Special Section on Corpora

Guest editors: NICHOLAS OSTLER and ANTONIO ZAMPOLLI

Introduction to Part One

NICHOLAS OSTLER
Linguacubun Ltd, London

Abstract

In the 1990s the empirical study of language from large bodies of recorded documents has assumed a new importance, and this was reflected by the European Commission's decision to support the project NERC, the Network of European Research Corpora, led by Antonio Zampolli. Its aim was to study the need for, and possible provision of analysed corpora for European languages.

NERC's first action was to organize an International Workshop, held at Pisa in January 1992, and attended by invited scholars from Europe and North America, to gather and cross-fertilize a variety of experience and views on how to further the project's aims. Re-worked versions of some of the papers presented then, together with a small number describing related work by other scholars, are now published as a special supplement to this and the next number of *Literary and Linguistic Computing*.

They are presented in an order which corresponds to NERC's own structure. After a general statement from José Soler of the European Commission on the importance of this field of study, this follows a spectrum of interest: from examinations of the demand for corpora (McNaught) and the administrative complications in making them available (Hockey and Walker), through analysis of the conceptual (Biber) and practical (Crowdy, Part 1) problems in selection of texts, to the issues that arise when designing (Sampson) and applying (Leech) a system of categories for annotating the language in the texts. A particular problem here is treatment of spoken texts when reduced to written form, and Ballester *et al.* offers a solution for Spanish, Crowdy Part 2, for English. After these studies in annotation, the focus shifts to statistical techniques for exposing the semantics of uninterpreted text, sometimes known as 'knowledge acquisition' (Bindi *et al.* and Brown *et al.*). Finally, this supplement contains reports from some current projects which make essential use of large corpora and their annotation categories for particular applications: designing lexicons (Antoni-Lay *et al.*, Khatchadourian and Modiano), multilingual text processing (Cowie *et al.*), and speech technology assessment (Fourcin and Gibbon).

1. Cui Bono?

In this short introduction, I shall try to catch hold of some of the amazingly slippery issues which this collection of scholars' views has stirred up. It would seem at first sight a simple matter to formulate a sensible procedure for selecting and annotating a set of texts as a general-purpose body of evidence about a language. But it is not simple.

The first question that demands an answer is: why are corpora needed? Evidently, they are part of the 1990s *zeitgeist* in computational linguistics, and there might be stimulating things to say about why this should be so. Has there been a loss of faith in the *a priori* understanding of language built into 1980s language–processing systems, and so a flight back to the basics of real data? Or is there instead now a greater faith in the capabilities of our systems, and so a desire to put them to work on greater challenges, the unknown varieties of the Great Book of the World? Both possibilities are believed by some, and McNaught comes closest in this collection to discussing this side of the issue directly. There is often a more concrete issue here, however. That is to ask what the corpora actually being collected and formatted are to be used for.

In a recent article in this journal (7.1 (1992) 1–16) I distinguished three major types of user (language-, content-, and media-specialists) and two major types of use (as a source of data, and as a testbed for systems). The first differ in the aspect of the textual data that interests them: do they focus on the texts as evidence of the language they are written (or spoken) in, or of the subject-matter they treat, or of the kind of information representation they use? The second differ on whether the corpus is of interest in its own right, as something we can learn from, or is just being used instrumentally, as an input to the process of improving something else—usually an electronic system.

It is clear that different authors here have different users and uses in mind. Biber focuses on the language–specialist interested in a source data, although his concern for correlating internal (linguistic) criteria with external indices of the texts' context means that success will make corpora more useful to content-specialists. McNaught also has language-specialists in mind, but concentrates on how corpora can help them improve their processing systems. Brown *et al.*'s problem (predicting a word from previous words in a sample) makes sense only as a part of the task of building efficient systems to handle text using the quantitative clues it contains: i.e. for testbed use by a media-specialist. The last four articles, since they focus on applications, necessarily presuppose that the corpora are used as testbeds; but the systems using them differ in the degree to which their crucial problems require linguistic, semantic, or information–theoretic analysis of the corpus data.

There is a prior question that tends to be assumed away. Is it possible at all to build a useful general-purpose corpus? People naturally assume that the type of corpus to be constructed should reflect the nature of

Correspondence: Nicholas Ostler, Linguacubun Ltd, 17 Oakley Road, London N1 3LL, UK.

the data excerpted into it, i.e. that its nature is determined by its content. Then it may be judged as good or bad depending on how representative it is of that content. But, as Biber points out, adequate representativeness depends on the application in view: only when you know what you want to do with the corpus will you know whether this corpus has enough data of the right type to serve you well. He therefore advocates an iterative cycle of corpus refinement and supplementation, something which is quite impossible if the corpus being used is externally generated and beyond the user's control.

Ballester *et al.* propose a wide variety of specific decisions about the level or detail to record in transcribing spoken Spanish: but without any specific use in mind, how could one possibly decide if these decisions are the right ones?

To register this unease is not necessarily to undermine the whole enterprise of building large general-purpose corpora. It may be that many different uses end up imposing requirements on the corpus builder that are quite compatible. And as Hockey and Walker show, the task of building and distributing those corpora on a significant scale is quite intimidating: the sheer weight of work is likely to mean that much better average use is made of textual data in quantity if the work is done centrally.

2. Science or Engineering?

McNaught, who focuses on the natural language processing (NLP) community, sees corpora's *raison d'être* as to place demands of greater realism and fuller coverage on NLP systems. But this is not all that they can contribute.

Corpora may also play a role in stimulating fresh ideas about what the generalizations are that theoretical linguistics should be capturing, and computational linguistics exploiting. And perhaps the fuller range of unexpected data revealed by the corpus will give unexpected types of test to principles that were already in the systems. In brief, corpora may provide a new way for the world of experience to interact with the hypotheses that make up a science, as well as faithfully pointing out all the i's that the model has failed to dot and the t's it has failed to cross.

This contract is implicit, but not stated, in Sampson's article. He attributes the absence, hitherto, of a Linnæan scheme of text–annotation for English to the different goals of theoretical and computational linguistics. Only now, he argues, when we as computationalists are trying to take seriously the full range of data, do we feel the full need for a comprehensive, unambiguous classification scheme.

This is paradoxical. It seems to go against the fact that successful NLP systems are always limited in their coverage, and even often restricted to a sublanguage and a specific application, whereas theoretical linguistics quite explicitly makes universal claims.

It is as if the processing of corpora is in an awkward no man's land where high-flown theory is brought down to earth and made to attempt an honest job of work in the real world. It soon becomes clear that the theoretical entities will not fit simply on to the facts at the level of detail recorded, but somehow it is always too much trouble to reconcile the differences. It serves no *practical* purpose to analyse the full corpus in terms of the theory, for the corpus is in any case not the domain for a practical application. Nor does it serve any *theoretical* purpose to add to the theory all the odds and ends of particular partial applications, which might cope adequately with parts of the corpus.

Admittedly, there may be a practical purpose in trying to make different partial applications compatible with each other; and this is what Sampson has set about doing with his SUSANNE parsing schema. (It is a question unanswered here how close to completeness the SUSANNE schema actually comes.) But it is important to note that greater coverage and consistency, the sort of benefits that access to an appropriate corpus can offer to a partial NLP system, are inappropriate goals when a corpus is analysed in terms of linguistic theory.

From a specific point of view, a theory may benefit from the revelation of new phenomena, hitherto unobserved, in the corpus. But it is wrong to expect it to assume responsibility to generate or describe all the phenomena the corpus presents. For the corpus is not identical with any set of facts the theory is trying to characterize.

3. Annotation: How Valid is the Corpus as Evidence of Linguistic Reality?

One major attraction of corpus work is the underlying belief that it can give an authentically objective judgement on the adequacy of a theory or a technology. The corpus has resulted from a process over which the theoretician or language technologist had no power of selection. Therefore the ability of his system to come to terms with the corpus is a convenient index of its ability to come to terms with undoctored reality.

Hence Leech's concern to distinguish between representational and interpretative information in a corpus. The first is constitutive of it, whereas the second is a kind of theoretical overlay. Leech remarks on how much more difficult the distinction is to make when the subject-matter is a transcription of speech, and this gives the clue to the status of this distribution, which is one of degree rather than kind.

Transcription of speech is a partial representation of it, but also the first (and least controversial) layer of its interpretation. Immediately we see that there may be at least two sources of disagreement about a piece of transcription: factual—'that's not what she said'; and theoretical—'that's not how you spell what she said'.

As more layers of annotation are added to the corpus, the scope for both these kinds of fault becomes greater, and the claim of the corpus to represent an undoctored slice of life harder to support. Leech is frank about the temptation to devisers of annotation schemes to secure an easier life for annotators by diminishing the number of factual distinctions that need to be made, even where they are intuitively clear (e.g. positing a single syntactic tag for the various senses of the English word *one*). Skeleton parsing (assigning structures without classifying the constituent parts) is an attempt to maintain high factual reliability in an area where theoretical doubts are rife.

From this perspective, it is more important that an annotation scheme should be uncontroversial in its application than that it should be revealing. Indeed, the judgement towards which Leech tends at the end of his article is that theoretically clear-cut annotations cannot be applied categorically at all, but simply exist as prototypes, to which actual instances in a corpus just approximate more or less. The verdict of his experience, then, seems to be that it is not sensible to try to apply high-level categories directly to a corpus *en masse*.

From all that I have written above, it will be clear that I do not believe that the corpus or its parts can work as a general-purpose proxy for reality in judging theoretical constructs. However, the corpus's independence of whatever system is in view is important, and worth working hard to guarantee. An independent judge may make a useful contribution, even if he is neither infallible, nor perhaps always even impartial.

4. Detecting Correlations

A major goal for the corpus-user who is a language- or a content-specialist is to identify internal properties of text which correlate with features of the context in which it was produced. Biber, for example, considers the incidence of seven linguistic phenomena as they pattern across three different registers (conversations, general fiction, academic prose). If this could be carried out widely and systematically, it would diminish some of the chicken-and-egg regresses between corpus and application which we have seen to threaten the validity of gathering large-scale corpora for general use.

Crowdy, Part 1, notes how demographic representativeness, which is being scrupulously observed in the data-gathering for the British National Corpus spoken materials, fails to do justic to the full linguistic variety when texts are compared across the whole gamut of context types. This can be seen as a special case of Biber's statement of the need for empirical work to discover the frequency of linguistic features and the degree of variation present in language from different registers.

A comparable goal at a different level of linguistic focus is to find distributional correlates of syntactic and semantic classes. Brown *et al.* quote statistically identified classes which seem to correlate with directions, mineral names, nationalities, model verbs, body parts etc. Bindi *et al.*, on the other hand, use distributional facts to discriminate between apparent synonyms.

Given significant and reliable correlations of either kind (i.e. of linguistic features with context of use, and of distributional word-classes with semantic features) it would become possible to identify various features and contents of texts automatically. On the basis of formal analysis of their content, but without explicitly understanding them, it would be possible to classify them by discourse-type and subject-matter, and to extract some basic information automatically. It might even allow some large-scale properties of texts to come to light which hitherto have escaped human speakers and readers.

Reading the later articles in the second part of this special section gives some idea of the state of our progress towards these goals. It is probably still true that formal analysis of text-corpora has told us nothing about our languages that we did not already know. But the formal techniques employed are quite alien from our traditional ways of making sense of text: if we can really believe them when they tell us what we already know, there is good reason to believe that they will soon teach us a lot more.