Book Reviews

Theory and Practice in Corpus Linguistics

Jan Aarts and Willem Meijs, editors
(University of Nijmegen and University of Amsterdam)

Amsterdam: Editions Rodopi, 1990,
iii + 254 pp. (Language and Computers:
Studies in Practical Linguistics 4)
Paperbound, ISBN 90-5183-174-9, $37.50

Reviewed by
Kenneth Ward Church
AT&T Bell Laboratories

Corpus linguistics is a hot topic, and for good reason. Text is more available than ever before. And consequently it is easier to use corpus data more effectively than it was in the 1950s, the last time that empiricism was in fashion.

Corpus linguistics is such a hot area that it is already splitting up into a number of different sub-areas. *Theory and Practice in Corpus Linguistics* focuses on a direction practiced in much of the U.K. and Scandinavia. This work has produced a number of part-of-speech taggers and parsers based on probabilities derived from corpus data. These programs work on unrestricted texts, with reasonable accuracy and efficiency. A good example of this approach is Garside, Leech, and Sampson (1987); see Lesk (1988) for a very positive review of this book and a strong endorsement of the approach that it represents.

One might contrast the view(s) held by Garside, Leech, Sampson, Lesk, and others with the AI tradition of using semantic networks and knowledge representation techniques to address lexical questions. Evens (1988), a volume from the ACL book series, is a good example of the AI approach to computational lexicography. The knowledge-based approaches tend to assume that the representation is central, and that much of the knowledge has to be entered into the system by hand, in contrast with probabilistic-based approaches where much of the knowledge is acquired through various training procedures that fit certain parameters to corpus data.

Yet a third quite distinct approach can be found among lexicographers, who have recently become interested in computational issues because of the success of the *COBUILD* dictionary:

> For the first time, a dictionary has been compiled by the thorough examination of a representative group of English texts, spoken and written, running to many millions of words. This means that in addition to all the tools of the conventional dictionary makers — wide reading and experience of English, other dictionaries and of course eyes and ears — this dictionary [*COBUILD*] is based on hard, measurable evidence. (Sinclair et al. 1987)

The experience of writing the *COBUILD* dictionary is well documented in Sinclair (1987), a collection of articles from many of the participants of the COBUILD project; see Boguraev (1990) for a strong positive review of this collection.

Like Sinclair (1987), *Theory and Practice in Corpus Linguistics* is also a collection of papers by participants in an area of corpus linguistics that may be revolutionizing the way we think about language. However, I would not expect this collection to stand up as well to the test of time. Many of the articles are timely, interesting, and

controversial. But, on the other hand, much of the work is still very much in progress. I would strongly recommend the book to researchers who are actively involved in corpus linguistics and computational lexicography. However, I would not recommend it to people looking for a good overview of the field. This collection reads more like a conference proceedings (which it is) than like a book. The papers are presented in alphabetical order by first author's last name, after a short two-page preface that starts out: "Like its predecessors... this new volume presents a kaleidoscope of recent developments in this field."

It is really hard to come up with useful unifying trends among the 11 papers. They cover such a wide range of sub-areas. Nevertheless, eight of the papers can be assigned to four topics, with two papers in each topic.

1. Corpus analysis on a small computer (e.g., an Apple or an IBM PC);
2. Corpus analysis (Sampson) vs. theoretical linguistics (Briscoe);
3. Corpus analysis and collocation;
4. Corpus analysis and discourse structure.

1. Corpus Analysis on a Small Computer

Two papers advocate the use of small computers for analyzing corpus material. One of them points out that a small computer today has much better turnaround time than mainframe computers of yesterday. Both papers are extremely enthusiastic about the possibilities of interactive concordance programs and so forth. I think it is important to point out that it is now possible for almost anyone to use large corpora. You no longer need an expensive computer center to look at a concordance. It is easier than ever before to look at data. And it isn't even expensive.

On the other hand, I do wish that the papers were a bit less enthusiastic. While PCs do provide a lot of computer power for less than the price of a car, they do not solve all of the world's problems. I fear that PCs may be not the best way to deal with the larger corpora. I also fear that PCs may not be the best way to explore various statistical possibilities. I may be old-fashioned (and spoiled rich), but I still believe that it is easier to work with something a little bit more powerful than a PC.

2. Corpus Analysis vs. Theoretical Linguistics

There are also two papers debating corpus analysis as practiced at Leeds and Lancaster as compared with a more "traditional" view. In a previous conference, Sampson (Leeds) has observed that phrase structure rules have a very skewed distribution. In particular, if you look in a typical corpus, according to Sampson, you will find very many instances of a few phrase structure rules, and just a few instances of a large number of phrase structure rules.

I actually don't find this surprising at all. All kinds of language facts have "Zipf-like" distributions (in which the probability of a type is roughly inversely proportional to its rank). Words are the classic example. A few function words (e.g., *the, a, of*) are very common, and many words occur just once in any corpus that you look at. In

fact, Zipf's law holds for all kinds of type–token distributions, such as the allocation of people to cities and the allocation of income to people. It is an empirical law that has fascinated statisticians for decades. The distribution keeps coming up in nature, but unlike other distributions, such as the normal distribution, it is not clear just what randomness assumptions give rise to it. (In addition, the law, as stated, can't be exactly right, since probability should sum to 1, but the integral of $1/r$ doesn't converge.)

In the present discussion, Clive Souter (also at Leeds, but somewhat removed from Sampson) suggests that the Zipfian distribution implies the hopelessness of the "traditional" approach to writing grammars. Advocates of the Leeds school would argue that it will take a lot of effort to enumerate all of the phrase structure rules in the language, because there are a large number that rarely occur (assuming a Zipf-like distribution). They then turn to alternative methods such as *simulated annealing*, that have a bit of a self-organizing flavor (cf. Jelinek 1990). (Apparently, there are some important differences within the Leeds school; although Sampson and Souter both endorse simulated annealing, Souter is sympathetic to self-organizing and/or connectionist methods, while Sampson is not.)

> The formalizing of systemic functional grammars for use in parsing rather than generating natural language can be achieved by extracting rules from suitably annotated English corpora. The very large size of such grammars puts into question the real value of building small 'competence' grammars by hand, particularly if the grammar is to be used in the parsing of relatively unrestricted English, which is a long-term goal of the COMMUNAL project. The frequency distribution of extracted rules, as well as words, adheres to Zipf's law. This open-ended characteristic of the extracted grammars suggests that a probabilistic parsing technique which employs frequency data from the corpora may be most suitable for the parsing of unrestricted English. (Souter, p. 195)

Briscoe counters by denying the Zipfian assumption. He argues that Sampson's argument has a failure-to-find fallacy. While any particular grammar, such as Sampson's or Souter's, may have a Zipf-like distribution, it is possible (though maybe not very likely) that there is another grammar that does not. Technically, one cannot rule out the possibility of such a grammar just because one did not happen to find it. I'll grant Briscoe the technical point and admit that it has not yet been proven that grammar must have a Zipfian distribution. Nevertheless, I am basically convinced that grammar probably does have a Zipf-like distribution even though I don't know how to prove it. It just doesn't seem very likely to me that there could be a small, nicely distributed set of phrase structure rules that would adequately describe grammar as it is actually used in practice. The performance grammar will probably have to be at least as large as the concise version of Quirk and Greenbaum (1973), a book of 500 pages.

On the other hand, while I agree with Sampson and others at Leeds in granting that grammar probably does have a Zipfian distribution, I do not accept the rest of their argument. I believe that it may still be practical to describe grammar with traditional methods, even though the performance grammar may be large and the distribution may be skewed. Lexicographers have managed to do a fairly good job of describing words (without explicit probabilities), and the set of words is large and the distribution is skewed. Of course, it will require a lot of hard work and a lot of drudgery. But I believe it can be done, given a monumental effort like Murray's *Oxford English Dictionary* project. I would rather bet on hard work than speculate on a silver bullet like simulated annealing. I would be particularly suspicious of Souter's attempts to appeal to self-organizing and/or connectionist methods as an alternative to hard work.

I would draw a different conclusion from the observation that grammars are large and their distribution is skewed. I think this observation points out the need to adopt more efficient methods of collecting and evaluating evidence. Lexicographers have developed methods for writing dictionaries that resemble an expedition-style assault on Mt. Everest. The whole enterprise is considered development, not research. Milestones are set and taken very seriously. A lot of time is spent at the beginning of the project on time-and-motion studies to make sure that the procedures are reasonably efficient and that the milestones are realistic.

I suspect that computational linguists should build grammars with more or less the same procedures. So far, most of the discussion in the literature has been on the form of the grammar: should we use unification grammars or probabilities? I suspect that these issues may be a bit of a red herring. I am much more concerned with how we are going to speed up the process of collecting and interpreting the evidence. We need to work out more efficient procedures, since it is going to be a big job and we have very limited resources. In the long term, at least, we have the responsibility to deliver a large grammar with broad coverage for unrestricted text. We need to start thinking now about how we could ever hope to achieve this long-term goal.

3. Corpus Analysis and Collocation

There are also two papers on collocation, one of the central problems in corpus linguistics. The introduction to Kjellmer's paper describes the problem very well.

> It is a common observation that words... tend to occur in clusters.... [I]t is not surprising, therefore, that the large-scale study of... collocations, made possible by... computers, has come to be seen as more and more important over the last few decades.... This is shown by the number of projects... devoted to the study of collocations, and also... by the... emphasis... in recent... dictionaries. There is a world of difference... between... the *Concise Oxford Dictionary* and the recent *COBUILD, Longman,* and *Oxford Advanced Learner's* dictionaries.

Kjellmer then goes on to study the distribution of collocations in the Brown Corpus. It is very difficult, though, to study collocations with such a small corpus. A million words is simply not enough to see interesting word pairs. In a million words, one can find a few of the more common two-word collocations such as *of the* (e.g., page 13 of Altenberg and Eeg-Olofsson's paper), but you can't possibly see the full range of two-word noun phrases such as *red herring* and verb + prep combinations such as *give up*. I suspect that it will require a huge amount of text to learn simple facts such as these. It might require even more text to learn syntax.

All 11 papers in this volume used very small corpora, ranging from a few hundred words up to one million words. In contrast, the COBUILD project used a corpus of 20 million words. These days, a million-word corpus is small; a corpus has to be at least ten times larger to be considered large. And there are many corpora in use today that are a hundred times larger than the Brown Corpus. None of the papers in this collection discussed a corpus that could be considered large by today's standards.

This fact seriously undermines some of the papers. For example, consider the two papers that argue for the use of small computers (e.g., an Apple or an IBM PC) in corpus analysis work. It is clear that such a small computer is appropriate for a small corpus. But will a small computer suffice for a reasonably sized corpus? Neither paper even considers the question. Many readers are likely to be very interested in this question, since they will have access to large corpora such as the British National Corpus or the material to be distributed by the ACL's Data Collection Initiative. Can

they expect to work with so much text on a small computer? I'm not sure what the answer is, but I'm sure that it is an important one that will come up again and again.

And dialectal variation adds another dimension that will surely consume even more text. Ossi Ihalainen, for example, begins his paper by citing Nelson Francis's observation in 1983 that there has been very little work on dialectal syntax (e.g., distributions of peculiar progressives such as *she was sat in that chair*) because such a study would require larger samples of language than have been available so far. He then cites Labov as saying more or less the same thing in 1970. After reading this introduction, I was expecting him to come up with a much larger corpus of some tens of millions of words, and conclude that with a larger corpus he could make observations that were not possible in 1983 and in 1970. Unfortunately, he did not have such a rabbit to pull out of his hat. Rather he had to make do with several small corpora, no bigger than what was available in 1983 and in 1970. I must say that I was impressed with what he was able to do with these tiny corpora, but I would like to believe that there is a reason why empiricism is back in fashion. My personal hunch is that the approach is technically more feasible than it was just a few years ago, and that is why it is back in fashion.

Although I haven't discussed all 11 papers, let me stop here. While many of my remarks may sound critical, I do not intend them to sound negative. I found all of the papers extremely thought-provoking. I would not have enjoyed the book so much if I agreed with all of it.

Let me end with one final quibble: I don't like the title *Theory and Practice in Corpus Linguistics*. The term *theory* seems badly out of place. Perhaps, the volume should have been called *Practice and Practice in Corpus Linguistics*. The papers do a fairly good job of describing, by example, how corpus linguistics is practiced, at least in parts of the U.K. and Scandinavia. The absence of papers from my country (the United States) is striking, though perhaps appropriate, given our history of objecting so strongly to empiricism.

References

Boguraev, Branimir, reviewer (1990). Review of Sinclair (1987). *Computational Linguistics*, **16**(3), 184–186.

Evens, Martha, ed. (1988). *Relational Models of the Lexicon*. Cambridge University Press.

Garside, Roger; Leech, Geoffrey; and Sampson, Geoffrey, eds. (1987). *The Computational Analysis of English: A Corpus-Based Approach*. Longman.

Jelinek, F. (1990). "Self-organized language modeling for speech recognition." In *Readings in Speech Recognition*, edited by A. Waibel and K. Lee. Morgan Kaufmann Publishers.

Lesk, Michael, reviewer (1988). Review of Garside et al. (1987). *Computational Linguistics*, **14**(4), 90–91.

Quirk, Randolph and Greenbaum, Sidney (1973). *A Concise Grammar of Contemporary English*. Harcourt Brace Jovanovich.

Sinclair, John; Hanks, P.; Fox, G.; Moon, R.; and Stock, P., eds. (1987). *Collins COBUILD English Language Dictionary*. Collins.

Sinclair, John, ed. (1987). *Looking up: An Account of the COBUILD Project in Lexical Computing*. Collins.

Kenneth Ward Church received his Ph.D. in computer science from MIT in 1983, and then went to work at AT&T Bell Laboratories on problems in speech and natural language. Recently, he has been advocating the use of statistical methods for analyzing large corpora. He is a member of the ACL Data Collection Initiative. He is presently on sabbatical at: USC/Information Sciences Institute, 4676 Admiralty Way, Marina del Rey, CA 90292; e-mail: church@isi.edu. After September 1991, his address will be: Room 2d444, AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, NJ 07974; e-mail: kwc@research.att.com.