Revista Brasileira de Linguística Aplicada
ISSN: 1676-0786
rblasecretaria@gmail.com
Universidade Federal de Minas Gerais
Brasil

Kytö, Merja
Corpora and historical linguistics
Revista Brasileira de Linguística Aplicada, vol. 11, núm. 2, abril-junio, 2011, pp. 417-457
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil

Corpora and historical linguistics

Corpora e linguística histórica

Merja Kytö*
Uppsala University
Uppsala / Sweden

ABSTRACT: The present article aims to survey and assess the current state of electronic historical corpora and corpus methodology, and attempts to look into possible future developments. It highlights the fact that within the wide spectrum of corpus linguistic methodology, historical corpus linguistics has emerged as a vibrant field that has significantly added to the appeal felt for the study of language history and change. In fact, according to a historical linguist with more than fifty years of experience, "[w]e 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" (RISSANEN, forthcoming). Historical corpora and other electronic resources have also made the study of language history attractive: working on them engages students in an individual and interactive way that they find appealing (CURZAN 2000, p. 81).

KEYWORDS: Electronic historical corpora; corpus methodology; electronic resources; interdisciplinary collaboration.

RESUMO: Este artigo objetiva fazer um levantamento e avaliar o estado da arte dos corpora históricos eletrônicos e da metodologia de estudos de corpora, assim como sugerir possíveis desenvolvimentos futuros na área. Destaca-se que dentro do espectro metodológico da linguística de corpus, a linguística de corpus histórica emergiu como um campo de investigação vibrante que tem adicionado interesse ao estudo da história e da mudança linguística. De acordo com um pesquisador da área com mais de cinqüenta anos de experiência, "pode-se dizer que sem o apoio e o novo ímpeto trazidos pelos corpora, a linguística histórica baseada em evidências teria estado próxima ao fim de sua vida nesses tempos de rápidas mudanças de vida e de pesquisa, aumentando a competição na carreira acadêmica e nas atrações metodológicas oferecidas aos jovens pesquisadores (RISSANEN, no prelo). Corpora históricos e outros recursos eletrônicos têm também tornado o estudo da história da língua atraente: eles engajam a atenção dos estudantes tanto de forma individual quanto interativa (CURZAN 2000, p. 81).

PALAVRAS-CHAVE: Corpora históricos eletrônicos; metodologia de estudos de corpora; recursos eletrônicos; colaboração interdisciplinar.

* merja.kyto@engelska.uu.se

1. Introduction

The title of this article, "Corpora and historical linguistics", is likely to have meant something different to linguists some thirty to forty years ago than what it is taken to mean today. Similarly, "historical corpus linguistics" might well have been considered an instance of tautology, given that, apart from re-construction, all historical linguistics is in a wide sense corpus-based. If 'a corpus' is taken to be, as most would agree, "a collection of texts or parts of texts upon which some general linguistic analysis can be conducted" (MEYER, 2002, p. xi), 'a historical corpus' is "intentionally created to represent and investigate past stages of a language and/or to study language change" Claridge (2008, p. 242). These definitions apply to two types of historical corpora, pre-electronic ones that antedate the advent of the computer, and electronic ones that exploit computer technology, the difference accounting for the above change in the use of terminology.

The present article aims to survey and assess the current state of electronic historical corpora and corpus methodology, and attempts to look into possible future developments. To begin with, it is important to keep in mind that within the wide spectrum of corpus linguistic methodology, historical corpus linguistics has emerged as a vibrant field that has significantly added to the appeal felt for the study of language history and change. In fact, according to a historical linguist with more than fifty years of experience, "[w]e 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" (RISSANEN forthcoming). Historical corpora and other electronic resources have also made the study of language history attractive: working on them engages students in an individual and interactive way that they find appealing (CURZAN, 2000, p. 81).

Such corpus-based projects as biblical concordances, early grammars and early dictionaries bear witness to the painstaking nature of manual work involved in the use of pre-electronic corpora comprising one text or several texts (MEYER, 2008, p. 1). In the 1970s and 1980s, when it became possible to compile and analyse large-scale electronic corpora far more rapidly than had been the case with pre-electronic corpora (JOHANSSON, 2008, p. 33), historical linguists found themselves at the threshold of a new era. When describing this transitional stage in his introduction to the panel discussion devoted to "Issues in historical linguistics" at the 30th ICAME (International

Computer Archive of Modern and Medieval English) conference in May 2008, the convenor, Christian Mair (University of Freiburg), pointed out that "long before the advent of computers, monumental corpus projects were conceived which in some instances were later digitised and have continued into the present".[1] An example of such projects is the Corpus Inscriptionum Latinarum, which was started in 1853 and which "includes the Latin inscriptions from the entire area of the former Roman empire, arranged by region and by inscription-type" and which since its foundation has been "the standard edition of the epigraphic legacy of ancient Rome" (http://cil.bbaw.de/). On the other hand, the Thesaurus Linguae Graecae, a research centre at the University of California, Irvine, founded in 1972, set out to represent "the first effort in the Humanities to produce a large digital corpus of literary texts". It has so far "collected and digitized most literary texts written in Greek from Homer to the fall of Byzantium in AD 1453", with the goal "to create a comprehensive digital library of Greek literature from antiquity to the present era" (<http://www.tlg.uci.edu/>). Similarly, it was not until 1970's that we could also trace the first large-scale historical electronic corpus project aimed at documenting a period of the English language in toto (ca. 450-1100), that is, the Dictionary of Old English Corpus in Electronic Form, "a complete record of surviving Old English except for some variant manuscripts of individual texts" (<http://ota.ahds.ac.uk/headers/2488.xml>).

The ensuing tradition of English historical corpus linguistics has been particularly rich and has presented, a constantly growing family of historical corpora which documents periods extending from thirty years (or shorter spans of time) to a millennium. There is an increasing interest in historical corpora for many other modern languages, among them German and Mittelhochdeutsche Begriffsdatenbank, the Bonner Frühneuhochdeutsches Korpus and DeutschDiachronDigital, French and Textes de Français Ancien, Spanish and Corpus del Español, and Portuguese and Corpus do Português, to name just a few (for further examples and references, see CLARIDGE, 2008 and XIAO, 2008). There have also been signs in cross-linguistic historical corpus compilation projects as will be shown in the present article later on. Even though English historical corpora will serve as the basis for the discussion in the present article, it is hoped that the methodological issues raised, or most of them, can largely be taken to pertain to historical corpora in general.

[1] I am indebted to Christian Mair for permission to cite his script.

This article is organised as follows. After some preliminary remarks (section 1), resources and methodology in historical corpus linguistics will be discussed (section 2). The rationale of the approach will be examined (2.1), and the types of historical corpora available or underway (2.2.) will be surveyed along with the tools enabling historical corpus analysis (2.3). Section 3 will be devoted to an assessment of the developments in the field, with a discussion of recent advances and remaining bottleneck areas. The main themes will be resources and their potential for enhancement and new projects (3.1), prospects of searchability and corpus annotation (3.2), the need to enhance access to and information on historical electronic resources (3.3), and the need to promote interdisciplinary collaboration. A summary of future directions and desiderata will conclude the article (section 4).

2. Resources and methodology

2.1 The rationale of the historical corpus linguistic approach

There are a number of reasons why it makes sense to study the history of a language and language change using corpus linguistic methodology. These will be touched upon in the present section as they also tend to lie behind corpus compilation methodology and guide the developments in the field (see section 3).

A useful discussion of the benefits brought by the corpus linguistic approach to the study of language change can be found in Curzan (2008). In the study of language change, the aim is often to detect and substantiate general trends in language development. For this, one needs easy access to large amounts of data representative of different registers and levels of language use. Computerised corpora allow the study of stages of linguistic development from a contrastive or comparative perspective. They also facilitate the statistical analysis of relationships between linguistic phenomena and linguistic or extralinguistic factors at work in language change. By drawing attention to the influence of language use on language structure, and by offering access to often less well-known texts outside the literary canon, historical corpora and other related electronic resources have become of great interest to those working with functional linguistic approaches. They have also contributed to bringing the study of the past and present of a language together by serving as a testing ground for, for instance, modern sociolinguistic theory and by making those interested in present-day grammar look at recent and on-going change in systematic and empirical terms to avoid the pitfalls of anecdotal observation (MAIR, 2008, p. 1111-1112). Access to computerised data has also meant an increase in the awareness of the importance of language-theoretical considerations in linguistic research: it has become much less acceptable to simply collect examples and present them without paying attention to language theory or generalisation than it was in the days of pre-electronic historical language study (RISSANEN forthcoming). Finally, the fact that historical linguists seldom have access to stratified, balanced corpora that would cover the full range of diachrony and/or genres investigated has meant that more open-ended and unbalanced electronic data sources need to be resorted to in search for further materials. Indeed, the work done in the field has made many question the notion of all too restrictive a definition for a 'corpus' that may not serve the broad spectrum of linguistic research as well as a more generous definition often seems to do. Accordingly, in addition to traditional stratified corpora, the present article will consider further electronic resources such as large-scale electronic text collections, electronic text editions, linguistic atlases and dictionaries.

The increasing popularity of corpus linguistic methodology in the study of language change also obviously has to do with the kind of research questions that we can reasonably ask when using historical corpora. Attempts to answer these questions have also contributed to advances in the area. The use of extensive textual evidence was already a landmark of the research carried out on pre-electronic corpora, and changes in "the different ways of saying more or less one and the same thing" had been addressed by scholars back in time, with attention paid to factors taken to explain the loss or emergence of linguistic forms. However, with the advent of electronic corpora, it has been the process of change itself, and the transmission or implementation stages in it that have emerged as perhaps of major interest. To demonstrate how the rivalry of variant forms in, for instance, the development of second-person address pronouns proceeded across time, genre and different groups of language users requires a carefully selected dataset that enables generalisations (cf. WALKER, 2007). This line of research had already been fuelled by the interest felt in the 1970's and 1980's in the question of how language theory could best explain or account for change. Among the influential works in this respect can be mentioned, for instance, Weinreich, Labov and Herzog (1968), Samuels (1972), Lass (1980) and Romaine (1982), all of which paid attention to the importance of the empirical study of language variation and change. Examples of recent work in historical sociolinguistics include the study of the macro-level spread of language change in the Early Modern English period (e.g. NEVALAINEN; RAUMOLIN-BRUNBERG, 2003) and micro-level change with individual language users in focus (e.g. NURMI, NEVALA; PALANDER-COLLIN, 2009). This research has helped trace changes originating 'from below', an area of special interest in terms of actuation and spread of change. In register and genre studies, the development of genres has attracted attention, and the history of written English, for instance, has been approached as the history of registers showing shifting relationships to the more oral style that characterises at least less formal registers of spoken language (BIBER; FINEGAN, 1989; 1992; 1997).

Another boosting factor contributing to the interest felt for historical corpora was the emergence and consolidation of the historical pragmatics approach starting in the 1990s and onward. Since Jucker (1995), historical pragmaticians have found computerised data useful for systematic analysis of historical dialogue features and dialogues (JUCKER; FRITZ; LEBSANFT, 1999b, p. 17; FITZMAURICE; TAAVITSAINEN, 2007; cf. KYTÖ, 2010, p. 33-34). In this approach, pragmatic meanings and the changes in their realisations over time are of interest, as in the study of, for instance, speech acts (e.g. JUCKER; TAAVITSAINEN, 2000; 2008a; 2008b; TAAVITSAINEN; JUCKER, 2007; 2008a; 2008b), and grammaticalisation, pragmaticalisation, and lexicalisation phenomena in the history of English (e.g. Brinton, 1996, 2006). In historical socio-pragmatics, the focus is on pragmatic uses and their developments over time across male and female language users representative of various social ranks (e.g. LUTZKY, 2009; CULPEPER; KYTÖ, 2010). Yet another approach that has encouraged the use of historical corpora includes cognitive semantics and prototype semantics that study the emergence of meanings and their expressions in human cognition, central vs. more peripheral meanings, and changes in these relations over time (e.g. RISSANEN *et al.*, 2007). These are all examples of analytical frameworks where the use of historical corpora and corpus linguistic techniques enables large-scale and sophisticated analyses and adds to the coverage and reliability of results. The criteria adopted for the compilation of corpora also offer a convenient short-cut for investigating the possible influence of extralinguistic factors on developments. Among the texts, of special interest are those reflecting informal, everyday language, or offering access to 'non-standard' language use (CLARIDGE; KYTÖ, 2010). Corpus linguistic methodology also enables statistical analyses that are beyond the traditional manual approach (e.g.

collostructional and keyword analyses, n-grams; for problems in practical applications with historical data, see 3.2).

2.2 Types of historical corpora and other electronic resources

According to McEnery and Wilson ([1996] 2001, p. 123), computerised resources and tools used to analyse them have become part of most research on historical linguistics today. Regarding English, there are currently thirty to forty English historical corpora available or underway, amounting to more than 130 million words, excluding the 400-million-word Corpus of Historical American English and the 100-million-word Time Corpus; if we deduct from this figure the 52-million-word Old Bailey Corpus (see below), the materials amount to some 78 million words. In the literature, the available corpora have been deemed to give a fair picture of the development of English vocabulary and grammar from the earliest times to our own days (CLARIDGE, 2008; RISSANEN forthcoming). However, there are gaps in coverage, to be discussed in section 3.1 below. In addition to historical corpora, resources containing historical material come to us in other forms that enable us to use them as corpora. It is often necessary for historical linguists to use various types of electronic (and non-electronic) resources in their hunt for information. This section surveys some of the main resource types by way of a background to the discussion of future desiderata in the field. In addition to stratified multigenre and specialised corpora, attention will be paid to large-scale text collections, electronic text editions, linguistic atlases and dictionaries (for further discussion, see KYTÖ, 2010 and forthcoming).

Multigenre corpora aim at representing a wide variety of registers and language use across several centuries in order to allow investigations of long-term developments in usage. The first stratified electronic historical corpus of English was The Helsinki Corpus of English Texts. Extending from 700's to 1710, this corpus of 1.5 million words spans from the Old English through the Middle English to the Early Modern English period and contains samples of genres such as law, philosophy, history writing, science, handbooks, travelogues, (auto)biographies, fiction, drama, private and official correspondence, and the Bible. A good number of these are represented across the corpus (e.g. law, philosophy, science, handbooks) while others only appear for a certain period or periods (e.g. homilies for the Old and Middle English periods, romances for the Middle English period, and trial proceedings for the Early Modern English period). ARCHER (A Representative Corpus of Historical English

Registers) (1.7 million words) is another multigenre corpus, extending from 1650 to 1990 and containing partly the same genres as the Helsinki Corpus, for instance, science, fiction, drama and correspondence. While the Helsinki Corpus only contains British English texts, ARCHER contains both British and American English texts. Historical corpora are mostly associated with the written medium, and texts that have been taken to reflect past 'spoken' interaction, phonological spellings or orthoepists' comments have been used as a way of obtaining indirect evidence of past spoken language. However, there is an increasing interest in historical corpora containing spoken texts that could provide direct evidence of the spoken medium. The Diachronic Corpus of Present-Day Spoken English (800,000 words) is such a corpus: it contains samples of recent English, drawing from the ICE-GB (the British component of the International Corpus of English (ICE), collected in the early 1990s) and the London-Lund Corpus of Spoken English (late 1960s-early 1980s). This multigenre corpus contains genres such as face-to-face and telephone conversations, broadcast discussions and interviews, spontaneous commentary, parliamentary language, legal cross-examination, and prepared speech.

As the data yielded by multigenre corpora tend to break down across the genres and periods distinguished, multigenre corpora are typically suitable for diagnostic purposes, pointing to trends that can be verified with the help of further data found in specialised corpora, for instance. Specialised corpora tend to focus on a genre (or related genres), a period, a certain aspect of language use, or even a single text or author. Examples of the last-mentioned are the Electronic Beowulf and the Shakespeare Corpus. Other types of specialised corpora have often been compiled to facilitate observing language change from a specific analytical framework (or a number of them). Thus the Corpora of Early English Correspondence (5.1 million words, letters from the early 1400s to 1800) were compiled to allow historical sociolinguistic study; Corpus of Early English Medical Writing 1375-1800 (estimated 3.8 million words, medical texts of various types) for observing stylistic change in early medical English; A Corpus of English Dialogues 1560-1760 (1.2 million words, dialogic texts) to allow the study of early speech-related language; Zurich English Newspaper Corpus (1661-1791) (1.6 million words, newspapers), and the Lampeter Corpus of Early Modern English Tracts (1640-1740) (1.2 million words, pamphlets and other tracts) for studies of language use in the public domain. Examples of period-specific and/or genre-specific corpora are the above-mentioned Dictionary of Old English Corpus in Electronic Form; A Corpus of Nineteenth-Century English (1800-1900, 1 million words, seven genres, British English only); the Time Corpus (or Time Magazine Corpus of American English, 1923-2006, 100 million words); and A Corpus of Historical American English (400+ million words, 1810's-2000's, popular magazines, newspapers, and academic writing). The last-mentioned is also an example of specialised historical corpora that focus on transplanted regional varieties. Among other such corpora can be mentioned A Corpus of Irish English (14th-20th centuries, 550,000 words) and the (Corpus of Oz Early English (1788-1900, 2 million words).

Like present-day corpora, historical corpora can also contain parts-of-speech or other grammatical or textual annotation. Examples of such corpora are the Parsed Corpus of Early English Correspondence (2.2 million words), which is available in plain text files, part-of-speech tagged files, and syntactically parsed files, with metadata about the letters (date, authenticity, recipient classification) and correspondents (name, date of birth, gender, etc.). The annotation scheme used for this corpus had earlier been applied to Penn-Helsinki Parsed Corpus of Middle English (second edition) and the Penn-Helsinki Parsed Corpus of Early Modern English. A remarkably richly annotated and manually checked resource is the above-mentioned Diachronic Corpus of Present-Day Spoken English, which comes with the ICECUP search suite and allows one "to perform a variety of different queries, including using the parse analysis *in* the corpus to construct Fuzzy Tree Fragments *to search* the corpus" (http://www.ucl.ac.uk/english-usage/projects/dcpse/).

In addition to stratified historical corpora proper, electronic versions of early texts have been made available in the form of facsimile or plain text files in huge computerisation projects such as the Literature Online collection (Lion), the Early English Books Online (EEBO), and its chronological sequel the Eighteenth Century Collections Online (ECCO). The Lion collection "offers the full text of more than 350,000 works of poetry, drama and prose in English from the eighth century to the present day", and "more than 800 classic literary essays, from the sixteenth century to the early twentieth". Further, Lion also provides links to more than 8,000 additional electronic texts from third-party internet sites. Importantly, "[a]ll texts are reproduced faithfully from the original printed sources without silent emendation" (http://lion.chadwyck.co.uk/marketing/editpolicy2.jsp). EEBO comprises over 22 million digital page images from "virtually every work printed in England, Ireland, Scotland, Wales and British North America and works in English printed elsewhere from 1473–1700" (http://eebo.chadwyck.com/home). Similarly, ECCO is a large-scale collection, comprising more than 136,000 titles in 26 million digital facsimile pages. ECCO covers a wide range of subject areas, among them literature and language, law, history and geography, social sciences and fine arts, medicine, science and technology, and religion and philosophy (<http://mlr.com/DigitalCollections/products/ecco/>). (For limitations set to searchability, see 3.2.)

The above text collections provide useful material for the study of language change even though they were not compiled for primarily linguistic research. Other such very large-scale collections, although more specialised, include newspaper texts. Among these are the ProQuest Historical Newspapers collection (www.proquest.com) and the Times Digital Archive (www.gale.cengage.com). The former is a massive collection that offers "full-text and full-image articles for [36] significant newspapers dating back to the 18th Century [1764-2008]" and mostly comprises sources representing American English. The latter represents British English and contains over 7.6 million articles published in The Times starting in 1785 over a period of more than 200 years. There are also smaller collections such as North American Review (Library of Congress), Blackwood's Edinburgh Magazine (Bodleian Library online), The Collected Works of Abraham Lincoln (Humanities Text Initiative online, University of Michigan) and American Whig Review (Library of Congress) (for references and further information, see MacQUEEN, 2010). Another specialised large-scale collection is The Proceedings of the Old Bailey, London's Central Criminal Court, 1674 to 1913 (Old Bailey Corpus). The Old Bailey Corpus provides "[a] fully searchable edition of the largest body of texts detailing the lives of non-elite people ever published, containing 197,745 criminal trials held at London's central criminal court" (http://www.oldbaileyonline.org/). The web site provides access to 190,000 images of the original pages of the Proceedings and 4,000 pages of Ordinar's Accounts, in addition to historical, social and other support material. This resource was originally intended for the use of historians, but a project aiming at converting the digitised transcripts into a linguistic corpus is underway at the University of Giessen, Germany (HUBER, 2007): mark-up will be provided to distinguish direct speech from the rest of the text in a 134-million-word section of the full corpus; this section will also be tagged for parts of speech. Sociolinguistic mark-up will be entered for about half of the material qualifying as direct speech (i.e. for ca. 57 million words out of the 113 million words comprising direct speech) (<http://www.uni-giessen.de/oldbaileycorpus/index.php>).

In addition to ready-made large-scale text collections, it is also possible to look for electronic texts on internet sites, for instance at the Project Gutenberg site (<http://www.gutenberg.org/wiki/Main_Page>) or from distribution houses such as the Oxford Text Archive (note that such material may be of uneven reliability in terms of editions used, the accuracy of the text, etc.). The Corpus of Late Modern English Texts, Extended Version (1710-1920) (15 million words) was compiled using texts available in these sources (see De Smet, 2005).

The possibility of combining digital manuscript images with searchable transcriptions and textual annotation has increased the interest in electronic text editions, especially such as are intended to render the original manuscript text as faithfully as possible (for recent work, see e.g HONKAPOHJA; KAISLANIEMI; MARTTILA, 2009, and KYTÖ; GRUND; WALKER forthcoming, and references therein). These editions can be used as electronic corpora and they also lend themselves to further digital applications such as hypertext databases. Compared with most historical corpora based on imprint material, the time-consuming nature of transcription work generally limits the text length of electronic editions. Examples of electronic text editions include collections such as the Corpus of Scottish Correspondence (1500-1730, 256,000 words), An Electronic Text Edition of Depositions 1560-1760 (267,000 words) and The Middle English Grammar Corpus (1100-1500, 450,000 words), and single texts such as Electronic Beowulf and A London Provisioner's Chronicle, 1550-1563, by Henry Machyn. Manuscript-based digitised transcriptions of early texts are also available in linguistic atlases such as A Linguistic Atlas of Early Middle English 1.1 (1150-1325) (c. 650,000 words) and A Linguistic Atlas of Older Scots, Phase 1 (1380-1500), both follow-up projects to the hard-copy Linguistic Atlas of Late Modern English (LALME) (1350-1450), which is being revised and digitised into an e-LALME version.

Electronic dictionaries are powerful tools that facilitate looking up information on words and phraseology. They do not of course generally provide such contexts as full-text corpora do for individual search items, but the information extracted can be used for follow-up searches in historical corpora proper. Large-scale dictionaries, which aim at covering the history of a language's vocabulary, are long-term projects going far back in time. Among such projects are the Oxford English Dictionary Online (OED Online) for English, Der digitale Grimm for German, and Svenska Akademiens ordbok for Swedish.

More specialised electronic dictionaries focus on a certain period as, for instance, the Dictionary of Old English and the Middle English Dictionary, or are digitised versions of early dictionaries such as Samuel Johnson's Dictionary of the English Language (1773 [1755]) (McDERMOTT, 1996). A collection of digitised early dictionaries is available in the Lexicons of Early Modern English (1480-1702) database, a multilingual resource that currently comprises close to 580,000 word entries drawn from 168 searchable lexicons (e.g. monolingual, bilingual, and polyglot dictionaries, hard-word glossaries and spelling lists) digitised from early imprints or manuscripts (LANCASHIRE, 2006).

2.3 Tools for historical corpus analysis

The basic tools used by historical corpus linguists do not differ essentially from those used for searching present-day material. Among these tools are word lists and concordances, combined with more sophisticated methods such as collocate, keyword, or n-gram (or lexical bundle or multi-word expression) analysis. Search programs currently available on the market are WordSmith Tools, MonoConc Pro, Corpus Presenter and Xaira. The last-mentioned provides advanced graphical support for investigating results. The powerful statistical computing and graphics program R can also be used to process language data (<http://www.r-project.org/>). (For a useful discussion of data retrieval software, see WYNNE, 2008).

Among the resources that provide a search engine of their own are, for instance, the Penn Parsed Corpora of Historical English and the Parsed Corpus of Early English Correspondence, which have been annotated for the purposes of the CorpusSearch 2 program (<http://corpussearch.sourceforge.net/index.html>), or the Corpus of Irish English, the Middle English Medical Texts and the Early Modern English Medical Texts (parts of the above-mentioned Corpus of Early English Medical Writing), and An Electronic Text Edition of Depositions 1560–1760, which each come with a customised Corpus Presenter application. Another solution has been opted for in the Corpus of Historical American English which can be accessed via a search interface allowing one to investigate, for instance, changes in the frequency of words and phrases, parts of words, grammatical constructions and collocates. Large-scale text collections (Lion, the Old Bailey Corpus) most often provide a search engine of their own. As these collections were not primarily designed for linguistic searches, applying the search engines to solve linguistic research questions seldom works adequately. Overall, using search programs on historical data is not altogether unproblematic, especially as regards spelling variation, a feature characteristic of pre-standard varieties (see 3.2).

3. Assessing the field: recent advances and bottleneck areas

As shown above, significant progress has been made in the production of historical corpora and other electronic resources over the past few decades. However, there are still problems in various areas that would benefit from further attention. A number of these will be addressed in the following. To begin with, gaps in the present coverage will be discussed, with special reference to the field of English historical linguistics, again with the aim that similar problem areas could be identified for other languages. Attention will then be drawn to recent advances in the corpus compilation "philosophies" that often lie behind corpus projects and the potential they have for further advances. Related to this, the question of comparability between different corpora will be highlighted, and attention also paid to various linguistico-philological issues in corpus compilation (3.1). Issues with searchability, corpus annotation, and spelling variation, referred to above, will be discussed along with the ways in which problems in these areas hamper the full use of, for instance, statistical tools in the study of language change (3.2). The remaining points taken up pertain to corpus linguistics in general but are nevertheless worth considering as regards historical corpus linguistics, in particular. These include copyright questions, and how to inform the community of linguists and other potential users of the availability and properties of historical corpora (3.3). Finally, a call will be made for enhancing awareness among historical corpus linguists of the benefits brought about by the interdisciplinary framework (3.4).

3.1 Resources: potential for enhancement and new projects

Regarding gaps in textual coverage in English historical corpora, according to Rissanen (forthcoming), "[t]he 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, both as concerns search engines and annotation." Claridge (2008, p. 245-246) goes even farther saying that "[w]hile the textual situation becomes better after the Middle Ages with regard to both amount and variation, the historical corpus linguist will always face shortages of some nature before the late 19th century". Compilers and users of historical corpora need to accept the sad fact that a lot of valuable material has been lost in fires, floods, wars, or in other circumstances (for instance, only very little evidence of English is preserved from the Early Middle English period, 1250-1350, as a consequence of political circumstances that led to Anglo-Norman and French being the languages of the ruling ranks). Also, the time distance between the date of the original text and the copy preserved to us can cover several generations of language users, making it difficult to draw conclusions about usage in the time of the original. This can be the case not only with medieval texts but also even in the early modern period (for instance, many sixteenth-century trial proceedings survive in seventeenth-century copies only, see CULPEPER; KYTÖ, 2010, p. 50-51). Nor are early texts easily accessible, especially if available only in manuscript form. There are also socio-historical and cultural constraints such as poor levels of literacy and writing skills, and limited access to formal education, which hampered the production of early texts. The lower and middle segments of society, in particular, were subject to illiteracy, so the language of the social and educational elite, and especially male writers, tends to dominate in historical corpora leaving language of women and representatives of the lower echelons underrepresented (CLARIDGE, 2008, p. 248). Finally, nor do we always know for certain whether it was a scribe or the ascribed author who produced the text. This can be the case with early letters written in the Middle Ages or with even much later letters. For instance, we have valuable 'non-standard' material in the so-called 'pauper letters' from the eighteenth and early nineteenth century, written by ordinary people on the verge of poverty to their overseers (Sokoll, 2001). An electronic corpus of these letters is now underway (by Mikko Laitinen, see RAUMOLIN-BRUNBERG, 2003), but what will limit the use of the material is that it is often unclear whether a letter was written by the ascribed author or by another person hired to do the job.

It is important that compilers of future historical corpora pay attention to the above problems and that they document their compilation decisions in clear terms in user guides, corpus manuals and like material that will accompany the release versions of the corpora. It would be all too time-consuming and virtually impossible for end-users to replicate the research done to find out about the background of texts included in historical corpora. For instance, early imprints of one and the same work may differ in details owing to compositors having made changes to the type in individual copies. For later verification purposes, it is necessary for the respective corpus file or manual to contain bibliographical reference information on the specific copy used for the corpus. Overall, assessing the reliability and validity of source texts as evidence of language use from the past periods is of prime importance to any historical corpus compilation project. For instance, text editions come in varying quality and based on varying editorial policies. Careful attention needs to be paid to the relationship of text editions to the original texts, and to keeping end-users aware of the value of the evidence drawn from them (for further discussion, see KYTÖ; WALKER, 2003; KYTÖ; PAHTA, forthcoming).

Despite the above considerations, there is a lot of potential in the various corpus compilation "philosophies" to enhance extant historical corpora and to develop new ones. As mentioned above, the first structured historical corpora containing early English were multigenre corpora intended for the study of language variation and change across the centuries. The underlying hypothesis was that comparative analysis of written texts which stand at different distances from speech may help us in our attempts to envisage what past 'spoken' language might have been like and that it is also possible to extrapolate from informal writing about everyday language use (KYTÖ; RISSANEN, 1983; RISSANEN, 1986, 1999). Commendably, such corpora are still being compiled as, for instance, the Leuven English Old to New (LEON) corpus, which is intended to span from the 900's to the twenty-first century (PETRÉ, 2009). The earlier corpora are also being enhanced in view of more sophisticated use, as is the case with for instance ARCHER (YÁÑEZ-BOUZA, 2011).

At the same time projects focusing on specialised corpora have produced a growing body of innovative research in areas such as historical sociolinguistics, genre and register studies, and the study of 'spoken' interaction in the past. All these directions are to be encouraged as the research carried out within these frameworks has significantly added to our knowledge of language history and processes of change. The results obtained in historical sociolinguistics have helped evaluate and re-assess some of the findings presented in modern sociolinguistic research. Similarly, systematic evidence-based genre and register studies have helped map and account for stylistic and grammatical shifts in language use from medieval to modern times in a way that would hardly have been possible without the support of historical corpora. The study of 'spoken' interaction in the past is also of special interest: while dialogic face-to-face interaction has been considered relevant in actuation of change (MILROY, 1992; TRAUGOTT; DASHER, 2002; CULPEPER; KYTÖ, 2010), historical evidence of it has been preserved only in written form. Even though texts containing early speech-related or speech-like language, whether in the form of dialogues (e.g. trial proceedings, drama) or private correspondence, cannot be expected to have preserved speech with the accuracy that modern audio-recording devices do, they are valuable as they can be studied "as communicative manifestations in their own right" (JACOBS; JUCKER, 1995, p. 9). There is also an interest in this approach among those working on the history of other languages than English as can be seen in works such as Collins' 2001 study of speech-reporting strategies in a substantial corpus of medieval Russian trial transcripts, and in articles included in Journal of Historical Pragmatics.

The above-mentioned Diachronic Corpus of Present-Day Spoken English allows the systematic study of change in spoken English in real-time, but only for a relatively brief period of time. More than 130 years have passed since the Chicago Daily Tribune (9 May, 1877) reported on the 'talking-machine' that Thomas Alva Edison was working on and that he later on that year presented as a phonograph, the first device able to record and replay the sound. This leaves us with oceans of material for historical corpus compilers to explore. A fascinating example of a study based on extensive audio-recordings provided by New Zealand's 'mobile disk unit' gives us information on how the earliest New Zealand-born settlers spoke and how this new variety of English first spoken in the 1850s developed (GORDON *et al.*, 2009). Having access to structured sets of early audio-recorded materials would enable real-time and apparent-time research on language change based on direct spoken language evidence. Such corpus compilation projects would contribute to current resources in most valuable ways.

As has been shown above, historical corpora have widened the spectrum of texts beyond those, mainly literary, that have traditionally been considered by language historians. It is desirable that historical corpus compilers continue to explore such materials further. More resources containing women's language, and language of untutored writers, or writers with little formal education are on end-users' wish list. This also holds for resources containing evidence of early 'spoken' interaction, and dialectal, regional or other 'non-standard' usage.

Considering the spread of English as an international world language, there is plenty of room for corpus projects aimed at recording the historical stages of the emergence and subsequent development of various transplanted varieties. It would also be fascinating to have access to materials representative of the development of individual genres or genre families across time periods.

An example of such a project underway is the Corpus of English Religious Prose (KOHNEN, 2007), which aims at documenting the history of English religious writing. On the whole, genres of chronological continuity would merit better attention, among them legal language, history writing, handbooks, science, philosophy, travelogues, (auto)biography, fiction, drama, and verse. As a genre may also change across time as regards stylistic and other conventions, attention should be paid to genre definitions across the diachrony; it may be difficult to see whether what we have at hand is language change or only change in genre conventions (cf., e.g., BIBER; FINEGAN, 1989).

But there is also room for new areas of interest. One so far rather neglected an area is the historical cross-linguistic perspective. Only very little has been done to compile historical parallel corpora that would combine different languages. A step in that direction has been the GerManC project launched at the University of Manchester to compile a representative historical corpus of written German for the years 1650-1800. The project aims at providing "a basis for comparative studies of the development of the grammar and vocabulary of English and German and the way in which they were standardized". For this end, the GerManC corpus has been structured and designed "to parallel that of similar historical linguistic corpora of English, notably the ARCHER corpus". The compilation team are collaborating with representatives of the ARCHER team to maximise the degree of comparability between the corpora. Once complete, the GerManC corpus "will contain 2000-word samples from nine genres: drama, newspapers, sermons, personal letters and journals (to represent orally oriented registers) and narrative prose (fiction and biographies), academic, medical and legal texts (to represent more print-oriented registers)" (http://www.llc.manchester.ac.uk/research/projects/germanc/). Another example is the "Three centuries of drama dialogue: A cross-linguistic perspective" project underway at Uppsala University. In its current pilot stages, this project aims at an English-Swedish Drama Dialogue corpus containing drama texts in English and Swedish from the three periods, 1725-1750, 1825-1850 and 1925-1950. The North Sea area offers ample opportunities for the compilation of interesting cross-linguistic historical corpora that could provide material for comparisons with Germanic and Romance languages. There are also counterparts for comparisons in the form of parallel corpora containing present-day language.

A further neglected area in historical corpus compilation is language teaching. There has been an increasing interest among historical pragmaticians in dialogues found in language teaching books (e.g. HÜLLEN, 1995; WATTS, 1999; for these and further references, see CULPEPER; KYTÖ, 2010, p. 45). A Corpus of English Dialogues 1560-1760 contains didactic works, a subsection of which is devoted to language teaching manuals. Language teaching texts have been separated from the other didactic works in this corpus owing to their special characteristics and socio-historical background. On the one hand, these texts are realistic in their display of language use they aim to teach. On the other hand, they also contain features uncharacteristic of authentic language use situations such as long vocabulary lists (CULPEPER; KYTÖ, 2010, p. 46-48). The target language may also have influenced to varying degrees the dialogues in which the teaching materials are couched (KYTÖ; WALKER, 2006, p. 23; CULPEPER; KYTÖ, 2010, p. 48). The texts included in this corpus were intended to teach English to the French and French to the English, with one text aimed at teaching German to the English. However, the material remains scanty in view of in-depth studies, and given the interest in present-day language teaching materials, more historical texts in searchable form would be welcome. Related to this, one new avenue would be the compilation of corpora containing early grammarians' and orthoepists' works. These have always been of major interest to historical linguists as, among other things, they provide glimpses of contemporaneous views of language use.

Regarding other forms of electronic resources than structured corpora, electronic text editions are an area that would deserve much more attention than is the case today. Libraries, archives and record offices contain great amounts of valuable manuscript material which, if scanned or transcribed, provided with metadata annotation, and, ideally, accompanied by manuscript images or samples of them, would be of the utmost interest to the research community. Transcriptions aiming at rendering the language and other features of the original manuscripts as faithfully as possible within the limitations set by modern typography and electronic processing facilities are to be encouraged (for linguistic annotation, see 3.2). Electronic editions of early imprints would also be welcome, especially in areas such as science and handbooks, where images play an important role and multimodal applications would enhance the value of the material. As for linguistic atlases that contain the texts they are based on, such as A Linguistic Atlas of Early Middle English, the work is only in its infancy. As for the history of English, dialect maps of regions or localities from the Old English and the early modern period would be of great value, to complement the current Middle English atlas projects.

Gaps in coverage often necessitate looking for data from a number of corpora. The question is to what extent corpora compiled on varying principles are comparable. There are examples of corpora that represent as perfect a match as is possible considering that genres may also change in time and that sources such as newspapers may be discontinued. The family of 'Brown corpora' presents a case of a number of matching corpora designed to enable one-to-one comparisons. These corpora follow the one-million-word Brown Corpus (or A Standard Corpus of Present-day Edited American English, for Use with Digital Computers) released in 1964, and include the LOB corpus (or Lancaster-Oslo/Bergen Corpus) of British English (1978), and their counterparts Frown Corpus (Freiburg-Brown Corpus of American English) and F-LOB (Freiburg-LOB Corpus of British English) (1999 original versions, 2007 POS-tagged versions). These match in size and composition, with the only difference that while the Brown and LOB corpora were compiled to represent language from 1961, Frown and F-LOB include sources 30 years after, from 1991. Two further family members are underway, the BLOB-1931 corpus sampled from the period 1928-1934, with a focus on 1931, and another from 1901, to provide further sources for comparison on the British English axis. These corpora allow systematic study of for instance recent and on-going change in English grammar, and the linguistic and social factors that are influencing processes of change (see, e.g., LEECH *et al.*, 2009).

However, gaps in textual representation, differences in period divisions and classification of social strata, and other such features usually entail that comparisons across corpora can seldom be straightforward; instead, further consideration and adjustments are needed on the part of end-users. It is of course desirable that future corpus compilers pay attention to previous compilation plans when launching their projects in order to facilitate research across historical corpora. This is also of prime importance for future annotation projects.

3.2 Issues of searchability and corpus annotation

In addition to enhancing extant resources and creating new ones, compilers and end-users of historical corpora would need to collaborate with computational linguists to a greater extent than has been the case so far. There is a general lack of consensus on platforms, and searching historical corpora, large-scale text collections and electronic dictionaries is not always as unproblematic as one could wish.

As mentioned above, many of the search engines that come with large-scale collections are not primarily intended for linguistic study but rather for identifying quotations in literary works (e.g. Lion) or for extracting historical information (e.g. the Old Bailey Corpus). Similarly, the EEBO and ECCO images are searchable only in the sense that one can look for a word or phrase and get a list of the full-text contexts of all instances, with the possibility of clicking over to the facsimile of the page (the same goes for ECCO). On the other hand, the results cannot be concordanced, and one has to find ways to determine the approximate number of words in the corpus in order to approximate an incidence figure for the expression at hand (for such techniques applied to very large-scale historical newspaper collections, see MacQueen, 2010, chapter 5). However, the bibliographical information on the EEBO texts can be searched. In addition, the Text Creation Partnership (TCP) at the University of Michigan has so far stored some 25,000 books in the collection in the form of searchable plain texts. Further, the search engine accompanying a central source such as the Corpus of Middle English Prose and Verse ("at present, sixty-two texts are available; about eighty others will be added soon, with another 150 smaller texts in preparation", see http://quod.lib.umich.edu/m/mec/about/) lists occurrences text by text separately, as they are not given conveniently in one and the same file. This invaluable resource and many others such as the Dictionary of Old English Corpus would benefit from a retrieval program that would make it easier to sort the texts by date, dialect, and genre, and to create subcorpora according to these parameters (Rissanen forthcoming). As implied above, it is also often surprisingly difficult, if not altogether impossible, to obtain word counts for each text (needed for counting the incidence figures for a linguistic feature per a certain text length, for instance) or download them for further *in situ* annotation or other processing.

The search programs available can be used for many basic and even advanced search tasks, but depending on the research questions and the type of material one is working on, professional computer programming skills are often needed to extract the kind of data one is after. Interesting results can also be achieved by exploring methodologies applied in other fields. For instance, as there is generally no coding for pragmatic phenomena such as speech acts in historical corpora, historical pragmaticians will need to develop methodologies to locate their data. Accordingly, for their study of compliments and gender in the history of English, Taavitsainen and Jucker developed an "ethnographic" method: to pin down "what was considered proper and polite, particularly in association with gender", they collected speech-act labels such as 'compliment', 'compliments', 'complement', 'complements' and their spelling variants (TAAVITSAINEN; JUCKER, 2008b, p. 207, with reference to ROMAINE, 2003, p. 104-105). The aim of the searches was "to locate relevant passages for qualitative assessment"; TAAVITSAINEN; JUCKER, 2008b, p. 208; for methodology, see also JUCKER; SCHNEIDER; TAAVITSAINEN; BREUSTEDT, 2008). The method has also been applied successfully to the study of apologies (JUCKER; TAAVITSAINEN, 2008b).

The searchability of a corpus is crucially dependent on how the corpus has been annotated. Again, there is a lack of consensus on this point, and compilers of historical corpora have been slow or even reluctant to apply standards such as the Text Encoding Initiative (TEI) Guidelines (P5) (<http://www.tei-c.org/index.xml>). Many of the better known corpora are annotated for the main textual features but not all, and not as exhaustively as could have been the case. The features that an end-user would need to be able to learn about with little effort include, for instance, the title of the text, date(s) (if composition and copy diverge), text-type/genre, content description, level of formality, medium (written/spoken), language use (prose/verse; dialect; foreign languages etc.), authenticity of the document (autograph/copy etc.), references to established citation systems, the original/edition used for the corpus, and other bibliographical information. Certain author properties would also be useful information: age, gender, social rank/class, parentage, education, profession(s), residence, dialect, type of possible author-recipient relationship (if interactive) etc. Coding plans paying attention to both the writer/speaker and the addressee/interlocutors are to be encouraged. For instance, the Sociopragmatic Corpus, part of A Corpus of English Dialogues 1560-1760, has been annotated for both speaker and addressee properties, turn by turn. Interrogating this corpus for advanced searches requires a customised search engine; a similar approach was adopted when coding the speaker turns for the above-mentioned English-Swedish drama corpus.

Enhancing the searchability of historical electronic resources is not a straightforward task. There are a number of factors complicating annotation efforts, and it is no surprise that the amount of grammatically annotated historical material is still relatively scant in comparison to corpora containing annotated present-day material. There are historical corpora that have been tagged completely by manual means, for instance, the German Bonner Frühneuhochdeutsch Korpus (CLARIDGE, 2008, p. 254-255), but resorting to automatic tagging and manual checking to correct tagging errors has also been attempted. As tagging systems and software have mostly been developed for present-day standard varieties, they run into problems when trying to deal with historical varieties that tend to vary internally and present unanticipated language structure and spelling variation. Compared with modern texts that can be tagged automatically at the rate of about 96-97%, Early Modern English material presents lower rates, from 80% to 95%, depending on the date of the text (CLARIDGE, 2008, p. 254). Manual checking and correction is usually required to produce more reliable results; for instance, a considerable amount of manual labour was needed to annotate the York-Helsinki Parsed Corpus of Old English Poetry, the York-Helsinki Parsed Corpus of Old English Prose, the Penn-Helsinki Parsed Corpus of Middle English, the Penn-Helsinki Parsed Corpus of Early Modern English and the Penn Parsed Corpus of Modern British English (1700-1914, close to 1 million words). Syntactic annotation (parsing) in the three Penn Parsed Corpora of Historical English "permits searching not only for words and word sequences, but also for syntactic structure" (<http://www.ling.upenn.edu/hist-corpora/>). In addition to syntactic annotation, the Parsed Corpus of Early English Correspondence contains parts-of-speech tagging.

Examples of semantic tagging of historical data are few. A notable exception is the Mitterhochdeutsche Begriffsdatenbank (Middle-High German Conceptual Database), which "provides very powerful **search functions** for a large number of the most important works of Middle-high German literature, with linguistic and semantic search criteria" and "a **Wordindex with Concepts** for the lemmas and words in the database" (http://mhdbdb.sbg.ac.at:8000/index.en.html). There has also been pilot work on Early Modern English newsbooks (613,000 words) by (re)training the UCREL Semantic Analysis System (USAS) to cope with this historical variety with the help of the web-based corpus tool Wmatrix (ARCHER; MCENERY; RAYSON; HARDIE, 2003). This tool, and the subsequent Wmatrix2, was originally developed for modern varieties, so the mismatch between the tags adopted for modern texts and those required by the historical material caused some problems. Similarly, the tool had difficulties in dealing with automated grammatical annotation and variant spellings. By way of remedy, the historical validity of the semantic tag set will be improved in future work with the help of the Historical Thesaurus of English (<http://libra.englang.arts.gla.ac.uk/historicalthesaurus/aboutproject.html>) and by pre-processing the texts to be tagged with a variant spelling detector (VARD, see below) (ARCHER, forthcoming). Semantic tagging of historical texts is clearly a field full of promise and in need of further work.

As seen above, spelling variation presents a problem for automatic annotation and searching of historical texts, and there has been some tension between the respect felt by historical linguists for the source text and the demands set by searchability. Only a little over a decade ago, we could read that "[i]n English studies, normalization and/or regularization have never been popular. As to their role in machine-readable corpus compilation, the common opinion seems to be that compilers ought to reproduce the specific features of their source text and not smooth them away. In line with this common understanding, hardly any studies concerning normalization or regularization can be found" (MARKUS, 1997, p. 211). To normalise or not to normalise, that was the hotly debated question for quite some time, with those remaining in the minority who advocated the need for normalised versions of the text. Over the past few years, interest in techniques such as keyword and n-gram analyses has certainly promoted the awareness of the value of texts displaying regularised spelling. One way out of the faithfulness *vs.* ease of retrievability dilemma is to represent both original and regularised spelling versions of the corpus, through an annotation system (as in the Lancaster Newsbook Corpus), or through a multi-level architecture, or through a link to a normalised index.

Also, over the past few years, significant advances have been made in variant spelling research with the help of the Variant Detector (VARD) computer program (<http://ucrel.lancs.ac.uk/VariantSpelling/>; see, also, RAYSON *et al.*, 2007). The current version, VARD2, "is intended to be a pre-processor to other corpus linguistic tools such as keyword analysis, collocations and annotation (e.g. POS and semantic tagging), the aim being to improve the accuracy of these tools" (<http://www.comp.lancs.ac.uk/~barona/vard2/>) (see BARON; RAYSON, 2008). The approach is to produce a list of variant spellings, which are manually matched to normalised forms. The variant detector computer program inserts modern equivalents of these forms when they appear in a given text, while preserving the original variant. This approach proved to be very effective. So far over 50,000 variants have been identified from analysis of different historical texts, and empirical studies of spelling variation across the sixteenth to the nineteenth centuries have been carried out. Even though the tool was designed specifically to deal with Early Modern English spelling variation, it has the potential to work on any form of spelling variation and in any language after training the program with a relevant dictionary and spelling rules. The program has already been applied to for instance A Corpus of English Dialogues 1560-1760, the Corpora of Medical Writing, ARCHER, the Innsbruck Computer Archive of Machine-Readable English Texts, the Lampeter Corpus, the Shakespeare Corpus, and EEBO texts to quantify the level and development of spelling variation in the history of English, and to identify spelling patterns across periods and genres (BARON; RAYSON; ARCHER, 2009a, 2009b; BARON; RAYSON, 2009). Clearly, tools such as VARD2 show the way to future development of software and have great potential to enhance the searchability of historical texts.

Having access to normalised spelling versions of historical corpora would thus facilitate the use of sophisticated statistical analyses. For instance, keyword analyses can be used to study the various ways in which texts function, their related semantic spaces and collocational patterns (WYNNE, 2008, p. 730-734; ARCHER, 2009). Similarly, n-gram analyses based on multi-word sequences located by the computer can be used to study recurrent phraseology across the history of a language (for the principle, see WYNNE, 2008, p. 734-735; on lexical bundles in Early Modern vs. Present-day English trials and play texts, see CULPEPER; KYTÖ, 2010, chapter 5). Further, by using a data-driven bottom-up clustering method Gries and Hilpert (2008) identified historical stages in the data based on differing quantitative distributions. The data, originally collected and exploited for Hilpert (2006), had been drawn from the Penn-Helsinki Parsed Corpus of Early Modern English and the Corpus of Late Modern English Texts, with the different spelling variants harmonised to their present-day counterparts (Gries and Hilpert, 2008: 65). The study showed that, for instance in the case of the verbal complementation of 'shall', the three consecutive 140-year periods that had been distinguished as a result of pooling together the original six successive 70-year periods in the corpora did not tally with the way in which the data actually distributed, falling instead into two 180-year groups in quantitative terms. Discoveries such as these are important in that they enable language historians to gain fresh insights and approach language change from a novel perspective. Clearly, developing such techniques, and providing versions of historical corpus texts that enable their use, are among the top priorities in historical corpus linguistics.

3.3 Access to and information on historical electronic resources

Copyright restrictions are an unquestionable bottleneck in the corpus compilation effort, and historical corpora are no exception in this respect. Applying for permission to use and distribute texts in electronic form can be a time-consuming and costly enterprise. Libraries and archives may sometimes be much more forthcoming than publishing houses. Some improvement has been shown recently by, for instance, the Wellcome Library in London, where a generous approach has been adopted for granting permission to use text and images; the British Library and local archives also tend to be generous, apart from requests concerning images, whose use and distribution usually cost considerable sums. Historical corpus compilers are fortunate in that a lot of material has fallen out of copyright. One solution might be to work with editions that are out of copyright, but a potential drawback is that such sources may reflect out-dated linguistic evidence. Also, even though early imprints have fallen out of copyright, libraries usually stipulate that no material from them be distributed to a third party without due application for permission. Compilers of historical corpora have adopted various solutions to the copyright problem, and some of them are worth discussing in the present context.

One way has been, if perhaps only for a transitional period, to publish those parts of the corpus for which copyright is available, as has been done with the Corpus of Early English Correspondence Sampler, which contains half a million of the overall 2.6 million words included in the original Corpus of Early English Correspondence; the rest of the materials could be consulted on an in-house basis. This was also the method applied to the sampler versions of the Innsbruck Computer Archive of Machine-Readable English Texts corpora. A further solution has been to aim at international collaboration within which resources can be shared on a collaborative basis; an example of this is the ARCHER consortium, which pools a number of scholars in many countries in Europe and in the U.S. and, even though no material can be distributed, the consortium is able to offer access to the materials on an in-house basis (YÁÑEZ-BOUZA, 2011). Yet another way is the one chosen for the Time Corpus and the Corpus of Historical American English: the corpus texts are made searchable via a web-based interface that enables a wide range of queries with KWIC displays showing the hit word(s) surrounded by 40 to 60 words or 180 to 200 words in expanded view. This solution is allowed by U.S. copyright law when no more than a certain percentage of each text is displayed to the end-user and when the original text cannot be cut and pasted together from the concordance lines. Even though the raw texts have not been made available, there is great search potential in the solution adopted (DAVIES, 2010, p. 414). Efforts to solve copyright problems will continue to be an important part of the historical corpus compilation initiative.

It is not always easy to obtain accurate and up-to-date information on electronic resources regarding whether the work on them has been completed or is still underway, for example. A recent tool designed to distribute information on English language corpora is the Corpus Resource Database (CoRD) web site at the VARIENG research unit at the University of Helsinki (<http://www.helsinki.fi/varieng/CoRD/index.html>). All descriptions have been submitted or approved by the compilers of each corpus. Each entry contains a set of core information, including a brief description of the corpus, its contents and structure, the names of the compilers, recommended reference line, copyright details, and availability. Other useful information is also offered, including the principles followed in the compilation of the corpus, its annotation conventions, and a bibliography of research conducted using a particular corpus. Compilers of English language corpora can be encouraged to send descriptions of their corpora to the site, and one would welcome similar initiatives for other languages.

3.4 Interdisciplinary considerations

There has been an increasing interest in corpus linguistic techniques among, for instance, literary scholars, discourse analysts, historians and ethnographers. This interdisciplinarity is natural in view of the present trend in historical linguistic research which emphasizes the influence of extralinguistic factors on variation and change in the history of a language. Large-scale text collections have proved useful especially for literary scholars to work on but smaller corpora can also be useful objects of study. Corpora containing full texts, such as the Corpus of Middle English Prose and Verse and the Innsbruck Computer Archive of Machine-Readable English Texts offer valuable material for literary and socio-historical research. Electronic editions such as the An Electronic Text Edition of Depositions 1560-1760 (ETED) which make available transcriptions of early official documents are of interest not only to historical linguists but also to legal and social historians.

Overall, the use made of electronic historical texts is diversifying, and it would benefit the research community if collaboration were increased and efforts pooled across disciplinary borders (WYNNE, 2010, p. 425). For instance, historical linguists have a lot to learn from the methodologies applied by social, political, legal, cultural, and other historians, and from the results they have obtained in their research. In terms of software and other developments, historical corpus linguists should perhaps be more active about reaching out and making their voices heard (CURZAN forthcoming). This would make it easier to make innovative use of even resources that have not necessarily been developed for linguistic research in the first place.

4. Outlook: prospects of historical corpus linguistics

The future prospects of historical corpus linguistics look favourable. As for the English language, there are already vast amounts of digitised material enabling the study of not only the history of the language and literature but also of various aspects of social, political and cultural history in the English-speaking parts of the world. There is also a growing interest in corpus compilation and exploitation and there are also many other areas for further work are many. These aspirations are becoming increasingly felt for many other languages as well. Such positive developments in the field are very much the result of a large body of inspiring research carried out on the extant resources so far. But there is nevertheless plenty of room for further work. Historical corpus linguistics is still very much in the stage where new and exciting discoveries are made but less attention is being paid to the synergetic effects that will become manifest only when resources and research agendas are pooled, and collaboration is extended across interdisciplinary borders.

By way of summary, the proposed list of desiderata for future developments in historical corpus linguistics is here divided into three overarching categories: i) enhancing and adding to the resources and methodologies for studying long-term and recent change, ii) ensuring comparability and links across corpora, other electronic resources, and software, and iii) increasing our knowledge of the sociohistorical and cultural context of corpus texts, with special reference to interdisciplinary considerations. We would benefit from creating further resources that contain everyday, colloquial, utilitarian or non-standard language, spoken and speech-related language, language of women and lower social ranks, language representative of early transplanted varieties and their pidgin and creole-based off-shoots, cross-linguistic material, and early manuscript material in transcriptions faithful to their respective source texts. Further, the present wish list also includes developing linguistically and historically responsible corpus compilation strategies and new corpus compilation "philosophies" aiming at novel explanatory models. This means paying special attention to extralinguistic and linguistic annotation, handling spelling variation, and developing search tools and statistical approaches well suited for interrogating and analysing early texts.

References

Corpora and other electronic resources

ARCHER (A Representative Corpus of Historical English Registers), version 3.1. 2006. See http://www.llc.manchester.ac.uk/research/projects/archer/.

BLOB-1931 corpus. In progress. Compiled by Geoffrey Leech, Paul Rayson and Nick Smith. See http://www.helsinki.fi/varieng/CoRD/corpora/BLOB-1931/index.html.

Bonner Frühneuhochdeutsch Korpus. See http://korpora.zim.uni-duisburg-essen.de/Fnhd/.

Brown Corpus (A Standard Corpus of Present-day Edited American English, for Use with Digital Computers). 1964 (original version). 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 Inscriptionum Latinarum. See http://cil.bbaw.de/.

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

Corpus of Early English Correspondence Sampler. 1998. Compiled by Jukka Keränen, Minna Nevala, Terttu Nevalainen, Arja Nurmi, Minna Palander-Collin and Helena Raumolin-Brunberg (Department of English, University of Helsinki).

Corpus of Early English Medical Writing 1375–1800. In progress. Compiled by Irma Taavitsainen, Päivi Pahta et al. (University of Helsinki). See Middle English Medical Texts and Early Modern English Medical Texts.

A Corpus of English Dialogues 1560–1760. 2006. Compiled under the supervision of Merja Kytö (Uppsala University) and Jonathan Culpeper (Lancaster University). See http://www.helsinki.fi/varieng/CoRD/corpora/CED/index.html.

Corpus of Historical American English. 2010. Compiled by Mark Davies (Brigham Young University). See http://corpus.byu.edu/coha/.

A Corpus of Irish English. 2003. Compiled by Raymond Hickey (University of Duisburg-Essen). See http://www.uni-due.de/CP/CIE.htm.

Corpus of Late Modern English Texts (Extended Version). 2006. Compiled by Hendrik De Smet (Department of Linguistics, University of Leuven). See http://www.helsinki.fi/varieng/CoRD/corpora/CLMETEV/.

Corpus of Middle English Prose and Verse. See http://quod.lib.umich.edu/m/mec/about/.

A Corpus of Nineteenth-century English. Compiled by Merja Kytö (Uppsala University) and Juhani Rudanko (University of Tampere). See Kytö, Merja, Juhani Rudanko and Erik Smitterberg (eds.), Nineteenth-century English: Stability and Change. Cambridge: Cambridge University Press, 2006.

Corpus of Oz Early English. 1998–2004. Compiled by Clemens Fritz (Free University of Berlin). See http://www.helsinki.fi/varieng/CoRD/corpora/COOEE/.

Corpus of Scottish Correspondence, 1500–1715. Compiled by Anneli Meurman-Solin (University of Helsinki). See http://www.helsinki.fi/varieng/CoRD/corpora/CSC/index.html.

The Diachronic Corpus of Present-day Spoken English. 2010. See http://www.ucl.ac.uk/english-usage/projects/dcpse/index.htm.

Dictionary of Old English. See http://www.doe.utoronto.ca/.

Dictionary of Old English Corpus in Electronic Form. 2004. Compiled by Antonette diPaolo Healey, Dorothy Haines, Joan Holland, David McDougall, Ian McDougall and Xin Xiang (University of Toronto). See http://www.doe.utoronto.ca/pub/corpus.html; for earlier versions, see http://www.doe.utoronto.ca/pub/pub.html; see, also, http://www.doe.utoronto.ca/.

Der digitale Grimm. See http://www.lehrer-online.de/digitaler-grimm.php.

Early English Books Online (EEBO). See http://eebo.chadwyck.com/home.

Early Modern English Medical Texts. 2010. Compiled by Irma Taavitsainen, Päivi Pahta, Turo Hiltunen, Ville Marttila, Martti Mäkinen, Maura Ratia, Carla Suhr and Jukka Tyrkkö. CD-ROM with software by Raymond Hickey included in Irma Taavitsainen and Päivi Pahta (eds.), Early Modern English Medical Texts: Corpus Description and Studies. Amsterdam: John Benjamins. See http://www.helsinki.fi/varieng/CoRD/corpora/CEEM/EMEMTindex.html. See, also, Corpus of Early English Medical Writing.

ECCO, see Eighteenth Century Collections Online.

EEBO, see Early English Books Online.

Eighteenth Century Collections Online (ECCO). See http://www.gale.cengage.com/pdf/facts/ECCO.pdf.

e-LALME. See http://www.ling.ed.ac.uk/research/ihd/projectsX.shtml.

Electronic Beowulf. 2003. Edited by Kevin Kiernan. See http://www.uky.edu/~kiernan/eBeowulf/guide.htm.

An Electronic Text Edition of Depositions 1560–1760. Forthcoming (2011). See Kytö, Merja, Peter J. Grund and Terry Walker, Testifying to Language and Life in Early Modern England. Including a CD containing An Electronic Text Edition of Depositions 1560–1760 (ETED). Amsterdam/Philadelphia: John Benjamins.

English Books Online (EEBO). See http://lion.chadwyck.co.uk.

English-Swedish Drama Dialogue. In progress. Compiled by Linnéa Anglemark, Merja Kytö, Ulla Melander Marttala and Mats Thelander (Department of English and Department of Scandinavian Languages, Uppsala University).

F-LOB (Freiburg-LOB Corpus of British English). 1999 (original version), 2007 (POS-tagged version). Original version compiled by Christian Mair (Albert-Ludwigs-Universität Freiburg); POS-tagged version compiled by Christian Mair (Albert Ludwigs-Universität Freiburg) and Geoffrey Leech (Lancaster University). See http://www.helsinki.fi/varieng/CoRD/corpora/FLOB/index.html.

Frown Corpus (Freiburg-Brown Corpus of American English). 1999 (original version), 2007 (POS-tagged version). Original version compiled by Christian Mair (Albert-Ludwigs-Universität Freiburg); POS-tagged version compiled by Christian Mair (Albert Ludwigs-Universität Freiburg) and Geoffrey Leech (Lancaster University). See http://www.helsinki.fi/varieng/CoRD/corpora/FROWN/index.html.

GerManC. In progress. Compiled by Martin Durrell, Paul Bennett, Silke Scheible and Richard J. Witt. See http://www.llc.manchester.ac.uk/research/projects/germanc/.

The Helsinki Corpus of English Texts. 1991. 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) (Department of English, University of Helsinki). See http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/index.html.

Historical Thesaurus of English. See http://libra.englang.arts.gla.ac.uk/historicalthesaurus/aboutproject.html.

Innsbruck Computer Archive of Machine-Readable English Texts. Compiled by Manfred Markus (University of Innsbruck). See, e.g., http://www.anglistikguide.de/cgi-bin/ssgfi/anzeige.pl?db=lit&ew=SSGFI&nr=000753.

International Corpus of English (ICE). See http://ice-corpora.net/ice/.

Johnson, Samuel, see McDermott 1996.

Lampeter Corpus of Early Modern English Tracts. See http://www.helsinki.fi/varieng/CoRD/corpora/LC/index.html and http://khnt.hit.uib.no/icame/manuals/LAMPETER/LAMPHOME.HTM.

Lancaster Newsbook Corpus. See http://www.lancs.ac.uk/fass/projects/newsbooks/.

Lexicons of Early Modern English. See http://leme.library.utoronto.ca/public/.

LOB (Lancaster-Oslo/Bergen Corpus). 1976 (original version), 1986 (POS-tagged version). Compiled by Geoffrey Leech, Stig Johansson, Knut Hofland and Roger Garside. See http://www.helsinki.fi/varieng/CoRD/corpora/LOB/index.html.

A Linguistic Atlas of Early Middle English (1150–1325), version 2.1. 2008. Compiled by Margaret Laing and Roger Lass (Edinburgh: The University of Edinburgh). See http://www.lel.ed.ac.uk/ihd/laeme1/laeme1.html.

A Linguistic Atlas of Older Scots, Phase 1 (1380–1500). 2008 (current version). © 2007 Edinburgh: The University of Edinburgh. See http://www.helsinki.fi/varieng/CoRD/corpora/LAEME/index.html and http://www.lel.ed.ac.uk/ihd/laos1/laos1.html.

Lion, see Literature Online.

Literature Online (Lion). See http://lion.chadwyck.co.uk/.

A London Provisioner's Chronicle, 1550–1563, by Henry Machyn: Manuscript, Transcription, and Modernization. 2006. Edited by Richard W. Bailey, Marilyn Miller and Colette Moore. Ann Arbor, Michigan: University of Michigan Press; Scholarly Publishing Office of the University of Michigan University Library. See http://quod.lib.umich.edu/m/machyn/.

London-Lund Corpus of Spoken English. 1980 (original version). Compiled by Jan Svartvik. See http://khnt.hit.uib.no/icame/manuals/LONDLUND/INDEX.HTM.

McDermott, Ann (ed.). 1996. Samuel Johnson: Dictionary of the English Language on CD-ROM. Cambridge: Cambridge University Press. See http://xml.coverpages.org/cup-johnson.html.

Middle English Dictionary. 2001. See http://quod.lib.umich.edu/m/med/.

The Middle English Grammar Corpus. See http://www.arts.gla.ac.uk/SESLL/EngLang/ihsl/projects/MEG/meg.htm.

Middle English Medical Texts. 2005. Compiled by Irma Taavitsainen, Päivi Pahta and Martti Mäkinen. CD-ROM with software by Raymond Hickey. Amsterdam: John Benjamins. See http://www.helsinki.fi/varieng/CoRD/corpora/CEEM/MEMTindex.html. See, also, Corpus of Early English Medical Writing.

Mitterhochdeutsche Begriffsdatenbank. See http://mhdbdb.sbg.ac.at:8000/index.en.html.

Old Bailey Corpus. See The Proceedings of the Old Bailey, London's Central Criminal Court, 1674 to 1913 available at http://www.oldbaileyonline.org/. See, also, http://www.uni-giessen.de/oldbaileycorpus/index.php.

Oxford English Dictionary Online (OED Online). See http://www.oed.com/.

Oxford Text Archive. See http://ota.ahds.ac.uk/.

Parsed Corpus of Early English Correspondence, parsed version. 2006. Annotated by Ann Taylor, Arja Nurmi, Anthony Warner, Susan Pintzuk, and Terttu Nevalainen. Compiled by the CEEC Project Team. York: University of York and Helsinki: University of Helsinki.

Parsed Corpus of Early English Correspondence, tagged version. 2006. Annotated by Arja Nurmi, Ann Taylor, Anthony Warner, Susan Pintzuk, and Terttu Nevalainen. Compiled by the CEEC Project Team. York: University of York and Helsinki: University of Helsinki.

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.

Penn Parsed Corpus of Modern British English. See http://www.ling.upenn.edu/hist-corpora/.

Penn-Helsinki Parsed Corpus of Early Modern English. See http://www.ling.upenn.edu/hist-corpora/.

Penn-Helsinki Parsed Corpus of Middle English (second edition). See http://www.ling.upenn.edu/hist-corpora/.

Project Gutenberg. See http://www.gutenberg.org/wiki/Main_Page.

ProQuest Historical Newspapers. See www.proquest.com.

Shakespeare Corpus. See, e.g., http://www.lexically.net/downloads/corpus_linguistics/shakespeare_corpus_readme.txt.

Sociopragmatic Corpus, a Specialised Sub-Section of A Corpus of English Dialogues 1560–1760. 2007. Compiled by Jonathan Culpeper and Dawn Archer (Lancaster University).

Svenska Akademiens ordbok. See http://g3.spraakdata.gu.se/saob/.

Thesaurus Linguae Graecae. See http://www.tlg.uci.edu/.

Time Corpus (Time Magazine Corpus of American English). 2007. Compiled by Mark Davies (Brigham Young University). See http://corpus.byu.edu/time/.

Times Digital Archive. See www.gale.cengage.com.

York-Helsinki Parsed Corpus of Old English Poetry. 2001. Compiled by Susan Pintzuk and Leendert Plug (University of York). See http://www-users.york.ac.uk/~lang18/pcorpus.html.

York-Helsinki Parsed Corpus of Old English Prose. 2003. Compiled by Ann Taylor, Anthony Warner, Susan Pintzuk, Frank Beths (University of York). See http://www-users.york.ac.uk/~lang22/YcoeHome1.htm.

Zurich English Newspaper Corpus. Version 1.0. 2004. Compiled by Udo Fries, Hans Martin Lehmann, Beni Ruef, Peter Schnieder, Patrick Studer, Caren auf dem Keller, Beat Nietlispach, Sandra Engler, Sabine Hensel and Franziska Zeller (University of Zurich). See http://www.helsinki.fi/varieng/CoRD/corpora/ZEN/index.html.

Other references

ARCHER, D. (Ed.). *What's in a word-list?* Investigating word frequency and keyword extraction. Farnham: Ashgate, 2009.

ARCHER, D. Data retrieval in a diachronic context: The case of the historical English courtroom. In: NEVALAINEN, T.; TRAUGOTT, E. (Ed.). *A handbook to the history of English*. Oxford: Oxford University Press, forthcoming.

ARCHER, D.; McENERY, T.; RAYSON, P.; HARDIE, A. Developing an automated semantic analysis system for Early Modern English. In: ARCHER, D.; RAYSON, P.; WILSON, A.; McENERY, T. (Ed.). Corpus Linguistics 2003 Conference. *Proceedings*... (UCREL technical paper number 16). Lancaster: UCREL, Lancaster University, 2003.

BARON, A.; RAYSON, P. VARD 2: A tool for dealing with spelling variation in historical corpora. In: Postgraduate Conference in Corpus Linguistics. *Proceedings*... Aston University, Birmingham, 22 May 2008. Retrieved May 5, 2011 from http://acorn.aston.ac.uk/conf_proceedings.html.

BARON, A.; RAYSON, P. Automatic standardization of texts containing spelling variation, how much training data do you need? In: MAHLBERG, M.; GONZÁLEZ-DIAZ, V.; SMITH, C. (Ed.). Corpus Linguistics Conference, CL2009. *Proceedings*... University of Liverpool, UK, 20-23 July 2009. Available at: <http://ucrel.lancs.ac.uk/publications/cl2009/>. Retrieved: May 5, 2011.

BARON, A.; RAYSON, P.; ARCHER, D. Word frequency and key word statistics in historical corpus linguistics. *Anglistik: International Journal of English Studies*, v. 20, n. 1, p. 41-67, 2009a.

BARON, A.; RAYSON, P.; ARCHER, D. Automatic standardization of spelling for historical text mining. In: 'Digital Humanities 2009'. *Proceedings*... University of Maryland, USA, 22-25 June 2009. 2009b.

BIBER, D.; FINEGAN, E. Drift and the evolution of English style: A history of three genres. *Language*, v. 65, n. 3, p. 487-517, 1989.

BIBER, D.; FINEGAN, E. 1992. The linguistic evolution of five written and speech-based English genres from the 17th to the 20th centuries. In: RISSANEN, M.; IHALAINEN, O.; NEVALAINEN, T.; TAAVITSAINEN, I. (Ed.). *History of Englishes*: new methods and interpretations in historical linguistics. Berlin/New York: Mouton de Gruyter, 1992. (Topics in English Linguistics 10)

BIBER, D.; FINEGAN, E. Diachronic relations among speech-based and written registers in English. In: NEVALAINEN, T.; KAHLAS-TARKKA, L. (Ed.). *To explain the present*: studies in the changing English language in honour of Matti Rissanen. Helsinki: Société Néophilologique, 1997. (Mémoires de la Société Néophilologique 52)

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

BRINTON, L. J. Pathways in the development of pragmatic markers in English. In: van KEMENADE, A.; LOS, B. (Ed.). *The handbook of the history of English*. London: Blackwell, 2006.

CLARIDGE, C.. Historical corpora. In: LÜDELING, A.; KYTÖ, M. (Ed.). *Corpus linguistics*: an international handbook. Berlin/New York: Walter de Gruyter, 2008. (Handbooks of Linguistics and Communication Science / Handbücher zur Sprach- und Kommunikationswissenschaft 29.1-2.)

CLARIDGE, C.; KYTÖ, M. 2010. Non-standard language in earlier English. In: HICKEY, R. (Ed.). *Varieties of English in writing*. The written word as linguistic evidence. Amsterdam/Philadelphia: John Benjamins, 2010. (Varieties of English Around the World G41)

COLLINS, D. E. *Reanimated voices*. Speech reporting in a historical-pragmatic perspective. Amsterdam/Philadelphia: John Benjamins, 2001. (Pragmatics & Beyond New Series 85)

CORPUS PRESENTER. Available at: <http://www.uni-due.de/CP/>.

CORPUS RESOURCE DATABASE (CoRD). Available at: <http://www.helsinki.fi/varieng/CoRD/index.html>.

CULPEPER, J.; KYTÖ, M. *Early Modern English dialogues*: spoken interaction as writing. Cambridge: Cambridge University Press, 2010.

CURZAN, A. English historical corpora in the classroom: the intersection of teaching and research. *Journal of English Linguistics*, v. 28, n. 1, p. 77-89, 2000.

CURZAN, A. Historical corpus linguistics and evidence of language change. In: LÜDELING, A.; KYTÖ, M. (Ed.). *Corpus linguistics*: an international handbook. Berlin/New York: Walter de Gruyter, 2008. (Handbooks of Linguistics and Communication Science / Handbücher zur Sprach- und Kommunikationswissenschaft 29.1-2.)

CURZAN, A. The electronic life of texts: Insights from corpus linguistics for all fields of English. In: KYTÖ, M. (Ed.). *English corpus linguistics*: crossing paths. Amsterdam: Rodopi, forthcoming.

DAVIES, M. More than a peephole: using large and diverse online corpora. In: POPE, C. W. (Ed.). Special issue on the bootcamp discourse and beyond. *International Journal of Corpus Linguistics*, v. 15, n. 3, p. 412-418, 2010.

DE SMET, H. A corpus of Late Modern English texts. *ICAME Journal*, v. 29, p. 69-82, 2005.

FITZMAURICE, S. M.; TAAVITSAINEN, I. (Ed.). *Methods in historical pragmatics*. Berlin/New York: Mouton de Gruyter, 2007. (Topics in English Linguistics 52)

GORDON, E.; CAMPBELL, L.; HAY, J.; MACLAGAN, M.; SUDBURY, A.; TRUDGILL, P. *New Zealand English*. Its origins and evolution. Cambridge: Cambridge University Press, 2009.

GRIES, St. Th.; HILPERT, M. The identification of stages in diachronic data: variability-based neighbor clustering. *Corpora*, v. 3, n. 1, p. 59-81, 2008.

HILPERT, M. Distinctive collexeme analysis and diachrony. *Corpus Linguistics and Linguistic Theory*, v. 2, n. 2, p. 243-256, 2006.

HONKAPOHJA, A.; KAISLANIEMI, S.; MARTTILA, V. Digital editions for corpus linguistics: representing manuscript reality in electronic corpora. In: JUCKER, A. H.; SCHREIER, D.; HUNDT, M. (Ed.). Corpora: pragmatics and discourse. 29th International Conference on English Language Research on Computerized Corpora (ICAME 29). *Papers...* Ascona, Switzerland, 14-18 May 2008. Amsterdam: Rodopi, 2009.

HUBER, M. The Old Bailey Proceedings, 1674-1834: evaluating and annotating a corpus of 18th- and 19th-century spoken English. In: MEURMAN-SOLIN, A.; NURMI, A. (Ed.). *Annotating variation and change*. Helsinki: Research Unit for Variation, Contacts and Change in English (VARIENG), University of Helsinki, 2007. (Studies in Variation, Contacts and Change in English 1). Available at: <http://www.helsinki.fi/varieng/journal/volumes/01/huber/>. Retrieved: May 5, 2011.

HÜLLEN, W. A close reading of William Caxton's Dialogues: "… to lerne Shortly frenssh and englyssh". In: JUCKER, A. H. (Ed.). *Historical pragmatics*: pragmatic developments in the history of English. Amsterdam/Philadelphia: John Benjamins, 1995. (Pragmatics & Beyond New Series 35)

JACOBS, A.; JUCKER, A. H. The historical perspective in pragmatics. In: JUCKER, A. H. (Ed.). *Historical pragmatics*. Pragmatic developments in the history of English. Amsterdam/Philadelphia: John Benjamins, 1995. (Pragmatics & Beyond New Series 35)

JOHANSSON, S. Some aspects of the development of corpus linguistics in the 1970s and 1980s. In: LÜDELING, A.; KYTÖ, M. (Ed.). *Corpus linguistics*: an international handbook. Berlin/New York: Walter de Gruyter, 2008. (Handbooks of Linguistics and Communication Science / Handbücher zur Sprach- und Kommunikationswissenschaft 29.1-2.)

JUCKER, A. H. *Historical pragmatics*: pragmatic developments in the history of English. Amsterdam/Philadelphia: John Benjamins, 1995. (Pragmatics & Beyond New Series 35)

JUCKER, A. H.; FRITZ, G.; LEBSANFT, F. *Historical dialogue analysis*. Amsterdam/Philadelphia: John Benjamins, 1999a. (Pragmatics & Beyond New Series 66)

JUCKER, A. H.; FRITZ, G.; LEBSANFT, F. Historical dialogue analysis: roots and traditions in the study of the Romance languages, German and English. In: JUCKER, A. H.; FRITZ, G.; LEBSANFT, F. (Ed.). *Historical dialogue analysis*. Amsterdam/Philadelphia: John Benjamins, 1999b. (Pragmatics & Beyond New Series 66)

JUCKER, A. H.; SCHNEIDER, G.; TAAVITSAINEN, I.; BREUSTEDT, B. Fishing for compliments: precision and recall in corpus-linguistic compliment research. In: JUCKER, A. H.; TAAVITSAINEN, I. (Ed.). *Speech acts in the history of English*. Amsterdam/Philadelphia: John Benjamins, 2008. (Pragmatics and Beyond New Series 176)

JUCKER, A. H.; TAAVITSAINEN, I. Diachronic speech act analysis: the case of insults. *Journal of Historical Pragmatics*, v. 1, n. 1, p. 67-95, 2000.

JUCKER, A. H; TAAVITSAINEN, I. (Ed.). *Speech acts in the history of English*. Amsterdam/Philadelphia: John Benjamins, 2008a. (Pragmatics and Beyond New Series 176)

JUCKER, A. H.; TAAVITSAINEN, I. Apologies in the history of English: routinized and lexicalized expressions of responsibility and regret. In: JUCKER, A. H.; TAAVITSAINEN, I. (Ed.). *Speech acts in the history of English*. Amsterdam/Philadelphia: John Benjamins, 2008b. (Pragmatics and Beyond New Series 176)

KOHNEN, T. Text types and the methodology of diachronic speech act analysis. In: FITZMAURICE, S. M.; TAAVITSAINEN, I. (Ed.). *Methods in historical pragmatics*. Berlin/New York: Mouton de Gruyter, 2007. (Topics in English Linguistics 52)

KYTÖ, M. Data in historical pragmatics. In: JUCKER, A. H.; TAAVITSAINEN, I. (Ed.). *Historical pragmatics*. Berlin/New York: Walter de Gruyter, 2010. (Handbooks of Pragmatics 8)

KYTÖ, M. Corpus linguistics. In: BERGS, A.; BRINTON, L. (Ed.). *Historical linguistics of English*: an international handbook. Berlin/New York: Walter de Gruyter, forthcoming. (Handbooks of Linguistics and Communication Science / Handbücher zur Sprach- und Kommunikationswissenschaft).

KYTÖ, M.; GRUND, P. J.; WALKER, T. *Testifying to language and life in Early Modern England*. Amsterdam/Philadelphia: John Benjamins, forthcoming (2011). Including a CD containing An Electronic Text Edition of Depositions 1560-1760 (ETED)

KYTÖ, M.; PAHTA, P. Evidence from historical corpora up to the twentieth century. In: NEVALAINEN, T.; TRAUGOTT, E. (Ed.). *A handbook to the history of English*. Oxford: Oxford University Press, forthcoming.

KYTÖ, M.; RISSANEN, M. The syntactic study of Early American English: the variationist at the mercy of his corpus? *Neuphilologische Mitteilungen*, v. 84, n. 4, p. 470-490, 1983.

KYTÖ, M.; WALKER, T. The linguistic study of Early Modern English speech-related texts: how "bad" can "bad" data be? *Journal of English Linguistics*, v. 31, n. 3, p. 221-248, 2003.

KYTÖ, M.; WALKER, T. *Guide to A Corpus of English Dialogues 1560-1760*. Uppsala: Acta Universitatis Upsaliensis, 2006. (Studia Anglistica Upsaliensia 130)

LANCASHIRE, I. *Introduction*. 2006. Available at: <http://leme.library.utoronto.ca/public/intro.cfm>. Retrieved: May 5, 2011.

LASS, R. *On explaining language change*. Cambridge: Cambridge University Press, 1980.

LEECH, G.; HUNDT, M.; MAIR, C.; SMITH, N. *Change in contemporary English*: a grammatical study. Cambridge: Cambridge University Press, 2009.

A LINGUISTIC ATLAS OF LATE MEDIAEVAL ENGLISH. Compiled by McINTOSH, A.; SAMUELS, M. L.; BENSKIN, M.; with the assistance of Margaret Laing and Keith Williamson. 4 v. Aberdeen: Aberdeen University Press, 1986.

LUTZKY, U. *Discourse markers in Early Modern English*: The case of 'marry', 'well' and 'why'. 325 p. 2009. (PhD thesis) – Vienna University.

MacQUEEN, D. S. *The integration of MILLION into the English system of number words*: a diachronic study. Frankfurt am Main/Berlin, etc.: Peter Lang, 2010. (English Corpus Linguistics 11)

MAIR, C. Corpora and the study of recent change in language. In: LÜDELING, A.; KYTÖ, M. (Ed.). *Corpus linguistics*: an international handbook. Berlin/New York: Walter de Gruyter, 2008. (Handbooks of Linguistics and Communication Science / Handbücher zur Sprach- und Kommunikationswissenschaft 29.1-2.)

MARKUS, M. Normalization of Middle English prose: possibilities and limits. In: LJUNG, M. (Ed.). Corpus-based studies in English. Seventeenth International Conference on English Language Research on Computerized Corpora. *Papers*... Stockholm, May 15-19, 1996. Amsterdam/Atlanta, GA: Rodopi, 1997.

McENERY, T.; WILSON, A. *Corpus linguistics*: an introduction. Second edition. Edinburgh: Edinburgh University Press, 2001 [1996].

MEYER, C. F. *English corpus linguistics*: an introduction. Cambridge: Cambridge University Press, 2002.

MEYER, C. F. Pre-electronic corpora. In: LÜDELING, A.; KYTÖ, M. (Ed.). *Corpus linguistics*: an international handbook. Berlin/New York: Walter de Gruyter, 2008. (Handbooks of Linguistics and Communication Science / Handbücher zur Sprach- und Kommunikationswissenschaft 29.1-2.)

MILROY, J. A social model for the interpretation of language change. In: RISSANEN, M.; IHALAINEN, O.; NEVALAINEN, T.; TAAVITSAINEN, I. (Ed.). *History of Englishes*: new methods and interpretations in historical linguistics. Berlin and New York: Mouton de Gruyter, 1992. (Topics in English Linguistics 10)

MONOCONC PRO. Available at: <http://www.athel.com/mono.html>.

NEVALAINEN, T.; RAUMOLIN-BRUNBERG, H. *Historical sociolinguistics*: language change in Tudor and Stuart England. London/New York/Toronto: Pearson Education, 2003.

NURMI, A.; NEVALA, M.; PALANDER-COLLIN, M. (Ed.). *The language of daily life in England* (1400-1800). Amsterdam/Philadelphia: John Benjamins, 2009. (Pragmatics & Beyond New Series 183)

PETRÉ, P. *Leuven English Old to New (LEON)*: some ideas on a new corpus for longitudinal diachronic studies. Paper given at the Middle and Modern English Corpus Linguistics conference, University of Innsbruck, 5-9 July 2009.

THE R PROJECT FOR STATISTICAL COMPUTING. Available at: <http:/ /www.r-project.org/>.

RAUMOLIN-BRUNBERG, H. Review of SOKOLL, T. (Ed.). Essex pauper letters 1731-1837. Records of Social and Economic History, New Series 30. Published for The British Academy by Oxford University Press, 2001. *Historical Sociolinguistics and Sociohistorical Linguistics*, v. 3, 2003. Available at: <http:// www.let.leidenuniv.nl/hsl_shl/sokoll.htm> Retrieved: May 5, 2011.

RAYSON, P,; ARCHER, D.; BARON, A.; SMITH, N. Tagging historical corpora – the problem of spelling variation. In: Digital Historical Corpora. Dagstuhl-Seminar 06491, International Conference and Research Center for Computer Science. *Proceedings*... Schloss Dagstuhl, Wadern, Germany, December 3rd-8th 2006. 2007. Available at: <http://drops. dagstuhl.de/opus/ volltexte/2007/1055/. Retrieved: May 5, 2011.

RISSANEN, M. Variation and the study of English historical syntax. In: SANKOFF, D. (Ed.). *Diversity and diachrony*. Amsterdam/Philadelphia: John Benjamins, 1986. (Current Issues in Linguistic Theory 53)

RISSANEN, M. Syntax. In: LASS, R. (Ed.). *The Cambridge history of the English language*, v. III, 1476-1776. Cambridge: Cambridge University Press, 97-109. p. 187-331.

RISSANEN, M. Corpora and the study of the history of English. In: KYTÖ, M. (Ed.). *English corpus linguistics*: crossing paths. Amsterdam: Rodopi, forthcoming.

RISSANEN, M.; KAHLAS-TARKKA, L.; HINTIKKA, M.; McCONCHIE, R. (Ed.). *Change in meaning and the meaning of change*: studies in semantics and grammar from Old to Present-day English. Helsinki: Société Néophilologique, 2007. (Mémoires de la Société Néophilologique de Helsinki 72)

ROMAINE, S. *Socio-historical linguistics*: its status and methodology. Cambridge: Cambridge University Press, 1982. (Cambridge Studies in Linguistics 34)

ROMAINE, S. Variation in language and gender. In: HOLMES, J.; MEYERHOFF, M. (Ed.). *The handbook of language and gender*. Malden, MA: Blackwell, 2003. (Blackwell Handbooks in Linguistics 13)

SAMUELS, M. L. *Linguistic evolution, with special reference to English*. Cambridge: Cambridge University Press, 1972. (Cambridge Studies in Linguistics 5)

SOKOLL, T. (Ed.). *Essex pauper letters*, 1731-1837. Published for The British Academy by Oxford University Press, 2001. (Records of Social and Economic History, New Series 30)

TAAVITSAINEN, I.; JUCKER, A. H. Speech act verbs and speech acts in the history of English. In: FITZMAURICE, S. M.; TAAVITSAINEN, I. (Ed.). *Methods in historical pragmatics*. Berlin/New York: Mouton de Gruyter, 2007. (Topics in English Linguistics 52)

TAAVITSAINEN, I.; JUCKER, A. H. Speech acts now and then: towards a pragmatic history of English. In: JUCKER, A. H.; TAAVITSAINEN, I. (Ed.). *Speech acts in the history of English*. Amsterdam/Philadelphia: John Benjamins, 2008a. (Pragmatics and Beyond New Series 176)

TAAVITSAINEN, I.; JUCKER, A. H. "Methinks you seem more beautiful than ever": compliments and gender in the history of English. In: JUCKER, A. H.; TAAVITSAINEN, A. H. (Ed.). *Speech acts in the history of English*. Amsterdam/Philadelphia: John Benjamins, 2008b. (Pragmatics and Beyond New Series 176)

TEXT ENCODING INITIATIVE (TEI). Available at: <http://www.tei-c.org/index.xml>.

TRAUGOTT, E. C.; DASHER, R. B. *Regularity in semantic change*. Cambridge: Cambridge University Press, 2002. (Cambridge Studies in Linguistics 96)

VARD2. Available at: <http://www.comp.lancs.ac.uk/~barona/vard2/>.

WALKER, T. *Thou and You in Early Modern English dialogues*: trials, depositions, and drama comedy. Amsterdam/Philadelphia: John Benjamins, 2007. (Pragmatics & Beyond New Series 158)

WATTS, R. J. Refugiate in a strange countrey: learning English through dialogues in the 16th century. In: JUCKER, A. H.; FRITZ, G.; LEBSANFT, F. (Ed.). *Historical dialogue analysis*. Amsterdam/Philadelphia: John Benjamins, 1999. (Pragmatics & Beyond New Series 66)

WEINREICH, U.; LABOV, W.; HERZOG, M. I. Empirical foundations for a theory of language change. In: LEHMANN, W. P.; MALKIEL, Y. (Ed.). *Directions for historical linguistics*: a symposium. Austin/London: University of Texas Press, 1968.

WORDSMITH TOOLS. Available at: <http://www.lexically.net/wordsmith/>.

WYNNE, M. Searching and concordancing. In: LÜDELING, A.; KYTÖ, M. (Ed.). *Corpus linguistics*: an international handbook. Berlin/New York: Walter de Gruyter, 2008. (Handbooks of Linguistics and Communication Science / Handbücher zur Sprach- und Kommunikationswissenschaft 29.1-2.)

WYNNE, M. Interdisciplinary relationships. In: POPE, Caty Worlock (Ed.). Special issue on the bootcamp discourse and beyond. *International Journal of Corpus Linguistics*, v. 15, n. 3, p. 425-427, 2010.

XAIRA. Available at: <http://www.oucs.ox.ac.uk/rts/xaira/>.

XIAO, R. Well-known and influential corpora. In: LÜDELING, A.; KYTÖ, M. (Ed.). *Corpus linguistics*: an international handbook. Berlin/New York: Walter de Gruyter, 2008. (Handbooks of Linguistics and Communication Science / Handbücher zur Sprach- und Kommunikationswissenschaft 29.1-2.)

YÁÑEZ-BOUZA, N. ARCHER past and present (1990-2010). *ICAME Journal*, v. 35, p. 205-236, 2011.

Recebido em 21/01/2011. Aprovado em 08/05/2011.