Studia Anglica Posnaniensia XXXII, 1997

CORPUS LINGUISTICS: SOME BASIC PROBLEMS

RUDOLF EMONS

University of Passau

0. Introduction and listing of problems

The words *authenticity, authentic* etc. play an important role in the self-estimation of people working with linguistic corpora. Let me just refer you to the recent announcement of John Benjamin's publishing company (Book Gazette, Winter 1996/97) of the new *International Journal of Corpus Linguistics:* "This journal seeks to publish research that uses language as a social phenomenon that can be investigated empirically on the basis of authentic written and spoken texts." What is roughly meant by this is that the linguist is not supposed to produce his own data. Against this background here is a quotation from an authentic English text: "Yet his determined cajoling towards economic, monetary and ultimately – in the Kohl view – political union is provoking rumblings, inside Germany and outside, that this time Mr Kohl going too far." This is quoted from *The Economist*, the January 13th 1996 edition. We all know of course that something is wrong with this authentic quotation; obviously the word *is* is missing from the last clause. The fact that quotations like the above can be found at all poses some general problems for corpus linguistics which I shall presently be going into. Since the use of corpora is playing a more and more important role, particularly for the compilation of dictionaries, these general questions seem to be worth considering. There are two fundamentally different positions about what kind of empirical linguistic data constitutes a legitimate basis for linguistic research. Jon D. Ringen (1975: 24) has put these two positions like this:

Position 1 is: "Reports on clear cases to be used as linguistic data must be of judgements shared by all (or a significant sample of) reliable informants."

Position 2 can be put as follows: "Reports on clear cases to be used as linguistic data must be of judgements reported by at least one reliable informant".

What Ringen seems to mean is "reported of at least one reliable informant", but that is only a minor point. The phrase "clear cases" refers back to Noam Chomsky's *Syntactic structures* (1957: 14): "A certain number of clear cases, then, will provide us with a criterion of adequacy for any particular grammar." I shall not discuss at this point if it is really clear what a clear case is but will take for granted, just as Ringen does, that this can be found out without difficulty. A problem concerning both positions is of course: "For whom must the cases be clear?" (Ringen 1975: 24), or more precisely: "For how many speakers must the cases be clear?" The answers will be radically different for Positions 1 and 2 respectively.

Geoffrey Sampson (1980: 64) adds a third so-called descriptivist position to the picture: "Accept everything a native speaker says in his language and nothing he says about it." Thus a descriptivist would reject positions 1 and 2 because both permit judgements of speakers on linguistic material. As far as the necessary quantity of native speakers is concerned Sampson seems to regard one speaker as sufficient. So all three positions have to do with the same problem: How many speakers are regarded as a sufficient empirical basis for linguistic work?

The next question is: How do these three positions relate to corpora? Before trying to answer this question let me draw your attention to three problems in this field: First we have – vaguely speaking – an "aboutness" – problem. Some linguists refuse to accept judgements of native speakers about their communication, their language and their speech. The reasons for this are not self-evident and ultimately relate to what any empirical science is supposed to describe.

The second problem is the problem of representativity. Is one native speaker sufficient as an empirical basis? Is he only sufficient if he is in some way representative? Are two speakers more representative than one? Is there an optimum of representativity?

The third problem is the problem of correctness. This problem is solved in a peculiar way in Sampson's quoted words "Accept everything a native speaker says in his language..." According to this a native speaker is equipped with a certain kind of immunity. You could put it like this: "The native speaker can do no wrong." The model for this phrasing is the sentence "The King can do no wrong", which again refers to the constitutional position of a king or queen in England's Early Modern Age. It does not mean of course that the king cannot make mistakes, but what it does mean is that there is no constitutional institution which can formally charge the king's acts. Thus, the native speaker in some sense is also a king, because there is no legitimate institution that can prove him wrong.

For a descriptivist this holds true at least until the native speaker does not metaphorically speaking usurp the competence of the linguist himself since he is – as we saw from Sampson's quotation – regarded as unfit to judge his own utterances.

1. Characteristics of lingistic corpora

Against this background let me try to get the relationship between native speakers and lingustic corpora into sharper focus. A corpus in the widest sense is any kind of accumulation of linguistic material that must be fixed in such a way that a linguist can use it at any given point of time.

1.1. Parameters

There are numerous and different ways to collect such corpora; to give a certain structure to these possiblilities, let me name two parameters:

1. the producer of the text
2. the kind of text

The two extreme positions in the case of a text producer are: On the one hand, that one producer is sufficient to establish a corpus, on the other hand, that ideally every native speaker is monitored to produce a corpus. Within the second parameter – the kind of text – we can also distinguish between two extreme positions. On the one hand it is just one word that can be seen as sufficient to be a linguistic corpus, on the other hand a corpus comprises all kinds of texts and this again – another extreme – in an exhaustive way so that such a corpus would comprise each and every utterance ever made by human beings. Against this background we can measure and categorize existing corpora, bearing in mind that the following will be a rather rough classification and that there remain of course numerous other possibilities to categorize corpora. So here is a brief introduction of three different kinds of corpora.

1.2. The "one-man" corpus

The first type of corpus could be called the one-man-corpus. Leonard Bloomfield's work on Tagálog (Bloomfield 1967) comprises 56 pages of written text; this text was compiled by Bloomfield with the help of a single informant. His way of working is characterized by John Wolff (1987: 174): "The text is treated as sacred. Bloomfield never edits, never normalizes, never adjusts an unexpected form to the expected. The result is a set of texts in which one can find no wrong Tagálog (although not everything is necessarily elegant Tagálog – Bloomfield took everything down as his informant dictated it, whether or not it was stylistically well-formed)."

1.3. The *sample*-corpus

The second type of corpus is the so-called *sample*-corpus, an expression used by John Sinclair (1982). Well-known examples are: The Lancaster-Oslo-Bergen corpus (LOB), the London-Lund corpus and the American Brown corpus. The main characteristics of this kind of corpora are well described by Sinclair for the Brown Corpus (Sinclair 1982: 2): "The limitation on continuous text is 2,000 words, and so any study of largish patterns is likely to be inappropriate. Its vocabulary is controlled only indirectly via the gènre classifiation also, so a study of the patterning of infrequent words is doomed, and indeed the only words for which it is reliable are those which occur frequently in a good range of genres." Such corpora comprise roughly one million words from different kinds of texts.

1.4. The monitor-corpus

The last type of corpus is what Sinclair calls the monitor-corpus. With modern computers everything that has ever been printed is available to the "determined researcher." "Text is available in staggering quantities from journalism alone", as Sinclair puts it. The consequence is – and this is what really constitutes a monitor-corpus – in Sinclair's words: "Looking through the computer, the whole state of a language can be passed before one's eyes. Sampling can be done to order on gigantic, slowly changing stores of text, and detailed evidence of language evolution can be held efficiently" (Sinclair 1982: 4).

Thus the new edition of Collins Cobuild Dictionary from 1995 has a database of roughly 20 million words, and Longman's Dictionary of Contemporary English from 1996 comprises 25 million words.

2. Discussion of problems

Against this background of distinguishing roughly three different kinds of corpora let me get back to the three problems mentioned before.

2.1. The problem of "aboutness"

First problem: Native speakers say something **about** their language. Examples of the unreliability of judgements of native speakers on their own linguistic behaviour are often quoted. So Ursula Oomen reports about a teacher who used the sentence "The car runs real good" in her own spontaneous speech, but when explicitely confronted with the question of whether this was an acceptable sentence said: "That's not English" (cf. Oomen 1982: 15). At first glance the descriptive principle "accept everything a native speaker says in his language, but nothing he says about his language" seems to be reasonable here. Judgements of native speakers about their language cannot be reliable.

I have a very general objection to this kind of approach. The alleged discrepancy **between** linguistic reality and the norm-guided judgements of native speakers **about** this reality does indeed exist, but it seems to be a methodological mistake to regard this fact as a serious problem for linguistic work. Rather it is one thing to describe the linguistic behaviour of native speakers and it is quite another to describe judgements of native speakers about their language, the latter belonging to the field of the sociology of linguistics. So these discrepancy problems are no real problems at all.

The "aboutness"-problem leads us to a more general problem, namely: How do the linguistic investigation and the investigating linguist influence the facts that are to be described? The anthropologist Charles L. Briggs describes very clearly and very strikingly that the interview situation itself had a great influence on the results of his work:

> *The communicative structure of the entire interview affects the meaning of each utterance.* To cite one instance: My initial interviews with Silvianita and George López were strained and relatively unproductive, and it was many years before I was able to appreciate why this had been the case. I had never considered the possibility that the Lópezes might not accept my definition of our interactions as interviews. The Lópezes viewed these sessions as pedagogical encounters between two elders and a young person with little knowledge of the community, *Mexicano* culture, or New Mexican Spanish.
> Even after I published a volume on the Lópezes and other carvers (1980), the couple told their visitors that I had come to learn how to carve. (They noted that I had indeed become a proficient carver but then, for some reason, had given up the work) (Briggs 1986: 102-ff).

Returning to linguistic work, one thing seems to be clear. The data that are collected by way of one-man-corpora à la Bloomfield are definitely influenced by the interview situation itself. Of course, one can regard this as irrelevant like: 'The main thing is that I get any linguistic data whether they are influenced or not', but I still think Briggs' general criticism of this method also holds true for the field of linguistics.

2.2. The problem of representativity

Now a certain version of corpus linguistics – and here I come to our second problem, that of representativity – seems to offer a solution here. With sample-corpora and certainly with monitor-corpora one can be quite sure that the influence of the collector of the data itself is negligibly small. But unfortunately that does not solve the problem of representativity, because we still do not know how large a corpus has to be to be representative. Antoinette Renouf, who works at the University of Birmingham in the Unit for Research and Development in English Language, gives a very brief answer to the question what the ideal size of a corpus should be: "It is impossible to say." (Renouf 1987: 130). We can draw here a scientific parallel between linguists and particle physicists. Linguists ask for larger and larger corpora, particle physicists ask for larger and larger and ever more expensive machines to find more and more particles. There are, however, in spite of this comparison two basic differences between what physicists do and what linguists do.

First, linguists cannot give a similar promise, as the physicist can, to get a principally better knowledge of the structure of language by using ever larger corpora; you just get more data but you do not get any new and unheard of insights into the nature and structure of a language.

Second, if you really want to monitor ideally every native speaker of English you would get thousands of billions of words every day, a gigantic amount of material and the question would be – is it really worth it? Is it really necessary to produce a good or better dictionary or a good or better grammar of a language?

2.3. The problem of correctness

An answer to this can be given by a close look at our third problem, the problem of correctness. The "authentic" example from *The Economist* I gave at the beginning of this paper, which is plainly wrong, is meant to illustrate the dialectics of variants vs. mistakes. These dialectics constitute an essential force for linguistic change. Whether linguistic utterances are correct or incorrect is **not** a matter of unanimity. What is regarded by one native speaker as a mistake is regarded as a mere variant by another. These differences are necessarily based on **judgements** by native speakers of a language. These differences are not properties of a language itself. They cannot be properties of a language itself, because correctness is not part of language structure. So if you try to describe language without considering the judgements of native speakers about their language you have no possibility whatever to distinguish between variants and mistakes.

3. Conclusion

Now what corpus linguistics proposes to do among other things is to disregard the judgements of native speakers. This aim is valid for all three types of corpora. The speaker of the one-man corpus is supposed to produce language and not judge his language. Also the millions of speakers of a prospective monitor corpus are supposed to produce linguistic data and nothing else. This method is applied to guarantee objectivity and its standards are drawn from the sciences. One important consequence of this is that the corpus linguistics method in its strict sense is principally unable to detect a simple mistake as in the introductory quotation. This again means that a linguistic corpus cannot be more than something like a photograph. Even the most gigantic monitor corpora will in this sense always remain photographs, and of course one would have to discuss the use of such gigantic "photographs". A road map on a scale of 1:1 is an absurdity, which entirely misses the purpose of maps. It is certainly much more difficult to know how large a corpus must be to become a similar absurdity, but certainly there is such a limit. Even if you compare different synchronic corpora through time in order to get a diachronic picture from historically different corpora you may for instance by statistical counting of frequencies of occurrencies be able to describe certain aspects of linguistic change. But what you will never be able to catch hold of is the dynamic of linguistic change resulting from the dialectic relationship between mistakes and variants in the minds of native speakers; and you cannot get hold of this dynamic because you are principally unable catch a mistake as what it is, namely a mistake.

This does not mean, of course, that the use of corpora is quite useless or impossible. There are many different purposes where corpora can be and are used to good effect as the rich literature shows, but the criterion of authenticity must not be stretched as to become a kind of fetish which prevents a good judgement of linguistic mistakes.[1]

REFERENCES

Aktins B.T. Sue – Antonio Zampolli (eds.)

1994 *Computational approaches to the lexicon.* Oxford: OUP.

1996/1997 *Benjamins Book Gazette.* Amsterdam: Benjamins.

Bloomfield, Leonard (ed.)

1917 *Tagalog texts with grammatical analysis.* (University of Illinois Studies in Language and Literature 111.) 2-3.

[1967] [Reprinted Urbana: Johnson Repr.]

[1] I have not said anything about the commercial background of corpus linguistics and its implications for the work with corpora but it is obvious that representativity and authenticity can be used as arguments in a commercial competition to influence a potential user to buy for instance a certain dictionary. Compare Fillmore – Atkins (1994: 375) remarks on the commercial influence on the compilation of dictionary entries.

Briggs, Charles

1986 *Learning how to ask. A sociolinguistic appraisal of the role of the interview in social science research.* Cambridge: CUP.

Chomsky, Noam

1957 *Syntactic structures.* The Hague: Mouton.

Cohen, David – Jessica Wirth (eds.)

1975 *Testing linguistic hypotheses.* New York: Wiley.

Collins Cobuild English language dictionary = Sinclair, John M. (ed.)

Fillmore, Charles – B.T. Sue Atkins

1994 "Starting where the dictionaries stop: The challenge of corpus lexicography", in: B.T. Sue Aktins – Antonio Zampolli (eds.), 349-393.

Greenbaum, Sidney – Randolph Quirk

1966 *Investigating linguistic acceptability.* The Hague: Mouton.

Hornby, Albert S. – Jonathan Crowther (eds.)

1995 *Oxford advanced learner's dictionary of current English.* (5th edition.) Oxford: OUP.

Johansson, Stig (ed.)

1982 *Computer corpora in English language research.* Bergen: Norwegian Computing Centre for the Humanities.

Longman dictionary of contemporary English = Summers, Della (ed.)

Meijs, Willem (ed.)

1987 *Corpus linguistics and beyond.* Amsterdam: Rodopi.

Oomen, Ursula

1982 *Die englische Sprache in den USA: Variation und Struktur.* Tübingen: Niemeyer.

Oxford advanced learner's dictionary of current English = Hornby, Albert S. – Jonathan Crowther (eds.)

Renouf, Antoinette

1987 "Lexical resolution", in: Willem Meijs (ed.), 121-131.

Ringen, Jon D.

1975 "Linguistic facts: A study of the empirical scientific status of generative grammars", in: David Cohen – Jessica Wirth (eds.).

Sampson, Geoffrey

1980 *Schools of linguistics.* London: Longman.

Sinclair, John M.

1982 "Reflections on computer corpora in linguistic research", in: Stig Johansson (ed.), 1-6.

Sinclair, John M. (ed.)

1995 *Collins Cobuild English language dictionary.* London: Harper Collins.

Summers, Della (ed.)

1987 *Longman dictionary of contemporary English.* (2nd edition.) London: Longman.

Wolff, John

1987 "Bloomfield as an Austronesianist", *Historiographia Linguistica* 14: 173-178.