Title: Brains "versus" software: new possibilities and limitations of computer assisted historical studies of English syntax

Author: Rafał Molencki

Rafał Molencki

Brains *versus* Software: New Possibilities and Limitations of Computer Assisted Historical Studies of English Syntax

Although the article was written fourteen years ago, its main idea is still valid: despite great progress in computer technology, also applied in linguistic studies, successful interpretation of the corpus material can only be performed by the human brain. The problem of identifying multiple meanings has been partially solved by better and better automatic disambiguation techniques. Upgraded search engines can now retrieve words in their original spellings. Many more diachronic English databases have become available since the end of the 20th century. Libraries all over the world have been supplying digitalized facsimiles of more and more medieval manuscripts, which enables scholars to have their own interpretations of ancient texts without referring to the printed 'emended' editions. This has already resulted in the publication of several important studies which changed generally accepted views. However, we were too optimistic concerning the Dictionary of Old English project: while the complete DOE Corpus is ready, the compilers of the dictionary itself are currently working on the entries beginning with the letter H. The Third Edition of the Oxford English Dictionary is still rather far from completion. Not only in this respect we have a long way to go…

The application of computer technology has had a tremendous impact on linguists' work in the last decade. The most obvious advantages are speed and access to very large text corpora of medieval and later English texts. The traditional linguist could only study a few texts, thus most typical dissertations published in the 1960s and 1970s were "descriptive syntaxes" of individual medieval texts (cf. the Mouton series of the late 1960s/early 1970s, e.g. Shannon 1964, Palmatier 1969, Brown 1970). Also my own Ph.D. dissertation of 1988 dealt with verb complementation in Old English and as illustration material I used only two Anglo-Saxon translations, which are attributed to King Alfred the Great, viz. of Bede's *Historia Ecclesiastica Gentis Anglorum* (*Ecclesiastical History of the English Nation*) and Gregory's *Cura pastoralis* (*Pastoral Care*). On the other hand, the Habilitationschrift of 1999 (i.e. in the computer era) on *A History of English Counterfactuals* used samples of dozens of different medieval and modern English texts. Nowadays **all** Old English and most Middle English texts are available in machine readable forms. Thus every single

use of every single Old English word is now recorded and concordanced (cf. the Toronto *Microfiche Concordance of Old English Words* and *A Microfiche Concordance of Old English High-Frequency Words*). Currently the Toronto specialists are finishing the new comprehensive electronic *Dictionary of Old English*, a very useful and timely replacement of the more than a century old Bosworth and Toller's *Anglo-Saxon Dictionary*.

For diachronic and comparative studies an indispensable tool is the Diachronic Part of the Helsinki Corpus of English texts, which provides us with fragments of numerous texts from the millennium between c700-1710, tagged for periods, sub-periods, type of text, style and other codings. Linguists can also make use of the CD-ROM edition of all the Bible translations into English, starting from the three Anglo-Saxon Gospels (written in different Old English dialects in the late 10th century), through 14th century Wycliffe's Bible, three 16th century translations, the landmark Authorised Version of 1611 and finishing with the latest revised translations. An additional program makes cross-referencing easy, so that the researcher has quick access to different versions of the same sentences. Other corpora, including complete works of individual authors (e.g. *The Complete Works of William Shakespeare* on CD-ROM), are obviously available, too. Many publishers make the works of (also classical) English writers accessible to general public in the electronic form. Some of them can be downloaded directly from various websites in the Internet. The Internet also enables us to exchange ideas in private contacts or taking part in various newsgroups like LINGUIST, HISTLING, ENGLISC, etc. Numerous publications become available long before they are printed as books or articles. The information about important publications and events such as seminars or conferences can be quickly spread to hundreds of linguists all over the world.

Another significant contribution was the publication of the second edition of the *Oxford English Dictionary* on CD-ROM in 1989, where 20 huge volumes were compressed into a single disk. The scholar cannot help being overwhelmed by the wealth of data — each entry is provided with definitions, pronunciation, etymologies, numerous examples of usage from all periods of the language's history, first attestations and in the case of obsolete words — their last occurrences. The reliability of this dictionary is very high and it does not seem to have a match in the lexicography of any language. An important advantage of such electronic dictionaries is the fact that they can be revised and updated quickly and at any time. A special browser program (*OED Browser*) enables us to find not only entries but also any occurrences of the sought items in the whole dictionary.

All those searching, browsing and concordancing programs have saved linguists long years of arduous and tedious hunt for language examples and, what is also important, a great deal of space earlier taken up by countless slips of paper, folders, drawers, filing cabinets. Saving search results in computer files makes them immediately available for quick reference at any time and the risk of losing precious data

is smaller, provided one is careful enough to make multiple copies of data in case of a major breakdown.

It goes without saying that search programs have started a new epoch in language studies, especially in morphology and syntax. Within a matter of minutes we can have all the instances of a particular item in a huge corpus. We can also look for groups of words, phrases, collocations, etc.The program will provide us with the linguistic context of a few thousand signs (i.e. several sentences) before and after the requested item. One can limit the search to the instances of some item only when it is accompanied by something else, e.g. a linguist looking for the instances of the so-called third conditional will ask the computer to find all the examples of the word *had* preceded by the context word *if* with the left horizon of, say, four words. If one needs all the inflectional variants of a word, e.g. *play*, it suffices to make a request for *play** and the search program will find all the instances of *play, plays, played, playing*. Likewise, if one is interested with some prefix (e.g. *omni-*), one should ask for *omni** and very soon all the words like *omnipotent, omnipresent, omniscient, omnivorous* will be picked out.

These are just some of the new possibilities of computers, unimaginable two or three decades ago, which have made historical linguists' work so much quicker and easier. Nevertheless one should not become overenthusiastic. One of the major drawbacks of computers is that they simply cannot think and as such will never be able to replace the human brain. Some of the apparent advantages mentioned above can actually make research more difficult. Very frequently we are faced with insurmountable *embarras de richesses*. We simply get much more than we would like to. More often than not as a search result one can be provided with so many instances that one simply gets lost in the wealth of unnecessary data. This usually happens when one looks for the occurrences of some high frequency item. Even in a relatively small corpus one can find thousands of matches.

Multifunctional items give rise to another type of difficulties. Obviously no search program is able to distinguish between the demonstrative pronoun *that*, the relative pronoun *that* and the conjunction *that* or between the conditional *if* and the interrogative *if*. The linguist has to go over all the instances by hand, which will probably take him or her longer than if they read the 'analogue' book and marked the relevant examples in it, quickly eliminating all the others. With the more commonly used words, it might be better if the program found us, say, only every 10th or every 20th instance, which should not be too difficult. Yet none of the four search programs I have been using has this option.

Similarly, the computer will not tell the difference between homonymous nouns and verbs, so searching for the verb *mark* we will also have to process all the noun instances including the proper name *Mark*. As is well-known, the English language is notorious for such homonyms. Looking for all the inflectional forms of the verb *let* a linguist may find to his horror that on asking for *let** , apart from the requested *let, lets* and *letting* he is dumped with words like *letter, lettuce, lethal* etc. Neither can

syntacticians and typologists rely on computers when they study word order. There is no way of finding examples of subject-verb inversion, objects preceding verbs or adjectives following nouns, to mention but a few examples of constructions unavailable electronically. If we want all the instances of the third conditional protasis from a corpus, we have to read all of it anyway.

Then there is an important issue of working with samples. This is the case with the majority of diachronic corpora, whose compilers usually take only 10—20 pages out of voluminous ancient books as a supposedly representative sample. The absence of a particular lexical item or a grammatical construction in the sample that has been chosen at random may be a pure coincidence. Quite often one can read articles where prudent linguists who do not fully rely on their electronic databases discover that the structure which is absent in the, say, Helsinki corpus sample does occur in the same text two pages later and surprisingly not only once.

Another major problem for the computer search are variant spellings of the requested word, especially plentiful during the most interesting period in the history of English, i.e. Middle English. Due to the considerable dialectal differentiation, no standard forms, scribal inconsistency and errors, many words have more than a dozen variant forms, some of them never recorded in either the *Oxford English Dictionary* or the *Middle English Dictionary*. Even if we asked the search program to find all the variants that we know of, without having read the whole corpus ourselves we cannot be certain whether some unusual form will not occur in it. Looking for all the Middle English conditional protases, one has to search for all the clauses beginning not only with canonical *if*, but also with *yif, ef, yef, gif, gef, iffe, yf, yffe, giffe, gife, gefe, geffe, gyfe, gyffe, 3if, 3yf, 3yffe, 3iffe* and perhaps some others, which for the specialist (but never for the machine on its own) are obvious markers the moment that they are found in the right context.

The problem of spelling also concerns the choice of fonts. Numerous programs are often incompatible with one another and working even with only two different computers we may have problems in reading medieval English texts, as there may be different renderings of Old English eths, thorns, yoghs or ashes. A scanner will not read these symbols, either. Since most search programs will work only in the ASCII environment, in the electronic corpora of early English texts we are forced to use rather awkward substitutes (commonly accepted now), e.g. "+a" for the lower case ash, "+T" for the upper case thorn, etc. This makes texts look very strange and very un-Anglo-Saxon in comparison with the original manuscripts, even worse than the early printed editions of the Early English Text Society. Since the mid-19th century the EETS has already published several hundred books and numerous objections have been raised to their former practice of 'emending' medieval texts, 'correcting' what was called scribal errors, imposing modern punctuation by adding non-existent commas, full stops, which often led to all kinds of misinterpretations. It is only in the recent publications of the series that the editors try to preserve the original form as much as possible. Even so, if a historical syntactician wants to be

on the safe side, he or she should necessarily consult the original manuscript or its facsimile. The electronic versions cannot render many of the medieval scribal practices such as specific punctuation or very common abbreviations. Over centuries manuscripts have suffered deletions brought about by flooding or fire, which are usually 'reconstructed' arbitrarily, when one would prefer to have the case open for various options. The human eye may take longer to notice something crucial, but the computer will never do so unless it is told. And even the younger linguists of the computer generation will admit that no electronic 'book' can replace the contact with the real medieval manuscript or its facsimile.

All in all, despite great facilitation of a linguist's work in terms of speed and access to huge text corpora, the computer cannot replace the human brain. One must bear in mind Mitchell's (1985:§3957) important caveat:

> Future syntacticians of OE will have to take care that they do not spend happy years programming a computer to produce detailed analyses of OE texts only to find themselves in complete agreement with the computer when it tells them what they have told it. At their present stages of development at any rate, the brain of the scholar is both more speedy and more sensitive for OE syntactical analysis than any computer.

Thus, a computer supplied with efficient software may only prove helpful in quick finding usage examples. However, it remains a linguist's sole responsibility what he or she does with the search results. Only the linguist — and never the computer — can be blamed if the analysis goes wrong.

References

Bosworth, J. and Toller T.N. (1898): *An Anglo-Saxon Dictionary*. London: Oxford University Press.

Brown, W.H. (1970): *A syntax of King Alfred's 'Pastoral Care.'* The Hague: Mouton.

Mitchell, B. (1985): *Old English Syntax.* Vol. 2. Oxford: Clarendon Press.

Palmatier, R.A. (1969): *A Descriptive Syntax of the Ormulum*. The Hague: Mouton.

Shannon, A. (1964): *A Descriptive Syntax of the Parker Manuscript of the Anglo-Saxon Chronicle from 734 to 891*. The Hague: Mouton.

Source

Molencki, R. (1999): "Brains *versus* software: new possibilities and limitations of computer assisted historical studies of English syntax." In: Banyś, W., Bednarczuk, L. and Karolak, S. (eds.): *Studia lingwistyczne ofiarowane Profesorowi Kazimierzowi Polańskiemu na 70-lecie Jego urodzin*. Katowice: Wydawnictwo Uniwersytetu Śląskiego, 280—284.