Per Anker Jensen, Finn Sørensen, Carl Vikner

Data, Corpora, and Linguistic Research

Introduction

In recent years the establishing and processing of computer-readable corpora have drawn much attention from all quarters where human language is under scrutiny. Also, projects involving corpus work have received fairly extensive funding. This is not to be wondered at since, unquestionably, computer-based corpora can be useful and important sources for the linguist in that they place at his disposal huge amounts of easily retrievable information about usage - current or obsolete, as the case may be.

This article is not calling into question that the research process may usefully involve reference to information culled from a corpus. Rather, it addresses a notable reappearance in the wake of the creation of large corpora of the inductivist misconception that such large collections of texts provide a secure basis for producing scientific results[1].This view is possibly inspired by the impressive number of items that can be stored in a computer, a circumstance which would seem to strengthen the inductive line of reasoning that true general statements can be derived from true particular statements: "Surely", the inductivist would claim, "the chance of a general statement being false if inferred from such colossal quantities of evidence is infinitesimal!" Seductive though this prospect may seem, it is demonstrably illusory and, if taken seriously, leads to either the barren quest for 'the scientific method', or, equally unfruitfully, to the production of tonnes of quantitative statements which, as they stand, have no bearing at all on our understanding of qualitative aspects of language.

The gist of the argument in this paper is the inviability of the inductivist doctrine as the basis for scientific work. Due to the space available, many points which follow naturally from the acceptance of inductivism, and which are therefore as untenable as the doctrine itself, are not taken on explicitly in this paper, however. Among such are the belief that 'grammaticality', a theoretical construct, can somehow be 'extrapolated' from a linguistic corpus if only the corpus is large enough, and the belief that 'grammaticality' can

---

1 See for instance Lenders/Willée (1986, 31-35), Sinclair (1987), Bergenholtz (1988, 8,11).

somehow be determined statistically. Closely related to such beliefs is the view that the production of frequency lists is an end in itself. This view may be rooted in the assumption that apparent numerical correlations represent scientific discoveries or explanations as they stand, whereas the fact of the matter is that any statistical correlation - if at all relevant to the problem under investigation, and, crucially, this must be justified by the linguist for each individual case - is a datum calling for explanation by an appropriate theory. Thus, the absence of any explicit discussion of such beliefs should by no means be taken to imply that we endorse them.

Before turning to a more detailed discussion of the shortcomings of the inductivist doctrine, we shall give a brief outline of some core components of linguistic research.

Linguistic theories

We assume that all kinds of linguistic research is undertaken explicitly or implicitly within a certain research program. Typical research programs known from the field of linguistics are the structuralist program which was proposed by Saussure, the generative enterprise initiated by Chomsky, and the systemic approach advocated by Halliday. A research program functions as a stabilizing factor which need not be revised all the time, a fact which should not be taken to mean that a research program makes the theories conceived within its bounds immune to falsification.

We further assume that linguistic research comprises at least four kinds of activities: problem appreciation, observation, theorizing, and testing. Problem appreciation is the activity in which the linguist discovers and formulates a problem. Observation is the research activity in which the linguist seeks to come up with descriptive statements of particular facts related to the problem. The outcome of this activity is thus a set of statements describing the facts to be explained. Theorizing is the activity in which the linguist constructs and presents a theory in response to a particular problem. Testing is the activity in which the linguist tries to find possible errors or shortcomings in his theory by confronting it with facts. These four activities are not necessarily individual steps following each other in time, but are often inextricably interwoven. For instance, observation will often occur as part of the problem appreciation or testing activities. On the other hand, a particular piece of linguistic research need not involve all four activities. Thus, a revision of an existing theory or a new formulation of an old problem may constitute major scientific advances.

The term 'theory' refers to a set of general statements which have the status of empirical hypotheses. The primary function we assign to a theory is to predict facts, to explain why the facts are as they are, and to unify as wide a variety of facts as possible. The predictive,

explanatory and unifying requirements that a theory should meet are not meant to make life hard on the linguist, but are epistemological criteria which allow him to assess the power of the theory he is working with as compared to other theories. They are part of a battery of criteria which, if he is lucky, tell him whether the knowledge expressed in the theory does in fact capture the structuring principles that govern the facts.

The general statements constituting a theory have the status of hypotheses about the facts under investigation. In principle then, every bit of an existing theory might be corrected some day in the future, and every bit of a theory should be corrected as soon and as often as necessary. Thus a theory cannot prove anything; nor does it allow the linguist to prove anything. In that respect, a linguistic theory is like any other empirical theory: prone to error and subject to revision. To some extent, then, linguistic research is nothing but the continuous attempt to correct results obtained in the past, some corrections being more pervasive and more profound than others.

The problem of method

So far we have characterized a theory with reference to what we find are its most central features: it is a set of general statements; it should be conceived of as a set of hypotheses; and it should be as powerful as possible in relation to its domain and with respect to its capacity to predict, to explain, and to unify.

We have not presented - and we will not present - any kind of method which allows the linguist to construct a theory. We simply do not believe that such a method exists. The quest for what is called 'the scientific method' is bound to fail for a host of reasons which have been advanced repeatedly and which have slowly come to be accepted by an increasing number of researchers and philosophers of science since the doctrine of inductivism was first challenged by Popper in the 1930's, see Popper (1959, 27-30)[2]. The kind of procedure traditionally assumed by inductivists to constitute a scientific method is an inductive inference procedure allegedly enabling the scientist to infer true general statements from a set of true particular empirical statements. But inductivism is intuitively untenable, incompatible with the practice of linguists and logically speaking simply false. This we would like to illustrate with some examples.

---

2 For a brief update on this problem the reader is referred to the afterword of Suppe (1977, 624-632), where it is concluded that the philosophy of science relegates issues of induction to a minor role in scientific reasoning and repudiates the central importance ascribed to induction by the positivistic program, cf. Suppe (1977, 331). For an overview of the many aspects of the problem of induction, see Popper (1983, 11-158). For a brief but reasonable discussion of the problem of induction, see Harré (1967, chap.5).

The inductive approach is intuitively untenable because known theories invariably make use of theoretical constructs (or concepts), which are in principle unobservable, cf. Lyons (1981,43). For instance, it is impossible to observe, in any reasonable sense of this word, what Saussure called la langue (the language system). For this reason he could not have derived his statements about the system from a set of data. Since, on the one hand, this theoretical construct is not based on induction, and, on the other hand, has engendered much fruitful research, it is unreasonable to insist on the inductive approach.

The doctrine of induction is incompatible with the practice of linguists. Linguists do not come up with their hypotheses (or theories) by applying logical inferences to a set of data, not even those linguists who are professed partisans of induction. Hjelmslev, for example, assumed an inductive procedure, cf. Hjelmslev (1943,13-14). But when he says that "les mots ont ceci de particulier d'être en principe illimités et incalculables" ("words are peculiar in that they are in principle unlimited and incalculable"), cf. Hjelmslev (1957,97), he simply cannot have inferred this statement from a finite set of data. For if the statement is true, then there does not exist a finite set of data which proves it. So the method proposed by Hjelmslev is incompatible with his own practice.

Finally the doctrine of induction is simply logically false, a fact which seems to have been realized as early as Hume, cf. Popper (1963,189). In its strongest form, inductivism claims that, from a set of true data, true generalizations can be derived by applying valid inference rules. But this is false for a very simple reason. No matter how large the set of data is, there will always be instances of the generalizations which are not actually in it. Hence, logically speaking, instances not actually in the set are possible counterexamples to the generalizations. In other words, there are no logical inference rules which allow the linguist to derive true general statements from a set of data, hence the so-called method outlined above is non-existent. It is clear, then, that the doctrine of induction should be abandoned.

Before leaving the doctrine of induction we would like to stress that its appeal might continue for ideological reasons or for technological reasons, cf. Popper (1983,255 ff). But such reasons do not establish it as a fruitful scientific approach.

Not only is the idea of a so-called inductive method due to a misunderstanding of what is actually going on in linguistics and other empirical sciences, but quite generally, there exists no scientific method, no method which guarantees success in scientific research. Research is not programmable. This may be a discouraging result, especially to those who had hoped for some advice about how to carry out scientific research. Luckily, it is possible to give some

actually do, and Popper, for one, has done so repeatedly, e.g. (1979,260, 265-66).

The whole thing starts with a problem, some sort of difficulty. At first we do not have any clear idea of the problem. We do not really understand it. This creates a seemingly insurmountable obstacle, for how can we find a satisfactory solution to a problem if we do not understand the problem? The answer is: We start with an unsatisfactory solution, and then we criticize this solution, i.e. we try to detect possible errors in the solution. When we have found out that a solution does not work, we try to come up with a new guess, and go on to criticize this new solution. In this way we gradually come to understand our problem better. We understand a problem by understanding why it is difficult to solve, and what kind of solutions will not work. It is very important that we criticize our own solutions, because this is the only way we will find their weak points and may be able to produce better solutions. Now, in the case of scientific investigations, our solutions or guesses are theories, and we criticize them by testing them against relevant data, by checking them for internal inconsistencies and by comparing them to other theories, maybe alternative theories which we construct ourselves just for the sake of comparison.

Our theories are bound to contain errors. A systematically critical attitude is the only effective way to find these errors, to learn from them, and to eliminate them. Thus the key word of this Popperian methodology is criticism. Crucially, criticizing our own solutions is not a method which ensures true results, it is a strategy which allows us to obtain better and better results, and that is all we can hope for.

Facts and data

The execution of a research program also involves the preparation of data. In this process the linguist interprets the facts he considers relevant for his field of study. His interpretation of the facts is represented in a series of observation statements, each of which serves as a datum of his investigation. We thus make a sharp and principled distinction between facts, which are out there in the real world, and data, which are the linguist's representation of the facts. The set of data is finite, while the set of facts is infinite. Together these properties of the two sets imply that we will never know all the relevant facts.

Another important property of the set of data is that each datum has the status of a hypothesis, i.e. it either corresponds to a fact or it does not (i.e. it is true or it is not). For some subset of the data we might have even very good reasons to assume that its members do indeed correspond to the facts. But in principle we have to construct a series of arguments and tests which point in the direction of a correspondence, a partial correspondence or no correspondence at all. For this very reason, what we said about the quest for a

scientific method in connection with theorizing also applies at the level of observation. That is, there is no method which allows the linguist to construct a set of data and which, at the same time, guarantees that the constructed set corresponds to the facts.

Thus we take facts to be part of reality. But facts do not just present themselves to the linguist so that he need only observe them and give them some kind of representation. Data collecting - or fact gathering as it is sometimes called - involves a high degree of selection.

The step from facts to data can be compared to the step from a naturally occurring object to a constructed model of it. From this point of view, the set of data, which formally speaking is a database, is a model of the part of the world referred to. The construction of any model - and thus of any set of data - involves simplification, abstraction and selection - or in short, just selection. So, when we are looking at the world in order to find some fact we want to represent in the set of data, we are not interested in all the facts we happen to come across. We want to select only those facts which are relevant to the problem under investigation. Given that it is impossible to construct a model (or a set of data) containing a representation of all the facts, and given that model construction is a selective process governed by higher order principles, it follows that the data construction activity called observation has as its primary goal to single out the relevant facts and thus the relevant data. As should be evident, the relevance of an observed fact is determined by evaluating its effect on our current knowledge (i.e. the theory on which we are working). But relevance cannot be determined without reference to a framework against which it can be measured. And at no point in the research process can we be sure that we have the full set of relevant data available. The crucial point in connection with any talk of data or facts is their relevance. And relevance is not and cannot be a function of either the facts or the data, but only of the framework constructed by the working linguist.

All this means that - contrary to a popular misconception about what scientists do, or ought to do - scientific research does not start with data collecting or observation, it starts with the awareness of a problem. Popper has made this point in a beautiful passage (1963,129):

> a young scientist who hopes to make discoveries is badly advised if his teacher tells him, 'Go round and observe,' and [...] he is well advised if his teacher tells him: 'Try to learn what people are discussing nowadays in science. Find out where difficulties arise, and take an interest in disagreements. These are the questions which you should take up.'

In other words, you should study the problem situation of the day.

Data and corpora

Now, how do corpora fit into this conception of linguistic research? Let us first stress a well-known triviality. Any corpus is just a collection of unanalyzed linguistic material. So by selecting a corpus the linguist has neither a set of facts, nor a set of data. All he has is undifferentiated linguistic material from which to construct pertinent data. In this respect, his situation is comparable to the one of a geologist facing a geological formation. Whether a corpus is a good one or not, depends on its capacity for producing relevant data, i.e. data which correspond to the facts and which contribute to enhancing our knowledge of the language under investigation. Thus a particular set of data does not correspond to the facts just because it has been constructed on the basis of a corpus. This type of data is subject to tests as much as any other proposed set of data. Corpora simply do not constitute a privileged source for the construction of data, nor do they represent in any direct way the facts we want to discover. Hence, corpora are neither necessary nor sufficient for the execution of a research program at the level of observation. However, whether they are used as an object of analysis or as a point of departure for tests of data and theories, they can be and have been very useful instruments to the linguist in his systematic investigations of language.

The final point we want to stress is that even if we have rather strong intuitions as to what would constitute relevant data, such intuitions do not provide us with a method for finding them. In fact, we may never find them. And the construction of a corpus is not a method for the construction of relevant data. There might be innumerable relevant data which cannot be constructed from a corpus no matter how large it is. Or, there might be a lot of potential relevant data in a certain corpus, but we may fail to notice them because our current knowledge of language does not bring them into focus. Last but not least, any corpus will allow the construction of an infinite set of irrelevant data, either because we interpret the facts wrongly or because the data constructed are inconsequential for the current state of our knowledge of language. Like any other material which must be analyzed in order to state what the facts are, a corpus can be a useful instrument. But it does not indicate where the data corresponding to relevant facts can be found. The inductivist belief that a corpus constitutes a secure base for the construction of relevant data or of theories is thus a methodological mistake which leads to results that do not give any hints at all as to the further execution of the adopted research program.

Bibliography

Bergenholtz, H. (1988): Empiriske metoder i sprogvidenskabelig forskning, i: *Hermes 1*, 1988, 7-23.

Harré, R. (1967): *An Introduction to the Logic of the Sciences*. New York: St. Martin's Press 1967.

Hjelmslev, L. (1943): *Omkring Sprogteoriens Grundlæggelse*. København: Akademisk Forlag 1966.

Hjelmslev, L. (1957): "Pour une sémantique structurale", i: Louis Hjelmslev: *Essais Linguistiques*. København: Nordisk Sprog- og Kulturforlag 1959 (deuxième édition 1970).

Lenders, W./G. Willée (1986): *Linguistische Datenverarbeitung. Ein Lehrbuch*. Opladen: Westdeutscher Verlag 1986.

Lyons, J. (1981): *Language and Linguistics. An Introduction*. Cambridge: Cambridge University Press 1981.

Popper, K.R. (1959): *The Logic of Scientific Discovery*. London: Hutchinson 1972.

Popper, K.R. (1963): *Conjectures and Refutations*. London: Routledge and Kegan Paul 1972.

Popper, K.R. (1979): *Objective Knowledge. An Evolutionary Approach*. Revised Edition. Oxford: Clarendon Press 1979.

Popper, K.R. (1983): *Realism and the Aim of Science*, Hutchinson, London.

Sinclair, J.M. (ed.), (1987): *Looking Up. An account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English Language Dictionary*, Collins ELT, London.

Suppe, F. (ed.), (1977): *The Structure of Scientific Theories*, University of Illinois Press, Urbana, Second Edition 1979.