UNIVERSITY OF MURCIA

International Journal
of
English Studies

IJES
www.um.es/ijes

A descriptive approach to computerised English historical corpora in the 21st century

NILA VÁZQUEZ, LAURA ESTEBAN-SEGURA & TERESA MARQUÉS-AGUADO*

University of Murcia

Received: 10 April 2011 / Accepted: 6 November 2011

ABSTRACT
Historical corpora offer many potentialities for linguistic research. Thus, the present article provides an overview of the major English historical corpora compiled or being compiled both in Spain and abroad. They include different types such as tagged and parsed corpora, and their main features will be outlined. As for the organisation of the article, after the introductory section, the historical corpora created abroad will be presented. Then, those being constructed in Spain (Coruña, Las Palmas, Málaga, Salamanca, Santiago and Sevilla) will be discussed. Some final remarks and the references close the article.

KEYWORDS: historical corpora, Old English, Middle English, Modern English, diachrony, annotated corpus, lemmatisation.

RESUMEN
Los corpus históricos ofrecen múltiples posibilidades para la investigación lingüística. En el presente artículo se proporciona una visión general de los corpus históricos ingleses más importantes, compilados o que están en proceso de compilación, a nivel nacional e internacional. Los corpus considerados incluyen distintos tipos, entre los que se encuentran los analizados morfológica y sintácticamente, de los que se esbozan las principales características. En cuanto a la estructura del trabajo, tras la introducción, se presentan los corpus creados en el extranjero y seguidamente se tratan aquellos que se están recopilando en España (La Coruña, Las Palmas, Málaga, Salamanca, Santiago y Sevilla). Cierran el estudio algunas consideraciones finales y las referencias bibliográficas.

PALABRAS CLAVE: corpus históricos, inglés antiguo, inglés medio, inglés moderno, diacronía, corpus anotado, lematización.

****Address for correspondence:*** Dr. Nila Vázquez, Dr. Laura Esteban-Segura and Dr. Teresa Marqués-Aguado. Departamento de Filología Inglesa, Facultad de Letras, Universidad de Murcia, Campus de La Merced, 30071. Murcia, Spain. E-mails: nilavg@um.es, lesteban@um.es, tmarques@um.es.

ISSN: 1578-7044

1. INTRODUCTION

The number of scholars working on the field of English historical corpus linguistics has increased in the last decades, and so has the need for revised and/or new material to be analysed. Since the *Helsinki Corpus of English Texts* (the first, and pioneering, computerised English diachronic corpus) appeared in 1991, a wealth of historical corpora has been created in response to that need, reflecting the growing interest and development within this specific linguistic area in a relatively short period of time.

Historical corpora can be divided into synchronic and diachronic corpora; the former take into account particular and defined periods such as, for instance, the *Century of Prose Corpus* (COPC; Milic 1995), which concentrates on the specific century 1680-1780. Diachronic corpora, on the other hand, cover longer periods; an example would be the *Helsinki Corpus*, ranging over ten centuries (ca. 730-1710). However, this distinction is not strict, as subsections of diachronic corpora may be regarded as synchronic historical corpora (Claridge 2008: 242).

Another classification, which may be valid for historical corpora as well, is into dynamic and static corpora, depending on whether they have a finite size (static) or, on the contrary, are open to the addition of further texts (dynamic).[1] Nevalainen (2008: 23) has pointed out that the majority of historical corpora are made up of text selections, so as to get a glimpse of different kinds of writing, such as scientific and legal documents, handbooks, drama, fiction, diaries, personal letters, etc. Accordingly, they can be designed as single- or multi-genre corpora. Historical corpora can also be annotated, including part-of-speech (POS) tagging (grammatical categories) and parsing (syntactic structure), and can contain background information about the texts and their authors that may, in turn, be of use from a social, historical or cultural standpoint.

The advantages that computerised historical corpora present for research are manifold. They allow looking into linguistic phenomena quantitatively, since language change can be a quantitatively verifiable process (from lists of frequencies to lists of collocational patterns, for instance). Corpora can also be used qualitatively, for example, by retrieving occurrences in a quick and efficient manner. Other benefits are the reliability furnished by computer/machine-processed data and the possibility of applying scientific criteria to linguistic studies (for example, for statistical measure), which result in objectivity and transparency. Moreover, historical corpora provide evidence of structures and forms of the language that were in use in specific periods. These structures and forms may account for linguistic change, which can occur at a variable pace in different registers and genres, and for linguistic usage; the array of corpora described in this article, as presented below, can cater for investigation in several periods, registers and genres.

However, several disadvantages of historical corpora should be mentioned. One of them has to do with the material available which, as opposed to contemporary corpora, is restricted to what has remained or come down to us. Moreover, the sources are only written. Chronological gaps may also be encountered, with some periods being more represented than others.

The aim of this article is to present a survey of the most important computerised English historical corpora available up to the present moment. The information has been arranged into two different sections. In the first one, the corpora compiled abroad are dealt with; they include the *Helsinki Corpus of English Texts*, the Parsed Corpora of Historical English, the *Corpus of Early English Correspondence* and the *Corpus of Historical American English*, among others, together with lesser known corpora, such as the *Leuven English Old to New Corpus* or the *Lampeter Corpus of Early Modern English Tracts*. The second section focuses on the historical corpora which are being currently compiled in Spain, and attention is paid to the work undertaken at the Universities of Coruña, Las Palmas, Málaga, Salamanca, Santiago and Sevilla. The output of some of these projects, particularly those in progress, will help to throw light on the dialectology of early periods of the English language, in addition to offering new data for different kinds of diachronic and synchronic research (morphosyntactic, lexical, sociolinguistic, etc.).

2. CORPORA COMPILED ABROAD

In this section, the leading English historical corpora compiled abroad are presented, focusing on their features, main advantages or achievements, and possible shortcomings (if any). Besides, other minor corpora are briefly explained. Those related by means of joint projects are listed and discussed together, and, by the same token, those derived (whether directly or indirectly) from the *Helsinki Corpus* are also dealt with together. Besides, reference is made to the software tools associated to particular corpora for retrieving linguistic information. The more relevant corpora are described in greater detail.[2]

2.1. *Helsinki Corpus of English Texts* (HC)[3]

Compiled by the team led by Prof. Matti Rissanen (University of Helsinki) between 1984 and 1991 and released that year, HC is without any doubt the pioneering work in the field of historical corpora compilation, which, as commented below, has led to subsequent compilation and work on other corpora. It is also fairly comprehensive, inasmuch as it collects data from Old English (OE) to early Modern English (eModE) (ca. 730-1710), with texts of different genres containing more than 1.5 million words. The timespan covered by this corpus is further divided into periods and subperiods, as shown in Table 1:

Subperiod	**Dates**	**Words**
Old English		
OE1	-850	2,190
OE2	850-950	92,050
OE3	950-1050	251,630
OE4	1050-1150	67,380
Subtotal		413,250
Middle English		
ME1	1150-1250	113,010
ME2	1250-1350	97,480
ME3	1350-1420	184,230
ME4	1420-1500	213,850
Subtotal		608,570
Early Modern English, British		
EModE1	1500-1570	190,160
EModE2	1570-1640	189,800
EModE3	1640-1710	171,040
Subtotal		551,000
Total (Basic corpus)		**1,572,820**

Table 1. The diachronic part of HC: size and period divisions[4]

Each text is headed by a short description including twenty-five parameters,[5] which provide pragmatic and sociolinguistic information such as gender and social position of the author, text-type, type of interaction, etc. In order to select the texts, the main objective was to achieve coverage in terms of chronology, regional distribution, sociolinguistic information and genres. As a result, this corpus lends itself well to sociolinguistic studies, together with those aimed at testing certain language structures with real data, especially because it is truly diachronic.

Its main shortcoming (as put forward by Kohnen 2007) is the lack of lemmatisation and/or tagging, which may be taken as a significant difficulty when dealing with OE and Middle English (ME) texts due to the morphological and orthographical variation existing in those periods. Bearing in mind this constraint, searches can be carried out by using general software tools such as *WordSmith Tools* (Scott 1998), or the *Oxford Concordance Programme* or *WordCruncher* (see Fraser 1996-1998a and 1996-1998b), which are word-based. Another drawback of HC is that some periods are under-represented, as observed in the data in Table 1.

2.2. The Parsed Corpora of Historical English

Kohnen (2007) has reflected on the need to change from 'long and thin' corpora, as Rissanen called them (2000: 9), to 'short and fat' ones. In other words, a gradual shift from corpora extending over a long period of time with few texts in each period, to corpora focusing on particular timespans and provided with more information is required. This is the reason why several projects have emerged to parse corpora belonging to different stages of development of the English language. They not only allow for searching for words, but also for syntactic structures. The texts come under three different formats: a) plain text; b) POS-tagged; and c) parsed.[6]

CorpusSearch, developed by Beth Randall (2000), is the software tool employed to retrieve data from most of these corpora.[7]

2.2.1. Penn Parsed Corpus of Middle English (PPCME2)[8]

The original version of the *Penn Parsed Corpus of Middle English* (PPCME) amounted to roughly half a million words and contained quite a simple annotation scheme because, for instance, there was no POS-tagging. It was prepared by Prof. Anthony Kroch and Dr. Ann Taylor, and so has been the second version (PPCME2; 2000), which is POS-tagged and for the most part based on the ME section of HC, although some additions and deletions have been made, the size of the samples being slightly larger.

2.2.2. York-Helsinki Parsed Corpus of Old English Poetry[9]

This corpus has been compiled with OE poetry texts, resulting in almost 72,000 words (Pintzuk & Plug 2001). It is syntactically and morphologically annotated, and the annotation scheme (developed by Prof. Susan Pintzuk, Dr. Ann Taylor, Prof. Anthony Warner, Dr. Leendert Plug and Frank Beths) is based on that of PPCME2. The materials for the corpus were taken from HC.

2.2.3. York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE)[10]

This project, based at the University of York, has been undertaken by Dr. Ann Taylor, Prof. Anthony Warner, Prof. Susan Pintzuk and Frank Beths (2003). It contains 1.5 million running words that have been parsed following the same scheme as PPCME2. It is largely based on the Toronto *Dictionary of Old English Corpus* (Healey *et al*. 2009).

This corpus has superseded the previous *Brooklyn-Geneva-Amsterdam-Helsinki Parsed Corpus of Old English* (Pintzuk *et al*. 2000)[11], which was drawn from the OE section of HC and whose texts came under four different formats, to be used with different search tools.

2.2.4. Penn Parsed Corpus of Early Modern English **(PPCEME)**[12]

Compiled by Prof. Anthony Kroch, Dr. Beatrice Santorini and Ariel Diertani between 1999 and 2004, PPCEME includes texts totalling almost 1.8 million words. It is made up of three subcorpora: one contains the Helsinki directories, and the other two comprise additional data, gathered from material by the same authors and the same editions as the HC directories.

2.2.5. Penn Parsed Corpus of Modern British English (PPCMBE)[13]

Released in 2010, this is the result of a joint project between the Universities of York and Pennsylvania, led by Prof. Anthony Kroch. Around one million words have been compiled following the genre structure of PPCEME.

2.3. The *Corpus of Early English Correspondence* (CEEC) family of corpora[14]

In his review of historical corpora, Kohnen (2007) highlights that CEEC can be taken as one of the "innovative solutions to the problems created by defective historical data". This ongoing project, started by Prof. Nevalainen (University of Helsinki) in 1993, came into being with the aim of testing the applicability of sociolinguistic methods to historical data. For this purpose, the selection of material to be compiled was made on the basis of the following criteria: a) their similar size; b) the availability of background information about the informants; c) their representing private writing (which is usually closer to spoken language); d) their easy access; and e) their allowing for diachronic comparisons. Accordingly, the corpus is mostly based on personal letters (besides other text-types, such as official and business letters), which permit the evaluation of diastratic variation, since details of the informants' social background are also known.

The project is structured into five daughter projects: the original corpus (Nevalainen *et al*. 1998), the *Sampler* (Keränen *et al*. 1998), the *Parsed Corpus* (Nevalainen *et al*. 2006, Nurmi *et al*. 2006 & Taylor *et al*. 2006), the 18th-century *Extension* and the *Supplement* (Kaislaniemi *et al*. forthcoming a and b), which will cover the period 1400-1800 once they are completed. The *Parsed Corpus* also belongs to the project of parsed corpora dealt with in 2.2. and, thus, uses the same annotation scheme.[15] As for text-coding, it employs the same parameters as HC. The remaining two products (i.e. the *Extension* and the *Supplement*) were started in 2000 (Kaislaniemi 2006) and are nearing completion, as illustrated in Table 2[16]:

Project	**CEEC**	**CEECS**	**PCEEC**	**CEECE**	**CEECSU**
Dates	(?)1410-1681	1418-1680	1410-1681	1681-1800	1410-1681
Number of running words	ca. 2.6 million	ca. 450,000	ca. 2 million	ca. 2.2 million	ca. 440,000
Number of informants	778	194	666	310	94
Additional information	1. Information on social background kept in a separate database 2. Plain-text format	Two parts: 1. CEECS1: 15th-16th centuries 2. CEECS2: 17th century	Three format files: 1. Plain text 2. POS-tagged 3. Parsed		Additional material

Table 2. Main data about the CEEC daughter projects

The problems encountered in the compilation of this corpus are different from those found by the HC compilers. On the one hand, finding letters written by the lower social strata has been a difficult task due to the socially stratified patterns of literacy. On the other, the sources have also been somewhat problematic, since not all editions have been considered suitable for the corpus. Consequently, only those preserving the original manuscript spelling have been used.

2.4. Corpus of *Early English Medical Writing* (CEEM)[17]

This corpus, subdivided diachronically into three smaller corpora, aims at compiling medical texts from 1375 to 1800. Hence, this is another genre-based corpus like CEEC (see 2.3). It is being compiled at the University of Helsinki by the team led by Prof. Taavitsainen (University of Helsinki) and Prof. Pahta (University of Tampere), and estimations have been made to reach 3.75 million words.

This far, only the ME and eModE subcorpora have been released. The first one (MEMT; Taavitsainen *et al*. 2005) contains almost half a million running words coming from 86 texts. Short texts are included in full, while extracts of 10,000 words from longer ones have been selected. They are mostly taken from editions, which have been processed so that editorial intervention (visible in explanatory notes, variant readings in parallel manuscripts, etc.) has been excluded. When coding the texts, expansions of abbreviations are not italicised, although line and paragraph division have been preserved. Likewise, emendations and modernised punctuation are duly marked.

The eModE subcorpus (EMEMT; Taavitsainen *et al.* 2010) comprises 2 million words drawn from around 450 texts from the period 1500 to 1700. The criteria taken into account for its compilation have been the availability of texts and the representativeness of all subperiods and genres (namely, general treatises, treatises on specific topics, recipe collections, health guides, surgical and anatomical treatises, and scientific journals). The size of the extracts is exactly the same as that established for MEMT.

It is envisaged that the late ModE (lModE) subcorpus (LMEMT) will be released by 2013 and will include around 2 million words drawn from a still unspecified number of texts from the 18th century.

Specific software tools, named *MEMT Presenter* and *EMEMT Presenter* (for the ME and eModE periods, respectively), have been designed for the retrieval of information. These are adaptations of *Corpus Presenter* (2010), a software tool developed by Raymond Hickey that allows extracting concordances from a word list, for instance, although it only enables word-based searches.[18]

2.5. *Middle English Grammar Project Corpus* (MEG-C)[19]

The *Middle English Grammar Project*, led by Dr. Merja Stenroos (University of Stavanger), intends to produce a full account of ME usage for the period 1100-1500 by using a lemmatised corpus of localised texts, which consists of 664,543 tokens. Therefore, this corpus is not an end in itself, but rather the means for the production of a ME grammar. The levels of orthography and phonology will be first accounted for, followed by the morphological one.

The input used for its compilation are the localised texts used in the *Linguistic Atlas of Late Mediaeval English* (LALME; McIntosh *et al*. 1986), as well as those in its early ME (eME) counterpart (LAEME; Laing & Lass 2007). Some drawbacks regarding this corpus may also be mentioned. First, LALME does not include a sample of all late ME (lME) texts and so leaves aside "non-localizable language varieties" (Stenroos 2007). Second, texts localised in Scotland in LALME have been excluded from MEG-C. Third (and concerning the method of data collection), a machine-readable corpus has been preferred over manual counting, as opposed to LALME's method. It may be added that in order to study syntax or morphology, the translated texts used in LALME may not be useful either, insofar as scribes did not usually 'translate' structures or morphology.

As a result, as for their requirements, the texts selected cover most linguistic features with medium to high frequency (be them orthographic, morphologic or syntactic). This has been further constrained by factors such as the legibility of the manuscripts where they are found or the availability of certain texts in microfilm format. Short texts have been included *in toto*, while 3,000-word samples from longer texts have been chosen, a size considered sufficient to show the usage of the features under study. This has been done with a view to achieving "as complete a coverage as possible of as much material as possible" (Stenroos 2007).

Annotation deals with information below word level (e.g. morphemes, and textual and manuscript context), since orthographic and morphemic concerns are of paramount importance in this corpus (Stenroos, 2007). Lemmatisation, in turn, has been drawn from the *Oxford English Dictionary* (OED; Simpson & Weiner 2004), which is an authoritative source.

The downside is that the samples being relatively small, the corpus will "not be able to provide a full basis for the study of each individual text […]". Likewise, "it will also be of limited use for the study of features with a low density of occurrence" (Stenroos 2007). Moreover, some periods are under-represented, with few texts available (as with HC).

2.6. *Corpus of Middle English Prose and Verse*[20]

As part of the *Middle English Compendium* (McSparran *et al*. 2001-), this collection of ME texts was assembled by the HTI (Humanities Text Initiative) from a variety of sources with the intent of including all the editions used for the *Middle English Dictionary* (MED),[21] together with other more recent scholarly editions.

The main advantage of this corpus is its availability online, as well as the possibility of carrying out different types of searches: simple (which look for words or phrases), proximity (which look for the co-occurrence of several words or phrases), Boolean (which find combinations of words in a section) and citation/bibliographic searches (which identify works by author and title). There is also an option called "bookbagging", which permits restricting a search to a finite set of works. Proximity searches may be of particular interest, since they allow the user to establish the span within which the search is to be performed.

2.7. A Representative Corpus of *Historical English Registers* (ARCHER)[22]

This diachronic corpus is one of the principal collections of historical English that comprehend an ample stretch of time and genres (Lee 2010: 113). It is a multi-genre corpus created to complement the diachronic part of HC. At the moment, it consists of 1.7 million words and approximately 1,000 texts, although it is undergoing expansion: a new enhanced and updated version (3.2) is currently under preparation. It encompasses the eModE period (both of British and American English) up to the present (1650-1990) with the inclusion of different text-types, hence allowing for the analysis of historical change in written and speech-based registers (Biber *et al*. 1994: 3). The original corpus, compiled in the early 1990s by Prof. Douglas Biber (Northern Arizona University) and Prof. Edward Finegan (University of Southern California) will be augmented to the first half of the seventeenth century, incorporating legal British and American texts, as well as texts related to the field of advertising. At present, ARCHER can only be consulted *in situ* at one of the fourteen departments belonging to the ARCHER Consortium.[23] Further information about this corpus is supplied in section 3.5.

2.8. Corpus of *Historical American English* (COHA)[24]

Another major corpus of historical English is COHA (Davies 2010), which contains more than 400 million words from American English texts spanning from 1810 to 2009. According to Lee (2010: 113), it "is 'balanced' in each decade for the genres of fiction, popular magazines, newspapers and academic prose". It has in common with ARCHER their being multi-purpose (broad focus) and their extension into the immediate present.

2.9. Corpus of *Late Modern English Texts* (CLMET)[25]

Compiled at the University of Leuven by Dr. Hendrik De Smet, CLMET is a collection of texts drawn from the *Project Gutenberg*[26] and the *Oxford Text Archive*.[27] It contains some 10 million words of running text, divided over three 70-year subperiods, from 1710 to 1920. There is also an extended 15-million-word version with additional texts from the above-mentioned sources and from the *Victorian Women Writers Project*.[28]

The texts included range from personal letters to literary fiction and scientific writing, written by both men and women who belong to different social classes.

2.10. Corpus of *English Novels*[29]

Also compiled at the University of Leuven by Dr. Hendrik De Smet, as CLMET, it contains some 25 million words of late-19th- and early-20th-century English drawn from the *Project Gutenberg*. It consists entirely of British and North American novels, written by 25 novelists between 1881 and 1922.

2.11. Corpus of *Late 18th-Century Prose*[30]

This project was directed by Prof. David Denison (University of Manchester) in collaboration with Dr. Linda van Vergen (University of Edinburgh) and Dr. Joana Soliva (University of Manchester). The corpus is a compilation of letters from the *Leghs of Lyme Collection*. The timespan covered is 1761-1790 and the size of the corpus is about 300,000 words. The material includes letters from educated and uneducated writers, and examples from plain business English to heavily dialectally-marked texts can be found.

2.12. *Leuven English Old to New Corpus* (LEON)[31]

The LEON Corpus is being compiled by Dr. Peter Petré, a member of the research group *Functional Linguistics Leuven: Grammar, Diachrony, Typology* at the University of Leuven. The idea behind this project is to put together all the English texts available in different corpora, including a 400,000-word corpus for each HC period, and after 1710 for the periods 1710-1780, 1780-1850, 1850-1920, 1920-1990 and post-1990. The compiler's intention is to present it without tagging but offering the possibility of external manipulation.[32] After facing some problems with copyright issues, his immediate aim is to design a tool which would allow researchers to access the material only if they can prove that they own an original copy of the different corpora used as sources for the compilation of his corpus.

Dr. Petré is also the compiler of the *Corpus of Middle English Quotations*, which is an ordered list of all quotations in MED in a single text file. These two corpora are only accessible to Leuven staff.

2.13. Other corpora

Space constraints make it impossible to describe all the historical corpora available, but a few more need to be mentioned. One of them is ICAMET (*Innsbruck Computer Archive of Machine-Readable English Texts*; Markus 1992-), which is comprised of the *Middle English Prose Corpus*, the *Letter Corpus* and the *Varia Corpus*, all amounting to more than 6 million words and providing complete versions of the texts. Another one is the *Corpus of English Religious Prose* (COERP; Kohnen *et al*. forthcoming).

As regards those useful for research in ModE, we find the *Lampeter Corpus of Early Modern English Tracts* (LC; Schmied *et al*. 1999), containing non-literary prose texts (pamphlets) and subdivided into six domains (ca. one million words); the *Corpus of English Dialogues 1570-1760* (CED; Kÿto & Culpeper 2006); and the *Zurich English Newspaper Corpus* (ZEN; Fries *et al*. 2004). The latter, as its name suggests, gathers most types of texts from the emerging newspapers genre (1.2 million words). It is also worth referring to the *Corpus of Nineteenth-Century English* (CONCE; Kytö & Rudanko forthcoming), a multi-genre corpus composed of one million words, as it addresses a rather neglected period in historical corpus linguistics.

3. CORPORA COMPILED IN SPAIN

In the following subsections, several English historical corpora compiled at some Spanish universities are described. For the sake of clarity, and the number of corpora being not as high as that of those compiled abroad, this section is organised according to the universities leading the projects.[33]

3.1. University of Coruña: *Coruña Corpus of English Scientific Writing* (CC)[34]

Dr. Isabel Moskowich-Spiegel is the main researcher of the team MUSTE (*Multidimensional Corpus-Based Studies in English*), whose members are working on the compilation of a corpus of English scientific writing from 1700 to 1900. CC is intended to complement other corpora (such as ARCHER, MEMT and HC) that share some characteristics with it, particularly its diachronic nature and the specificity of the samples contained. In addition, CC includes periods not covered individually by these other corpora. A summary of the disciplines and subcorpora in CC is presented in Table 3:

Field	UNESCO disciplines	CC Discipline	Subcorpora
Natural Sciences	Astronomy	Astronomy	CETA
	Biology, Botanics, Zoology, Horticulture, Veterinary, Medicine	Life Sciences	CELiST
	Physics	Physics	CETePH
	Chemistry, Biochemistry	Chemistry	CECheT
	Mathematics	Mathematics	CEMaT
Humanities	Philosophy, History of science and technology	Philosophy	CEPHiT
	History, Archaeology, Numismatics, Palaeography, Genealogy	History	CHET
	Modern languages	Linguistics	CETeL

Table 3. UNESCO/CC disciplines and subcorpora

Astronomy was the first discipline selected for the compilation of scientific texts and, consequently, the first subcorpus is the *Corpus of English Texts on Astronomy* (CETA). The texts contained here, together with information about the authors' lives, provide a rich perspective on the increasing degree of the institutionalisation of science during the period. CETA, which includes samples of texts on modern astronomy together with one metadata file per sample, is already available contacting Dr. Moskowich.[35]

Two 10,000-word text files have been compiled per decade, with the two centuries represented containing approximately 200,000 words each. Thus, this first subcorpus in CC comprehends 400,000 words, more or less equally distributed. The team is now working on the subcorpora dealing with Philosophy, Chemistry, Life Sciences, Linguistics and Mathematics, as well as on the completion of their research tool, the *Coruña Corpus Tool*, which will enable the user of CC to retrieve information easily.

3.2. University of Las Palmas de Gran Canaria: *Corpus of Early English Recipes* (CoER)[36]

The research group *Tecnologías Emergentes Aplicadas a la Lengua y Literatura* (TeLL), with Dr. Francisco Alonso-Almeida as its main researcher, is working on the compilation of CoER, a collection of recipes written in English from 1375 to 1750, with a future addition of recipes from 1750 to 1850. Some 1.5 million words have been compiled by now, mostly from eModE medical, gardening and culinary texts, out of the 3 million words the corpus will offer at the end of the process.

The classification of the recipes is carried out chronologically, taking also into account the topic dealt with. The topical organisation is shown in Figure 1.

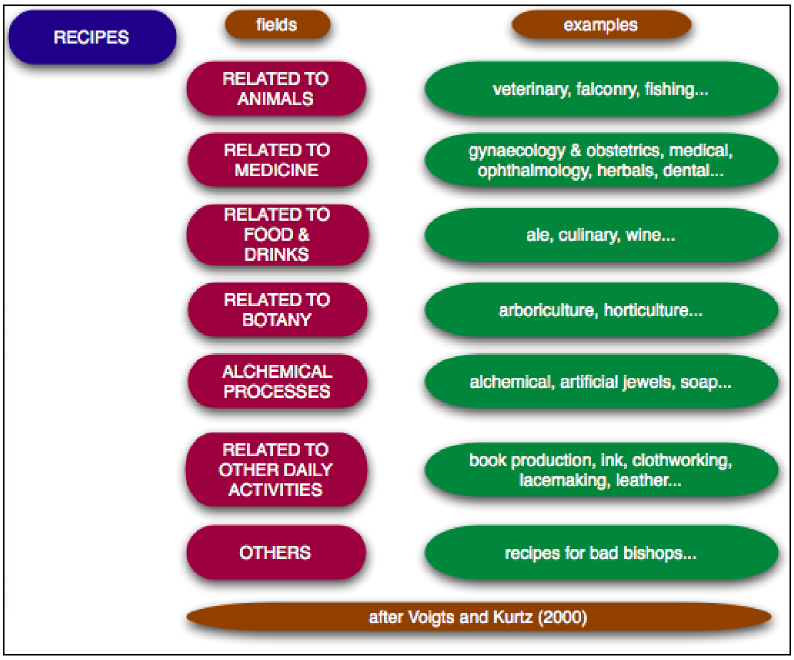
RECIPES
fields
examples
RELATED TO ANIMALS
veterinary, falconry, fishing...
RELATED TO MEDICINE
gynaecology & obstetrics, medical, ophthalmology, herbals, dental...
RELATED TO FOOD & DRINKS
ale, culinary, wine...
RELATED TO BOTANY
arboriculture, horticulture...
ALCHEMICAL PROCESSES
alchemical, artificial jewels, soap...
RELATED TO OTHER DAILY ACTIVITIES
book production, ink, clothworking, lacemaking, leather...
OTHERS
recipes for bad bishops...
after Voigts and Kurtz (2000)

Figure 1. Topical organisation in CoER

Besides the text in files, a header presents information regarding documentation (bibliographic information of the text in question), file length and word count. Encoding is kept to a minimum, and only for the purpose of indicating the use of marginal notes, the use of languages other than English, the beginning and end of complementary text (text not labelled as recipes), page numbers and non-keyable characters. All this information will be encoded by specific use of XML metalanguage. CoER will not include morphological tagging. It is designed to be used in the *Corpus Presenter* suite platform (Hickey 2000; 2010), although other software tools might be used to retrieve data.

At the moment, the group is preparing a computing tool to be used online to produce searches and generate word lists besides many other retrieval options. The tool is called OnICoM (*Online Interface for Corpus Management*) and it will be freely available on demand. This interface will host other compilations, and so the user will have to select the corpus to be used before enquiries start.

3.3. University of Málaga: *Corpus of Late Middle English Scientific Prose*[37]

The *Corpus of Late Middle English Scientific Prose* is a research project, currently in progress, undertaken at the University of Málaga in collaboration with the Universities of Murcia, Oviedo, Jaén and Glasgow. Led by Dr. Antonio Miranda-García and part of the work carried out by the research group *Cambio Lingüístico y Edición Filológica de Textos* (CLEFT), it was conceived as a corpus of unedited scientific, primarily medical, 15th-century English texts from the Hunterian Collection at Glasgow University Library. However, it has been extended so as to include texts from other collections (in particular, the Wellcome Library). At present, its size amounts to ca. 1.2 million tokens.

The method of compilation involves several stages, which include: a) transcription of the texts, using the manuscripts themselves or their digitised images; b) lemmatisation;[38] c) annotation, involving POS tagging, reference to folio and line, and meaning. The main advantage of the resulting lemmatised and annotated corpus is the possibility of retrieving information automatically, thus allowing the user to draw word and lemma lists, to generate concordances or to build up glossaries.

In conjunction with the corpus, a software application, *Text Search Engine* (TexSEn) (Miranda-García & Garrido-Garrido forthcoming), has been specifically designed for the extraction of morphosyntactic, lexical and statistical data. The application enables, for instance, both Boolean and non-Boolean searches and, as just mentioned, retrieving concordances and building customised glossaries, which may be yielded in several formats (Excel spreadsheets, plain text files, eXtensible Markup Language or Rich Text Format).

3.4. University of Salamanca: *Salamanca Corpus* (SC)[39]

Some years ago, a group of researchers from this university, led by Dr. García-Bermejo Giner, started the compilation of SC. This includes texts representing the dialectal speech of the different counties in England from the 1500s to the 1950s. Most of these texts have not been re-edited since their publication, nor have they appeared in other corpora dealing with this timespan.

To date, the group DING (*Dialectología Inglesa y Diacronía Inglesa*) has digitised and transcribed some 400 novels and dialogues and they are now working on the drama and poetry sections. All the texts are available in Word format to allow for different types of computer-assisted work. They are organised in terms of the type of dialect representation, namely Literary Dialect and Dialect Literature;[40] a further classification includes the period they were written in, the county which they are ascribed to, and their genre and authorship.

3.5. University of Santiago de Compostela: COLMOBAENG and CHELAR[41]

The research group *Variation, Linguistic Change and Grammaticalization* (VLCG) has been working on three different projects, one of them completed and the other two under way. On the one hand, the finished work comprises the corpus known as COLMOBAENG (*Corpus of Late Modern British and American Prose*), compiled by Prof. Teresa Fanego (VLGC's main researcher), which is available contacting the compiler via email. On the other, the team has additional projects in progress, one related to the ARCHER project (see section 2.7) and the other to the compilation of the *Corpus of Historical English Law Reports* (CHELAR).

COLMOBAENG (1700-1879) is a 1.17-million-word database comprising texts taken from printed and electronic sources. Among the printed texts, works such as those by Baym (2003) and Poirier (1990) are found, whereas electronic texts are drawn from the *Century of Prose Corpus*, the *Project Gutenberg* and the Electronic Text Center at the University of Virginia,[42] among others. The structure of the corpus is shown in the following table:

British English	American English
BrE1 1700-1726 (200,000 words)	–
BrE2 1732-1757 (200,000 words)	AmE2 1732-1759 (50,000 words)
BrE3 1761-1797 (200,000 words)	AmE3 1774-1804 (120,000 words)
BrE4 1850-1879 (200,000 words)	AmE4 1851-1879 (200,000 words)

Table 4. Dialects, periods and number of words in COLMOBAENG

This corpus has a number of words similar to traditional historical corpora such as HC (although the eModE section comprises half a million words), CONCE (one million 19th-century English words) and ARCHER (1.7 million words from the 1650s to 1990). Thus, COLMOBAENG can be regarded as a useful tool to supplement these corpora.

Concerning ARCHER, the contribution of the VLGC team will include law reports drawn from the Incorporated Council of Legal Reporting for England and Wales, through Justis Publishing Limited,[43] an online resource for legal texts. When finished, this subcorpus of British English legal texts will comprise some 160,000 words of running text.

As for CHELAR,[44] the texts selected by this research unit include British English law reports. These records of judicial decisions, which explain the main facts of a case, are cited by lawyers and judges to be used as precedent in subsequent cases. The timespan covered is from 1500 to 2000 and the corpus will contain about half a million words once completed.

3.6. University of Sevilla: *Seville Corpus of Northern English* (SCONE)[45]

The team led by Dr. Julia Fernández-Cuesta and Dr. Gabriel Amores, in cooperation with the National University of Distance Education in Madrid (UNED) and the University of Westminster, London, are compiling an electronic corpus of texts written in Northern English (SCONE). The timespan covered goes from the 7th/8th to the 16th centuries, and their main aim is to produce an e-corpus which contains both the edition of the manuscripts and information about the language at different linguistic levels, including spelling/phonology, morphosyntax and lexis. All the extant texts from Old Northumbrian and the majority of texts from eME have been annotated and included in the database, as well as some legal texts from eModE. Although the main purpose is to highlight the features that characterise these texts as northern, the interface will also allow users to perform searches which are not necessarily focused on dialectal variation.

Currently, the texts are available in TEI-conformant XML and HTML[46]. A search engine allows users to look up texts which contain words annotated with a set of pre-defined diagnostic features. Moreover, annotated words in the text provide information as regards their dialectal provenance, some information on their orthographical, phonological and morphological features, and a translation into Present-Day English (PDE).

4. FINAL REMARKS

The potential that historical corpora offer nowadays for research is not only limited to morphosyntactic questions, but also allows for investigating on pragmatic, cultural, historical or sociolinguistic aspects. However, there are several drawbacks affecting this specific type of corpora, which may not be present in contemporary ones. Firstly, material is limited when compared to that existing for PDE, since availability depends on what is extant or has survived. Secondly, historical linguists have had to depend on canonical sources for their searches, such as older philological studies or historical dictionaries (Curzan & Palmer 2006: 23). This situation seems to be changing in the light of the new corpora being compiled (scientific and technical texts, official records, letters, etc.), as described in the previous sections.

The reliability of editions as input material may represent another problem, as editors generally follow different methods or conventions, which may be suitable for certain studies but not for others. Thus, those editions in which spelling variants have been standardised may hinder dialectal studies. Editions can also contain transcription mistakes, lack of homogenisation, etc. In this regard, Lass (2004: 31) has argued that they cannot be readily trusted given that "an emended text is a falsehood, if as so often happens it's taken unreflectively as a witness for a past language state". As a consequence, "the ideal model for a corpus or any presentation of a historical text is an archaeological site or a crime-scene: no contamination, explicit stratigraphy, and an immaculately preserved chain of custody" (Lass 2004: 46). Nonetheless, editions also have some advantages, such as the fact that texts are easily available, thus saving time for researchers to focus on other steps of their work.

Another limitation for historical corpora is the set of sources available, which are only written and restricted, to a greater or lesser extent, depending on the period. In OE, for example, materials are basically reduced to glosses of Latin texts, written records and the annals collected in the Anglo-Saxon Chronicle. Religious works predominate in ME, although the range of genres (literary and non-literary, religious and secular) widens from 1350 onwards. EModE sees a broadening of the spectrum with the increase of private writing; there are also trial records, secular genres, drama and popular literature (Taavitsainen *et al*. 2008).

In spite of these constraints, the spread and richness of historical corpora, and the possibility of using and exploiting them, bears witness to the fact that a bright future for research lies ahead.

ACKNOWLEDGEMENTS

The present research has been funded by the Spanish Ministry of Science and Innovation (FFI2008-02336/FILO, FFI2011-26693-C02-01 and HUM2007-60706/FILO "CONSOLIDER") and by the Autonomous Governments of Andalusia (Proyecto de Excelencia P07-HUM-02609) and Murcia (0862929/PHCS/08). These grants are hereby gratefully acknowledged.

NOTES

[1] However, it must be mentioned that at some point they may become static when no "new" material (supplied by, for instance, transcribing and electronically editing/digitising manuscripts, records, letters, etc.) may be added; finiteness specially applies to Old English. Yet, the amount of unedited works in Middle English and Modern English found in collections and libraries bodes well for the enhancement of historical corpora. As Claridge (2008: 248) remarks, "the selection of written texts is broad enough for most periods to construct varied corpora".

[2] The data offered about these corpora have been compared to that appearing at http://www.helsinki.fi/varieng/CoRD/corpora/index.html.

[3] A detailed introduction to this corpus can be found in Rissanen (2005) and in the manual of the corpus (Kytö 1996). See also http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/index.html.
[4] This table has been adapted from http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/period.html.
[5] These parameters are introduced in the COCOA format (http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/generalintro.html; Kytö 1996: 40-56), which helps to set the limits on the scope of the searches to be performed.
[6] This information is available at http://www.ling.upenn.edu/hist-corpora/annotation/index.html.
[7] This software tool can be accessed at http://corpussearch.sourceforge.net/CS.html.
[8] Further information can be found at http://www.ling.upenn.edu/ hist-corpora/PPCME2-RELEASE-3/index.html and http://www.helsinki.fi/varieng/CoRD/corpora/PPCME2/index.html.
[9] Further information available at http://www-users.york.ac.uk/~lang18/pcorpus.html.
[10] Further information available at http://www-users.york.ac.uk/~lang22/YcoeHome1.htm and http://www.helsinki.fi/varieng/CoRD/corpora/YCOE/index.html.
[11] Further information available at http://www-users.york.ac.uk/~sp20/corpus.html.
[12] Further information available at http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-2/index.html and http://www.helsinki.fi/varieng/CoRD/corpora/PPCEME/index.html.
[13] Further information available at http://www.ling.upenn.edu/hist-corpora/PPCMBE-RELEASE-1/index.html and http://www.helsinki.fi/varieng/CoRD/corpora/PPCMBE/index.html.
[14] Further information available at http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/index.html; http://www.helsinki.fi/varieng/domains/CEEC.html.
[15] Further information available at http://www-users.york.ac.uk/~lang22/PCEEC-manual/index.htm.
[16] The latter (CEECSU) incorporates various types of materials, which partially complement those already present in CEEC (Kaislaniemi 2006).
[17] Further information available at http://www.helsinki.fi/varieng/CoRD/corpora/CEEM/index.html.
[18] *Corpus Presenter* can be accessed at http://www.uni-due.de/CP/.
[19] Further information available at http://www.uis.no/research/culture/the_middle_english_grammar_project/meg-c/ and http://www.helsinki.fi/varieng/CoRD/corpora/MEG-C/index.html.
[20] Available at http://quod.lib.umich.edu/c/cme/.
[21] Available at http://quod.lib.umich.edu/m/med.
[22] Further information available at http://www.llc.manchester.ac.uk/research/projects/archer/.
[23] These are the following: Department of English, Northern Arizona University; Department of Linguistics, University of Southern California; Englisches Seminar, Albert-Ludwigs-Universität Freiburg; Anglistisches Seminar, Ruprecht-Karls-Universität Heidelberg; Department of English, University of Helsinki; Department of English, Uppsala University; Department of English, University of Michigan; Department of Linguistics and English Language, University of Manchester; Department of Linguistics and English Language, Lancaster University; Lehrstuhl für Englische Sprachwissenschaft einschließlich Sprachgeschichte, Otto-Friedrich-Universität Bamberg; Englisches Seminar, Universität Zürich; Fachbereich Anglistik, Universität Trier; School of English, Sociology, Politics & Contemporary History, University of Salford; and Research Unit on Variation, Linguistic Change and Grammaticalization, Departamento de Filología Inglesa y Alemana, Universidad de Santiago de Compostela.
[24] Available at http://corpus.byu.edu/coha/.
[25] Available at https://perswww.kuleuven.be/~u0044428/clmet.htm.
[26] Available at http://www.gutenberg.org/wiki/Main_Page.
[27] Available at http://ota.ahds.ac.uk/.
[28] Available at http://webapp1.dlib.indiana.edu/vwwp/welcome.do.
[29] Available at https://perswww.kuleuven.be/~u0044428/cen.htm.
[30] Available at http://www.llc.manchester.ac.uk/subjects/lel/staff/david-denison/corpus-late-18th-century-prose/.
[31] Further information can be found at http://wwwling.arts.kuleuven.be/fll/ppetre/.
[32] The compiler describes it as a kind of ‘wiki corpus’ (Petré 2009).
[33] Further information on these particular corpora can be found in Vázquez (Forthcoming).
[34] Available at http://www.udc.es/grupos/muste/corunacorpus/index.html.

[35] It will be published in the course of 2011 in CD-ROM format, together with a volume edited by Moskowich & Crespo (2011).

[36] Further information available at http://www.gi.ulpgc.es/tell/.

[37] Further information available at http://hunter.uma.es/.

[38] The main headword in MED has been taken as the lemma.

[39] Further information available at http://salamancacorpus.usal.es/SC/index.html.

[40] Literary Dialect refers to the representation of vernacular speech in novels, poems and plays written in the standard language, whereas Dialect Literature is that wholly composed using a particular variety. Thus, we can speak, on the one hand, of the Midland Literary Dialect used by some of George Eliot's characters and, on the other, of the rich 19th-century Dialect Literature of Yorkshire, represented, for instance, by the poetry of John Castillo (Dr. García-Bermejo Giner, personal correspondence, 2011).

[41] Further information available at http://www.usc-vlcg.es/team.htm.

[42] Further information available at http://www2.lib.virginia.edu/etext/history.html

[43] Available at http://www.justis.com.

[44] A detailed description of CHELAR is found in Rodríguez-Puente (Forthcoming).

[45] A full account of the compilation process can be found in Fernández-Cuesta *et al*. (Forthcoming). Further information available http://www.helsinki.fi/varieng/CoRD/corpora/SCONE/index.html.

[46] They are also stored in a database. Future versions of the corpus will furnish the possibility of obtaining concordances and dictionary entries from the texts.

REFERENCES

Baym, N. (Ed.). (2003). *The Norton Anthology of American Literature* (Vol. I: *Literature to 1820*). New York: W.W. Norton & Company.

Biber, D., Finegan, E., Atkinson, D., Beck, A., Burges, D. & Burges, J. (1994). The Design and Analysis of the ARCHER Corpus: A Progress Report. In M. Kytö, M. Rissanen and S. Wright, *Corpora across the Centuries: Proceedings of the First International Colloquium on English Diachronic Corpora* (pp. 3-6). Amsterdam: Rodopi.

Claridge, C. (2008). Historical Corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook* (Vol. 1) (pp. 242-259). Berlin, New York: Walter de Gruyter.

Curzan, A. & Palmer, C. C. (2006). The Importance of Historical Corpora, Reliability, and Reading. In R. Facchinetti & M. Rissanen (Eds.), *Corpus-based Studies of Diachronic English* (pp. 17-34). Frankfurt am Main, etc.: Peter Lang.

Davies, M. (Comp.). (2010). *COHA: Corpus of Historical American English*. Available at http://corpus.byu. edu/coha/

De Smet, Hendrik. (2005). A Corpus of Late Modern English Texts. *ICAME Journal, 29*, 69-82.

Fernández Cuesta, J., García García, L. & Amores, G. (Forthcoming). Compilation of an Electronic Corpus of Northern English Texts from Old to Early Modern English. In R. Dance & L. Wright (Eds.), *Current Concerns in Middle English Language and Literature.* Bern: Peter Lang.

Fraser, M. (1996-1998a). *Oxford Concordance Programme. Guide to Digital Resources*. Available at http://users.ox.ac.uk/~ctitext2/resguide/resources/o125.html

Fraser, M. (1996-1998b). *WordCruncher. Electronic Text Viewer. Version 7.1*. Available at http://www.wordcruncher.com/wordcruncher/default.htm

Fries, U., Lehmann, H. M., Ruef, B., Schnieder, P., Studer, P., auf dem Keller, C., Nietlispach, B., Engler, S., Hensel, S. & Zeller, F. (Comps.). (2004). *ZEN: Zurich English Newspaper Corpus. Version 1.0*. Available at http://es-zen.unizh.ch. Zurich: University of Zurich.

Healey, A. diPaolo, Holland, J., McDougall, I, McDougall, D. (Comps.). (2009). *Dictionary of Old English Corpus*. Toronto: University of Toronto.

Hickey, R. (2000). Processing Corpora with *Corpus Presenter*. *ICAME Journal*, *24*, 65-84.

Hickey, R. (2010). *Corpus Presenter*. (12th ed.). Available at http://www.uni-due.de/CP/

Kaislaniemi, S. (2006). The Corpus of Early English Correspondence. Extension and Supplement. Poster presented at the 27th ICAME (International Computer Archive of Modern and Medieval English) Conference. University of Birmingham, United Kingdom, May 24-28.

Kaislaniemi, S., Laitinen, M., Nevala, M., Nevalainen, T., Nurmi, A., Palander-Collin, M., Raumolin-Brunberg, H. & Sairio, A. (Comps.). (Forthcoming a). *Corpus of Early English Correspondence. Extension*. Department of English, University of Helsinki.

Kaislaniemi, S., Laitinen, M., Nevala, M., Nevalainen, T., Nurmi, A., Palander-Collin, M., Raumolin-Brunberg, H. & Sairio, A. (Comps.). (Forthcoming b). *Corpus of Early English Correspondence. Supplement*. Department of English, University of Helsinki.

Keränen, J., Nevala, M., Nevalainen, T., Nurmi, A., Palander-Collin, M. & Raumolin-Brunberg, H. (Comps.). (1998). *Corpus of Early English Correspondence. Sampler*. Department of English, University of Helsinki.

Kohnen, T. (2007). From Helsinki through the Centuries: The Design and Development of English Diachronic Corpora. In P. Pahta, I. Taavitsainen, T. Nevalainen & J. Tÿrkko (Eds.), *Studies in Variation, Contacts and Change in English* (Vol. 2). Retrieved from http://www.helsinki.fi/varieng/journal/volumes/02/kohnen/

Kohnen, T., Rütten, T., Marcoe, I., Gather, K. & Groeger, D. (Comps.). (Forthcoming). *Corpus of Early English Religious Prose*. Cologne: University of Cologne.

Kroch, A., Santorini, B. & Diertani, A. (Comps.). (2004). *Penn-Helsinki Parsed Corpus of Early Modern English*. Available at http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-2/index.html

Kroch, A., Santorini, B. & Diertani, A. (Comps.). (2010). *Penn Parsed Corpus of Modern British English*. Available at http://www.ling.upenn.edu/hist-corpora/PPCMBE-RELEASE-1/index.html

Kroch, A. & Taylor, A. (Comps.). (2000). *Penn-Helsinki Parsed Corpus of Middle English. Second edition*. Available at http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-3/index.html

Kÿto, M. (1996). *Manual to the Diachronic Part of the Helsinki Corpus of English Texts: Coding Conventions and Lists of Source Texts*. (3rd ed.). Helsinki: Helsinki University Printing House.

Kÿto, M. & Culpeper, J. (Comps.). (2006). *A Corpus of English Dialogues 1560-1760*. Uppsala: University of Uppsala.

Kytö, M. & Rudanko, J. (Comps.). (Forthcoming). *CONCE: A Corpus of Nineteenth-Century English*. Uppsala University and University of Tampere, Finland.

Laing, M. & Lass, R. (Comps.). (2007). *A Linguistic Atlas of Early Middle English, 1150-1325. Version 2.1*. Edinburgh: University of Edinburgh. Available at http://www.lel.ed.ac.uk/ihd/laeme1/ laeme1.html

Lass, R. (2004). Ut custodiant litteras: Editions, Corpora and Witnesshood. In M. Dossena & R. Lass (Eds.), *Methods and Data in English Historical Dialectology* (pp. 21-48). Bern: Peter Lang.

Lee, D. Y. W. (2010). What Corpora Are Available? In A. O'Keeffe and M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 107-121). London: Routledge.

Lewis, R. E. *et al.* (Eds.). (1952-2001). *Middle English Dictionary*. Ann Arbor: University of Michigan Press. Online version in F. McSparran *et al.* (Eds.). (2001-). *Middle English Compendium*. University of Michigan Digital Library Production Service. Available at http://quod.lib.umich.edu/m/med

Markus, M. (Comp.). (1992-). *ICAMET: Innsbruck Computer Archive of Machine-Readable English Texts*. Innsbruck: University of Innsbruck.

McIntosh, A., Samuels, M. L. & Benskin, M. (1986). *A Linguistic Atlas of Late Mediaeval English* (4 vols.). Aberdeen: University of Aberdeen Press.

McSparran, F. *et al.* (Eds.). (2001-). *Middle English Compendium*. University of Michigan Digital Library Production Service. Available at http://quod.lib.umich.edu/m/mec/

Milic, L. T. (1995). The Century of Prose Corpus: A Half-Million Word Historical Data Base. *Computers and the Humanities*, *29*, 327-337.

Miranda-García, A. & Garrido-Garrido, J. (Forthcoming). *Text Search Engine* (*TexSEn*). Málaga: Servicio de Publicaciones de la Universidad de Málaga.

Moskowich, I. & Crespo, B. (Eds.). (2011). *Astronomy 'playne and simple': The Writing of Science between 1700 and 1900.* Amsterdam: John Benjamins.

Nevalainen, T. (2008). Corpora, Historical Sociolinguistics and the Transmission of Linguistic Change. In A. M. Hornero, M. J. Luzón & S. Murillo (Eds.), *Corpus Linguistics: Applications for the Study of English* (pp. 23-37). Bern: Peter Lang.

Nevalainen, T., Raumolin-Brunberg, H., Keränen, J., Nevala, M., Nurmi, A. & Palander-Collin, M. (Comps.). (1998). *Corpus of Early English Correspondence*. Department of English, University of Helsinki.

Nevalainen, T., Raumolin-Brunberg, H., Keränen, J., Nevala, M., Nurmi, A., Palander-Collin, M. & Taylor, A. (Comps.). (2006). *Parsed Corpus of Early English Correspondence. Text version.* Helsinki, York: University of Helsinki and University of York.

Nurmi, A., Taylor, A., Warner, A., Pintzuk, S. & Nevalainen, T. (Comps.). (2006). *Parsed Corpus of Early English Correspondence. Tagged version*. York, Helsinki: University of York and University of Helsinki.

Petré, P. (Comp.). (2009). *Leuven English Old to New (LEON): Some Ideas on a New Corpus for Longitudinal Diachronic Studies*. Conference delivered at the MMECL Conference. University of Innsbruck, Germany, July 6-9.

Pintzuk, S., Haeberli, E., van Kemenade, A., Koopman, W. & Beths, F. (Comps.). (2000). *The Brooklyn-Geneva-Amsterdam-Helsinki Parsed Corpus of Old English*. Available at http://www.users.york.ac.uk/~sp20/corpus.html

Pintzuk, S. & Plug, L. (Comps.). (2001). *The York-Helsinki Parsed Corpus of Old English Poetry*. Available at http://www-users.york.ac.uk/~lang18/pcorpus.html

Poirier, R. (Ed.). (1990). *The Oxford Authors: Ralph Waldo Emerson*. Oxford: Oxford University Press.

Randall, B. (2000). *CorpusSearch*. Available at http://corpussearch.sourceforge.net/

Rissanen, M. (2000). The World of English Historical Corpora: From Cædmon to Computer Age. *Journal of English Linguistics*, *28*, 7-20.

Rissanen, M. (2005). The Development of *till* and *until* in English. In J. Fisiak & H. K. Kang (Eds.), *Recent Trends in Medieval English Language and Literature in Honour of Young-Bae Park* (Vol. I) (pp. 75-92). Seoul: Thaehaksa.

Rissanen, M., Kytö, M., Kahlas-Tarkka, L., Kilpiö, M., Nevanlinna, S., Taavitsainen, I., Nevalainen, T. & Raumolin-Brunberg, H. (Comps.). (1991). *The Helsinki Corpus of English Texts*. Department of English, University of Helsinki.

Rodríguez-Puente, P. (Forthcoming). Introducing the Corpus of Historical English Law Reports: Structure and Compilation Techniques. *Revista de Lenguas para Fines Específicos, 15*.

Schmied, J., Claridge, C. & Siemund, R. (Comps.). (1999). *The Lampeter Corpus of Early Modern English Tracts*. In K. Hofland, A. Lindebjerg & J. Tunestvedt (Eds.), *ICAME Collection of English Language Corpora* (CD-ROM, version 2). Bergen: The HIT Centre.

Scott, M. (1998). *WordSmith Tools. Version 3.0*. Oxford: Oxford University Press.

Simpson, J. A. & Weiner, E. S. C. (Eds.). (2004). *The Oxford English Dictionary. Second edition on CD-ROM. Version 3.1*. Oxford: Oxford University Press.

Stenroos, M. (2007). The Middle English Grammar Project: Aims and Plans. In A. Meurman-Solin & A. Nurmi (Eds.), *Studies in Variation, Contacts and Change in English* (Vol. 1). Retrieved from http://www.helsinki.fi/varieng/journal/volumes/01/stenroos/

Stenroos, M., Mäkinen, M., Horobin, S. & Smith, J. (Comps.). (2011). *MEG-C. The Middle English Grammar Corpus. Version 2011.1*. Available at http://www.uis.no/research/culture/the_middle_english_grammar_project/meg-c_base/

Taavitsainen, I., Pahta, P. & Mäkinen, M. (Comps.). (2005). *Middle English Medical Texts* (CD-ROM). Amsterdam: John Benjamins.

Taavitsainen, I., Pahta, P., Hiltunen, T., Mäkinen, M., Marttila, V., Ratia, M., Suhr, C. & Tyrkkö, J. (Comps.). (2010). *Early Modern English Medical Texts* (CD-ROM). Amsterdam: John Benjamins.

Taavitsainen, I., Tyrkkö, J. & Vartiainen, T. (2008). Current Issues in Corpus Linguistics: Introducing VARIENG Research. Material provided for the course. Department of English, University of Helsinki, Finland.

Taylor, A., Nurmi, A., Warner, A., Pintzuk, S. & Nevalainen, T. (Comps.). (2006). *Parsed Corpus of Early English Correspondence. Parsed version*. York, Helsinki: University of York and University of Helsinki.

Taylor, A., Warner, A., Pintzuk, S. & Beths, F. (Comps.). (2003). *The York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE)*. York: University of York. http://www-users.york.ac.uk/~lang22/YcoeHome.htm

Vázquez, N. (Ed.). (Forthcoming). *Creation and Use of Historical English Corpora in Spain*. Newcastle: Cambridge Scholars.