Corpus Linguistics - Linguistics or Language Engineering?*

Tamás Váradi

Linguistics Institute, Hungarian Academy of Sciences
Benczúr u 33, 1068 Budapest, Hungary
varadi@nytud.hu

Abstract

The use of large corpora is increasingly seen as indispensable in many areas of human language technology. It is also widely assumed that they represent the most objective empirical evidence about the language. This paper sets out to examine the aims and methods of corpus linguistics from a linguistic point of view. After a brief historical perspective, it focuses on some key issues in corpus design, in particular it subjects to a critical analysis a recent influential proposal for the notion of a representive corpus by Douglas Biber.

1. Introduction

The purpose of this paper is to review some of the linguistic principles underlying current practice in Corpus Linguistics. The interest of course is not to provide a precise classification *per se* of where the field of Corpus Linguistics belong. Given the explicit focus of the present conference on language engineering, I think it would be useful to examine common assumptions underlying Corpus Linguistics with a view to seeing how far the claims stand up to basic linguistic principles.

It is customary to question the relevance and indeed the utility of linguistics in some areas of natural language processing. By choosing to review Corpus Linguistics in terms of the dichotomy given in the title, I do not wish to widen the divide, noted by Abney (1996) between the linguistic and the natural language processing communities. Nor do I wish to sound denigratory about LE when phrasing the dichotomy in the above way.

The idea behind the title is simply this. Engineering is defined in the Concise Oxford Dictionary as "the application of science to the building, design and use of machines, constructions etc." Whichever discipline is meant to be applied here, linguistics does have a claim to be at least considered.

2. Corpus Linguistics as an "emerging" field

2.1. Structural linguistics

Although, Corpus Linguistics is rightly considered a dynamically growing field, the use of corpora for linguistic research goes back to a long tradition. In fact, studying language through a collection of texts is the only possible approach when native language intuitions, either of one's own or that of an informant, are not available – such as in historical linguistics or in the study of extinct languages. This was the ruling paradigm in American structural linguistics till the mid-fifties. Positivism was the ruling philosophy of the day and linguists rejected any appeal to intuitions as unscientific. Their objective was to compile a corpus of utterances and to derive the grammar underlying the data by rigorously applying some set of so called discovery procedures. The data was usually obtained from informants through various elicitation techniques. Thus, the aim of compiling the corpus was to gather enough evidence to construct the grammatical system of one informant or at most a small group of informants. The data gathering (corpus compilation, in our terminology) was done peace-meal and in an interactive way. It is perhaps surprising to find that one of the most distinguished advocate of corpus based structural linguists, Harris, suggested going back to the informants for more data whenever the corpus was found lacking (Harris, 1951 p. 12). The other point to note is that tthe corpus was gradually extended to cover an increasingly wider set of speakers. In the process, the corpus was to be regularly checked for consistency and completeness.

2.2. The generative enterprise

The rise of generative grammar in the fifties produced a paradigm shift in linguistics, which had profound implications for the issues we are examining here, the aims and methods of using data in linguistics. Many of the well-known tenets introduced by Chomsky have a direct bearing on corpus linguistics as practiced then and indeed as is practiced now. A grammar is a theory of the implicit knowledge of the language (competence) that each speaker has in their minds. It cannot be derived from a corpus of data (performance) through some mechanical procedure for several reasons. First, the grammar must account not only for the whole set of actual but the potential sentences as well. In fact, any grammar is severely underspecified by the language data, barring any mechanical procedure to derive the former from the latter. Second, performance data is not only patchy but degenerate as it is produced under a variety of practical constraints deflecting the surface data from their ideal, grammatical form. Third, along with the downgrading of empirical data as of secondary importance at best, generative linguistics brought about the reappraisal of evidence based on intuition and even introspection.

* The work on this paper was supported by Országos Tudományos Kutatási Alapprogramok (grant number T026091).

All these points argued against the use of corpora and the corpus-based methodology of the structural linguistics.

Another tenet that has generated a lot of controversy is the view that the postulated grammar exists in the mind of the ideal speaker/hearer who lives in a homogeneous speech community. The idealization of the messy data to a homogeneous speech community was presented as a necessary expedient but many people argued that it stemmed from an inherent lack of interest in the social dimensions of language. In fact, the very notion of language has become suspect in Chomsky´s recent writings. Earlier he customarily defined language as the set of sentences defined by a grammar, which of course is music to the ear of any corpus linguists who are in the business of dealing with sentences. Recently, (Chomsky 1986) he introduced the dichotomy of E-language and I-language. The latter covers the notion of an internalized grammar of the language residing in the mind of its speakers. The concept of E-language, which corresponds to the everyday notion of language is declared a dubious concept, 'an epiphnomenon at best'. It is knowledge of the language in the form of an I-language that is solely of interest. (For more information on Chomsky vs. corpus linguistics see McEnery & Wilson 1996).

2.3. Corpus Linguistics emerges

It is against this context that Corpus Linguistics as we know it today emerged in the early sixties. The aim from the very beginning was to compile a corpus that was representative of a language. In terms of the concepts introduced above, this means nothing less than to design a corpus that models the totality of language use of a speech community. This is certainly a tall order, given the complexity and the scope of the phenomena that it undertakes to cover. In practice, though, the task was attempted from the outset with some reasonable limitations in the temporal and geographic dimensions of the data. The pioneering Brown corpus set out to capture the written language of the United States of the year 1963. It was intended to be a general purpose, balanced corpus of American English of the period. We'll see how representativeness was achieved later.

Another claim that Corpus Linguistics makes is that it shows up 'language as is spoken', real language in its rawness and richness. This intention in obviously inherent in the whole corpus linguistic enterprise of capturing vast amount of actual data. Apart from marketing purposes, it only needed emphasizing in contraposition to the ruling generative linguistic school.

3. Basic design issues

It is clear that the key issue for Corpus Linguistics to make good its promises lies in the scope and composition of the data that it provides. This will be the focus of our attention for the rest of the paper. It is widely agreed that a corpus is not simply an archive of texts but rather a principled collection of texts. One of the first and most important principles referred here concern the selection of texts to go into the corpus.

The first question that arises in examining this issue is whether we should care too much about the composition of the corpus. Accordingly there developed two kinds of schools of thought supporting two kinds of corpora: the so-called opportunistic and the balanced corpora.

3.1. Monitor corpus vs. balanced corpus

It is fairly easy to deal with the opportunistic kind as it denies that there is any principled way to balance a corpus and it makes recourse to the law of large numbers. Perhaps size will automatically sort out all questions of 'balance' in the structure of the data. This approach is vigorously represented by John Sinclair (1991) who proposes instead the idea of a monitor corpus – a very large corpus, which after reaching some sort of a saturation point will undergo a partial self-recycling: the new material flowing in will be subjected to an automatic monitoring process which will only retain those parts of the incoming data which show some significantly different features than the stable part of the data.

Once it is decided that some sort of scheme will be set up to compile a corpus in some principled way, the question that confronts us is whose job is it to do so. John Sinclair considers it is a task that should not belong to the students of culture rather than corpus linguists. They only undertake it as a matter of necessity. The use of language, Sinclair seems to argue, should be studied in the wider cultural context, which goes beyond the competence of the corpus linguist.

3.2. Units of sampling

Another important sampling issue to decide is the units of the overall population in terms of which the sample will be compiled. Should the sample be compiled in terms of the speakers or language? If the latter is chosen, as it was originally done, (without apparently considering any alternative) what are to be the linguistic units in terms of which the population is sampled: words, sentences, texts, speech situations etc.?

In the first generation of balanced corpora, the Brown and the LOB corpus, this issue was decided by a panel of experts who designed a scheme where different varieties of language, called *genres*, are represented in specific proportions. Table 1 shows how the 1 million word corpus is divided into 15 genres and how many texts of 2000 word length each are allocated into each category. Note how despite the professed intention to develop a replica of the pioneering American corpus for British English, the internal composition of the LOB corpus was slightly changed in categories E,F and G). These subtle changes were introduced so as to accommodate the structure of the corpus to the peculiarities of British culture. As for the selection of the particular texts, apparently, a great deal of effort was spent into making sure that the texts within each category were chosen at random but I am not aware of any public arguments offered in justification for the particular ratios used *between* the categories.

3.3. Methods of sampling

Choosing things at random suggests itself as a safe procedure to eliminate any bias or skewing in the result. However, purely random sampling works against the selection of items that are relatively rare in the population,

out of which the sample is made. An important principle that a sample should meet in order to be representative of the population is that the sample should show the same ratios between elements within the sample as they have in the population. Samples are, as it were, severely scaled down versions of the population. The more frequent an item is in the population, the better chance it stands of being selected at random. Therefore, it may easily happen that items which occur pretty rarely in the population, will not be selected by the random process at all. Alternatively, if for some reason or other, we would like to see the rare items included in the sample, we would have to increase the size of the sample, perhaps *out of all manageable proportions*.

	Genres	No of texts	
		Brown	LOB
A	Press:report	44	44
B	Press:editorial	27	27
C	Press:reviews	17	17
D	Religion	17	17
E	Trades, hobby, leisure	36	38
F	General lore	48	44
G	*Belles lettres*, biography, essays	75	77
H	Misc. government. documents, public reports, university catalogues	30	30
J	Scientific journals	80	80
K	General fiction	29	29
L	Crime fiction	24	24
M	Science fiction	6	6
N	Adventure and Western	29	29
P	Romance	29	29
R	Humour	9	9
~	~	500	500

Table 1 Composition of the BROWN and the LOB corpus

One solution that is devised to overcome the above difficulty is to use *stratified random sampling*. Under this procedure the population is first divided into a number of categories (strata) and random sampling is only applied to fill up the chosen categories with items selected at random. The question of how many categories to set up into which the population is arranged and how much data should be collected for each category is decided beforehand. (These are indeed the figures shown in Table 1 for the BROWN and the LOB corpus.) The taxonomy of the categories is established independently of statistical considerations. Yet, it has a direct bearing on the quantitative results as well. Once a category is established, it is bound to be represented in the sample. For example, if we have a general category for reviews, chance will decide whether the random sampling will select any articles on reviews of early twentieth century travel books. (Chance will be helped by the number of such articles in the whole population in that the more there are the higher the chances that a purely random method will select them.) If on the other hand a special category is adopted to cover travel books, this is taken as a target to be met and the selection procedure is considered incomplete until data is selected for that category as well. Hence, the granularity of the classification scheme will effect the structure of the sample as well.

An even more direct intervention in the workings of chance is the setting of target figures for the amount of data to be collected within each category (i.e. the figures against the categories in Table 1 representing the number of texts, each about 2000 words long). In order for a sample to be representative of the population for the set of categories in terms of which the sample is compiled, the sample should conform to the principle of *proportionality*. This requires the various categories in the sample to be represented in the same ratio as they are in the total population. For the BROWN corpus to qualify as a representative sample of the totality of written American English for 1963 for humorous writing, it would have be established that humorous writings did make up 1.8 % of all written texts created within that year in the US. This single requirement serves to illustrate the enormous difficulty if not impossibility of the task. Surely, it is simply not feasible to put a figure on the amount of text within the various genres in the totality of texts produced by a speech community. Yet, this is what the statistical concept of a representative sample calls for.

Note that the difficulty is not necessarily that of dealing with an infinite set. Although theoretical linguistics chooses as its domain the set of all *possible* sentences, corpus linguistics handles *actual* sentences. Given that speech production or writing is an activity taking place in real time, one could make a theoretical assessment of the maximum number of sentences produced by even a whole speech community. Of course, even the written output of a literate speech community of any size is way beyond the capacity of current technology to handle. Hence the inevitable need for using a sample. The difficulty noted above, however, is inherently a logical one. If sampling is done in terms of text type, a representative sample would require knowledge about the whole population that is simply not available. If it were, we would hardly need a sample, and in order to find out about proportions obtaining in the population, one would obviously like to examine a sample of it.

3.4. Demographic vs. context-based sampling

How can we break this vicious circle? One lesson obviously is that one can only provide a representative sample of the population in terms of its features about which one has reliable knowledge from some independent source. One such source of outside knowledge is indeed available in data about the *speakers.* One could consult National Census figures to find out about chief characteristics of speakers such as age, gender, schooling, type of settlement they live in etc. It is then feasible to compile a representative sample of speakers *for such selected features*. This type of *demographic* sample of informants is a well established procedure in opinion poll surveys, psychological or socio-linguistic research. For corpus linguistics, the total output of such representative group of speakers would *ipso facto* amount to a representative corpus of the speech population.

This procedure was indeed used by the spoken component of the British National Corpus (cf. Burnard

1995 p. 20-25). 124 adults were selected so that, as far as practical limitations allowed, they would be represented in equal numbers in terms of sex, age (divided into six age groups) and social class (defined in four main categories). The recruited informants were asked to record their speech conversations, unobtrusively whenever possible, for a period of up to a week. Approximately four million words were collected in this manner, a little under half of the spoken component of the BNC, which in turn, for obvious practical constraints, made up one tenth of the 100,000 word corpus. The rest of the spoken component, termed the *context-governed part*, was selected by "*a priori* linguistically motivated categories" defined in terms of a hierarchy of categories with the four context categories educational, business, public/institutional and leisure at the top and three regional and two interaction type categories providing further subdivisions.

Despite the undeniable practical difficulties of implementing it, the demographic sampling technique was applied in a limited way on purpose. The Reference Guide notes that

> 'many types of spoken text are produced only rarely in comparison with the total output of all "speech producers": for example, broadcast interviews, lectures, legal proceedings and other texts produced in situations where – broadly speaking – there are few producers and many receivers. A corpus constituted solely on the demographic model would thus omit important spoken text types'. (Burnard op. cit. p. 20)

4. Biber's notion of representativeness

The issues reviewed so far are certainly nothing new to practitioners of the field. With predictable regularity a discussion flares up on the Corpora List around the notion of the balanced corpus. Newcomers to the discussion are often referred to Douglas Biber's article "Representativeness in Corpus Design" (Biber 1993), which is indeed one of the most comprehensive discussions of the topic available in print[†].

Biber distinguishes three possible approaches to corpus design depending on whether they are aimed at covering text production, text reception and texts as products. The first two are basically different from the third in that they both define the population in terms of the agents (i.e. speaker/hearer) of language use, while the third covers it in terms of the output i.e. language. Accordingly, the first two approaches would call for a demographic sample. However, Biber also rejects demographic samples on the grounds that "they would not represent the range of text types in a language, since many kinds of language are rarely used, even though *they are important on other grounds.*' It would thus be difficult to stratify a demographic corpus in such a way that it would insure representativeness of the range of text categories. Many of these categories are very important, however, in defining a culture" [emphasis added] (op. cit. p. 243).

This revealing passage spells out some assumptions that may be difficult to reconcile with some basic assumptions about the role of corpus linguistics. One of the fundamental aims of Corpus linguistics as I understand it is to show up language as is actually attested in real life use. However, Biber seems to argue that in designing a corpus one should apply a notion of importance that is derived from a definition of culture. For lack of any means of operationalizing this criterion of relative importance in culture, this throws the door wide open to subjective judgement in the compilation of the body of data that is expected to provide solid empirical evidence for language use.

Biber seems to think very little of the value of a corpus assembled on demographic criteria. "Such a corpus would permit summary descriptive statistics for the entire language represented by the corpus. These kinds of generalizations, however, are not typically of interest for linguistic research.", "… it is not necessary to have a corpus to find out that 90% of the texts in a language are linguistically similar (because they are all conversations)"; rather, we want to analyse the linguistic characteristics of the other 10% of the texts since they represent the large majority of the kinds of registers and linguistic distributions in a language." (op. cit. p. 248)

Biber concedes that there is no a priori way to establish the relative proportions of the different genres obtaining in the population hence a representative sample would have to be demographic by definition. This impasse leads Biber to conclude that the notion of representativeness as we know it from statistics do not apply in corpus linguistics. What lies at the root of the problems to implement representativeness is the principle of proportionality that has been discussed above. Biber not only considers proportional sampling difficult or unfeasible to implement in any other way than the demographic approach but goes as far as to simply reject the notion of a proportional sample as an appropriate concept. In justifying his position he makes the following curious argument: "proportional samples are representative *only* (sic!) in that they accurately reflect the relative numerical frequencies of registers in a language – they provide no representation of relative importance that is not numerical. Registers, such as books, newspapers, and news broadcasts are much more influential than their relative frequencies indicate." [emphasis added] (op. cit. p. 248)

First, it is disingenuous to find fault with proportional sampling for something it is not intended for i.e. to reflect this non-numerical relative importance. Second, there is no suggestion how this kind of importance can be established, let alone quantified in any objective manner. No attempt is made to show how to measure and accommodate the extent of the influence of the above registers. Earlier, we already noted the potential methodological danger for arbitrary decisions creeping in the corpus design principles. One cannot avoid feeling that once recourse is made to non numerical factors such as importance in compiling the corpus, this makes the whole enterprise of corpus design so vulnerable to subjective value

[†] There is one unwritten item that comes to mind: there was a live debate held in Oxford between prominent advocates of the two corpus design philosophies Quirk aided by Leech speaking up for the balanced corpus vs. Sinclair and Meijs arguing for the open-ended monitor corpus. Oral tradition has it that the debate was decided by the audience in favour of the Sinclair team.

judgements that any amount of methodological rigour applied in the random selection of the items for categories looks like the farcical effort of searching for the lost key where there is light.

Rejecting the traditional notion of representative sampling based on the principle of proportionality, Biber blandly declares that "language corpora require a different notion of representativeness", "researchers require language samples that are representative in the sense that they include the full range of linguistic variation existing in a language." (op. cit. p. 247) First of all, one must voice serious misgivings about any attempt to divest such a key term of its well-established meaning, which has a clear interpretation to statisticians and the general public alike. Of course, any self-respecting corpus would like to advertise itself as a representative corpus. There is such a strong and unanimous expectation from the public and scholars alike for corpora to be representative that it is an assumption that is virtually taken for granted. To meet this demand by the semantic exercise of redefining the content of the term is a move that hardly does credit to the field.

5. Conclusions

My aim with this brief overview of the issues in corpus design has been to highlight the linguistic implications of the choices that are made. It is clear that many of the concerns examined are of limited relevance for the purposes of many practical applications. By highlighting on the uncertainties, inconsistencies and methodological fudges employed by linguists my intention was to take stock of the state of the art and show up where further effort is needed. The picture that emerges helps to dispel the unintended disparity in scientific rigour: in order to live up to its expectations corpus linguistics must put its methodology on more solid footing and language engineering would do well to be aware of the linguistic issues at stake and the extent to which they can expect ready solutions.

6. References

Abney, S. (1996) Statistical Methods and Linguistics in Klavans, J. and Resnik, Ph. (eds.) *The Balancing Act.* Cambridge, Mass., MIT Press

Biber, D. (1993). Representativeness in corpus Design. *Literary and Linguistic Computing 8/4:243-257* Oxford: Oxford University Press.

Burnard, L. (ed.) (1995) *British National Corpus. Users Reference Guide for the British National Corpus.* Oxford University Computing Service.

Chomsky, N. (1986) *Knowledge of Languague Its Nature, Origin and Use.* New York, Westport, London, Praeger.

McEnery, T. and Wilson, A. (1996) *Corpus Linguistics.* Edinburgh, Edinburgh University Press.

Harris, Z. S. (1951). *Methods in Structural Linguistics.* Chicago, University of Chicago Press.

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.* Department of English, University of Oslo

Sinclair, J. (1991). *Corpus, Concordance, Collocation.* Oxford, Oxford University Press.