of his hypothesis that future *be going to* derives not from the motion with a purpose construction, as is usually assumed, but from an inceptive 'turning or preparing to do an action/to be about to or on the verge of'. Garrett cites *He is fumbling with his purse-string, as a school-boy with his points when he is going to be whipped, till the master weary with long stay forgives him* (1628 Earle, *Microcosmography* [Mossé 1938: 166]). Note also that the example in question could be construed as involving motion (the boy may have to go somewhere, e.g. the headmaster's room or some public space, for the whipping).

15 Hilpert (2008: Chapter 3) interestingly shows that by the early eighteenth century *be going to* cooccurs with *marry* and speech act verbs like *say, answer, observe* ('say'), rather than with verbs of motion such as auxiliary *gaan* in Dutch collocates with.

16 This is the published, slightly revised version of Higgins's 1971 dissertation.

17 Higgins (1979: 309–315) treats this as a morphosyntactic criterion generalizable to most specificational sentences; it is better understood as a semantic criterion.

18 There is one earlier one, dated 1640, cited in (20a), but WH-clefts in general did not appear with any frequency till the 1660's and later. (20a) is in a translation from French, as are several other examples. The extent to

which French models may have been influential remains to be investigated.

19 While *to* could be omitted early with ALL-clefts very early (see (18c)), it was apparently not omissible until the twentieth century with WH-clefts. The 1980's are a period when the loss of *to* in both constructions became wide-spread.

20 Thanks to Gabriele Diewald for pointing this out to me.

21 This is consistent with Tuggy's demonstration that words can be polysemous or vague in different contexts. His example is *paint*, as in 'paint artistically', and 'paint strips on the pavement'; in some cases these allow ellipsis (*When I'm painting I try to get the color on evenly, and so does Jane* is felicitous under either interpretation; here *paint* is vague), in others it does not (*I am a painter and so is Jane* is not felicitous if one paints artistically and the other paints stripes on a road; here *paint* is polysemous) (Tuggy 1993: 277).

Sources

CEEC *Corpus of Early English Correspondence*, Oxford Text Archive, http://ota.ahds.ac.uk (accessed June 2008)

COBUILD *Collins Birmingham University International Language Database*, http://www.collins.co.uk/Corpus/CorpusSearch.aspx

ICAME *ICAME Corpus Collection*, http://nora.hd.uib.no/corpora.html (accessed June 2008)

LION *Literature Online*, http://lion.chadwyck.com (accessed June 2008)

UVa University of Virginia, Electronic Text Center, *Modern English Collection*, http://etext.lib.virginia.edu/modeng/modeng0.browse.html (accessed June 2008)

References

Arber, Edward (ed.) (1869), *The revelation of the Monk of Evesham* (English Reprints 18). London: Murray.

Bach, Kent (2004), 'Pragmatics and the philosophy of language', in: Horn and Ward (eds.), 463–487.

Bisang, Walter, Nikolaus Himmelmann and Björn Wiemer (eds.) (2004), *What makes grammaticalization – a look from its fringes and its components* (Trends in Linguistics: Studies and Monographs 158). Berlin/New York: Mouton de Gruyter.

Bobrow, Daniel G., Robert D. Cheslow, Cleo Condoravdi, Lauri Karttunen, Tracy Holloway King, Rowan Nairn, Valeria de Paiva, Charlotte Price and Annie Zaenen (2007), 'PARC's Bridge question answering system', in: Tracy Holloway King and Emily M. Bender (eds.) *Proceedings of the Grammar Engineering Across Frameworks (GEAF07) Workshop*. Stanford University, July [13th–15th] 2007. Available at: http://www2.parc.com/istl/members/karttune/publications/bridge.pdf (accessed May 4th 2010).

Bonelli, Elena Tognini (1992), '"All I'm saying is …": the correlation of form and function in pseudo-cleft sentences', *Literary and Linguistic Computing*, 2: 30–41.

Brinton, Laurel J. and Elizabeth Closs Traugott (2005), *Lexicalization and language change* (Research Surveys in Linguistics). Cambridge: Cambridge University Press.

Bybee, Joan (2006), 'From usage to grammar: the mind's response to repetition', *Language*, 82: 711–733.

Bybee, Joan L. and William Pagliuca (1987), 'The evolution of future meaning', in: Anna Giacalone Ramat, Onofrio Carruba and Giuliano Bernini (eds.) *Papers from the 7th International Conference on Historical Linguistics,* (Current Issues in Linguistic Theory 48). Amsterdam/Philadelphia: Benjamins, 108–122.

Bybee, Joan, Revere Perkins and William Pagliuca (1994), *The evolution of grammar: tense, aspect, and modality in the languages of the world.* Chicago: University of Chicago Press.

Clark, Herbert H. (1975), 'Bridging', in: *TINLAP '75: proceedings of the 1975 Workshop on Theoretical Issues in Natural Language Processing*, 169–174.

Clark, Herbert H. (1996), *Using language*. Cambridge: Cambridge University Press.

Clark, Herbert H. and Susan E. Haviland (1977), 'Comprehension and the given-new contract', in: Roy O. Freedle (ed.) *Discourse production and comprehension* (Discourse Processes: Advances in Research and Theory 1). Norwood, N. J.: Ablex, 1–40.

Cooper, Robin and Ruth Kempson (eds.) (2008), *Language in flux: dialogue coordination, language variation, change and evolution.* (Communication, Mind & Language 1). *London:* Kings College.

Croft, William (2006), 'Typology', in: Mark Aronoff and Janie Rees-Miller (eds.) *The handbook of linguistics* (Blackwell Handbooks in Linguistics). Oxford: Blackwell, 337–368.

Danchev, Andrei and Merja Kytö (1994), 'The construction *be going to* + *infinitive* in Early Modern English', in: Dieter Kastovsky (ed.) *Studies in Early Modern English* (Topics in English Linguistics 13). Berlin/New York: Mouton de Gruyter, 59–77.

Detges, Ulrich (2006), 'From speaker to subject. The obligatorization of the Old French subject pronouns', in: Hanne Leth Andersen, Merete Birkelund and Maj-Britt Mosegaard Hansen (eds.) *La linguistique au coeur. Valence verbale, grammaticalisation et corpus. Mélanges offerts à Lene Schøsler à l'occasion de son 60e anniversaire* (University of Southern Denmark Studies in Literature 48). Odense: University Press of Southern Denmark, 75–103.

Detges, Ulrich and Richard Waltereit (2002), 'Grammaticalization vs. reanalysis: a semantic-pragmatic account of functional change in grammar', *Zeitschrift für Sprachwissenschaft*, 21: 151–195.

Diessel, Holger (1999), *Demonstratives: form, function, and grammaticalization* (Typological Studies in Language 42). Amsterdam/Philadelphia: Benjamins.

Diewald, Gabriele (2002), 'A model for relevant types of contexts in grammaticalization', in: Wischer and Diewald (eds.), 103–120.

Diewald, Gabriele (2006), 'Context types in grammaticalization as constructions'. *Constructions* SV1–9. Available at: http://www.constructions-online.de/articles/specvol1/686 (accessed May 4th 2010).

Diewald, Gabriele and Gisella Ferraresi (2008), 'The diachronic rise of modal particles in German', in: López-Couso and Seoane (eds.), 77–110.

Dufter, Andreas (2008), 'On explaining the rise of *c'est*-clefts in French', in: Ulrich Detges and Richard Waltereit (eds.) *The paradox of grammatical change* (Current Issues in Linguistic theory 293). Amsterdam/Philadelphia: Benjamins, 31–66.

Eckardt, Regine (2006), *Meaning change in grammaticalization: an enquiry into semantic reanalysis* (Oxford Linguistics). Oxford/New York: Oxford University Press.

Enfield, Nicholas J. (2005), 'Micro- and macro-dimensions in linguistic systems', in: Sophia Marmaridou, Kiki Nikiforidou and Eleni Antonopoulou (eds.) *Reviewing linguistic thought: converging trends for the 21st century* (Trends in Linguistics. Studies and Monographs 161). Berlin/New York: Mouton de Gruyter, 313–325.

Evans, Nicholas and David Wilkins (2000), 'In the mind's ear: the semantic extensions of perception verbs in Australian languages', *Language*, 76: 546–592.

Ewing, Michael G. (2005), *Grammar and inference in conversation: identifying clause structure in spoken Javanese* (Studies in Discourse and Grammar, 18). Amsterdam/Philadelphia: Benjamins.

Fischer, Olga (2007), *Morphosyntactic change: functional and formal perspectives*. Oxford: Oxford University Press.

Garrett, Andrew (2012), 'The historical syntax problem: reanalysis and directionality', in: Jonas, Whitman and Garrett (eds.), 52–72.

Geeraerts, Dirk (1993), 'Vagueness's puzzles, polysemy's vagaries', *Cognitive Linguistics*, 4: 223–272.

Geis, Michael L. and Arnold M. Zwicky (1971), 'On invited inferences', *Linguistic Inquiry*, 2: 561–566.

Grice, H. Paul (1989 [1975]), 'Logic and conversation', *Studies in the way of words*. Cambridge, MA/London: Harvard University Press, 22–40.

Hansen, Maj-Britt Mosegaard (2008), *Particles at the semantics/pragmatics interface: synchronic and diachronic issues. A study with special reference to the French phasal adverbs* (Current Research in the Semantics/Pragmatics Interface 19). Oxford: Elsevier.

Harris, Alice and Lyle Campbell (1995), *Historical syntax in cross-linguistic perspective.* Cambridge: Cambridge University Press.

Haspelmath, Martin (1998), 'Does grammaticalization need reanalysis?', *Studies in Language*, 22: 315–351.

Heine, Bernd (2002), 'On the role of context in grammaticalization', in: Wischer and Diewald (eds.), 83–101.

Heine, Bernd, Ulrike Claudi and Friederike Hünnemeyer (1991), *Grammaticalization: a conceptual framework*. Chicago: University of Chicago Press.

Heine, Bernd and Tania Kuteva (2007), *The genesis of grammar: a reconstruction.* Oxford: Oxford University Press.

Heine, Bernd and Hiroyuki Miyashita (2008), 'Accounting for a functional category: German *drohen* "to threaten"', *Language Sciences*, 30 (1): 53–101.

Higgins, F. R. (1979), *The pseudo-cleft construction in English* (Outstanding Dissertations in Linguistics). New York: Garland.

Hilpert, Martin (2008), *Germanic future constructions. A usage-based approach to language change* (Constructional Approaches to Language 7). Amsterdam/Philadelphia: Benjamins.

Himmelmann, Nikolaus P. (2004), 'Lexicalization and grammaticalization: opposite or orthogonal?', in: Bisang, Himmelmann and Wiemer (eds.), 19–40.

Hopper, Paul J. (1991), 'On some principles of grammaticization', in: Traugott and Heine (eds.), vol. I: 17–35.

Hopper, Paul and Sandra A. Thompson (2008), 'Projectability and clause combining in interaction', in: Ritva Laury (ed.) *Crosslinguistic studies of clause combining: the multifunctionality of conjunctions* (Typological Studies in Language 80). Amsterdam/Philadelphia: Benjamins, 99–123.

Hopper, Paul J. and Elizabeth Closs Traugott (2003 [1993]), *Grammaticalization* (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press.

Horn, Laurence R. (1985), 'Metalinguistic negation and pragmatic ambiguity', *Language*, 61: 121–174.

Horn, Laurence R. (1996), 'Exclusive company: *only* and the dynamics of vertical inference', *Journal of Semantics*, 13: 1–40.

Horn, Laurence R. and Gregory Ward (eds.) (2004), *The handbook of pragmatics* (Blackwell Handbooks in Linguistics). Oxford/Malden, MA: Blackwell.

Jonas, Dianne, John Whitman and Andrew Garrett (eds.) (2012), *Grammatical change: origins, nature, outcomes*. Oxford: Oxford University Press.

Kay, Paul (2002), 'Patterns of coining'. Paper presented at the Second International Conference on Construction Grammar (ICCG), Helsinki, Sept.; see www.icsi.berkeley.edu/~kay/coining.pdf (accessed May 4th 2010).

Kearns, Kate (2006), 'Lexical semantics', in: Bas Aarts and April McMahon (eds.) *The handbook of English linguistics* (Blackwell Handbooks in Linguistics). Oxford/Malden, MA: Blackwell, 557–580.

Keller, Rudi (1994 [1990]), *On language change: the invisible hand in language*. Translated by Brigitte Nerlich. London: Routledge.

Kempson, Ruth M. (1996), 'Semantics, pragmatics, and natural-language interpretation', in: Shalom Lappin (ed.) *The handbook of contemporary semantic theory* (Blackwell Handbooks in Linguistics). Oxford/Malden, MA: Blackwell, 561–598.

Kiparsky, P. (2012), 'Grammaticalization as optimization', in: Jonas, Whitman and Garrett (eds.), 15–51.

König, Ekkehard and Elizabeth Closs Traugott (1982), 'Divergence and apparent convergence in the development of "yet" and "still"', in: Monica Macaulay, Orin D. Gensler, Claudia Brugman, Inese Civkulis, Amy Dahlstrom, Katherine Krile and Rob Sturm (eds.) *Proceedings of the Eighth Annual Meeting of the Berkeley Linguistic Society*. University of California, Berkeley: Berkeley Linguistics Society, 170–179.

König, Ekkehard and Letizia Vezzosi (2004), 'The role of predicate meaning and the development of reflexivity', in: Bisang, Himmelmann and Wiemer (eds.), 213–244.

Krug, Manfred (2002), 'A path to volitional modality', in: Teresa Fanego, María José López-Couso and Javier Pérez-Guerra (eds.) *English historical syntax and morphology: selected papers from 11 ICEHL, Santiago de Compostela, 7–11 September 2000* (Current Issues in Linguistic Theory 223). Amsterdam/Philadelphia: Benjamins, 131–147.

Kuteva, Tania (2001), *Auxiliation: an enquiry into the nature of grammaticalization.* Oxford: Oxford University Press.

Kytö, Merja and Suzanne Romaine (2005), '*We had like to have been killed by thunder & lightning*: the semantic and pragmatic history of a construction that like to disappeared', *Journal of Historical Pragmatics*, 6: 1–35.

Lehmann, Christian (1993), 'On the system of semasiological grammar', *Allgemein-Vergleichende Grammatik,* 1. Bielefeld: Universität Bielefeld, Unviersität München.

Lehmann, Christian (2004), 'Theory and method in grammaticalization', in Gabriele Diewald (ed.), *Grammatikalisierung.* Special issue of *Zeitschrift für Germanistische Linguistik 32*: 152–187.

Lehmann, Christian (2008), 'Information structure and grammaticalization', in: Seoane and López-Couso (eds.), 207–229.

Levinson, Stephen C. (1995), 'Three levels of meaning', in: F. R. Palmer (ed.) *Grammar and meaning: essays in honor of Sir John Lyons*. Cambridge: Cambridge University Press, 90–115.

Levinson, Stephen C. (2000), *Presumptive meanings: the theory of generalized conversational implicature* (Language, Speech, and Communication). Cambridge, MA: MIT Press, a Bradford Book.

Milroy, James (1992), *Linguistic variation and change: on the historical sociolinguistics of English.* Oxford: Blackwell.

Molenci, Rafal (1997), 'Albeit a conjunction, yet it is a clause: a counterexample to unidirectionality hypothesis?', *Studia Anglica Posnaniensia: International Review of English Studies*, 31: 163–178.

Moore, Colette (2007), 'The spread of grammatical forms: the case of *be* + *supposed to*', *Journal of English Linguistics*, 35: 117–131.

Mossé, Ferdinand (1938), *Histoire de la forme périphrastique* être + *participe présent en germanique.* Paris: Klincksieck.

Patten, Amanda (2010), Cleft sentences and grammaticalization. PhD dissertation, Edinburgh University.

Prince, Ellen F. (1978), 'A comparison of WH-clefts and *it*-clefts in discourse', *Language*, 54: 883–906.

Recanati, François (2004), 'Pragmatics and semantics', in: Horn and Ward (eds.), 442–462.

Seoane, Elena and María José López-Couso (eds.), in collaboration with Teresa Fanego (2008), *Theoretical and empirical issues in grammaticalization* (Typological Studies in Language 77). Amsterdam/Philadelphia: Benjamins.

Sorva, Elina (2007), 'Grammaticalization and syntactic polyfunctionality: the case of *albeit*', in: Ursula Lenker and Anneli Meurman-Solin (eds.) *Connectives in the history of English* (Current Issues in Linguistic Theory 283). Amsterdam/Philadelphia: Benjamins, 115–143.

Stubbs, Michael (1995), 'Collocations and semantic profiles: on the cause of trouble with quantitative studies', *Functions of Language*, 2: 23–56.

Sweetser, Eve E. (1990), *From etymology to pragmatics: metaphorical and cultural aspects of semantic structure* (Cambridge Studies in Linguistics 54). Cambridge: Cambridge University Press.

Timberlake, Alan (1977), 'Reanalysis and actualization', in: Charles N. Li (ed.) *Mechanisms of syntactic change.* Austin/London: University of Texas Press, 141–177.

Traugott, Elizabeth Closs (2008a), '"All that he endeavoured to prove was…": on the emergence of grammatical constructions in dialogic contexts', in: Cooper and Kempson (eds.), 143–177.

Traugott, Elizabeth Closs (2008b), 'Grammaticalization, constructions and the incremental development of language: suggestions from the development of degree modifiers in English', in: Regine Eckardt, Gerhard Jäger and Tonjes Veenstra (eds.) *Variation, selection, development – probing*

the evolutionary model of language change. (Trends in Linguistics 197). Berlin/New York: Mouton de Gruyter, 219–250.

Traugott, Elizabeth Closs (2010), 'Dialogic motivations for syntactic change', in: Robert A. Cloutier, Anne Marie Hamilton-Brehm and William A. Kretzschmar (eds.) *Studies in the history of the English language V. Variation and change in English grammar and lexicon: contemporary approaches* (Topics in English Linguistics 68). Berlin: De Gruyter Mouton, 11–27.

Traugott, Elizabeth Closs and Richard B. Dasher (2002), *Regularity in semantic change* (Cambridge Studies in Linguistics 97). Cambridge: Cambridge University Press.

Traugott, Elizabeth Closs and Bernd Heine (eds.) (1991), *Approaches to grammaticalization* (Typological Studies in Language 19). Amsterdam/ Philadelphia: Benjamins.

Traugott, Elizabeth Closs and Ekkehard König (1991), 'The semantics-pragmatics of grammaticalization revisited', in: Traugott and Heine (eds.), vol. 1: 189–218.

Traugott, Elizabeth Closs and Susan Pintzuk (2008), 'Coding the *York-Toronto-Helsinki Parsed Corpus of Old English Prose* to investigate the syntax-pragmatics interface', in: Donka Minkova and Susan Fitzmaurice (eds.) *Empirical and analytical advances in the study of English language change* (Topics in English Linguistics 61). Berlin/New York: Mouton de Gruyter, 61–80.

Traugott, Elizabeth Closs and Graeme Trousdale (2010), 'Gradience, gradualness and grammaticalization: how do they intersect?', in: Elizabeth Closs Traugott and Graeme Trousdale (eds.), *Gradience, gradualness and grammaticalization* (Typological Studies in Language 90). Amsterdam/ Philadelphia: Benjamins, 19–44.

Trousdale, Graeme (2010), 'Issues in constructional approaches to grammaticalization in English', in: Ekaterini Stathi, Elke Gehweiler and Ekkehard König (eds.), *Grammaticalization: current views and issues.* Amsterdam/Philadelphia: Benjamins, 51–72.

Tuggy, David (1993), 'Ambiguity, polysemy, and vagueness', *Cognitive Linguistics*, 4: 273–290.

Viberg, Åke (1984), 'The verbs of perception: a typological study', in: Brian Butterworth, Bernard Comrie and Östen Dahl (eds.) *Explanations for language universals*. Berlin/New York: Mouton, 123–162.

Waltereit, Richard and Ulrich Detges (2007), 'Different functions, different histories. Modal particles and discourse markers from a diachronic point of view', in: Maria Josep Cuenca (ed.) *Contrastive perspectives on discourse markers*, Special issue of *Journal of Catalan Linguistics*, 6: 61–80.

Weinreich, Uriel, William Labov and Marvin Herzog (1968), 'Empirical foundations for a theory of language change', in: W. P. Lehmann and Yakov Malkiel (eds.) *Directions for historical linguistics: a symposium*. Austin: University of Texas Press, 97–195.

Wischer, Ilse and Gabriele Diewald (eds.) (2002), *New reflections on grammaticalization* (Typological Studies in Language 49). Amsterdam/ Philadelphia: Benjamins.