The linguistic relevance of Corpus Linguistics

Tamás Váradi
Department of CL
Research Institute for Linguistics
Hungarian Academy of Sciences
varadi@nytud.hu

1. Introduction

The present paper is intended as a review of some basic principles of corpus linguistics (CL) from a linguistic point of view. In particular, it examines the aims and methods of CL in terms of how it stands up to some basic linguistic tenets regarding the nature and use of data for linguistic analysis. Central to the discussion will be the problem of representativeness, which will be discussed with reference to an influential paper by Douglas Biber (1993).

The perspective from which CL is examined in this paper is its utility and relevance for linguistic description. This is just one of the contexts in which CL is used today and indeed it may not be the bulk of its applications. It can hardly be doubted that CL has proved its utility in areas of human language technology too numerous to mention. Within fields more close to linguistics, its role in lexicography is nothing short of revolutionizing the whole practice of the discipline, introducing a new technology that produced a new generation of dictionaries. These are undeniable achievements that has established CL as a thriving and dynamically growing field.

Amidst this burgeoning activity it is perhaps opportune to take this occasion to reflect on some underlying assumptions and methodological principles characterizing current practice in CL.

2. The concept of language and corpus linguistics

Fundamentally, CL undertakes an empirical study of language. As Leech (2000: 685) rightly points out, this means that CL is 'patently' concerned with documenting and analysing *performance*, or the Saussurean notion of *parole*. The notion *performance* grammar is somewhat misleading in that CL is obviously concerned with the actual *product* of language use as against *processes* involved in speech production. The main theoretical claim that CL can make as a contribution to linguistics is that through the compilation and analysis of masses of data corpus linguistics can provide a solid objective empirical foundation on which to build a grammar of a language and indeed the technological facility to sift through large amounts of data leads to new insight into the structure of the language. One key aspect of language use that CL is particularly well suited to reveal is its quantitative characteristics.

It is to be noted, though, that the very concept of language that CL aims to capture is most forcefully rejected by Chomsky. Earlier, he consistently defined language as a set of sentences (Chomsky 1957, 1965), which, at least on the face of it, seemed to be congruous with the concerns of CL in the sense that, typically, CL also approaches language in terms a set of sentences. However, in later works he came to regard the traditional concept of language, which he now prefers to call *External-language*, as "an epiphenomenon at best" (Chomsky 1986: 25) a notion so laden with "complex and obscure sociopolitical, historical and normative-teleological elements (Chomsky 1991: 31) that he is doubtful if the concept can be given a coherent interpretation at all. The primary object of linguistic investigation in Chomsky's view is the *Internalized-language*, the mental grammar that each native speaker has internalized and which they draw upon in their linguistic communication.

Even if one does not share Chomsky's dismissive opinion on the relative standing of *E-language* vis-a-vis *I-language*, the dichotomy is one which CL also has to face. The issue boils down to the need to account for *I-language* from the facts of *E-language* as displayed in a corpus. It seems reasonable to take the view that language primarily exists in the minds of the speakers. Hence the starting point and focus of the linguistic enquiry should be the individual speakers with their *I-language*, their communicative competence and the actual products of their language use. Focusing on the actual products of actual language use by individual speakers of a language should, on the face of it, be quite amenable to the methods of CL.

Even at the level of the idiolect, the *I-language* of individual speakers, the research program raises some inherent theoretical problems stemming from the need to extrapolate from finite data to an infinite system. The difficulties are compounded, however, by the attempt to deal with the linguistic output of a group of speakers, let alone that of a whole language community. The hypothesis of an ideal hearer-speaker of a homogeneous speech community (Chomsky 1965) does save the theoretical linguist from a host of complications but CL cannot operate from this perspective.

The ambition of CL was from the very beginning to provide an empirical record of a language at the level of the whole speech community. As an interesting historical aside, we should note that this ambition was absent from what McEnery and Wilson (1996) term early phase of CL i.e. prior to Chomsky. Without the technological support ensuring a semblance of feasibility, one could not even entertain the idea of capturing and processing enough data for the whole language community. Nor was this the intent. Harris (1951:12), advocates a method of corpus compilation that proceeded in close cooperation with informants. This piecemeal, interactive procedure was driven by the expectations of the field workers, based on experience, about the completeness and consistency of the grammar that they were seeking to build.

3. E-language as a sampling problem

As noted above CL deals with the realm of actual language use. The primary raw data it encounters is a set of utterances produced by the language community. At first sight, this seems to be an infinite set. However, the number of words produced either in speech or writing are limited by obvious human physiological limits. Biber et al. (1999: 27) quote the figure of around 7000 words per hour as the average speech rate observed in the conversational part of the Longman Spoken and Written English Corpus. Therefore, once the inevitable fuzziness in the geographical, social and temporal boundaries of the notion of the language community is somehow resolved, one can put a number to the totality of language produced by any set of speakers over a given period of time.

Nevertheless, even if, as the above thought experiment suggested, the set of utterances produced by a language community in a given time interval is finite in size, it is unrealistic to expect that the totality of language production could ever be captured on electronic media. The bottleneck is not necessarily storage or technological capacity. While the incredible rate of advancement of computerization and the exponential spread of the Internet may eventually make the bulk of written language output at least accessible, capturing spoken output in a corresponding manner is not only unfeasible and even imponderable. This fact leaves us to conclude that any corpus, however large, is and will necessarily remain a mere sample of the totality of language output. That is how the issue of representativeness in the design of the corpus assumes key importance.

From the very beginning the aim of CL 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 (Kuèera and Francis 1967) 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. 4.2 discusses how representativeness was achieved and Table 1 displays the corresponding figures.

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, CL only needed emphasizing this in contraposition to the ruling generative linguistic school, which tended to base its findings on introspective evidence.

4. 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.

4.1 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 Sinclair (1991 pp. 23-24) 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. Sinclair (op. cit.: 13) holds that it is a task that should belong to the students of culture rather than corpus linguists. They should 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.

4.2 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.

4.3 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 to 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. It is, rather, 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.

4.4 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 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: 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.

It should be noted that the demographical sample used by the BNC cannot be considered representative in the sense of the sample being proportional to the population. Curiously, for reasons not disclosed in the manual, the BNC did not make use of actual demographic figures about the relative proportions of sex, age and social class obtaining in the UK population. Instead, "the intention was, as far as possible, to recruit *equal numbers* of men and women, *equal numbers* from each of the six age groups, and *equal numbers* from each of four social classes" [emphasis added] (Burnard op.cit.: 21). This methodological laxness in the almost single area where the required information to compile a representative sample was, in fact, available is hardly compensated by the care to use "established random location sampling procedures" to select individual members within the groups.

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.: 20)

5. 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[1]. The rest of the paper will concentrate on this as a canonical text, particularly as it reflects views that are upheld by Biber in essentially the same form in more recent works (Biber 1998, Biber et al 1999).

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.: 245).

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

[1] 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.

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 judgment 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.: 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 also goes as far as to simply reject the notion of 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. : 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 judgments 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.: 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. However, 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.

6. 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. By highlighting on the uncertainties, inconsistencies and methodological fudges currently employed in corpus linguistics, my intention was to 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 users of corpus linguistics would do well to be aware of the linguistic issues at stake and the extent to which they can expect ready solutions.

References

Biber D 1993 Representativeness in corpus design. *Literary and Linguistic Computing* 8(4):243-257.

Biber D, Conrad S, Reppen R, 1998 *Corpus Linguistics. Investigating Language Structure and Use.* Cambrdige, Cambridge University Press.

Biber D, Johansson S, Leech G, Conrad S, Finegan E 1999 *Longman Grammar of Spoken and Written English.* London, Longman.

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

Chomsky N 1965 *Syntactic Structures.* Cambridge Mass., MIT Press.

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

Chomsky N 1991 Linguistics and cognitive science: Problems and mysteries. In Kasher A (ed) *The Chomskyan Turn.* Oxford, Blackwell, pp. 26-53.

Kuèera H, Francis W 1967 *Computational Analysis of Present-Day American English.* Providence RI, Brown University Press.

Leech G N 2000 Grammars of Spoken English: New Outcomes of Corpus-Oriented Research *Language Learning* 50(4):675-724.

McEnery T, 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, 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.