ENGLISH CORPUS LINGUISTICS AND HISTORICAL RESEARCH

0. INTRODUCTION

The aim of this paper is twofold. On the one hand, we intend to show an overview of what has been and is being done with respect to so-called Corpus Linguistics as far as the English language is concerned. On the other, special attention will be paid to the possibilities of using computerised textual corpora when doing historical research. The former goal will comprise a quick overview of the history of English Corpus Linguistics (§1) and a brief account of technical features such as the systems of incorporated annotations (§2), related software (§3), and so on. Updated lists of institutions (§4), collections of corpora (§5) and completed or in-progress projects in this field (§6) will also follow. With regard to the historical dimension, which this paper also intends to cover, section 7 shows a panorama of different products consisting of electronic English texts previous to the present-day standard. More specifically, in section 8 the authors concentrate on the *Helsinki Corpus of English Texts*, of which they make habitual use. The final section (§9) is devoted to illustrating two popular software tools, namely the *Oxford Concordance Program* and the *WordCruncher* applied to random searches in the *Helsinki Corpus*, which will prove extremely useful in order to draw some initial conclusions concerning the success of using computerised corpora and related electronic items as part of linguistic research.

1. INTRODUCTION TO CORPUS LINGUISTICS

From the times of the American structuralists of the 60's, corpora of texts have been an extremely useful tool for linguists as a means of obtaining data which cannot be gathered through introspection. This latter alternative, based on the linguist's intuitions, gained success during the times of the explosion of Transformational-Generative Linguistics, even though attention to corpus studies was never abandoned. Even nowadays, despite the fact that the Chomskyan tradition takes corpora as irrelevant, it is not difficult to find works, couched in the generative tradition, which make use of standard corpora of texts.

The first text-corpora consisted of data taken from books, newspapers, television, conversations, and the like. These, together with tests made to native speakers, were used in similar ways. We will have to wait till the late 60's to find a modern-style corpus of texts: the well-known *Survey of English Usage*, which consisted of a large file of cards containing speech transcribed by a group of researchers led by Prof. Randolph Quirk. From that moment onwards, taking advantage of the possibilities offered by personal computers, three famous

Some Sundry Wits Gathered Together 1996: 73—98
Emma Lezcano — Javier Pérez — Elena Seoane

packages appear: the *Brown University Corpus* (1,000,000 words of texts published in 1961 in the United States) *[Figure 1: Brown untagged]*, the *Lancaster / Oslo-Bergen Corpus* (British version of the Brown Corpus), known as the *LOB Corpus [Figure 2: LOB untagged]*, and the *London-Lund Corpus of Spoken English* (codified version of the *Survey of English Usage).* Despite the contemporary techniques of selection, classification and annotation in the corpora just mentioned, we are still far from vast projects like the *Collins Birmingham University International Language Database (COBUILD),* which comprises 20 million words, and which has been used in dictionaries, grammars, concordance lists, etc. Undoubtedly, 'language corpora A. C.' (after computers) were born.

This corpus-fever gave even way to the appearance of a specific newsgroup or electronic discussion list, namely, CORPORA, which deals with everything related to text corpora (availability, compiling, tagging, parsing, bibliographical lists, etc.) [CORPORA is run by NCCH <knut. hofland@hd. uib. no>], and even corpus-particular lists, like recent BNC-DISCUSS on the *British National Corpus* [<natcorp@oucs. ox. ac. uk>].

With respect to the size of the corpora, a few years ago, one million words was regarded as large. The aim of some of the current projects I shall summarise below is to gather 100 million words. This is possible, on the one hand, because of the availability of texts already formatted for computer use, and on the other, because of the improvement of the so-called OCR scanning packages (Optical Character Readers) which facilitate the computerisation of texts. Corpora of spoken texts, on the contrary, have to be specially gathered and prepared for electronic use, which complicates their extension.

Size is not, however, the only factor which makes a given corpus useful for scientific purposes; the selection of the texts is more important. In connection to this, it is common to distinguish between corpora and textbanks. As pointed out by Edwards (1993: 282-83),

> [c]orpora are intended to be representative of some specified population or genre. Textbanks tend to be collections of available data with looser connection to each other, or focus on a restricted number of genres (including perhaps only one).

A special type of corpus is the «monitor corpus», which is unbounded in extension and / or time, and which is constantly improved and enlarged.

2. ANNOTATION

The possibilities of having a large number of texts which can be easily and quickly consulted, organised, selected and printed, etc. are enormous for different fields related to the linguistic science itself: Lexicography, language teaching / learning, language acquisition, computer-aided translation, speech synthesis and speech recognition, etc. Besides, a corpus is not only a means directed towards a given purpose but also an end in itself. The science dealing with the codification, improvement and analysis of computerised corpora is called *Computational Corpus Linguistics.*

More specifically, two fields interest researchers at present: *automatic tagging* and *automatic parsing* of the texts, both included in the general label 'corpus annotation.' The former refers to the codification of the grammatical categories of the words and phrases contained in the corpus (adjective, noun phrase), and the latter to the way functional labels can be attached to actual utterances (subject, object). In Leech's (1993: 275) words,

> corpus annotation is the practise of adding interpretative (especially linguistic) information to an existing corpus of spoken and / or written language, by some kind of coding attached to, or interspersed with, the electronic *representation* of the language itself.

With regard to computer annotation, three types of corpus can be distinguished:

(i) 'raw' or 'pure' corpora, which are not annotated,
(ii) tagged corpora *[Figure 3]*, and
(iii) parsed corpora, known as 'treebanks.' *[Figures 4 and 5]*

Tagging is the usual alternative of annotation because of its simplicity, since in most cases it does not require contextual information. There are two well-known schemes of tagging a corpus: (i) the one used in the Brown Corpus, which consists of 87 word-tags, and (ii) the tagging system of the LOB Corpus, which makes use of 132 tags. In new corpora, a tendency towards a smaller number of tags can be appreciated. For example, only 57 word-tags have been incorporated in the British National Corpus. Other widely used tagsets are those of the TOSCA project at Nijmegen, the ASCOT project at Amsterdam, and the Lund (158 tags) and PENN (just 25 word-tags) *[Figure 6], inter alia.*

On the other hand, tagging can be done manually and automatically. Among the automatic tagging systems, CLAWS is most noteworthy (166 word-tags). Automatic parsers are not perfect, and human correction is always required, especially to avoid any possible ambiguous readings. To give a few examples, the level of successful automatic parsing in the Lancaster Parsed Corpus is 60%, and the one in automatically parsing the British National Corpus with CLAWS 97%.

A minority of corpora are semantically, discoursively or prosodically analysed. Figure 7 shows a type of semantic tagging. Figure 8 illustrates a text which is discoursively annotated. The prosodic notation is exemplified in Figure 9.

3. RELATED SOFTWARE

(i) Word-processors: By means of a word-processor a corpus-user can read, cut, copy, paste, delete, find, save, print, search sections of the text.

(ii) Concordance programs: They offer alphabetical-ordered lists of words with or without context, that is, KWIC (keyword in context, or «concordances») *[Figure 10]* and KWOC

(keyword out of context) *[Figure 11]*. The most famous tools are: the Oxford Concordance Program (*OCR*), WordCruncher and Lexa, commented below.

(iii) Lexa is in fact a comprehensive tool developed at the University of Essen which comprises several programs:

® Lexa Compare: it compares two texts byte for byte.
® Lexa Text: text-editor.
® Dbtrans: this program normalises the spellings of a text according to the output-language field (Middle English, Modern English, etc).
® Lexa (main program):
- tagger: automatic, semi-automatic and manual modes
- lexical databases
- concordance files

® Cocoa: this software is able to retrieval information according to the COCOA headers (see Figure 12).
® Lexa Words: it produces lists of unique words from source text files.

(iv) TACT is a text retrieval program, working on a textual database. Possible outputs are full texts, concordance lists in KWIC format, etc.

(v) WordCruncher is a textual analysis program, comprising two tools: WCIndex, a program for indexing texts, and XCView, which allows us to analyse already-indexed texts, doing searches on the coding, on actual words, producing concordance outputs.

(vi) Hypercard Stack and Free Text Browser for Macs.

(vii) The Linguistic Database (LDB), developed by the TOSCA group at Nijmegen University, is a phrase-marker analyser, specially designed for the Nijmegen Corpus of Modern British English.

4. INFORMATION INSTITUTIONS AND CENTRES

(i) The Norwegian Computing Centre for the Humanities (NCCH), which houses the ICAME archive (see §5) and the CORPORA discussion list (see §1) [<icame@hd. uib. no>].

(ii) The Computers in Teaching Initiative Centre for Textual Studies (CTI) at the University of Oxford [<ctitext@vax. oxford. ac. uk>].

(iii) The Center for Electronic Texts in the Humanities (CETH) at Rutgers and Princeton Universities [<ceth@zodiac. rutgers. edu>].

(iv) The Association for Computers and the Humanities (ACH) at Carnegie-Mellon University, which sponsors the Rutgers / Princeton National Text Archive and the HUMANIST electronic discussion group [<rudman@cmphys. bitnet>].

(v) The Association for Literary and Linguistic Computing (ALLC) at Oxford University Press in the USA.

(vi) The Association for Computational Linguistics (ACL), USA [<walker@bellcore.com>].

5. AVAILABILITY: MAIN COLLECTIONS OF CORPORA

These are the main institutions which act as distributors of linguistic corpora:

(i) ACL / DCI (Association of Computational Linguistics Data Collection Initiative), which houses the DARPA-funded Linguistic Data Consortium [<myl@unagi. cis. upenn. edu>].

(ii) CETH [see §4iii].

(iii) CHILDES (Child Language Exchange System) [<brian+@andrew. cmu. edu>].

(iv) ICAME (The International Computer Archive of Modern English *[not only 'modern' English in fact]*) [see 43i].

(v) Library of the future: it distributes a set of CD-ROMs containing novels, stories, plays and historical documents.

(vi) OTA (The Oxford Text Archive) [<archive@vax. oxford. ac. uk>].

(vii) Project Gutenberg: it makes available literary works on electronic media [<hart@vmd. cso. uiuc. edu>].

Apart from the distributors just listed, full details on other existing corpora can be found in the following Internet addresses:

- University of Lancaster Survey (via anonymous ftp to <nora. hd. uib. no> [filename: pub / icame / survey. corpora]).
- The Georgetwon University Catalog of Projects in Electronic Text (<pmangiafico@gu-vax. georgetown. edu>).

6. CORPORA OF CONTEMPORARY ENGLISH

Neither multilingual corpora nor those dealing with 'Englishes' for specific purposes have been included[1]

ACL Data Collection Initiative Disc

620 MBytes of texts from *Wall Street Journal* (1987-89), the *Collins Dictionary*, scientific abstracts and a variety of tagged and parsed materials from the Treebank project.

AMERICAN HERITAGE INTERMEDIATE CORPUS (American Heritage Dictionary Division)

[1] Other main projects are: ACL / DCI (USA): collection of several hundred million words of text and speech; BANK OF ENGLISH (UK): dynamic corpus to be constantly updated (monitor corpus); BROWN UPDATE; FLOB CORPUS: LOB updated; INTERNATIONAL CORPUS OF ENGLISH (*ICE*) (several countries): collection of national and regional varieties —each regional sub-corpus will contain one million running words—, tagged and parsed; INTERNATIONAL CORPUS OF LEARNER ENGLISH (*ICLE*) (several countries): collection of essays written by EFL students.

c. 5 million words of written American English.

AMERICAN NEWS STORIES (via Oxford Text Archive)

period: 1979; 250,000 words of 1979 news stories from the Associated Press network.

BELLCORE LEXICAL RESEARCH CORPORA (Language and Knowledge Resources Research, Bellcore)

c. 200 million words of American texts taken from newspapers, c. 50 million words of magazine articles, a collection of English machine-readable dictionaries and reference books, electronic-mail digest, and other assorted texts.

BERGEN CORPUS OF LONDON TEENAGER LANGUAGE *(COLT)*

BERKELEY CORPUS (University of California, Uppsala University)

spoken and written American English.

BIRMINGHAM CORPUS (University of Birmingham)

written (90%) and spoken (10%) English; c. 20,000,000 words; the total Birmingham Collection of English Texts *(BCET)* will comprise c. 40 million words.

BRITISH NATIONAL CORPUS *(BNC)* (Oxford U. Press)

c. 100 million words of spoken and written British English; tagged; available now.

BROWN CORPUS (Brown University)

texts of American English published in 1961; 1,014,294 words; untagged (2 formats), KWIC concordance, WordCruncher version. It was the fist computer-corpus and is said to be «the most analyzed corpus of English to date» (Edwards 1993: 283).

CHILDES DATABASE (Carnegie-Mellon University, Pennsylvania)

English parent-child interactions.

CORNELL CORPUS (available via CHILDES)

c. 1.6 million words of written and spoken British and American English texts.

CORPUS OF ENGLISH-CANADIAN WRITINGS (Queen's University)

textbank of c. 3 million words of Canadian English from magazines, books and papers.

CORPUS OF SPOKEN NORTHERN IRELAND ENGLISH (The Queen's University of Belfast)

c. 400,000 words.

GÖTHEBORG CORPUS

c. 128,000 words from the Brown Corpus; American English; parsed by hand; reparsed with Surface and Underlying Structural Analyses of Naturalistic English (SUSANNE)

KOLHAPUR CORPUS (Shivaji University, Kolhapur)

c. 1 million words of printed Indian English; untagged; WordCruncher.

LANCASTER-LEEDS TREEBANK

c. 45,000 words from the LOB; parsed by hand.

LANCASTER-OSLO / BERGEN CORPUS *(LOB)*

texts of British English published in 1961; c. 1,000,000 words; untagged, KWIC concordance, tagged (horizontal and vertical formats), WordCruncher.

LANCASTER PARSED CORPUS

c. 140,000 words of the LOB Corpus; tagged and parsed.

LANCASTER / IBM SPOKEN ENGLISH CORPUS *(SEC)*

c. 52,000 words of spoken British English of 1984-1987; two parsing systems (UCREL and CCALAS); orthographic and prosodic transcription.

LEXIS (Mead Data Central)

commercial archive of legal codes.

LONDON-LUND CORPUS OF SPOKEN ENGLISH *(LLC)*

c. 435,000 words of educated spoken British English; prosodic notation; untagged, KWIC concordance (2 versions), WordCruncher (2 versions).

LONGMAN / LANCASTER ENGLISH LANGUAGE CORPUS (Longman Group Ltd., Lancaster University)

c. 30-50 million words of varieties of English; in progress.

MACQUARIE (UNIVERSITY) CORPUS

c. 1 million words; Australian English.

MEDIS (Mead Data Central)

commercial archive of medical literature.

MELBOURNE-SURREY CORPUS

period: 1980-81; c. 100,000 words of Australian newspaper texts.

NEXIS (Mead Data Central)

commercial archive of newspapers, newsletters and other periodicals.

NIJMEGEN CORPUS

c. 1.5 million words of educated written British English; part is tagged and hand parsed.

PARSED LOB CORPUS

automatic parsing with UCREL.

PENN TREEBANK (U. of Pennsylvania)

4,885,798 words of written and spoken English; tagged and 'skeletal' parsing.

PIXI CORPORA (via Oxford Text Archive)

conversations in bookshops in Britain and Ireland.

POLYTECHNIC OF WALES CORPUS (PoW)

c. 61,000 words of children's spoken English; parsed by hand according to Hallidayan systemic-functional grammar; orthographic transcription.

SURVEY OF ENGLISH USAGE (University College, London)

written British English; c. 500,000 words.

SUSANNE CORPUS (U. of Leeds)

c. 128,000 words of the Brown Corpus.

TIPSTER Information Retrieval Text Research Collection (ARPA / SISTO, U. S.)

3 gigabytes of documents of all-style American English.

TOSCA CORPUS

c. 1,000,000 words.

TREEBANK OF WRITTEN AND SPOKEN AMERICAN ENGLISH (University of Pennsylvania)

2.6 million words together with part of speech tags, skeletal hand-made syntactic parsing (it contains the first fully parsed version of the Brown Corpus) and intonational boundaries (for spoken language); in progress.

WARWICK CORPUS (available via OTA)

c. 2.5 million words of written British English.

WESTLAW (West Corporation)

commercial archive of legal codes.

7. ENGLISH HISTORICAL CORPORA[1]

AUGUSTAN PROSE SAMPLE (Cleveland State University)

c. 80,000 words of samples of Augustan prose.

ARCHER (Northern California University, University of Southern California)

period: 1650-1990; British and American English; c. 1.7 million words; its tagging is now being completed at the U. of Bergen and U. of Helsinki; available through ICAME later this spring.

BROOKLYN - GENEVA - AMSTERDAM - HELSINKI PARSED CORPUS OF OLD ENGLISH

Old English section of the Helsinki Corpus; glossing, morphological and syntactic tagging, bracketing in progress.

CAMBRIDGE-LEEDS CORPUS OF EARLY MODERN ENGLISH (University of Leeds, University of Cambridge):

period: c. 1600-1800; format: WordCruncher, Oxford Concordance; parsing and tagging in the future; in preparation.

CENTURY OF PROSE CORPUS (Cleveland State University)

period: 18th century; c. 0.5 million words; in preparation.

CORPUS OF EARLY AMERICAN ENGLISH (University of Helsinki)

period: 1600s-1700s; 305,500 words.

CORPUS OF IRISH ENGLISH (University of Essen)

period: 14th to the 20th century; in preparation.

[1] Non-corpus historical computerised databases: HISTORICAL THESAURUS OF ENGLISH (U. of Glasgow) —in preparation; Old English Thesaurus is near completion—; LINGUISTIC ATLAS OF EARLY MEDIAEVAL ENGLISH (*LAEME*) (U. of Edinburgh) —period: c. 1150-1300; parsing and tagging in progress; EARLY MODERN ENGLISH RENAISSANCE DICTIONARIES CORPUS (U. of Toronto); CORPUS OF SOURCE TEXTS FOR JOHNSON'S *DICTIONARY* (U. of Birmingham); OLD ENGLISH THESAURUS (King's College, London).

CORPUS OF LATE MODERN ENGLISH (University of Manchester)
period: 1861-1919; informal private letters and journal entries by British writers; c. 100,-000 words; Helsinki format.

CORPUS OF MODERN ENGLISH TEXTS *(COMET)* (Glasgow University)
19th and 20th century novels and drama; in preparation.

CORPUS OF 19TH-CENTURY ENGLISH
c. 250,000 words; format: WordPerfect 5.0 / 5.1.

HELSINKI CORPUS OF EARLY AMERICAN ENGLISH (University of Helsinki)
period: 1620-1720; in preparation.

HELSINKI CORPUS OF ENGLISH TEXTS (University of Helsinki)
c. 1,600,000 words; COCOA-header parameters; untagged, WordCruncher version; being tagged and syntactically bracketed in 1994 (Old & Middle English parts already done).

HELSINKI CORPUS OF OLDER SCOTS (University of Helsinki)
period: 1450-1700; 579,380 words; in distribution during 1995; tagging is finished (University of Edinburgh).

HELSINKI CORPUS ON EARLY ENGLISH CORRESPONDENCE (University of Helsinki)
period: 1420-1680; personal letters; c. 2 million words.

INNSBRUCK COMPUTER ARCHIVE OF MIDDLE ENGLISH TEXT *(ICAMET)*
in preparation; 100 books already compiled.

LAMPETER CORPUS OF EARLY MODERN ENGLISH TRACTS (German Research Association, Technical University of Chemnitz-Zwickau)
period: 1640-1740; texts from the Tract Collection of St David's University College at Lampeter (Wales); c. 1 million words; format: Helsinki; in preparation.

PENN-HELSINKI PARSED CORPUS OF MIDDLE ENGLISH *(PPCME)* (Univ. of Pennsylvania)
c. 500,000 words (extended version of the prose Middle English section of the Helsinki Corpus plus some additional texts; in preparation.

ZURICH CORPUS OF ENGLISH NEWSPAPERS *(ZEN)* (Zürich)
period: 1671-1791; c. 300,000 words; London newspapers; in preparation.

8. THE HELSINKI CORPUS OF ENGLISH TEXTS: AN EXAMPLE OF A HISTORICAL CORPUS

The aim of this section is describing the main characteristics of the computerized corpus that serves as a database for our research, and to outline the criteria behind the text selection.

8. 1. SIZE AND STRUCTURE OF THE HELSINKI CORPUS OF ENGLISH TEXTS

The Helsinki Corpus of English Texts: Diachronic and Dialectal (henceforth The Helsinki Corpus) is a computerized collection of early English texts[1] comprising about two million

[1] It is available from the Norwegian Computing Centre for the Humanities and the Oxford Text Archive. The different formats of the corpus are listed in the order forms from the distributors.

words of running text. It consists of two parts: a diachronic part containing texts from 750 to 1700 and a dialectal part, based upon transcripts of interviews with speakers of British rural dialects from the 1970's.

This corpus is the result of a rather laborious project launched under the supervision of Dr. Matti Rissanen and Dr. Ossi Ihalainen in 1984. The aim of this project was to provide representative material for the historical and dialectal study of English: a computerized database which promoted and facilitated the study the development of morphology, syntax and vocabulary in the history of English as well as the different varieties of English. Therefore, it is obvious that our corpus is multi-purpose[1], that is to say, it has not been compiled for a particular, clearly-defined research topic; on the contrary, it was meant to provide the basis for a variety of studies over an extended period of time.

This section will focus on the diachronic part of the corpus (completed in 1991), the one we use in our research for statistical data and examples. The historical part of the Helsinki Corpus comprises 400 samples of texts covering the period from the 8th century to the beginning of the 18th century. The extracts from long texts vary from 2,500 to some 20,000 words; shorter texts are included *in toto*. Two supplementary corpora are in preparation at present: the corpus of Older Scots and the corpus of Early American English. They will be integrated into the basic corpus soon.

The Helsinki Corpus is made up of three sections: Old English, Middle English and Early Modern English, as illustrated in Table 1, listing the total number of words in each subperiod.

TABLE 1

Subperiod		*Words*
OLD PERIOD		
I	-850	2,190
II	850-950	92,050
III	950-1050	251,630
IV	1050-1150	67,380
		413,250
MIDDLE ENGLISH		
I	1150-1250	113,010
II	1250-1350	97,480
III	1350-1420	184,480
IV	1420-1500	213,850
		608,570
EMODE, BRITISH		
I	1500-1570	190,160
II	1570-1640	189,800
III	1640-1710	171,040
		551,000

Kytö and Rissanen (1992: 9) are aware of the fact that a corpus of 1.5 million words is probably too small for the purposes of exhaustive diachronic research on a concrete aspect of the history of English, especially if we take into account that, when dividing the different

[1] In fact, it is the first multi-purpose corpus of English compiled so far which covers the time span of several cc.

periods, we arrive at subcorpora no larger than c. 150,000 words each. This fact is determined by practical, technical and financial factors.[1]

It is obvious, therefore, that the Helsinki Corpus does not completely represent the English language of the past, and so it would be convenient to complement the study of the texts in this Corpus with analogous material.

8. 2. FRAMEWORK AND SELECTION OF SAMPLES

As Rissanen (1992: 188) points out, in compiling the Helsinki corpus heuristic considerations prevail over theoretical ones. Nevertheless, we can assert that the compilation work, in broad terms, has its basis in the theories of modern sociolinguists such as Labov, Bailey, Milroy and Romaine. According to them, as is well known, linguistic variation and change can be conditioned not only by linguistic factors but also by extralinguistic factors. The latter have been relevant to the compilation of texts.

Among the criteria that contributed to the selection of the material for this diachronic multipurpose corpus, the following can be mentioned: chronology, type of text, dialect and sociolinguistic factors.

8. 2. 1. CHRONOLOGICAL COVERAGE

The date of the text has been of decisive importance. The aim of a diachronic corpus is to give researchers the opportunity to compare structures and paradigms in the successive synchronic stages in the past.

In applying this criteria, a number of problems had to be faced. On the one hand, there is a lack of texts from the early periods of the English language: the amount of available text from the periods OE1, OE2, OE4 and ME1 and ME2 is scanty and these subperiods remain, therefore, under-represented. We must not forget either the risky question of dividing the time span in coherent subperiods, for which a certain degree of arbitrariness cannot be avoided.

On the other hand, problems of dates have also been an important obstacle in the selection of texts: unknown dates and the difference between the dates of the original and the manuscript version of a text[2] made things difficult for the compilers.

Of course, it is absolutely impossible to achieve a balanced and symmetrical chronological coverage in the first period. Yet, in the last section (EModE), the subdivision in periods of 70 years is not arbitrary or based on practical factors but reflects changes in society and the different stages in the evolution of English in this period. Thus, as Rissanen (1992: 191) puts it,

[1] Kytö and Rissanen (1992: 9) welcome any kind of suggestion for addition to the current material. In fact, they intend to produce an improved version briefly.

[2] Finally they decided to follow the date of the manuscript when grouping the text into subsections. But they offer two code values when necessary (Kytö and Rissanen 1992: 13).

The first sub-period, EModE1, is in many respects indicative of the Middle English heritage. EModE2 marks the process of rapid and radical change, while the third subperiod and the last subperiod reflects the gradual establishment of the present-day structural system of English.

Chronology is the most important criterion to be taken into account by diachronic corpus compilers since this kind of corpus should be representative of all the different synchronic stages of the period it is intended to cover.

8. 2. 2. TYPE OF TEXT

This relevant criterion is difficult to apply coherently when constructing a corpus. Since a clear-cut satisfactory model of textual classification is not available, once more practical reasons rather than logical principles were followed for the selection of texts. Extralinguistic criteria such as subject matter, purpose, discourse situation and relations between writer and reader were taken into account in order to avoid or diminish the risk of circularity in results. Table 2 shows the typological division that was eventually reached.

TABLE 2

Old English	*Middle English*	*EMod English*
LAW	LAW	LAW
DOCUM	DOCUM	---
HANDB ASTRONOMY	HANDB ASTRONOMY	---
HANDB MEDICINE	HANDB MEDICINE	---
---	HANDB OTHER	HANDB OTHER
SCIENCE ASTRONOMY	---	---
---	SCIENCE MEDICINE	SCIENCE MEDICINE
---	---	SCIENCE OTHER
---	---	EDUC TREAT
PHILOSOPHY	PHILOSOPHY	PHILOSOPHY
HOMILY	HOMILY	---
---	SERMON	SERMON
RULE	RULE	---
REL TREAT	REL TREAT	---
PREFACE / EPIL	PREFACE / EPIL	---
---	PROC DEPOS	---
---	---	PROC TRIAL
HISTORY	HISTORY	HISTORY
GEOGRAPHY	---	---
TRAVELOGUE	TRAVELOGUE	TRAVELOGUE
---	---	DIARY PRIV
BIOGR LIFE SAINT	BIOGRLIFE SAINT	---
---	---	BIOGR AUTO
---	---	BIOGR OTHER
FICTION	FICTION	FICTION
---	ROMANCE	---
---	DRAMA MYST	---
---	---	DRAMA COMEDY
---	LET PRIV	LET PRIV
---	LET NON PRIV	LET NON PRIV
BIBLE	BIBLE	BIBLE

They tried to ensure diachronic continuity in the corpus.[1] That is the reason why eight text types occur in all the three sections of the Corpus (law, handbooks, science, philosophy, history, biography, fiction and the Bible). However, there are types of text which only occur in two periods (religious treatises and private letters) or even in only one period (geography). To solve the problem of the lack of full generic continuity over the individual subperiods within the three main periods, they suggest resorting to what they have called 'diachronic text prototypes.' Thus, the different texts could be divided into the following six generic prototypes: statutory (which would include laws and documents in the different periods), secular instruction (handbooks, scientific, educational and philosophical treatises), religious instruction (homilies, rules, sermons), expository, imaginative narration and non-imaginative narration.

When classifying the texts according to text-type, they decided to include primarily non-literary texts. In the same way, most texts are in prose although they allowed room for relevant verse texts as well. They have also tried to define the relationship of some texts to spoken language and have included speech-based types of text such as sermons, private letters, trial records, etc. which are supposed to reflect certain characteristics of spoken language. This is useful if we consider the total lack of oral evidence at that time.

Obviously, the compilers do not claim that these texts are absolutely uniform and homogeneous as far as distribution of linguistic or discourse features is concerned. They are aware that this is only one of the many possible generic classifications and that much more research is needed in this field.

8. 2. 3. REGIONAL COVERAGE

It is not possible to conceive a diachronic corpus without making reference to dialectal distribution in the periods preceding the establishment of a standard language. All the samples of the Corpus have been given localization parameter values. In Early Modern English, all texts are selected as representing the Southern standard, so the dialectal criterion is not taken into consideration. As mentioned above, two supplementary corpora which do observe geographical variety distributions in this period: Older Scots (1450-1700) and Early American English (1620-1720) are in preparation.

The different dialects in OE and ME[2] as coded in the Corpus are given in Table 3.

---

[1] Another way to ensure continuity in the Corpus has been the inclusion of samples of translations of the *Bible* and Boethius' *De Consolatione Philosophiae* dating from several centuries.

[2] The dialect coding of most earlier Middle English texts are based on the definitions found in the Middle English Dictionary; for later Middle English texts, the *Linguistic Atlas of Late Mediaeval* English has been consulted (Kytö and Rissanen 1992: 18).

TABLE 3

Old English	*Middle English*	*EMod English*
A / X AM AM / X AN K K / X WS WS / K WS / A WS / AM WS / X	EML EML / NL EMO WML WMO NL NO NO / EMO SL SO KL KO X	ENGLISH
A = Anglian K = Kentish NL, NO = Northern X = 'Unknown'	AM = Anglian Mercian WS= West Saxon SL, SO = Southern ENGLISH = Southern British Standard	AN = Anglian Northumbrian EML, EMO = East Midland KL, KO = Kentish

When trying to provide the Corpus texts with a reliable dialectal code, the compilers had to face many problems:

(i) It is difficult to follow the transmission history of a manuscript.
(ii) Extralinguistic data about the author and his origins are not reliable in the earlier periods of the English language.
(iii) The information given in reference works on the background of texts and authors is vague and confusing.

As a consequence of this, the Corpus-user must allow a certain degree of circularity in dialect coding since it had to be based mainly on linguistic features rather than on extralinguistic evidence. We must bear in mind that the Helsinki Corpus is not absolutely perfect, and it is therefore advisable to resort to further sources to get more information on the dialect of the text.

8. 2. 4. SOCIOLINGUISTIC AND DISCOURSIVE FACTORS

Just as the dialectal parameter was only taken into account in OE and ME texts, the sociolinguistic criterion is only relevant from the ME period onwards. Systematic information on sex, age and social rank is only given in the final section of the corpus, since the information we have from the earlier periods is too inconsistent to be reliable. Sociolinguistic factors distinguish our corpus, which is not an isolated and arbitrary collection of texts but a representative compilation of texts from a certain sociohistorical context. Thanks to the extralinguistic information provided in the Corpus, we can study how the same structures or words are used in a different way according to sex, age or rank.

Discoursive factors have also been taken into account. Thus, the relationship between sender and receiver is indicated in letters (intimate vs. distant or equal vs. up or down), and the degree of formality is also defined according to the extra-linguistic factors involved in the discourse situation (official letters and sermons are regarded as formal while private letters or comedy have been coded as informal).

One of the innovations of the Helsinki Corpus is that all this extralinguistic information is clearly indicated at the beginning of each of the samples. The codes are introduced in COCOA format and they aim at providing the user with all the possible relevant data about the text in question. Several retrieval programs can use this code system in order to search for examples which fulfil a predefined set of requirements.

As Rissanen (1992) points out, the parameter coding enables the researcher to choose between two different approaches to the corpus: he can collect all the occurrences of a structure in the whole corpus or a part of it and observe its distribution according to the different parameters (descriptive approach, closer to inductive methods), or he can restrict his searches only to those texts with certain constraints and contrast the instances found with others occurring in texts with different variables (deductive approach, more dynamic).

We must realize, though, that just as the Helsinki Corpus does not completely reflect the language of the past, nor can parameter coding provide an absolutely complete and exhaustive description of each sample. It is difficult to find information about authors and the general status of the texts especially at the early stages of the language, when most of the material from which the sampling can be made is scanty and one-sided.

At present, one of the main problems of the Helsinki Corpus both for compilers and users is the high number of variant spellings in certain periods of the English language. The detailed list that is included at the end of the Corpus with all the different spellings of all the words in the Corpus is not enough to avoid this inconvenience. It would be useful to produce a version of the corpus in which each word is equipped with a lexical tag. In fact, grammatical tagging in general is one of the most important future developments of the Helsinki Corpus.

In the light of all that has been said above, we can conclude that despite all the limitations, the Helsinki Corpus has helped to make the access to textual evidence much easier than before. It has also enabled scholars to tackle linguistic problems and carry out ambitious projects which would have been impossible without this useful tool. Yet, we must not forget that the corpus only offers a material basis for the analysis, and additional material should be used for the correct interpretation of the evidence. And, of course, as Rissanen (1992: 202) remarks, text corpora should never alienate scholars from the study and love of original texts, but should increase their curiosity and imagination.

9. USING COMPUTERISED CORPORA: SOME EXPLORATIONS IN THE HELSINKI CORPUS

In this section we illustrate the pros and cons that the researcher faces when using computerized linguistic corpora and related software. Obviously, the possibilities the latter offer

vary depending on the type of research. As will be duly explained, related software is fairly helpful for lexical and morphological studies, but this is not the case as concerns syntactic searches through corpora. For our illustration of lexical, morphological and syntactic searches, we use the Helsinki Corpus (HC) and the Oxford Concordance Program (OCP),[1] with some occasional mention of the WordCruncher (WC).

9. 1. MORPHOLOGICAL SEARCHES

With the aim of exemplifying a morphological search through a corpus, we looked up the suffix *-able* in the early Modern English section of the HC with OCP. The purpose for this study would be to find out the productivity of such a suffix in the formation of deverbal and denominal adjectives in EModE.

Before dealing with the results of the search themselves, we shall briefly describe the commands for the search. An OCP command file consists of various sections, and an asterisk marks the beginning of a new section. The first is the '*input' section. Here you tell OCP to interpret the characters in the 'input' file, i. e., the file to be searched, correctly, and also which parts of it to select for search. In this 'input' section we have the following commands:

- *References* COCOA "<" to ">". This command tells OCP to interpret everything between "<"and">" as references rather than text. As was mentioned before, COCOA is an established format used to enter references.
- *Text to 80.* This command tells OCP up to which column in the text file the text to be searched can be found. In this particular case, the command tells OCP that the text spans up to 80 columns (or characters) per line. The reason for this specification is that, in some cases, though not in the HC, the first characters are saved for references.
- *Comments* between "(\" to "\)", "[^" to "^]", "[}" to "}]", "[\" to "\]". This command tells OCP to ignore everything between these symbols because they are editorial comments of some sort.

The next section is the '*word' section, where you clarify which characters your text consists of and how they ought to be interpreted by OCP.

- *Alphabet* "A=a +A=+a B=b C=c D=d +D=+d E=e +E=+e F=f G=g +G=+g H=h I=i J=j K=k L=l M=m N=n O=o P=p Q=q R=r S=s T=t +T=+t U=u V=v W=w X=x Y=y Z=z 0 1 2 3 4 5 6 7 8 9 &: = +". In the alphabet command, you explain to OCP what the alphabet used in the coded text file looks like. 'A=a' means that capital and regular letters are equivalent. In this case, most characters are regular alphabet characters, except for "+A", representing ash, and "+G" representing yogh, among others.
- *Padding "` ~: ' - # [ ] { } () ^ / \ |: " ".* The padding command gives OCP the list of characters it is likely to find but should ignore in searches and for sorting purposes.
- *Punctuation ",;::?!".* This command tells which characters are punctuation markers.

[1] Both the Helsinki Corpus and WordCruncher were described above. As for OCP, it is a program that analyses raw texts and provides lists of words, indexes and concordances.

The next section is the '*action' section, where you tell OCP what to do. The basic options are wordlist, index and concordance.

- *Do wordlist.*
- *References Q = 20, P = 8.* This command specifies which references you want to have printed for the items OCP finds. *Q* identifies the texts in the HC and P the page number.
- *Pick words "*able".* This command tells OCP to pick up the words ending in —that is what the asterisk means— *-able.*

Finally, the section '*go' simply tells OCP to start its search.

The list we obtain shows all the words ending in *-able* and how frequently they appear in the relevant section of the HC. Obviously, this list is not perfect, in the sense that some of the words, namely the adjectives *able* and *unable,* must be excluded from the total count.

The result of the search is only partially satisfactory. The positive aspect is that not only have we saved time —OCP reads more than 500,000 words in just one hour in an average 386SX IBM-compatible computer— but we can also rely on the resulting figures, for no occurrence is missing, a distinct possibility if the search is manual. However, the information obtained is not enough for our morphological study, because the productivity of *able* can only be found if the number of its occurrences is compared with the number of occurrences of all adjectives, simple and derived. OCP provides the total number of words, which is not relevant for this study, but it cannot provide the number of adjectives. Therefore, we cannot obtain reliable scientific results. In this respect, raw texts such as those in the HC are limited. In tagged corpora, such as, for instance, the Helsinki Corpus of Older Scots, this kind of search would be complete and satisfactory.

An example of a real morphological investigation through the HC is that conducted by Dr. Dalton-Puffer (University of Vienna). She studied the productivity of the suffixes of Romance origin as compared with other suffixes. She used WordCruncher and the results of her search led to the conclusion that these suffixes are not productive.

9. 2. LEXICAL SEARCHES

The search we carried out to exemplify a lexical investigation concerned the use of the possessives *thyne* and *mine* in the early Modern English period. The commands are similar to the ones described above, but in the 'action' section a concordance, and not a list, is asked for. In this search we also include a 'stats' command referring to a basic statistical calculation, a type-token count.

The resulting concordance provides information about the text, page and some context where each *thyne / mine* occurrence is found. The results are in keeping with those obtained in the morphological search: despite the fact that it has to be revised, it is time-saving and precise.

A real lexical search is that carried out by Dr. Moskowich and E. Seoane, from the University of A Coruña (forthcoming). The words of Scandinavian origin were searched through the EModE section of the HC with WC. The said search delivered favourable results with re-

markable celerity. Previous to the search itself, however, we had to look for all the possible spelling variants of the words in question in the additional files provided by the HC, in order not to miss any occurrence of the relevant words. With a lexically tagged corpus this previous search would not be necessary, and there is also the advantage that we would not have to specify, and search for, every spelling variant.

9. 3. SYNTACTIC SEARCHES

Finally, we exemplify the use of a corpus and related software for a syntactic investigation. The case illustrated here is a research still in progress dealing with the passive construction in the EModE section of the HC.

As was mentioned earlier, syntactic investigations find many limitations in the available software, and with these we shall start. The most important drawback is that this software cannot identify passives. WC can locate and count the occurrences of the different forms of be, for instance. Among these, however, we obtain the occurrences of *be* not only as passive auxiliary, but also as primary auxiliary and main (copulative) verb. The selection of the different uses of *be* takes too long for this search to be worthwhile. Another possibility, available with WC, is to look for the verb *be* in combination with words ending in *—ed*, in order to obtain the passives of regular verbs. What you get, however, are non-passives like "He was tired" or "It is red", and even occurrences where the *—ed* form occurs before the verb *be*, as in "This bed is new". A further disadvantage is that the passive constructions in which words appear between the auxiliary and the participle, as in "He was by his mother killed", which are so common in EModE, would not be identified.

The results of the search with OCP, which is illustrated below, are more accurate but not satisfactory. The command file in this particular case is a bit more complicated. Some new commands have to be inserted in the different sections, and they are the following:

*Input

- *Select where C="E1", C="E2", C="E3".* This command tells OCP to select only the three EModE subperiods for its search, instead of going through all the HC.

*Action

- Include only collocates "was" upto 3 "*d", "were" upto 3 "*d", "am" upto 3 "*d", "are" upto 3 "*d", "is" upto 3 "*d", "been" upto 3 "*d", "being" upto 3 "*d", "wer" upto 3 "*d", "bee" upto 3 "*d". In the search for passives we do not look for particular words but for particular collocates, and for that reason we must use the 'include' and not the 'pick' command. In 'include' we tell OCP to look for occurrences of the listed forms of *be* followed by any word ending in "d" with up to three words in between.
- *Maximum context left upto ". ;:: ? !" and right upto ". ;:: ? !" and span L.* This command tells OCP that it should include, in the context that it will print out, everything that is between the token and the listed punctuation markers, both to the left and the right of

the token. 'Span L' tells OCP to go beyond line boundaries, if necessary, when selecting the context that is to be printed out.

*Format
In this section we specify the kind of format we want the concordance to have.
- *Headwords left.*
- *Context size 2 and left aligned and indent 3.*
- *References right.*

Too many errors are found in the resulting concordance. Firstly, the passives of irregular verbs are not identified. Secondly, there are cases where the selected tokens are not passives. In this search, for instance, the occurrences of "am and" and "am glad" are picked up. And thirdly, the context provided by the concordance in those cases where the token is a passive is not enough to extract the relevant data, such as the presence or absence of an agent, or its length. Therefore, the basic search, that for the location of passives, must be done manually, by reading on the screen or by reading a hard copy.

Another search required for this syntactic investigation has to do with active constructions. To determine the frequency and distribution of passives in the different text-types it is necessary to know the proportion of passives with respect to actives. Obviously, the search for active transitive verbs with a complement eligible to become passive subject cannot be done with either WC or OCP. This is possible only with a syntactically tagged corpus. This search also must be done manually.

Other obligatorily manual searches are, for instance, the count of agent and agentless passives and the type of preposition used to mark the agent (if you search for the EModE agentive prepositions, *by, of, through* and *from,* all their uses other than the agentive —locative, instrumental and so on— one will be identified).

Thus far, we have dealt with the limitations of untagged corpora and related software as regards syntactic studies. We shall turn now to illustrate the advantages of dealing with computerized corpora for this type of study.

At some point in our investigation, we had to find out the reasons behind the use of the passive. One of the prospects we considered is that the passive might be lexically determined; that is, that the presence of certain matrix verbs could trigger the use of the passive. To find out whether this was the case or not, we had to take each verb in our passive samples and count the number of its occurrences as an active transitive verb followed by a complement, and, then, compare the number of active and passive occurrences with relation to the general active / passive ratio. In this search for active transitive verbs we used WC, after having looked for all the possible spelling variants in the additional HC files. The results were amenable and time-saving because this particular search had lexical rather than syntactic characteristics.

A syntactic investigation carried out successfully with the HC and related software is that by Prof. Matti Rissanen (1991). He studies the occurrence of the high frequency verbs *say, tell,*

think and *know* followed by *that* and zero complementiser in the HC. Of course, he had to exclude from his count all the occurrences of such verbs in other constructions, such as the parenthetic ones, but he extracts relevant data from the whole HC, that is, more than 750,000 words, a task almost impossible if the corpus used was not computerized. Similarly, Merja Kytö studied the diachronic variation of modal verbs in the HC, and many other syntacticians find computerized searches time-saving and convenient.

All this, to say that computerized corpora and related software are necessary for the study of historical linguistics, and particularly helpful for investigations of lexical and morphological nature. For syntactic studies, however, it is urgent to have tagged and parsed corpora. For this reason, the work of Dr. Susan Pintzuk and Dr. Ann Taylor from the University of Pennsylvania, who are tagging the Old and Middle English sections of the HC respectively, pave the way for the diachronic study of English.

Emma Lezcano González (Universidade da Coruña)
Javier Pérez Guerra (Universidade de Vigo)
Elena Seoane Posse (Universidade de Santiago de Compostela)

FIGURES

1) BROWN UNTAGGED (TEXT FORMAT II)

```
A01 0010 1 The Fulton County Grand Jury said Friday an investigation
A01 0020 1 of Atlanta's recent primary election produced "no evidence"
A01 0020 9 that any irregularities took place.
A01 0030 5 The jury further said in term-end presentments that
A01 0040 3 the City Executive Committee, which had over-all charge
A01 0050 2 of the election, "deserves the praise and thanks of
A01 0050 11 the City of Atlanta" for the manner in which the election
A01 0060 11 was conducted.
A01 0070 1 The September-October term jury had been charged
A01 0070 9 by Fulton Superior Court Judge Durwood Pye to investigate
A01 0080 8 reports of possible "irregularities" in the hard-fought
A01 0090 6 primary which was won by Mayor-nominate Ivan Allen
A01 0100 5 Jr&.
```

2) LOB UNTAGGED

```
A01 1 **[001 TEXT A01**]
A01 2 *<*'*7STOP ELECTING LIFE PEERS**'*>
A01 3 *<*4By TREVOR WILLIAMS*>
A01 4 |^A *0MOVE to stop \0Mr. Gaitskell from nominating any more Labour
A01 5 life Peers is to be made at a meeting of Labour {0M P}s tomorrow.
A01 6 |^\0Mr. Michael Foot has put down a resolution on the subject and
A01 7 he is to be backed by \0Mr. Will Griffiths, {0M P} for Manchester
A01 8 Exchange.
A01 9 |^Though they may gather some Left-wing support, a large majority
A01 10 of Labour {0M P}s are likely to turn down the Foot-Griffiths
A01 11 resolution.
A01 12 *<*7*'ABOLISH LORDS**'*>
A01 13 |^*0\0Mr. Foot's line will be that as Labour {0M P}s opposed the
A01 14 Government Bill which brought life peers into existence, they should
A01 15 not now put forward nominees.
A01 16 |^He believes that the House of Lords should be abolished and that
A01 17 Labour should not take any steps which would appear to *"prop up**" an
A01 18 out-dated institution.
```

3) LOB TAGGED

```
A01 2 ^ *'_*' stop_VB electing_VBG life_NN peers_NNS **'_**'. _.
A01 3 ^ by_IN Trevor_NP Williams_NP. _.
A01 4 ^ a_AT move_NN to_TO stop_VB \0Mr_NPT Gaitskell_NP from_IN
A01 4 nominating_VBG any_DTI more_AP labour_NN
A01 5 life_NN peers_NNS is_BEZ to_TO be_BE made_VBN at_IN a_AT meeting_NN
A01 5 of_IN labour_NN \0MPs_NPTS tomorrow_NR. _.
A01 6 ^ \0Mr_NPT Michael_NP Foot_NP has_HVZ put_VBN down_RP a_AT
```

```
A01 6 resolution_NN on_IN the_ATI subject_NN and_CC
A01 7 he_PP3A is_BEZ to_TO be_BE backed_VBN by_IN \0Mr_NPT Will_NP
A01 7 Griffiths_NP, _, \0MP_NPT for_IN Manchester_NP
A01 8 Exchange_NP. _.
A01 9 ^ though_CS they_PP3AS may_MD gather_VB some_DTI left-wing_JJB
A01 9 support_NN, _, a_AT large_JJ majority_NN
A01 10 of_IN labour_NN \0MPs_NPTS are_BER likely_JJ to_TO turn_VB down_RP
A01 10 the_ATI Foot-Griffiths_NP
A01 11 resolution_NN. _.
A01 12 ^ *'_*' abolish_VB Lords_NPTS **'_**'. _.
A01 13 ^ \0Mr_NPT Foot's_NP$ line_NN will_MD be_BE that_CS as_CS labour_NN
A01 13 \0MPs_NPTS opposed_VBD the_ATI
A01 14 government_NN bill_NN which_WDTR brought_VBD life_NN peers_NNS into_IN
A01 14 existence_NN, _, they_PP3AS should_MD
A01 15 not_XNOT now_RN put_VB forward_RB nominees_NNS. _.
```

4) PARSING: SPOKEN ENGLISH CORPUS

```
[S[N Nemo_NP1, _, [N the_AT killer_NN1 whale_NN1 N], _, [Fr[N who_PNQS N][V 'd_VHD
grown_VVN [J too_RG big_]] [P for_IF [N his_APP$ pool_NN1 [P on_II [N Clacton_NP1
Pier_NNL1 N]P]N]P]J]V]Fr]N], _, [V has_VHZ arrived_VVN safely_RR [P at_II [N his_APP$
new_]] home_NN1 [P in_II [N Windsor_NP1 [ safari_NN1 park_NNL1 ]N]P]N]P]V], _, S ]
```

5) PARSING: PENN TREEBANK

```
((S
      (NP Mr. Vinken)
      (VP is
            (NP chairman
                  (PP of
                        (NP      (NP Elsevier N. V.)
,
                                 (NP the Dutch
                                       publishing
                                       group))))))
.)
```

6) PENN TAGGING

```
The / DT practice / NN of / IN state-owned / JJ vehicles / NNS for / IN use / NN of / IN employees
/ NNS on / IN business / NN dates / VBZ back / RP over / IN forty / CD years / NNS. /
At / IN least / JJS one / CD state / NN vehicle / NN was / VBD in / IN existence / NN in /
IN 1917 / CD. /
```

7) SEMANTIC NOTATION

PPHS1	She	Z8 (=pronoun)
VVD	laughed	E4.1+ (=happy and sad)

RR	disagreeably	O4.2- (=judgement of appearance, etc.)
,	,	,
VVG	squashing	A1.1.1. (=general actions)
APPGE	her	Z8 (=pronoun)
NN1	cigarette	F3 (=cigarettes and drugs)
II.	in	Z5 (=function word)
AT	the	Z5 (=function word)
NN1	butter	F1 (=food)

8) DISCOURSAL NOTATION

Numbers bind semantically-related constituent; <n indicates a pronoun with a preceding antecedent.
S. 1 (0) The state Supreme Court has refused to release (1 [2 Rahway State Prison 2] inmate 1) (1 James Scott 1) on bail.
S. 2 (1 The fighter 1) is serving 30-40 years for a 1975 armed robbery conviction.
S. 3 (1 Scott 1) had asked for freedom while <1 he waits for an appeal decision
S. 4 Meanwhile, [3 <1 his promoter 3], ((Murad Muhammed 3), said Wednesday <3 he netted only $ 15,250 for (4 [1 Scott 1]'s nationally televised light heavyweight fight against [5 ranking contender 5]) (5 Yaqui Lopez 5) last Saturday 4).

9) PROSODIC NOTATION: LONDON-LUND CORPUS

'!: features of stress; \/\/: tones; -: pause; #: tone unit boundary; @:: pause filler `er'; {}: subordinate tone unit; (()): uncertain material.
1 1 1 10 1 1 B 11 ((of ^Spanish)). graph\ology# /
1 1 1 20 1 1 A 11 ^w=ell#. /
1 1 1 30 1 1 A 11 ((if)) did ^y / ou _set _that# - /
1 1 1 40 1 1 B 11 ^well !J\oe and _I# /
1 1 1 50 1 1 B 11 ^set it betw\een _us# /
1 1 1 60 1 1 B 11 ^actually !Joe 'set the: p\aper# /
1 1 1 70 1 1 B 20 and *((3 to 4 sylls))* /
1 1 1 80 1 1 A 11 *^w=ell#. /
1 1 1 90 1 1 A 11 "^m / \ay* I _ask# /
1 1 1 100 1 1 A 11 ^what goes !\into that paper n / ow# /
1 1 1 110 1 1 A 11 be^cause I !have to adv=ise#. /
1 1 1 120 1 1 A 21 ((a)) ^couple of people who are !d\oing [dhi: @] /
1 1 1 130 1 1 B 11 well ^what you: d\ / o# /
1 1 1 140 1 2 B 12 ^is to - - ^this is sort of be: tween the: tw\ / o of /
1 1 1 140 1 1 B 12 _us# /
1 1 1 150 1 1 B 11 ^what *you*: d\ / o# /
1 1 1 160 2 1 B 23 is to ^make sure that your 'own. !c\andidate /
1 1 1 170 1 1 A 11 *^[\m]#* /
1 1 1 160 1 2(B 13 is. *. * ^that your. there`s ^something that your /
1 1 1 160 1 1(B 13: own candidate can: h\ / andle# - - /
1 1 1 180 2 1 B 21 ((I ^won`t)) /
1 1 2 190 1 1 A 11 *((^y\eah#))* /

10) KEYWORD IN CONTEXT, kwic-FILE

```
[. To.]
33. : . And. made. forward. erly. for. to. ryse,
34. : . To. take oure. wey. ther. as. I. yow. deuyse.
37. : . Me. thynketh. it. acordaunt. to. resoun
38. : . To. telle. yow. al. the. condicioun
[. Wel.]
24. : . Wel. nyne. and. twenty. in. a. compaignye
29. : . And. wel. we. weren. esed. atte. beste.
[. What.]
5. : . What. Zephirus. eek. with. his. sweete. breath
18. : . That. hem. hath. holpen. what. that. they. were. seeke.
40. : . An. whiche. they. were, . and. of. what. degree,
41. : . An. eek. in. what. array. that. they. were. Inne;
```

11) KEYWORD OUT OF CONTEXT, kwoc-FILE

```
[ To ]
33      : And made forward erly for {to} ryse,
34      : {To} take oure wey ther as I yow deuyse.
37      : Me thynketh it acordaunt {to} resoun
38      : {To} telle yow al the condicioun
[ Wel ]
24      : {Wel} nyne and twenty in a compaignye
29      : And {wel} we weren esed atte beste.
[ What ]
5       : {What} Zephirus eek with his sweete breath
18      : That hem hath holpen {what} that they were seeke.
40      : An whiche they were, and of {what} degree,
41      : An eek in {what} array that they were Inne;
```

12) COCOA HEADERS

1: <B = name of text file>	2: <Q = text identifier>
3: <N = name of text>	4: <A = author>
5: <C = part of corpus>	6: <O = date of original>
7: <M = date of manuscript>	8: <K = contemporaneity>
9: <D = dialect>	10: <V = verse or prose>
11: <T = text type>	12: <G = relation to foreign original>
13: <F = foreign original>	14: <W = relation to spoken language>
15: <X = sex of author>	16: <Y = age of author>
17: <H = social rank of author>	18: <U = audience description>
19: <E = participant relation>	20: <J = interaction>
21: <I = setting>	22: <Z = prototypical text category>
23: <S = sample>	24: <P = page>
25: <L = line>	26: <R = record>

REFERENCES

Aarts, J., De Hann, P. & Oostdijk, N. eds. 1993: *English Language Corpora: Design, Analysis and Exploitation.* Amsterdam, Rodopi.

Asher, R. E. ed. 1994: *The Encyclopedia of Language and Linguistics.* Oxford, Pergamon. [Entries: «Data Collection in Linguistics» and «Corpora»].

Dalton-Puffer, Ch. 1994: Productive or Non-productive? The Romance Element in Middle English Derivation. > Fernández, F. et al. eds. 1994: *English Historical Linguistics 1992.* Amsterdam, John Benjamins: 247-60.

Edwards, J. A. 1993: Survey of Electronic Corpora and Related Sources for Language Researchers. > Edwards, J. A. & Lampert, M. D. eds. 1993: *Talking Data: Transcription and Coding in Discourse Research.* London, Erlbaum: 263-310.

García-Miguel, J. Mª 1994: Lingüística de Corpus. Unpublished paper delivered at the 'Seminario de Industrias de la Lengua: Introducción a la Lingüística Informática.' Soria.

Kytö, M. 1991: *Variation and Diachrony, with Early American English in Focus.* New York, Peter Lang.

Kytö, M. 1993: *Manual to the Diachronic Part of the Helsinki Corpus of English Texts: Coding Conventions and List of Source Texts.* Helsinki, Helsinki U. P. (Department of English, University of Helsinki).

Kytö, M. & Rissanen, M. 1992: A Language in Transition: The Helsinki Corpus of English Texts. *ICAME Journal* 16: 7-27.

Kytö, M., Rissanen, M. & Wright, S. eds. 1994: *Corpora Across the Centuries. Proceedings of the 1st International Colloquium on English Diachronic Corpora.* Amsterdam, Rodopi.

Leech, G. 1993: Corpus Annotation Schemes. *Literary & Linguistic Computing* 8.4: 275-81.

Moskowich, I. & Seoane, E. forthcoming: The Scandinavian Lexical Element in EModE: Some Preliminary Considerations. *Neuphilologische Mitteilungen.*

Nevalainen, T. & Ramoulin-Brungerg, H. 1989: A Corpus of Early Modern Standard English in a Socio-Historical Perspective. *Neuphilologische Mitteilungen* 1 XC: 67-110.

Rissanen, M. 1991: On the History of *That / _* as Object Clause Links in English. > Aijmer, K. & Altenberg, B. eds. 1991: *English Corpus Linguistics. Studies in Honour of Jan Svartvik.* London, Longman: 272-89.

Rissanen, M. 1992: The Diachronic Corpus as a Window to the History of English. > Svartvik, J. ed. 1992: *Directions in Corpus Linguistics.* Berlin, Mouton.

Rissanen, M. 1994: The Helsinki Corpus of English Texts. > Kytö, M., Rissanen, M. & Wright, S. eds. 1994: *Corpora Across the Centuries. Proceedings of the 1st International Colloquium on English Diachronic Corpora.* Amsterdam, Rodopi.

Sampson, G. 1992: Analysed Corpora of English: A Consumer Guide. > Pennington, M. C. & Stevens, V. eds. 1992: *Computers in Applied Linguistics: An International Perspective.* Clevedon, Multilingual Matters: 181-200.

Svartvik, J. ed. 1992: *Directions in Corpus Linguistics*. Berlin, Mouton.

* * *