Estudios Ingleses de la Universidad Complutense
1999, n.º 7: 299-306
ISSN: 1133-0392

English historical linguistics online

Javier MARTÍN ARISTA
Universidad de La Rioja

1. INTRODUCTION

The World Wide Web is a natural medium for the advertising of new commercial products, the publication of pioneering projects, the divulgation of up-to-the-minute research results and the testing of materials whose elaboration is still in progress. English Historical linguistics has not been absent from the computer revolution and has taken advantage of the spectacular growth of the Internet in several respects: bibliography is easily accessible for reference and purchase; software can be tested, retrieved as shareware and, sometimes, used as freeware; historical texts (incluiding glossaries, some translations, images, and other related materials) can be downloaded; projects in progress are receiving critiques, suggestions, additional information and every kind of feedback from potential users of the final product; and dictionaries based on concordances are available (at least partially) which allow for different kinds of searches. In the following sections of this review I offer a selection and a critical revision of some of the English Historical Linguistics resources which are obtainable from the Internet. I have classified these resources under four headings: software, glossaries and dictionaries, non-annotated corpora and annotated corpora.

2. SOFTWARE

The Summer Institute of Linguistics (http://www.sil.org) is an organization specialising in the study of minority languages and cultures which is located at

the International Linguistics Center in Dallas. This institution offers resources for anthropology, translation, literacy, language learning, and computing, including a list of computer programmes of which I recommend the following: AMPLE (a morphological parser for Windows and Macintosh); PC-PATR (a syntactic parser for Windows and Macintosh in beta version); MacLex (a lexical database for Mac in beta version); ParaConc (a concordance generator for Windows and Macintosh); Shoebox (a database management programme for Windows and Macintosh developed for interlinarising text and developing lexicons); and IT (an interlinear text processor for MS-DOS and Macintosh). Many of the resources that I mention in the following sections include a concordance processor [1]. As a Macintosh user, I find the concordancer MacCONC (version 1.7.6) particularly useful: it works on plain text documents, can export results to a file and does not take up much memory; it can generate concordances (including concordances based on interlinearised text) and indexes and build glossaries. Moreover, MacCONC has no bugs and is freeware. On the other hand, MacCONC has file size limitations [2]. FreeText Browser (available from the *International Computer Archive of Modern and Medieval English, ICAME*, page: http://www.hd.uib.no/icame.html) poses no size problems, but cannot extract files. All the programmes I have reviewed in this section can be downloaded (at least in Macintosh version, most versions for Windows are commercial) from http://www.sil.org/computing/html [3].

3. GLOSSARIES AND DICTIONARIES

The ANXADAT (Archived discussion on Anglo-Saxon England, University of Newfoundland, Canada) *Project* maintains a Modern English to Old English Vocabulary site at http://www.mun.ca/Ansaxdat/vocab/wordlist.html. The glossary, which offers the Old English equivalents of Present-Day English lexical items and some grammatical information (gender, case and government), is currently being constructed.

The Glossarial DataBase of Middle English (http://www.hti.umich.edu/english/gloss) is being prepared by Larry Benson, from Harvard University. *The Canterbury Tales* text has already been morphologically annotated. The following sample line illustrates the markup of the text (quoted from *The Glossarial DataBase of Middle English* site):

(1) <DIV0 TYPE="frag" N="Frag1">
<HEAD>Fragment I, CT<HEAD>
<DIV1 TYPE="CT" N="GP">
<HEAD>General Prlogoue </HEAD>
<L N="GP:1">
<W LEMMA="whan" ANA="ADVC">Whan</W>

```
<W TYPE="gram" LEMMA="that" ANA="CONJ">that</W>
<W LEMMA="april" ANA="NOUN">Aprill</W>
<W LEMMA="with" ANA="PREP">with</W>
<W TYPE="infl" LEMMA="his" ANA="P1GN">his</W>
<W TYPE="infl" LEMMA="shour" ANA="NPL0">shoures</W>
<W LEMMA="sote" ANA="ADJ0">soote</W>
</L>
```

This annotation, which provides textual and grammatical information, allows for different types of searches: fulltext (all parts of the word, including descriptive information), lemma (lemma form), analysis (analytical markup), language (LAT-in/FR-ench/GR-eek/EN-glish) and form (inflected vs. grammatical word). Whole texts are accesible through hypertext links (the title preceding every line of the concordance). Each word of the corpus is provided with its grammatical description and the number of occurrences in *The Canterbury Tales* for the grammatical category under scrutiny [4].

The 2nd. Edition of *The Oxford English Dictionary* (http://www.oed.com) can be consulted (access is restricted) in the site of the Virtual Library of Virginia (http://etext.lib.virginia.edu/eng-on.html). The CD-ROM version can be ordered from http://www.oed.order.htm (Windows or Macintosh version, $395, £250). The *OED* is currently being revised in order to add new entries and improve the definitions, derivations, pronunciations and historical quotations. The first prototype version of the *OED Online* has been available for some time now. The final version will be ready by October 1999, eleven years before the third edition of the *OED* is due to be published. I refer the reader to http://www.com/inside/revision.htm for further details about the revision programme [5].

The *Middle English Compendium* is an interface maintained by the *Humanities Text Initiative* of the University of Michigan at http://www.hti.umich.edu/. It has been designed to provide interconnectivity between two Middle English electronic resources: the *Corpus of Midle English Prose and Verse*, henceforth *MEC*, (http://www.hti.umich.edu/english/mideng) and the *Middle English Dictionary*, hereafter *MED*, (http://www.hti.umich.edu/dict/med). The June 1998 release of the *MEC* includes 1,073 hyperbibliography and copies of Middle English texts (almost all entries for V-J). The *MEC* can be searched in several ways: simple searches (single words or phrases), Boolean searches (combinations of two or three words in a given paragraph, verse or gloss), proximity searches (co-occurrence of two or three words or phrases) and bibliography searches (author and title). The latest release of the *MED* (June 1998) comprises 29,095 lexical entries (all entries for I-U). Every entry of the *MED* consists of the lexical item, its category, morphology, etymology, meanings, syntactic complementation and semantic complementation. Entries are provided with quotations preceded by the title of the source

document, which is a hypertext link that allows the user to search the whole text.

4. NON-ANNOTATED CORPORA

The Electronic Text Center of the University of Virginia offers several English language online resources at http://etext.lib.virginia.edu/english.html. These resources include *The Old English Corpus* (University of Virginia only), *The Middle English Collection* (which contains forty publicly-accessible titles) and *The Michigan Early Modern English Materials*. Most publicly accessible texts have been provided by *The Oxford Text Archive* (http://ota.ox.ac.uk/tei/ota.html). When searching these corpora the whole text collection is examined, not only free-access texts. Results from queries include instances and concordances which contain, in turn, links to full texts and textual information.

The Helsinki Corpus of English Texts is a computerised collection of extracts of continuous texts of 1.5 million words from 850 to 1700. As is described by Kytö (1996: v), "the Old English section of the corpus is based on the material taken (...) from the machine-readable transcript (Release 1, October 1982) prepared for the *The Dictionary of Old English Corpus*" (University of Toronto). A number of Middle English texts have been obtained from *The Oxford Text Archive*, from where the whole corpus can be obtained (http://ota.ox.ac.uk/tei/ota.html). Although *The Helsinki Corpus* is not tagged, a number of conventions have been included in order to code orthographic and editorial practices. The beginning of a text contains information about the author, the date, the register and the dialect of the text. The typical beginning of a texts looks as follows (quoted from the *ICAME* page: http://www.hd.uib.no):

(2) <B COPREFCP>
<Q O2 XX PREF PRCP>
<N PREF CP>
<A ALFRED>
<C O2>
<O 850-950>
<M 850-950>
<K CONTEMP>
<D WS>
<V PROSE>
<T PREFACE/EPIL>
<S SAMPLE X>
[TEXT: ALFRED'S PREFACE TO CURA PASTORALIS. KING

ALFRED'S WEST-SAXON VERSION
OF GREGORY'S PASTORAL CARE, PART I.
EARLY ENGLISH TEXT SOCIETY, O.S. 45.
ED. H. SWEET.
LONDON, 1958 (1871).
[PP. 3.1 - 9.7]
[B9.1.1]
<P 3>
<R 1>
[}+DEOS BOC SCEAL TO WIOGORA CEASTRE.}]
+alfred kyning hate+d gretan W+arfer+d biscep his wordum luflice & freondlice;
<R 2>
& +de cy+dan hate +d+at me com swi+de oft on gemynd, hwelce wiotan iu w+aron giond Angelcynn, +ag+der ge godcundra
hada ge woruldcundra;
<R 4>

The new *ICAME Corpus Collection on CD-ROM* is available from February 1999. Along with *The Helsinki Corpus*, many other text corpora in machine-readable form (WordCruncher, TACT and Wordsmith versions) have been included. The largest are the following: *The Brown Corpus* (1 million of words of written American English), *The LOB Corpus* (1 million words of written British English), *The Kolhapur corpus* (1 million words of written Indian English), *The London-Lund Corpus* (half a million words of spoken British English), and the *Frown* and *FLOB* corpora, which replicate and update the *Brown* and *LOB* corpora and thus provide the basis for the diachronic analysis of Present-Day English [6].

The Dictionary of Old English Corpus in Electronic Form (*OEC*) is an online database consisting of all surviving Old English texts, which amount to 3,035 texts containing about three million words of Old English and another two million in Latin. Where differences of time or dialect are relevant, more than one copy of a given text has been collected. Texts are coded in fully conformant way to Sperberg, McQueen and Burnard's *Guidelines for Electronic Text Encoding and Interchange* (*TEI P3*). The *OEC* is freely available until February 1999 (http://www.hti.umich.edu/english/oec/.). Annual institutional site licenses from the University of Michigan Press (site-licenses@umich.edu) cost $200. Copies of the *OEC* for academic use can be ordered at $200 (corpus@doe.utoronto.ca). The *OEC* allows for simple searches (single words or phrases), Boolean searches (combinations of two or three words in a given paragraph, verse or gloss), proximity searches (co-occurrence of two or three words or phrases), bibliography searches (author and title) and Old English-Latin comparison. Searches turn out textual

instances linked to whole texts. Each corpus text is classified according to genre, period and language [7].

5. ANNOTATED CORPORA

Cathy Ball is responsible for the Old English pages of the University of Georgetown (http://www.georgetown.edu/cball/oe/old_english.html), where information concerning texts, software and bibliography can be found. Ball provides links to several Old English texts (mainly poetry), with audio recordings, manuscript images and translations where available. More relevant to this section are *The Metrically Scanned Anglo-Saxon Poetic Records*, which can be downloaded from ftp://ftp.csd.abdn.ac.uk (one can login as user ftp by giving one's e-mail address as password).

The Penn-Helsinki Parsed Corpus of Middle English (henceforth *PHPCME*) is a corpus of prose text samples of Middle English compiled at the University of Pennsylvania by Anthony Kroch and Ann Taylor. The texts which they have chosen come mainly from *The Diachronic Part of The Helsinki Corpus*, with some additions and deletions. The *PHPCME* and search tools can be downloaded at no cost (kroch@change.ling.upenn.edu). The annotation of the corpus involves clause syntax, grammatical relations and certain structure-changing operations. The following extract illustrates the syntactic tagging (quoted from the *PHPCME* page, http://www.ling.upenn.edu/mideng/):

(3) ([+ For] [f he 1[L [c-1 +tat] %s-1 r % [at wyll] [- not] [v trauayle] [l here] [p wyth men] , L]1] [p as 2[B %op-2% %d-2 r % [s Seynt Barnard] [vt sayth] , B]2] %[s he]% [at schall] [v trauayle] [t ay] [p wyth +te fendes of hell] .) (MIRK,2.26)

According to the analysis provided by the authors, the syntactic formalism specifies that the subordinating conjunction *for* is followed by the left-dislocated pronoun *he*, which is modified by the relative clause *that wyll not trauayle here wyth men*. Next comes the parenthetical *as Seynt Barnard sayth* and then the main clause, in which *he* is the subject, *schall* the auxiliary, *trauayle* the main verb, *ay* the temporal adverbial and *wyth ðe fendes of hell* the prepositional complement. The revised edition of the *PHPCME*, which will grow to more than 1 million words in size, offers part-of-speech tags, indicates the internal structure of the noun phrase and improves the annotation of interclausal syntax. The *PHPCME* 2 is scheduled for release in August 1999.

Other corpora using the annotation of *PHPCME* are currently being constructed. Of special interest for this review is the Old English corpus

which Susan Pintzuk (University of York) and Eric Haeberli (University of Geneva) are compiling (mainly from *The Helsinki Corpus*) and annotating.

6. BY WAY OF CONCLUSION

Even though English Historical Linguistics in the Internet is probably in its infancy if we consider what may appear in the near future, this discipline has responded to the challenge of new technologies and, as I have shown in this review, has clearly benefited from the existence of the Net. The moment is auspicious: the availability of new resources and the renewed interest in the history of the English language that the recent past has witnessed allow us to look ahead to a promising future.

NOTES

[1] For the combination of concordancing of texts and citation of instances, see Sinclair (1991: 40).

[2] Cathy Ball, from Georgetown University, maintains a concordance to Old English and Middle English *Apollonius of Tyre* at http://www.georgetown.edu/cball/oe/old_english.html. This site is a good illustration of what MacCONC can -and cannot- do.

[3] See Biber, Conrad and Reppen (1998) for a more detailed list of analytical tools available from the Web.

[4] See Pérez Guerra (1998: 33) on corpus annotation.

[5] I refer the reader to Stubbs (1996: vxii) for more information about computer-readable corpora of English.

[6] I should like to thank the Editor and the anonymous referees of *Estudios Ingleses de la Universidad Complutense* for their accurate corrections and useful suggestions, including the information about the *Frown* and *FLOB* corpora which I provide in this paragraph. I accept responsibility for all flaws in the final product.

[7] For full details of the compilation of the *OEC* I refer the reader to Cameron (1973) and Healey and Venezky (1980).

Departamento de Filologías Modernas
Universidad de La Rioja
Edificio de Filologías
C/ San José de Calasanz s/n
26.004-Logroño
javier.martin@dfm.unirioja.es

REFERENCES

Biber, D.; Conrad, S. and Reppen, R. (1998). *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Cameron, A. (1973). A List of Old English Texts. In Frank, Roberta and Angus Cameron (eds.) *A Plan for the Dictionary of Old English.* Toronto: University of Toronto Press. 29-267.

Healey, A. and Venezky, R. (1980). *A Microfiche Concordance to Old English.* Toronto: Centre for Medieval Studies, University of Toronto.

Kytö, M. (1996 [1993]). *Manual to the Diachronic Part of The Helsinki Corpus of English Texts. Coding Conventions and List of Source Texts.* Helsinki: Department of English, University of Helsinki.

Pérez Guerra, J. (1998). *Análisis computarizado de textos. Una introducción a TACT.* Vigo: Servicio de Publicacións da Universidade de Vigo.

Sinclair, J. (1991). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Stubbs, M. (1996). *Text and Corpus Analysis.* Oxford: Blackwell.