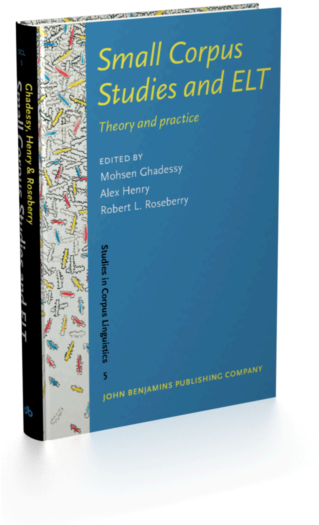
Preface

John McH. Sinclair

https://doi.org/10.1075/scl.5.01sin


Small Corpus Studies and ELT
Theory and practice
EDITED BY
Mohsen Ghadessy
Alex Henry
Robert L. Roseberry
Studies in Corpus Linguistics
5
JOHN BENJAMINS PUBLISHING COMPANY
Ghadessy, Henry & Roseberry
Small Corpus Studies and ELT

Pages vii–xvi of
Small Corpus Studies and ELT: Theory and practice
Edited by Mohsen Ghadessy, Alex Henry and Robert L. Roseberry
[Studies in Corpus Linguistics, 5] 2001. xxiv, 419 pp.

Preface

J.M.Sinclair

There are several reasons why it is a pleasure to write a preface to this collection of papers showing what can be done with small corpora. For one thing, many of the authors are personal friends, former colleagues and students who have chosen to work with corpora, and I hope that the encouragement of this work in Birmingham and the provision of the resources they found there may have helped them to focus on corpus linguistics. Another reason for wanting to identify with small corpora is that since my own main work over many years has been with the largest available corpora of the period, some commentators have concluded that I have a low opinion of small corpus work. I hope that not only this preface, but my co-authorship of one of the papers herein, will put paid to such nonsense. The type of work will certainly be constrained by the corpus size, but that has nothing to do with the quality.

Let me first of all review the way in which the dimensions of corpora have developed since they first became available for study; then deal briefly with large corpora, and why they are necessary at all, for I am unrepentant of my constant pressure to make more language data available. From this discussion it can be seen what jobs are well suited to small corpora, and what jobs are better left to large corpora. Then I will point to the differences between language data seen as a set of texts and seen as a corpus; and finally celebrate the genuine strengths of small corpora.

The incredible shrinking corpora

In the early sixties work began on both sides of the Atlantic on corpora to be held in electronic form and to be processed by computer. The two projects were unaware of each other until 1965. At Brown University in Rhode Island, USA, Nelson Francis and Henry Kuçera set out to compile a million-word sample of published American English of the year 1961, which was the last year for which publication records were available when they began their work (Francis and Kuçera 1979). Meanwhile at the University of Edinburgh in Scotland, a corpus was taking shape, consisting of transcriptions of informal spoken interaction among speakers of British English. Because of the greater amount of labour involved in spoken language collection, this corpus only reached some 300,000 words, of which 166,000 were regularly used in research. Edinburgh University did not possess a computer at the time, so arrangements were made with Manchester University to carry out preliminary processing; the project moved to Birmingham in 1965 (Sinclair et al 1970, Jones and Sinclair 1973).

These corpora were simultaneously the largest and smallest of their type, being the only ones, and they remained so for a number of years. In fact the Brown size and structure proved to be well-judged, and was followed by an equivalent corpus of British printed English for the same year (LOB — the Lancaster-Oslo-Bergen Corpus, Johansson et al. 1978). In 1982 Yang Hui-Zhong compiled JDEST (English for Science and Technology, Shanghai Jiao Tong University, Yang 1985) following the Brown size and structure, and several were compiled in Birmingham in the seventies, beginning with Peter Roe's (Roe 1977); these kept to the size of a million words but contained full texts, rather than the 2000-word samples of the Brown style.

The million-word tradition continues, and there are now modern equivalents of Brown and LOB compiled at the University of Freiburg in Germany, and called “Flob” and “Frown”. They are identical in size and structure with Brown. All of these corpora, and several more, can be found on a compact disk published by ICAME, the International Computer Archive of Modern English, which has pioneered the archiving of language data for over twenty years (Hofland et al. 1999).

But million-word corpora must now be regarded as small by today's standards. As the power and storage capacity of computers has risen sharply, and very large amounts of text have become available in electronic form, they are dwarfed by the current norms; what were large corpora have become small ones — yet the first seemed to contain unheard-of riches when it was first released to the world.

Now the largest available corpus is The Bank of English[1], some 410 times the size of the Brown family. Spoken corpora have also grown very fast despite the problems they pose; the first one to be widely used, the London-Lund corpus, was about 500,000 words in length (Svartvik and Quirk 1980); now the Bank of English contains 50 million words of transcribed speech, about 300 times the size of the original Edinburgh corpus. Nevertheless, one spoken corpus is living proof that small can be beautiful — the Longman/Lancaster Spoken English Corpus contains only 53,000 words (Taylor and Knowles 1988).

So there is a kind of relativity in corpus sizing — the dimensions of a "small" corpus vary with the date it is compiled; the apparently massive corpora of a few years ago are now perceived as tiny, and in another decade or two, anything less than a few billion words will count as a small corpus, because there is every reason to make bigger and bigger corpora, and the job becomes easier as the size goes up.

No data like more data

This slogan comes from the Linguistic Data Consortium (LDC), established in the USA a decade ago, and devoted to distributing language data, whether organised as corpora or not. The United States, after the brilliant start given by the Brown Corpus, then lagged behind Europe for twenty years (indeed the American National Corpus, cloned from the British one of ten years ago, is at the planning stage as I write), and the impetus to start LDC came because of pressure from statisticians, whose primary need was a large amount of language in order to advance research such as Speech Recognition.

LDC is not a corpus and does not claim to be one; it is an archive and a distribution service, and it contains whatever material, in any language, is made available to it. Its attitude and slogan were important a few years ago because many students of language were so accustomed to working with very small amounts of data, and were unfamiliar with the methods of handling large amounts, that there was a good deal of resistance to the arrival of the cornucopias of language that LDC proposed to gather. One has to remember, too, that the dominant attitude to language in USA over forty years has been concerned with language in the mind, and not language on paper or in the air; hence large quantities of it were simply not required.

Phrasemaking

What do you get from a large corpus that you do not get from a small one? Essentially you get repetitions of multi-word choices in combination. The large number of words in a language, and their characteristically uneven distribution (Zipf 1935) mean that despite the clear tendency of languages to practice coselection, that coselection is subject to so much variation that if one wants to study collocation or phraseology by automatic methods then even the 9-figure corpora are pitifully small.

Try the following test: first fix on a figure of the number of occurrences of a word that you need for a job — maybe fifty for personal study, 250 if the computer will be asked to detect significant patterns in the concordance. Then select a word that is frequent in the corpus but not one of the “grammatical” words which are frequent in any corpus. Note the frequency of your selected word, and also of its most frequent collocate — again leave the grammatical words on one side. It is very likely that the frequency of the pair in collocation with each other is an order of magnitude below the frequency of the single word, unless you have chosen a word like kith, which will not normally occur except with kin. If the frequency of the collocation is still substantial, then add a third word, namely the word which most frequently occurs with the pair. You should drop an order of magnitude again.

In this way you can make an assessment of which features of which words can be studied in any given corpus. For example, if you are studying the contextual patterns of words, then you will almost certainly be looking for repetitions. So it is bad news that around half of the vocabulary of any corpus (ie the number of different word-forms) consists of single occurrences. On the other hand, if you are studying an author’s use of single occurrences as an indication of style, then even a very small collection can be useful because the phenomenon is so common (Marcinkevičinė 1998).

Methodology

It is clear from the above discussion that factors other than being “small” or “large” must be used to distinguish two different kinds of corpus. Until recently they were not perceived as different in anything but size, but there has been a development in thinking, of which this book is the first monument; now "small" and "large" corpora are seen as in contrast with each other. So the difference must be methodological, because it cannot be just size, whether relative or absolute size.

The Editors hint at this towards the end of their Introduction, "the methodologies are … intuitive". A small corpus is seen as a body of relevant and reliable evidence, and is either small enough to be analysed manually, or is processed by the computer in a preliminary fashion, using the kinds of tools presented in Section II; thereafter the evidence is interpreted by the scholar directly. There is no need to collect the quantities of data needed in order to delay the direct participation of the human being.

There is thus a fairly sharp contrast in method; the so-called Small Corpora are those designed for early human intervention (EHI) while the Large Corpora are designed for late or delayed human intervention (DHI). (Of course in DHI the human being is indirectly controlling the process, and the process has probably been built up over many EHI sessions, and the human being must eventually participate in order to interpret the results.[2])

So this book is essentially a celebration of the EHI method. The researchers have a clear goal in mind, and they build a corpus for an investigation, or if they are lucky enough, use one that is already available. The processing is usually with standard tools, so packages like WordSmith are invaluable for EHI, but occasionally these are adapted, or special ones devised for the job.

Text and Corpus

There is another methodological point to be made, to clarify what is meant by calling a collection of texts a corpus. It is still, after all, a collection of texts, so what is different about it when it is seen as a corpus?

By calling a group of texts or text samples a corpus we are investing it with linguistic status. The corpus is gathered on the basis of *external* criteria (Clear 1992), to do with the sociocultural roles of the texts it contains, and the claim is implicitly made that an investigation into the *internal* patterns of the language used will be fruitful and linguistically illuminating. So if it is a general corpus, researchers expect to find in it information about the language as a whole, and if it is a more specialised corpus, then the characteristics of the genre will be discoverable.[3]

A text is a single, unified, meaningful event, an artefact; it is read and studied as such; a corpus is a multiple set of events, and is studied for the similarities and differences among its events, and the component parts of its events. So in principle the same piece of language can be treated as a text or a corpus, and different points will emerge (Tognini Bonelli forthcoming 2000, Introduction).

Working with small corpora, then, is not the same as working with texts, and the papers in this book make that clear, in their innovative research strategies and novel results.

Comparison

The main investigative technique that is used here, and in most EHI studies, is comparison; comparison uncovers differences almost regardless of size. Samples from different genres are typical of EHI work, where the computer detects proportional differences or the presence or absence of particular phenomena, and the researcher interprets this information. The Brown corpus was designed with this kind of study in mind, and although it was also used to establish very basic facts about American published English as a whole, the aspect of comparison has dominated research — the cluster of similar corpora that are mentioned above form various axes of comparison, US/UK English with Lob, 1960s/1990s English with Frown and Flob, etc. One of the most recent flowerings of comparative EHI research is the large Longman Grammar of Spoken and Written English (Biber et al. 1999). The core corpus that informs this work consists of just under 20 million words, in four components of 4–5 million words each, covering conversation, fiction, news and academic prose.

Across languages, the comparative method is a very good starting point, and in recent years *parallel* corpora, sets of translated texts, have become very popular. The texts are aligned with respect to each other, using a variety of simple aligners, or by hand if the corpora are very small, and translation equivalences can then be studied.

Language Teaching

The focus of this collection is the application of small corpus – EHI – research to the business of teaching and learning languages. As the papers show, corpus evidence can illuminate language teaching from many different angles; as well as the comparisons mentioned above there is the accurate description of structure, reliable models of usage, how words and phrases are actually translated, what are the essentials in a syllabus, what are the characteristic errors of learners, etc. Small corpora can be put together quickly for a classroom job or an individual need, and can be honed to very specific genres and sub-genres. Corpus resources can be placed under the control of the students, and "self-access" can take on a new and rich meaning.

The origin of this lively movement is not in corpus linguistics as such, but in the tiny computers of twenty years ago, which in UK were associated with the name of Clive Sinclair (no relation). The early models were quite unable to handle corpora, having memories as small as 1 kilobyte, but with the ingenuity of Tim Johns and others (eg Higgins and Johns 1984) could be given a highly motivating role in language learning. As the small computers gained power, and the microprocessor developed into the PC, then the notions of "classroom concordancing" and "data-driven learning" became popular. These were adaptations of mainframe computer routines to the smaller machines, and the miniaturisation gave rise to Microconcord (Johns 1986), the precursor of WordSmith. At around this time, in the middle eighties, the mainframe was going out of fashion for large-scale data processing, and distributed computing over a network of minicomputers was preferred; in this way, and within the broad heading of language pedagogy, the two communities of linguistic computing converged. There was, and is, still a demarcation line in the operating systems on which software is mounted; the "micro" tradition is dependent on DOS/Windows, whereas the "mini" group use Unix/Linux

The re-emergence, then, of the small corpus has a broader base in applications than just language teaching, and although this book is focused on the teaching side there is mention of other applications, in translation studies, literary studies etc. Indeed the application to literary stylistics is a rather separate line of development, with origins even earlier than the first "general" corpora.

There is no special virtue in being small, except that many scholars like to keep the dimensions of their studies modest in order to be manageable without requiring special expensive equipment or a high level of technical skill. As linguistic computing gets ever easier and more flexible and more powerful, the meaning of "small" will be frequently reinterpreted, and the only distinction that is here to stay is the methodological one, the type of human intervention and the timing of it.

Notes

1. For up-to-date information on The Bank of English, see the Cobuild Web-page, www.cobuild.collins.co.uk.
2. Some resource packages use fully automatic analysis and leave the interpretation to the user — eg Cobuild Collocations CD-ROM. The corpus was mildly pre-processed for reasons like the protection of anonymity, but the results are not edited in any way.
3. Some discussion of these and other relevant points can be found in the EAGLES Project Reports of 1996–see http://www.ilc.pi.cnr.it/EAGLES96/browse.html#wg1, the Corpus Working Group and the files corpustyp.ps and texttyp.ps.

References

Biber, D., Johansson, S., Leech, G.N., Conrad, S. and Finegan, E. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education Ltd.

Clear, J.H. 1992. "Corpus sampling". In *New Directions in English Language Corpus Methodology.* G.Leitner (ed), 21–31, Berlin: Mouton de Gruyter.

Francis, W.N. and Kuçera, H. 1979. *Manual of Information to accompany a Standard Sample of Present-day Edited American English, for use with digital computers*. Providence, R.I.: Department of Linguistics, Brown University.

Higgins, J. and Johns, T.F. 1984. *Computers in Language Learning.* London and Glasgow: Collins ELT.

Hofland, K., Lindeberg, A. and Thunestvedt, J. 1999. *ICAME Collection of English Language Corpora*, Second Edition. The HIT Centre, University of Bergen, Norway. (CD-ROM).

Johansson, S., Leech, G.N. and Goodluck, H. 1978. *Manual of Information to accompany The Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers*. University of Oslo, Department of English.

Johns, T.F. 1986. "Micro-concord: a language-learner's research tool". *System* 14(2): 151–162.

Jones, S. and Sinclair, J.M. 1973. "English lexical collocations: A study in computational linguistics". *Cahiers de Lexicologie* XXIII-II.

Marcinkevičinė, R. 1998. "Hapax legomena as a platform for text alignment". In *Translation Equivalence: Proceedings of the Third European Seminar*. W. Teubert, E. Tognini Bonelli and N. Volz (eds), 125–136, Mannheim: The TELRI Association e.V. and Pescia, The Tuscan Word Centre.

Roe, P. 1977. *Scientific Text*, ELR Monographs no 4. University of Birmingham, Department of English.

Sinclair, J.M., Jones, S. and Daley, R. 1970. *English Lexical Studies*. Final Report to OSTI on Project C/LP/08. University of Birmingham, Department of English.

Svartvik, J. and Quirk, R. (eds), 1980. *A Corpus of English Conversation*. Lund: Lund University Press.

Taylor, L. and Knowles, G. 1988. *Manual of Information to Accompany the SEC Corpus*. UCREL, University of Lancaster.

Tognini Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam and Philadelphia: John benjamins.

Yang Huizhong 1985. "The use of computers in English teaching and research in China". In *English in the World*, R. Quirk and H.G.Widdowson (eds), 86–100. Cambridge: CUP.

Zipf, G.K. 1935. *The Psychobiology of Language*. Houghton Mifflin, reprinted 1965, Cambridge, MA: MIT Press.