Revue québécoise de linguistique

Corpus-Based Studies in English: Papers from the Seventeenth International Conference on English Language Research on Computerized Corpora (ICAME 17, Stockholm, May 15-19, 1996), M. Ljung et al., Amsterdam, Rodopi. viii + 388 p.

Tom Cobb

Volume 27, Number 2, 1999

À propos des catégories syntaxiques

URI: https://id.erudit.org/iderudit/603180ar
DOI: https://doi.org/10.7202/603180ar

See table of contents

Publisher(s)

Université du Québec à Montréal

ISSN

0710-0167 (print)
1705-4591 (digital)

Explore this journal

Cite this review

Cobb, T. (1999). Review of [*Corpus-Based Studies in English: Papers from the Seventeenth International Conference on English Language Research on Computerized Corpora (ICAME 17, Stockholm, May 15-19, 1996)*, M. Ljung et al., Amsterdam, Rodopi. viii + 388 p.] *Revue québécoise de linguistique*, *27*(2), 187–191. https://doi.org/10.7202/603180ar

Revue québécoise de linguistique, vol. 27, n° 2, 1999, © RQL (UQAM), Montréal

CORPUS-BASED STUDIES IN ENGLISH: PAPERS FROM THE SEVENTEENTH INTERNATIONAL CONFERENCE ON ENGLISH LANGUAGE RESEARCH ON COMPUTERIZED CORPORA (ICAME 17, STOCKHOLM, MAY 15-19, 1996)

M. Ljung et al., Amsterdam, Rodopi, viii + 388 pp.

Tom Cobb
Université du Québec à Montréal

The role of the computer in modern science is well known. In disciplines like physics and biology, the computer's ability to store and process inhumanly large amounts of information has disclosed patterns and regularities in nature beyond the limits of normal experience. Similarly, in language study the computational analysis of large texts reveals facts about language that are not limited to what people can experience, remember, or intuit. In the natural sciences, however, the computer merely continues the extension of the human sensorium which began two hundred years ago with the invention of the telescope and microscope. But language study did not have its telescope or microscope. The computer is its first analytical tool, making feasible for the first time a truly empirical science of language.

The details of this new empiricism are still being worked out, mainly at conferences rather than in books or journals. *Corpus-Based Studies in English* contains selected papers from the seventeenth International Conference on English Language Research on Computerized Corpora (ICAME 17), held in Stockholm in 1996. This review of the conference's proceedings will sample from the fruits of the new empiricism, as well as its issues, procedures and problems, mainly for the benefit of applied linguists who are curious about the computational end of the field, but who do not follow what goes on there. I will not give a one-liner on each contribution, which is provided in the book's preface, but will rather explore a handful of themes in slightly more depth. These themes deal mainly with applied rather than theoretical questions (the series title is "Studies in Practical Linguistics"). The examples are almost entirely

from English, provided mainly by Germans and Scandinavians, and all the ideas and methodologies are ripe for adaptation to the study of French.

The practical focus of many of the contributions is language pedagogy. A corpus finding that has impacted strongly on English teaching is the importance of lexicalized phrases in language use and acquisition (also known as "formulaic expressions" or "chunks" (discussed in Nattinger and De Carrico 1992). In contrast to the slot-and-filler grammars attributed to Chomsky, where any noun can fill the slot wherever NP is indicated in the tree chart, it now seems clear that only one form of one noun will fill certain slots. For example, you can say "He got cold feet and refused to sigh", but not "He got cold legs and refused to sign" or "He got a cold foot and refused to sign". The analysis of large text corpora has shown that such restrictions are rather more common than unaided intuition would suggest. If language learners apply rules freely and productively, they will often end up with sentences that are grammatically acceptable but idiomatically unusual.

How do we produce an inventory of these lexical phrases in a language and determine the degrees of freedom that particular phrases have ? Barkema's piece discusses a procedure for doing this. Briefly, using a syntactically parsed corpus of adequate size, the linguist can examine the degree of syntactic flexibility of phrases like *red tape* or *wet blanket* by extracting from the corpus all instances of the following pattern: *Premodifying adjective (absolute form) plus singular noun as head of noun phrase.* This output, once alphabetized and viewed in concordance format (i.e. phrase plus immediate context) will show whether or not English speakers ever say *The project was tied up in miles of bright red tape* or *This news dropped two wet blankets over the dinner party.* The verdict for language learners is that while *bright red tape* and *two wet blankets* are possible, they are extremely infrequent, and best left to native speakers. Learners should treat *wet blanket* and *red tape* as fixed and immutable.

Another practical focus is the study of translation through parallel corpora (of translated texts, normally appearing in concordance format on a horizontally split screen). Schmied and Schäffler use the Chemnitz German-English translation corpus to look at the phenomenon of "translationese", whereby texts that have been translated show systematic differences from texts written in the target language. They argue that while some of these differences may stem from particular differences between source and target languages, others are universal features of the translation process. Two of these are explicitness and condensation or "showing more underlying elements on the surface" (or fewer). For instance, the large number of non-finite verb constructions possible in English are not available in German, and to be translated, must be broken down

into relative clauses specifying agency, tense and other information that is implicit in the English infinitive and participle. The writers find more instances of explicitness than condensation in their translation corpora and offer an information-processing account for why this should be so. Since all their translation data deals with only one pair of languages, they will presumably want to look at other pairs of languages before making a final commitment to their universal hypothesis. In the meantime, Chomskyans may take comfort from knowing that interest in deep and surface structures and various kinds of universalism are alive and well among corpus linguists.

Chomskyans will also find familiar the corpus linguists' occasional interest in linguistic invisibles. The invisibles, as ever, are implicit traces of deep structure, for example the elided relative pronoun in the sentence *The dog I bought died* (i.e. *The dog that I bought died*). Lehmann's piece describes a way of inserting a placeholder Ø between each two NPs in a corpus where there could have been a relative pronoun. This insertion is simple in a fully tagged corpus, of course, where every grammatical element and relation has been fully marked with a tag-set, as for example in *The_artDefdog_nounCommon Countable that_relElided, etc.).* In this case, a simple search for *relElided* would bring forth all instances of the phenomenon for inspection. But Lehmann is interested in working with more natural tests that have been parsed only with an automatic tagging system (which does not attempt to assign phrase-level tags).

What practical purpose is served by assigning all these Ø's ? Lehmann is interested in machine translation. A problem that has plagued automatic translation between English and several other languages is that elided relatives are permitted in English in certain conditions, but not permitted in French or German under any conditions. In French, one cannot say **Le chien j'ai acheté est mort*, nor in German **Der Hund ich kaufte ist gestorben*, but those are what the translation module will generate for *The dog I bought is dead.* Lehmann has worked out a way of searching through an English text for the eight condition-sets where relatives may be omitted, and inserting Ø in each. With Ø's inserted in the English text, it can be passed to a machine translation system which will replace Ø with *que* or *qui*, as appropriate.

Another echo from the past unexpectedly encountered in this volume is the grammaticality judgement task, so derided in the early days of corpus analysis (when it seemed hard facts would supplant soft intuitions entirely). No one is building any empires on whether subjects will grant grammaticality to *Colorless green ideas sleep furiously*, but there is still a role for grammatical judgement in certain cases. Mönnink describes a problem she has had in attempting to write a corpus-based descriptive grammar, which is that even in

corpora of substantial size there is a very low representation for some structures that native speakers would instantly judge to be grammatical. Taking as an example the NP, whose grammar-book formula is *optional determiner + Ø or more premodifying elements + obligatory head + Ø or more postmodifying elements*, Mönnink shows that several legitimate changes may be rung on this theme which are entirely legitimate and yet which will appear very infrequently in a corpus of reasonable size. Such changes include shifted premodifications ("I wouldn't give it *so romantic a name"*), discontinuous modifications ("We can do *as much* guessing about her *as we please*"), floating postmodification ("*Much evidence* has accumulated *concerning cytoplasmic* DNA"). These forms are clearly decent English, and yet a purely frequency-based approach to constructing a descriptive grammar might underplay or omit them.

The writer proposes supplementing corpus information about such NP's with information from a principled set of elicitation tasks. One such task might be evaluation (rate from 1= perfectly acceptable to 7= not at all acceptable this sentence: *I never saw so beautiful a person*). Another might be composition (give all possible sentences that can be constructed from these phrases: *was made, to win a medal, today, no effort*). A methodology for blending frequency and elicitation information is presented.

A small quibble with Mönnink's piece is that the corpus in which she finds legitimate structures under-represented is a smallish corpus of only about 120,000 words in four text genres. She might find less need for experimental supplementation, and all the vagueness that introduces, were she to consult a larger source such as the British National Corpus, currently weighing in at 100 million words and growing every day. On the other hand, smaller corpora and home-grown corpora have their uses (for an excellent example see Granger 1998, who works with a corpus of learner English produced by Belgian francophones), and in these cases a purely frequency-based approach will often be usefully complemented by more judgement-based elicitation data.

The three classics of corpus study are all represented in this volume: corpus comparisons of older English and newer English, of written English and spoken English, and of British English (BE) and American English (AE). Only the latter will be discussed. The point of BE-AE comparisons is to find evidence whether BE and AE are different, as they intuitively sound as if they are. A claim in need of support is, for instance, that the phrase *Steve Forbes, political neophyte* "is an apposition type more characteristic of AE than BE". However, a perceived problem with such studies, and more or less the opposite of the problem discussed just above concerning over-reliance on objective data, is that the data behind the AE-BE comparisons may not be objective enough.

Kretzschmar, Meyer, and Ingegneri argue that the sampling procedures by which corpora of AE are put together do not meet the standards necessary to allow the statistical inference of representativeness. Indeed, to support such inferences would require that linguists have the resources of large political polling organizations or indeed, of the US Federal government. Until then, all we know about *Steve Forbes, political neophyte* (linguistically speaking) is that it is a phrase produced at least once in an AE publication and widely understood by AE speakers (but then, by BE speakers, too).

The reader is invited to read the book itself for more details on the studies I have reported, and for all the details on the no less interesting studies I have not reported for lack of space. The sense I take away from this volume is that corpus investigation deals with extremely interesting questions and relies on hard evidence as much as possible. However, I also see that as the discipline matures, some of the problems with getting and using hard evidence are presenting themselves, and that some of the lines that once seemed so clear between old and new linguistics are blurring.

References

British National Corpus 1999 Online at *http://info.ox.ac.uk/bnc*

GRANGER, S. 1998 *Learner English on computer* London, Addison Wesley Longman.

NATTINGER, J. and J. DECARRICO 1992 *Lexical phrases and language teaching,* London, Oxford University Press.