Creating and Digitizing Language Corpora

Volume 1: Synchronic Databases

Edited by

Joan C. Beal
Sheffield University

Karen P. Corrigan and Hermann L. Moisl
Newcastle University

Foreword by Shana Poplack
University of Ottawa

Poplack, Shana. 2007. Foreword. In Beal, Joan, Corrigan, Karen & Moisl, Hermann (eds.), Creating and digitizing language corpora. Houndmills: Palgrave-Macmillan. ix-xiii.

palgrave
macmillan

Foreword

Only two or three decades ago, those of us who had the patience and the wherewithal to construct a computerized corpus of recorded speech, however clunky, were the envy of our colleagues. In those days, linguists interested in quantitative analysis simply slogged through their audio-tapes, extracting unfathomable quantities of data by hand. Cedergren, to name but one notable example, analyzed 53,038(!) tokens of phonological variables, culled individually from her tapes, in her 1973 analysis of Panamanian Spanish.

The gold standard for transcribed corpora at the time was the concordance, possessed by a fortunate few, and coveted by all who were doomed to manual extraction. Of course the vintage concordance was largely limited to lexically-based retrieval, but at least it was searchable. The papers that Joan Beal, Karen Corrigan and Hermann Moisl have assembled in these companion volumes are eloquent testimony to how far the field of corpus linguistics – now rife with electronic corpora – has come in so short a time.

Building a corpus arguably involves a greater investment in time, resources and energy than any other type of linguistic activity. Decisions are legion at every stage of the process: sampling, ensuring representativeness, collecting data, transcribing them, correcting, standardizing the transcription, correcting, tagging and markup, correcting, and facilitating retrieval. Adding to the challenge is the fact that at the outset of the project the researcher is often not even familiar enough with the materials to make the best decisions, and changing midstream is costly and time-consuming. What could possibly make such a huge front-end investment worthwhile? Dealing with corpora at every stage of development, from fledgling endeavours to large-scale, heavily exploited enterprises, these reports offer a state-of-the-art synthesis of the problems researchers have encountered and the solutions they have adopted to deal with them.

The focus of these volumes is on *unconventional* corpora, like the non-standard, regional and dialectal varieties of speech, creole texts and child language discussed in Volume 1. Each poses problems hardly imaginable to the early builders of more orthodox corpora based on written or standard materials. The unifying question is how to 'tame' them, in the editors' terminology. Taming, as understood here, is largely a question of representation: How to represent forms for which there is

no standard orthography, what to represent, how much to annotate, how much analysis to impose on the materials, how to represent ambiguities and indeterminacies, how to represent the finished product to the end-user. Noting the diversity, not only in the models underlying different corpora but also in their methods of encoding and analysis, the editors, themselves seasoned corpus builders, question whether it is reasonable or even feasible to aim for standardized protocols of the kind employed in traditional corpora for the collection, transcription, annotation and preservation of their less conventional counterparts.

Perhaps the first to grapple with the problem of taming unconventional data were the Sankoff-Cedergren team, whose *Montreal French Corpus* (Sankoff and Sankoff 1973) was built to elucidate a stigmatized variety previously widely believed to be an incorrect version of European French. Their goal was to show that the 'deviant' forms were part of a complex sociolinguistic structure, by tapping into different sources of speech variation: inter-individual, intra-individual and intra-linguistic. Chief among the problems inherent in such an endeavour was the issue of representativeness: How to guarantee representativeness of all the possible diversity in speech, while maintaining randomness in the selection of informants? They achieved this by implementing a detailed sampling frame, which, in contrast to their material procedures, has not yet been superseded. Their problems and solutions hark back to a simpler time, especially as compared with those corpus linguists face today. The transcription protocol – standard orthography – was dictated by the number of symbols on the punch keyboard for the IBM computer cards they used. Correction was effected by removing the card containing the error and inserting a correctly punched card in its place. The 100,000 cards containing the transcriptions then had to be converted into reams of computer printouts – and all without dropping a single card! In an era in which an entire corpus can be carried around on a memory stick or an iPod, it is worth noting that the print concordance of the 3.5 million-word *Ottawa-Hull French Corpus* (Poplack 1989), for example, occupies an entire wall – floor to ceiling – of the Ottawa Sociolinguistics Lab. The technology was primitive.

Since then, striking advances, not only in terms of hardware, but also in the area of annotation systems, have revolutionized corpus linguistics. No protocol has yet emerged as standard, though – as observed by the editors in initiating this project. So it's no surprise that the issue of annotation enjoys pride of place in these volumes, with researchers weighing in on what to annotate, how much detail to include, and whether it is preferable to replicate markup schemes of other corpora or

tailor them to one's own. It is clear that the old problem of finding the right balance of quantity, recoverability and faithfulness is still with us. Faithfulness at every linguistic level to data with much inherent variability (i.e. all speech, and many older and/or nonstandard written texts) inevitably results in diminished recoverability and less quantity. Without sufficient quantity, statistical significance is impossible to establish and full cross-cutting conditioning yields mostly empty cells. Optimum recoverability comes at the expense of less faithfulness to the many variant realizations of what is underlyingly a single form.

Each of the contributors to these volumes grapples with these problems in their own way. Some prefer to abandon one or more of the principles, others respond with complicated interfaces. As a result, the corpora described in this collection illustrate the full gamut of possibilities, from an annotation system so rich and complex that it already incorporates a good deal of the linguistic analysis, at one extreme, to virtually no markup whatsoever at the other. Linkage of transcripts to (audio and video) recordings and syntactic parsing will no doubt be the wave of the future.

The projected use of the corpus, as *end-product* or *tool*, is clearly the determining factor. Those for whom the corpus is a tool tend to advocate minimal annotation. These researchers are able to tolerate more indeterminacy and ambiguity, either because they have determined that it will not affect what they're looking for (e.g. a number of the corpora described here provide no detail on phonetic form or discourse processes), or because the sheer volume of data available allows them to omit the ambiguous cases or neutralize errors through large-scale quantitative analysis. Others, for whom the corpus is the end-product, tend to aim for consistency with guidelines for existing corpora, even if these do not seem immediately relevant to the proposed research. So what is the best annotation system? The amalgamated wisdom to be gleaned from these contributions: the one that works for you. At the moment, then, the answer to the editors' query regarding the feasibility of standardizing transcription protocols seems to be a qualified 'no'.

Comparatively less emphasis is placed on the issue of *representativeness,* the extent to which the sample of observations drawn from the corpus corresponds to the parent population. Achieving representativeness for (socio)linguistic purposes involves identifying the major sources of variation in the population (of speakers and utterances) and taking them into account while constructing the sample. Few corpora in these volumes, by necessity or design, claim to be representative in the sense of Sankoff (1988). Rather, in most of these contributions, (as in much

social science research more generally), the sample is opportunistic. This is an issue that every corpus must come to terms with, since even large numbers of observations cannot compensate for a sample frame from which the major sources of variation are missing. To the extent that the sample does not span the variant answers to the research question, pursuit of that question via that corpus can only be spurious.

Whether representativeness or annotation is more fundamental to the eventual utility of the corpus is a moot point. It is worth noting, however, that the awkward, and for some, simplistic, transcription protocols of early unconventional corpora did nothing to diminish their interest, value and current relevance. Hundreds of studies have been, and continue to be, based on them, perhaps because the research questions they were constructed to answer are still burning ones. The same is of course true of a number of the established corpora described in these volumes, and no doubt will be of many of the more incipient ones as well. The good news is that these repositories have an enduring value that far transcends our automated treatment and handling of them.

I end this foreword by returning to the question I posed at the beginning. What could possibly make the huge front-end investment required to build a corpus worthwhile? Obvious answers include the enormously enhanced speed of data collection, enabling consideration of ever greater quantities of data with relatively little extra effort. This in turn increases the chances of locating rare tokens, achieving statistical significance and determining which factors condition the choice between alternating forms. All of these are inestimable boons for quantitative analysis, but they pale in comparison to what for me remains the most exciting aspect of corpus work: the opportunity it affords to serendipitously discover what one wasn't looking for, to characterize the patterned nature of linguistic heterogeneity, and in particular the hidden, unsuspected or 'irrational' constraints that are simply inaccessible to introspection or casual perusal.

How much closer are we to the goal of agreeing on a standardized annotation? Well, we aren't there yet, though only time will tell. In the interim, anyone who has ever considered building a corpus or is engaged in doing so now will want to have a copy of this book close at hand. The wide variety of contributions convey much of the excitement of this burgeoning field. Despite inevitable differences in methods and projected end uses, the common thread is the shared goal of finding and implementing the best practices in corpus construction and preservation. These companion volumes, examining both synchronic and diachronic corpora, serve as a model for how to achieve them. For this,

we can only be grateful to the editors, who encouraged such stimulating dialogue.

SHANA POPLACK

References

Cedergren, Henrietta. 1973. 'Interplay of social and linguistic factors in Panama'. PhD dissertation, Cornell University.

Poplack, Shana. 1989. 'The care and handling of a mega-corpus'. *Language Variation and Change* (Current Issues in Linguistic Theory, 52), ed. by R. Fasold and D. Schiffrin, pp. 411-451. Philadelphia: John Benjamins.

Sankoff, David. 1988. 'Problems of representativeness'. *Sociolinguistics. An International Handbook of the Science of Language and Society*, Vol. 2, ed. by U. Ammon, N. Dittmar and K. J. Mattheier, pp. 899-903. Berlin: Walter de Gruyter.

Sankoff, David and Sankoff, Gillian. 1973. 'Sample survey methods and computer-assisted analysis in the study of grammatical variation'. *Canadian Languages in their Social Context*, ed. by R. Darnell, pp. 7-63. Edmonton: Linguistic Research Inc.