*ZAA* 54.2 (2006): 113-120 ©

VOLKER GAST

# Introduction*

## 1. Central questions addressed in this issue

Corpus linguistics has undoubtedly become one of the most important and most widely used empirical methods in English linguistics. The recent 'popularization' of corpus-based research is certainly related to the availability of large corpora and easy-to-handle software packages in the public domain. While the early pioneers of computer-aided linguistic research needed a significant amount of computer skills, corpus searches can now be easily carried out by any interested linguist. Obviously, there are chances and risks in this. Among the most important chances is, broadly speaking, that one branch of linguistics seems to be gradually developing into a well-behaved empirical science. The major risks of the recent surge in corpus-based studies relate to the methodological foundations: the ease with which corpus searches can be carried out seems to conceal a number of important methodological problems and complications. In some cases, it is not even immediately obvious why the use of corpus data and quantitative methods of analysis should be revealing at all. It is the objective of this special issue of ZAA to provide an overview of some central methodological problems of corpus linguistics, thus contributing to a discussion that has been going on for some time (see e.g. Granger and Petch-Tyson eds. 2003 and Tummers et al. 2005).

The present issue is intended both as a critical stock-taking and as an outlook into possible future developments and extensions of corpus methods. The questions addressed include the following: What resources are available? Where are the limits of corpus methods, and when should corpus research be supplemented by other empirical methods? How reliable are the results? In what respects do corpus methods and resources have to be developed further? How can the statistics be refined? Can new fields of application be opened up? Some of the contributions address these questions directly, while others discuss them in the context of some particular empirical study.

---

* This special issue has grown out of a workshop on corpus linguistics organized by Ekkehard König, Volker Gast and Norbert Schlüter on June 11, 2005 at the Free University of Berlin. This workshop was financed from the Max Planck Research Award for International Cooperation, given to Ekkehard König in 2003 by the Alexander-von-Humboldt Foundation and the Max-Planck Society. The financial support from these institutions is gratefully acknowledged.

## 2. Different ways of doing corpus linguistics

In philosophy – more precisely, in epistemology – the term 'empiricism' contrasts with 'rationalism' and refers to a philosophical tradition which regards knowledge as being primarily based on sensory experience, rather than intuition and *a priori* reasoning. In linguistics, the empiricism/rationalism debate is reflected in the contrast between 'data-driven' and 'theory-driven' research. However, such dichotomies are of course simplifying. Linguists differ in terms of the importance that they attribute to either observation or reasoning, but no one would seriously contend that either of those components can be dispensed with entirely. Strongly rationalist attitudes can be found in parts of theoretical linguistics, which often relies on introspection as the sole source of data (but then, even introspection is one way of observation, a point already made in J. Locke's *Essay Concerning Human Understanding*). On the other extreme, there are linguists with a radically empiricist attitude who aim to reduce the amount of theoretical premises and aprioristic classification to a minimum; but reducing such pre-empirical premises to zero is likewise impossible, since the very process of data-structuring requires the establishment of (*a priori*) criteria for classification (cf. Schmied 1993). In other words, 'empiricism' and 'rationalism' should be regarded as scalar rather than complementary opposites, and what we should really talk about is the way responsibility can be apportioned between observation and reasoning for an accurate description and analysis of language.

The range of research strategies in between the two (idealized) extremes of a 'radical empiricism' and a 'pure rationalism' is broad. As far as corpus linguistics is concerned, we can roughly distinguish three major paradigms (this distinction is partly sociological; for related discussion cf. Tognini-Bonelli 2001, Chs. 4 and 5; Mukherjee 2005, 37-47; Tummers et al. 2005, 233-242). The first approach – located towards the 'empiricist' end of the spectrum – is based on the assumption that information about language can be obtained by determining 'structural properties' of a corpus such as frequency distributions and co-occurrence patterns (where 'structural' refers to the structure inherent in the data itself). On such a view, all the relevant information is contained in the corpus itself, and the linguist's task is to *extract* that information and make it visible. This approach is particularly widespread in lexicography (cf. Sinclair ed. 1987). When applied to grammatical description, it is sometimes called a 'corpus-driven approach' or 'corpus-driven lexicogrammar'. A more explicit characterization of this method is given in Tognini-Bonelli (2001, Chapter 5). In this issue, the 'radically empiricist' way of doing corpus research is represented by Ilka Mindt's contribution.

The second group of corpus linguists – less 'radically empiricist' than the first – approaches the data from the perspective of moderate 'corpus-external' premises such as a model of grammatical description. Unlike in the corpus-driven approach, specific linguistic classifications are consequently taken for granted and used as a means of structuring the data. This type of research is sometimes called 'corpus-based'. Typically, the primary objective of corpus-

based studies is also the finding of frequency distributions, which are then – ideally – interpreted linguistically. Most of 'mainstream corpus linguistics', including descriptive work and variation studies, can be regarded as representing the 'corpus-based' approach (cf. Mukherjee 2005, 37-47 for an overview).

Finally, there is a third type of corpus studies which is located even further towards the centre of the empiricist/rationalist scale and which uses corpora to test *a priori* hypotheses or theories, i.e. to carry out experiments. The difference between this 'experimental' approach, as we may call it, and 'corpus-based' research as outlined above basically concerns the research objectives and the nature of the hypotheses investigated. While corpus-based studies aim to describe a language within a given model (say, a system of classifications and rules), 'experimental corpus linguistics' goes beyond describing the language system itself and uses corpus material as evidence for other – for instance, cognitive – processes underlying the production and processing of language. This approach can be illustrated using a study carried out by Gries (2001, 2003). Gries examines the principles underlying the distribution of verbal particles in English (*John picked the book up* vs. *John picked up the book*). He hypothesizes that the placement of such particles is determined by the 'processing effort' spent by language users. From this (very general) hypothesis a number of more specific ones can be derived which concern phonological, morphosyntactic and semantic properties of the relevant sentences, and it is these specific hypotheses that can be tested using corpus data. For instance, according to the 'Processing Hypothesis' particles will predominantly be verb-adjacent (*John picked up the book*) when the object is indefinite or complex. Since this procedure is compatible with highly abstract models of language, it has a particularly broad range of application, which has, however, not been widely made use of so far.

## 3. Some methodological problems

The three research strategies depicted above rely on various epistemological premises and face several types of challenges. We will only consider two central problems in the following which also figure prominently in some contributions to this issue. The first challenge concerns the status of *a priori* as opposed to empirically motivated classifications (Section 3.1), and the second relates to the role of statistical inference in corpus studies (Section 3.2).

### 3.1 Linguistic classification and objectivity

Doing linguistics means generalizing over natural language data; and generalizing over language, in turn, implies classifying linguistic items. There are two ways how linguistic items can be classified. The first is aprioristic: we can simply categorize natural language elements according to some criterion that seems useful to us. For instance, we can distinguish between 'vowels' and 'consonants' on the basis of articulatory properties (e.g. the degree of obstruction in the air

stream), and we can distinguish 'stative' from 'dynamic' (uses of) verbs according to their semantics. The second way of categorization is *empirically motivated* and usually relies on distributional criteria. For example, we can distinguish the class of speech segments that can occur in the nucleus of a stressed syllable (vowels) from all the others (consonants), or we can distinguish the class of verbs that occurs in the progressive aspect (dynamic verbs) from those verbs that do not have a progressive form (stative verbs). Empirically motivated classifications are generally preferred in linguistics, even in allegedly rationalist traditions like generative grammar.

The question arises as to what role *a priori* classifications should play in corpus linguistics. Given that corpus studies are typically concerned with matters of lexico-grammar and syntax, raising this question amounts to discussing the legitimacy of semantic classifications as a corpus structuring device. The problems associated with such classifications are obvious: since semantic classifications of corpus data are often a matter of interpretation, objectivity is compromised. Still, making semantic distinctions often seems inevitable if one aims to answer questions concerning the language system. For instance, if we aim to determine the portion of specific types of readings of the present perfect in English (cf. Schlüter this issue), we have to make a decision for each occurrence of the present perfect as to what type of reading it represents. Sometimes there are contextual clues, but in most cases the researcher has to rely on his/her intuition. Semantic classifications are also used in Mindt's contribution dealing with propositional complement-taking adjectives: Mindt determines distributional classes of adjectives on the basis of a strictly empirical method, but the environments used to identify distributional classes are partly defined semantically (e.g. 'intentional' vs. 'non-intentional' subjects, 'impersonal' vs. 'referring' *it*). In the research documented in my own paper, I faced the problem that I had to decide whether or not an attested occurrence of *also* contradicts one of my hypotheses, which can only be decided on the basis of a semantic (or pragmatic) interpretation. These problems are challenging because objectivity is one of the most important requirements on any empirical study.

The issue of *a priori* classifications even concerns the most basic and (apparently) most trivial level of classification, namely PoS-tagging. While PoS-tags are probably uncontroversial for major word classes, some minor classifications are less self-evident. For instance, all *self*-forms are tagged as 'PNX' (for 'reflexive pronoun') in the BNC even though there are clearly (at least) two different distributional classes of such forms: (i) a class of '(genuine) reflexive pronouns' (e.g. *He likes himself*; cf. Germ. *sich*), and (ii) a class of 'intensifiers' (e.g. *the president himself*; cf. Germ. *selbst*; see also König and Gast eds. 2002). Of course, we do not have to rely on PoS-tags when carrying out a corpus study, but the problem is of a more general nature: for a study to be 'objective', there should be general agreement on the underlying classifications. Relying on descriptive work such as Quirk et al. (1985) is certainly a legitimate heuristic move, but it seems to me that establishing an empirically motivated 'ontology of

corpus entities' (which may, but need not, be identical to an ontology of linguistic entities more generally) is nevertheless a major methodological desideratum.

### 3.2 Corpora and statistical inference

A second set of methodological questions that emerge in some contributions to this issue concerns the (internal and external) validity of corpus studies. One central challenge is of course the question of 'representativeness' (cf. McEnery and Wilson 1997, 21-22 and 63-66; Tognini-Bonelli 2001, 57-59). This perennial problem is particularly noteworthy because – in spite of its obvious importance – no satisfying solution has been found for it so far. Corpus builders usually take a 'hermeneutic' approach, arguing on the basis of sanity and reason and regarding representativeness "as an act of faith" (Leech 1991, 27; cf. also Clear 1992, Biber 1993). However, given some obvious problems it is hard to keep faith. One central issue concerns the proportion of different registers documented in a corpus, which – ideally – should reflect the proportion of these registers in actually produced language. As Evert (this issue) points out, the BNC contains only 10% of spoken language and 90% of written language, which certainly does not reflect the real proportions. This problem can of course be tackled by using specific sub-corpora (e.g. only spoken or only written language), but this will merely shift the problem, since we will then have to determine to what extent a given sub-corpus is representative of the register that it represents.

A second central challenge concerns the internal data structure of corpora and has important implications for the interpretation of the results, in particular with regard to the application of statistical methods. Even though a corpus is clearly not a random sample of words, it is usually treated as such (cf. Evert this issue). This means that the common statistical methods generally applied to random data are, strictly speaking, inaccurate. Corpora consisting of texts or text fragments are necessarily skewed because, as Evert puts it, the 'unit of sampling' (texts) differs from the 'unit of measurement' (words, sentences, constructions, etc.). Specifically, there are effects of 'term clustering', which have a particularly strong impact on the results for items with a low frequency. Moreover, the distribution of data is sometimes affected by 'persistence effects', a point emphasized by Gries (this issue). Such problems do of course not render statistical methods useless; however, the various effects of non-randomness inherent in corpora should be taken into account in the application of statistical methods. While most corpus studies (especially in the 'corpus-driven' and 'corpus-based' paradigms) still rely on raw frequency counts, the intricacies of corpus statistics require the application of more complex, often multi-factorial statistical models. Some relevant suggestions are made in the contributions by Gries and Evert.

A further aspect relating to the (external) validity of a corpus study is of course the size of the corpora used. As the above discussion has shown, the size of a corpus is not the only parameter determining its quality. However, it is clearly relevant, especially when it comes to the consideration of infrequent structures. This becomes particularly obvious when we consider more

specialized (and therefore smaller) corpora, for instance dialect corpora (cf. Hollmann and Siewierska this issue). Even the largest corpora of specific dialects are too small to allow for any reliable generalizations concerning certain structures such as the 'pronominal double object construction' (*I gave it him/gave him it/gave it to him*). Note that the problem of infrequent structures is also far from irrelevant for major corpora like the BNC. As is well known, more than 50% of word forms occur only once in the corpus, and less than 20% occur more than ten times (Leech et al. 2001, 9). Moreover, even fairly frequent items – such as the aforementioned intensifiers – occur too rarely if one aims to determine specific aspects of semantic or pragmatic appropriateness.

## 4. The contributions to this issue

In her paper on 'Pedagogical applications of corpora', Ute Römer provides an overview of the major corpus resources available and illustrates how these resources can be used in one central field of applied linguistics, namely language pedagogy. Römer identifies three desiderata for future developments: first, she points out that even more resources are needed, for instance corpora of spoken language. Second, she suggests that the inclusion of contrastive and learner corpora would be particularly fruitful for pedagogical purposes but is only rarely found. Finally, Römer calls for 'missionary work', in the sense that corpus resources should be more widely distributed among practitioners such as "teachers, students, materials writers, and syllabus designers" (p. 128).

Norbert Schlüter discusses one important standard of empirical research, namely 'reliability'. He compares the results arrived at by various researchers studying the present perfect in English ('How reliable are the results? Comparing corpus-based studies of the present perfect'). Schlüter takes three aspects of quantitative distribution into account: (i) the relative frequency of verb phrases in the present perfect, (ii) the relative frequency of 'temporally specified' instances of the present perfect, and (iii) the proportion of 'indefinite past' as opposed to 'continuative past' occurrences. The results seem to point to a considerable degree of convergence between the various studies, which correlates directly with the size of the corpora and the granulation of register specification.

In her contribution on 'Distributional data and grammatical structures: the case of so-called "subject extraposition"', Ilka Mindt aims to determine distributional classes of adjectives complemented by *that*-clauses (*happy that …*, *obvious that …*, etc.) using the statistical tool of 'hierarchical cluster analysis'. Mindt's radically empiricist approach has far-reaching consequences in some respects. For instance, Mindt strictly rejects the assumption of any transformational relationships between sentences and thus denies the existence of 'extraposition' in structures such as *It is clear that …* .

In my own contribution, I present preliminary results concerning 'The distribution of *also* and *too*'. Three hypotheses are tested, with different degrees of success. For two of the hypotheses the test procedure reveals considerable

methodological difficulties, and the question is raised to what extent corpus methods are adequate at all to test such hypotheses.

In his contribution entitled 'How random is a corpus? The library metaphor', Stefan Evert addresses some central problems of statistical inference in corpus linguistics. He points out that corpus procedures naturally rely on the 'random sample model', i.e. the assumption that natural language data is randomly distributed. But why is this assumption justified in the first place, given that language is clearly based on a non-random system? Evert shows that the randomness of a sample is not a property of the linguistic data itself but results from the choice of a specific selection of texts in the process of corpus building, which he illustrates using the 'library metaphor'. Still, corpora thus compiled will not be completely random, for the reasons outlined in Section 3.2 above.

The question of how corpus data are to be appropriately interpreted is also addressed by Stefan Gries, who makes 'Some proposals towards a more rigorous corpus linguistics'. A first issue concerns the organization of and perspective on the data contained in a corpus. Gries argues that a distinction between 'by-subjects analyses' and 'by-items analyses', commonly made in psycholinguistics, should also be introduced into corpus linguistics. A second problem, also addressed in Evert's contribution, concerns the role of 'dispersion', i.e. of an uneven distribution of elements over a corpus. Finally, Gries raises some very general points concerning the choice of interpretation methods in corpus linguistics, arguing for the use of more sophisticated methods than are commonly applied.

In their contribution on 'Corpora and (the need for) other methods in a study of Lancashire dialect', Hollmann and Siewierska consider the utility of corpora for the study of dialects. They discuss three variables in Lancashire dialect: (i) the 'pronominal double object construction', (ii) paradigm levelling in past tense BE, and (iii) the realization of the definite article. One important result is that corpus data alone proves too sparse for infrequent structures such as the 'pronominal double object construction'. The problem of 'sparsity of data' is also relevant to the other two variables investigated by Hollmann and Siewierska, which are both subject to intra-speaker variation. Hollmann and Siewierska tackle this issue by considering the degree of 'sociolinguistic salience' associated with each variable. They propose a way of measuring socio-linguistic salience which is based on the accommodation behaviour of speakers in interviews, thus illustrating how corpus data can be supplemented by other empirical methods.

In the last paper of this issue, Anke Lüdeling explores a novel application of corpus linguistics in the domain of comparative linguistics ('Using corpora in the calculation of language relationships'). She aims to determine similarity and genetic relationships between languages on the basis of parallel corpora. Using methods from bioinformatics, Lüdeling illustrates that parallel corpora can be used to generate 'similarity trees', i.e. diagrams showing the relative distance between languages at various levels of linguistic analysis. Such unrooted similarity tress can then be transformed into rooted trees, which in turn can be

regarded as hypotheses for genetic trees. In this way, languages are classified not by reconstructing earlier language stages and positing sound changes, but by determining degrees of linguistic similarity. Even though her results are not fully conclusive due to the insufficiency of her data sources, this method is undoubtedly highly appealing and may be a first step towards an implementation of corpus methods in comparative linguistics. It may turn out to be particularly valuable for languages for which no historical sources are available.

**Works Cited**

Biber, Douglas (1993). "Representativeness in corpus design." *Literary and Linguistic Computing* 8.4, 243-257.

Clear, Jeremy (1992). "Corpus sampling." Leitner, G., ed. *New Directions in English Language Corpora*. Berlin: Mouton de Gruyter, 21-31.

Granger, Sylviane and Stephanie Petch-Tyson, eds. (2003). *Extending the Scope of Corpus-Based Research: New Applications, New Challenges*. Amsterdam: Benjamins.

Gries, Stefan Th. (2001). "A multifactorial analysis of syntactic variation: particle movement revisited." *Journal of Quantitative Linguistics* 8.1, 33-50.

Gries, Stefan Th. (2003). *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement.* London and New York: Continuum Press.

König, Ekkehard and Volker Gast, eds. (2002). *Reflexives and Intensifiers – the Use of* Self-*Forms in English*. Special Issue of *ZAA*, 50.3.

Leech, Geoffrey (1991). "The state of the art in corpus linguistics." Aijmer, K. and B. Altenberg, eds. *English Corpus Linguistics. Studies in Honour of Jan Svartvik*. London and New York: Longman, 8-29.

Leech, Geoffrey, Paul Rayson and Andrew Wilson (2001). *Word Frequencies in Written and Spoken English: based on the British National Corpus.* Harlow etc.: Pearson Education.

McEnery, Tony and Andrew Wilson (1997). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Mukherjee, Joybrato (2005). *English Ditransitive Verbs. Aspects of Theory, Description and a Usage-Based Model*. Amsterdam: Rodopi.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Schmied, Josef (1993). "Qualitative and quantitative research approaches to English relative constructions." Souter, C. and E. Atwell, eds. *Corpus Based Computational Linguistics*. Amsterdam: Rodopi, 85-96.

Sinclair, John McH., ed. (1987). *Looking Up: An Account of the COBUILD Project in Lexical Computing.* London: Collins ELT.

Tognini-Bonelli, Elena (2001). *Corpus Linguistics at Work*. Amsterdam: Benjamins.

Tummers, José, Kris Heylen & Dirk Geeraerts (2005). "Usage-based approaches in Cognitive Linguistics. A technical state of the art." *Corpus Linguistics and Linguistic Theory* 1.2: 225-261.