DE GRUYTER OPEN

***ICAME Journal*, Volume 42, 2018, DOI: 10.1515/icame-2018-0002**
Reprinted from *ICAME Journal* 13 (1989): 16–19.

Three problems connected with the use of diachronic corpora[1]

Matti Rissanen, University of Helsinki

In compiling and testing the diachronic part of the Helsinki Corpus of English Texts, our project group has come across three problems which arise from the use of computer corpora in studies of syntax and vocabulary. While these problems are mainly associated with work on diachronic corpora, they may be universal enough to deserve somewhat more general consideration. They could be called "The philologist's dilemma", "God's truth fallacy", and "The mystery of vanishing reliability". The first could be described as pedagogical, the second methodological and the third pragmatic.

"The philologist's dilemma" pertains to the very essence of the use of text corpora in linguistic or philological research. Particularly in the historical study of language, there is a risk that corpus work and computer-supported quantitative research methods will discourage the student from getting acquainted with original texts, from being on really intimate terms with his material and thus acquiring a profound knowledge of the language form he is studying. In the extreme case, this might mean the wane of philologically oriented language studies and result in a great impoverishment in the field of the historical research of language. We would soon be missing the scholars who have a solid, semi-intuitive knowledge of Old and Middle English, based on an extensive reading of original texts. Unquestionably, scholars of this type are the best guarantee of the continuous advancement of our knowledge of the earliest stages of English.

The best way to avoid this risk of impoverishment is constantly to remind ourselves and our students of the importance of reading the texts which form the corpus. Students ought to be trained to see wider textual and extralinguistic contexts, to get a glimpse of the author and society behind the text. It should be our duty to emphasize that, first and foremost, the computer only stores sets of data and organizes and lists them rapidly and efficiently. In the analysis, synthesis and conclusions, the machine does not replace the human brain. We will be able to ask the right questions, draw inferences and explain the phenomena revealed by our data only if we develop a good overall mastery of the ancient language form we are studying.

In teaching our students to use computer corpora, either individually or in class, we should give sufficient attention to the description of the texts on which the corpus is based. Ideally, a set of these texts should be available in or near the location of the computer facilities, in paper copy and, if possible, in original editions. Some information on the source texts, their character and availability should be included in the corpus manual.

"God's truth fallacy" is, in fact, closely related to the problem discussed above, because it, too, pertains to the student's attitude to his corpus as a research tool. An authoritative corpus may easily create the erroneous impression that it gives an accurate reflection of the entire reality of the language it is intended to represent. This risk is particularly acute with a historical corpus as we are not intuitively aware of its limitations in the same way we are with corpora containing present-day language. If a corpus is intended for one research purpose only, the ill effects of this fallacy are not remarkable, but if it is intended to offer a basis for a variety of studies over an extended period of time – as most corpora are – we ought to be aware of this problem.

One way to avoid the "God's truth fallacy" is to keep the corpus open-ended – to structure it in a way that makes improvement and supplementation easy and uncomplicated. If this can be effected in a way that constantly reminds the user of the unfinished and unclosed state of the corpus, so much the better. Once again, a careful description of the texts, in the manual or in other appropriate contexts, may help to remind the user of the scope and necessary limitations of the corpus: what kind of genres and levels of language he may find in it and, even more significantly, what types of language are not included.

Inevitably, there are problems in keeping a corpus open-ended. The most obvious of these is that the results based on earlier and later versions of the same corpus are not directly comparable. But I regard this as a lesser evil in comparison to the idea of a (necessarily) limited and one-sided corpus giving skewed results and fettering research for decades. In this time of easy communication and ever-improving computer facilities with on-line services, updating the old version and distributing the new one is a simple task. Revised corpus versions would not, of course, be introduced every year; five-year intervals might be appropriate and realistic.

"The mystery of vanishing reliability" is connected with the detailed textual coding attached to, e.g., the Helsinki Corpus. Perhaps paradoxically, this fine-meshed coding, which we have considered an important aim in our corpus project, may also become a problem. The number of parameter values is, of course, inversely proportional to the amount of evidence in each information area sampled. For this reason, particularly in a corpus divided both according to chronology and text type, it may be difficult to maintain the reliability of the quantitative analysis of less frequent syntactic and lexical variants. The problem becomes even more obvious if attention is paid to sociolinguistic parameters.

This problem is discussed at some length by Merja Kytö and myself in an earlier report[2] and there is no need to repeat the details of that discussion in the present context. We point out that the best way to cope with this problem would be to compile very large corpora (cf. the success of the gigantic Birmingham University International Language Database), but the restrictions of the hardware and software available for linguists set certain limits to the size of the corpus.

Another solution we offer to this problem, applicable to a text-type-sensitive diachronic corpus, is to classify and code the texts according to text categories containing more than one type of text. These larger categories aim at diachronic representativeness and are still highly experimental. In our report, we enumerate nine "diachronic text prototypes". After further study and experiments, we have reduced the number of the categories to five; two of these are divided into two subcategories.[3] To give our diachronic prototypes some theoretical coherence, we are now using an application of Egon Werlich's text type division as the basis of our grouping.[4]

At the moment, our prototypical text categories are the following:

directive (laws, documents)
instructive:

– secular (handbooks, recipes, etc.)
– religious (homilies, sermons, rules, etc.)

argumentative (trials, etc.)
narrative:

– non-imaginative (chronicles, diaries, biographies, etc.)
– imaginative (fiction, romances, etc.)

expository (scientific treatises, philosophy, etc.)

All categories except the argumentative include texts dating from Old, Middle and early Modern English. There are, of course, interesting text types which have not been grouped under our prototypical categories: the most important of these are private and official correspondence and drama texts. Our diachronic corpus also contains samples of Bible translations from Old English to the Authorized Version and translations of Boethius' *De Consolatione Philosophiae* from King Alfred through Chaucer to Queen Elizabeth.

This categorization has, for better or worse, been included in the coding scheme of our corpus. We still regard it as preliminary and liable to further changes. We hope to find out, through pilot studies, whether it is useful synchronically and diachronically – in other words, whether the texts grouped under one and the same category label share relevant linguistic and textual features.

Notes

1. This is a revised version of a talk given at ICAME 9TH, Birmingham, May 1988.
2. "The Helsinki Corpus of English Texts: Classifying and Coding the Diachronic Part", *Corpus Linguistics, Hard and Soft: Proceedings of the Eighth International Conference on English Language Research on Computerized Corpora,* eds. M. Kytö, O. Ihalainen and M. Rissanen, 169–179, Amsterdam: Rodopi, 1988.
3. The group working particularly on this problem includes Merja Kytö, Anneli Meurman-Solin, Terttu Nevalainen, Helena Raumolin-Brunberg and Matti Rissanen.
4. E. Werlich, *A Text Grammar of English.* Heidelberg: Quelle and Meyer, 1983 (1982).