Machine-readable text corpora and the linguistic description of languages

Mair, Christian

Veröffentlichungsversion / Published Version
Konferenzbeitrag / conference paper

MACHINE-READABLE TEXT CORPORA AND THE LINGUISTIC DESCRIPTION OF LANGUAGES

CHRISTIAN MAIR

To understand the role of machine-readable text corpora in linguistics it is necessary to consider the four possible sources of data for the linguist, viz. (1) the analyst's own introspection/ intuition, (2) more or less systematically conducted elicitation experiments with groups of native speakers of the language studied, (3) collections of authentic spoken or written citations gathered unsystematically, and (4) evidence extracted systematically from a well-defined corpus of texts. After a discussion of the advantages and disadvantages of the various sources of data, I will briefly exemplify recent advances made in the corpus-based description of languages that have become possible as a result of the application of computer technology to linguistics and then go on to present the major databases currently available for the study of English and German.

1. The Linguist's Data

Linguists draw their primary data from four possible sources. These are: (1) the analyst's own introspection/ intuition - fully available only for the mother tongue, (2) elicitation experiments - ranging from informal polls among friends and colleagues to systematic test batteries used with representative samples of speakers, (3) authentic spoken or written citations collected randomly, and (4) well-defined text corpora from which the relevant evidence can be extracted exhaustively and systematically. The source of data chosen limits both the type of question that can be asked and the results likely to be obtained.

Consider, for example, the English "medio-passives" illustrated by (1):

(1) This book reads well also in translation.

Officials bribe easily in some countries.

* The figures will understand better if presented in a table.

For the benefit of an audience consisting chiefly of non-linguists, the "medio-passive" can be defined as a grammatical structure that is active in form (i.e. the verb *reads* has the same form as in *she reads lots of books*), but passive in meaning (i.e. the first example can be rephrased as *this book can be read well also in translation*). English mediopassives have aroused linguists' interest because they are a recent and spreading innovation and also because English affords a greater variety of such constructions than other European languages.

Of the three sentences given in (1), a native speaker consulting his linguistic intuition will very likely accept the first as normal, the second as possible but somewhat strange but reject the third (hence, following standard linguistic conventions, the asterisk in front of this example). If asked, the native-speaking informant could come up with a few dozen more examples of the construction, which is probably enough for a first approximative linguistic description of the English medio-passive.

However, here as in many other areas of the grammar, there is a grey zone in which native speakers' judgments are unreliable. Consider:

(2) Children educate less easily nowadays than they used to.

Most native-speaking informants will hesitate to give an immediate judgment on this example and resort to *ad-hoc* explanations of the kind: "well, I could imagine a journalist using it," "maybe it's American" (if they themselves are British) or "sounds British" (if they are American).

Obviously, it is precisely this type of phenomenon that lends itself easily to investigation on the basis of elicitation and corpora. Thus, a hundred British or American speakers could be asked for their opinions on:

(3) *East Enders* will screen weekly at 8 p.m. from next Monday.

If the results remain unclear, the number of informants could be increased, or they could be re-grouped according to level of education, age or other potentially significant variables. One problem, however, will remain: what is tested in such elicitation procedures is informants' metalinguistic judgment, which does not necessarily reflect their spontaneous language use. Tests designed to elicit performance are very difficult to construct - how, for example, could an informant be made to actually say or write the structure illus-

trated in (3)? And even in cases where this is possible, it is far from clear whether such elicited performance is the same as spontaneous language use in a natural communicative context.

In this particular case, it is thus the corpus-linguistic approach that will bring us closest to an answer. Analysing machine-readable newspaper corpora from the U.K., the US, Australia and New Zealand, Marianne Hundt (personal communication) finds that the medio-passive use of *screen* originated in the latter variety and is still largely specific to it. Note, however, that one thing which is still unclear is whether this usage is typical of New Zealand English usage in general or restricted to the professional jargon of a particular sub-group of speakers such as, for example, media professionals. At the end of our illustrative analysis, we are therefore thrown back to square one and would probably have to continue by asking native speakers of New Zealand English for their intuition on the matter.

If this little introductory example suggests that in linguistics as well as in many other fields the empirically most adequate descriptions result from methodological pluralism and eclecticism, this is a message which the present writer can endorse fully. In what follows, however, I shall confine myself to an exclusive discussion of the advantages and limitations of the corpus-based approach to the study of languages.

2. The Corpus-Linguistic "Take" on Language

Corpus-based linguistics had a first heyday in the late nineteenth and early twentieth centuries. The renowned Danish Anglicist Otto Jespersen, for example, counted - or, to paraphrase his own words, asked one of his pupils to count - continuous forms of verbs (*I am reading*, etc.) in extracts from two successive translations of the Bible in order to show how such forms had spread in the language from the Middle English period to the time of his writing (1909-49: IV, 177).[1] Of course, the "Shakespeare-corpus" and the "Chaucer-corpus" have long been concordanced and used for literary as well as linguistic research. Charles C. Fries, a prominent representative of the American structuralist school, wrote a grammar of English based on a corpus of letters and telephone conversations. Today, we are awed both by some of the results of such work and by the drudgery that must have gone into it. The inevitable tedium associated with work on them probably explains why corpora went out of fashion for a time.

It was the advent of computational storage and retrieval methods that gave corpus linguistics a new lease of life from the nineteen-sixties onwards. After relatively timid beginnings, we are now seeing a veritable boom in the field, which is due to rapid developments in hardware and retrieval software, the availability of more, larger, and more diversified corpora, and also to the fact that more and more corpus-linguists are beginning to realise that the statistics and examples they dig up are interesting to the linguistic community only to the extent that they shed light on important empirical and theoretical issues.

Today, corpus-linguists have gathered a formidable body of evidence that may eventually force a reorientation in linguistic theory. While in Chomskyan linguistics and related schools of thought which have dominated theoretical debates in the field for the past thirty years, grammatical structure is modelled as an autonomous formal algorithm, corpus linguistics emphasises the fuzziness of all grammatical categories, and the interdependence between structure, meaning and context. The distinction between underlying linguistic "competence" (Chomsky) - a neat and tidy system accessible to introspection - and "performance" - its imperfect realisation in communication - has become increasingly difficult to uphold. The analysis of large masses of data as they are available in machine-readable corpora reveals that there is patterning and order also on the level of performance, and that this needs to be taken into account if we want to understand the nature of language.

Where once there was poverty of data - the changes used to be rung on standard examples of the type *John sees Mary* - there is now an abundance of data. Consider, for example, the following authentic attestations of the verb *see* which were all culled from the 1995 electronic edition of the *Guardian* (1 - 31 March) with minimal effort. They are outside the scope of what general reference grammars and dictionaries have to say about this verb and thus pose an immediate challenge for linguistic analysis:

(4) Although the government is committed to seeing justice done, there are few qualified people to work in the judiciary to deal with the prisoners. (11 March 1995)

(5) Already there are signs that the electrical and machine tools sectors, two of the pillars of the German economy, are seeing their order books affected by the exchange rate. (17 March 1995)

(6) "I am very interested in seeing cable companies provide local channels for community access." (30 March 1995)

Here the verb *see* is not used in its established literal or metaphorical meanings but as a grammatical device. A government that wants to "see" justice done does not really want to see anything at all; it merely wants justice to be done. If someone sees their order books affected by a change in the exchange rate, the emphasis is not on the act of physical or mental perception as such: this is just another way of saying that their order books are affected. Likewise, "to be interested in seeing cable companies do something" is the same as saying "to be interested in their doing something." The use of the verb *see* merely results in a grammatical restructuring of the sentences in question that highlights the position of the grammatical subject as an affected entity, because the gratuitous use of the verb introduces a note of physical substance and human activity into an otherwise abstract state of affairs.

The process of "grammaticalisation" - to give it the name commonly applied to such processes in linguistics - has proceeded even further in the following two examples of nonliteral uses of the verb *see*:

(7) His almost obsessional devotion to horseracing is the more remarkable seeing that his beloved father had striven to keep him away from racing in any shape or form. (12 March 1995)

(8) Not even the Kennel Club, which had no comment to make. Perhaps that's just as well seeing what a hash it makes of things when it does deign to comment. (19 March 1995)

This is not the verbal participle *seeing* that we have in a sentence like *seeing a huge crowd approach, he withdrew.* Rather, *seeing* has become a conjunction, a grammatical operator introducing relevant background information and paraphrasable as "in view of the fact."

This is not a unique development. In the course of its history, English has seen many verbs turn into grammatical operators, for example *concerning*, *regarding*, or *if* (from the Old English imperative form for *giefan*, "give"). The difference between these examples and *seeing* is that the split between the lexical verb and the grammatical operator is not yet complete, and the meaning(s) of the emerging grammatical operator(s) are therefore difficult to pin down precisely. Grammaticalisation is a gradual process, in which the statistical proportions of lexical vs. grammatical uses of a word slowly shift from generation to generation. Data from individual native speakers' introspection are largely irrelevant to a description of the phenomenon, and it is therefore very likely that with the availability of more and more machine-readable corpora covering more and more lan-

guages and registers, the study of grammaticalisation processes will be raised to a qualitatively new level in the near future.

I have mentioned an example from the field of grammaticalisation because it makes a point already made above. The analysis of corpora yields the best results when the new data and technology are used to address existing theoretical issues, and grammaticalisation is just one such case. The fact that lexical items are worn down to grammatical operators in the course of language history is a linguistic universal and has never escaped serious linguists' notice. In fact, the term "grammaticalisation" was coined and defined by Meillet as early as 1912. What was lacking were the rich and easily accessible data bases that would have been required to refine and flesh out a promising hypothesis with the necessary empirical detail. What in the study of language change in Middle English will always remain a dream - matching corpora documenting the state of development in the language every 20 or 30 years for reliable quantitative and qualitative analysis - has now become a reality, and genuine advances in linguistic scholarship are possible.

To put it as briefly as possible, machine-readable corpora are a superior source of data for the linguist for four reasons. The first two are practical in nature, the remaining two concern a more central issue, namely the goals of linguistic research:

(1) Machine-readable corpora make it possible to retrieve large amounts of linguistic data with minimal effort, which allows the exploration of promising as well as not-so-promising working hypotheses in a reasonable span of time. A year's work may be wasted if a large corpus is scanned manually for data which later turn out to be useless; a few hours' worth of work are lost if a similar search is done by computer.

(2) Access to most corpora is not restricted to a few individuals but open to larger sections of the research community. Linguists working on the same material can thus build on each other's results, which leads to co-operation and cumulative progress of a kind not typical of the field as a whole.

(3) Machine-readable corpora are superior sources of data because they present data in their original textual and situational context, sharpening our awareness of the flexibility with which grammatical rules and categories are applied in practice and of the many interdependences between form and meaning or language and context. Close scrutiny of individual examples in context constitutes the qualitative aspect of corpus-based linguistics.

(4) Machine-readable corpora are superior sources of data because they make it possible to analyse the data statistically where, as in the case of grammaticalisation or in the study

of regional, social or text-type specific variation, such analysis is desirable. This it the quantitative aspect of corpus-based work.

In order to give interested outsiders a flavour of what corpora are typically used for these days in linguistics, I would like to conclude this section by briefly commenting on the seventeen contributions published in a recent volume of conference proceedings (Fries, Tottie, Schneider, eds. 1994).

Three contributors report on corpora they are themselves compiling. The first is ARCHER, a somewhat laboured acronym based on the project-title "A Representative Corpus of Historical English Registers." The texts in this corpus are divided into historical compartments of 50 years each, with the focus being on British English but three periods (1750-1799, 1850-1899, 1950-1990) also providing American material in order to make possible the systematic study of regional as well as historical variation. All told, the corpus contains about 1.7 million words and, by present standards, has thus to be included among the small specialised corpora in the field. As an illustration of the type of problem ARCHER could be used to investigate, the compilers show that there is a drastic increase in information-orientation in medical writing after the middle of the nineteenth century, which can be taken as a symptom of the discipline's redefining itself as a hard science. The second paper devoted to corpus building is a report on the progress of the Hong Kong component of the International Corpus of English (on which see below), and Johansson/ Hofland introduce their projected English-Norwegian parallel corpus, which is supposed to benefit not only theoretical linguists but also bilingual lexicographers, language teachers and translators.

It is not without apprehension that I move on to a selective survey of the fourteen papers following, because the type of problem studied by linguists often meets with baffled incomprehension outside the field. But here is the list of what we waste our time doing, for what it is worth.

As I have pointed out above, one strong point of corpora is that they show how artificial the dividing line between grammar and the lexicon really is and how much of language consists of habitually used recurrent word combinations. Henk Barkema pursues this line of inquiry and wants to find out which idioms are inflexible multi-word lexical items and which allow limited modification. *Cold war*, for example, turns out to be of the latter type, with attested variants including *not-so-hot wars*, *melting cold wars*, *periods of hot and cold civil wars*, and so on. Eeg-Olofsson and Altenberg study discontinuous recurrent word combinations (frames like *in_of* or *in_with*) in a corpus of spoken English and - predictably - show that the most frequent way of filling these slots is to produce more prefabricated building blocks (*in terms of, in touch with*) rather than creatively coined

novel expressions. Peters compares variant past tense and participle forms for verbs (e.g. *dreamed/ dreamt*) to determine whether Australian norms are closer to British or to American ones in this part of the grammar. A whole cluster of papers is given to attempts at statistical identification of language change (Nevalainen on adverb derivation, Raumolin-Brunberg on the placement of adjectival modifiers in Late Middle English), text-types or stylistic registers (for example Svartvik/ Ekedahl/ Mosey on "public speaking"). Somewhat surprisingly, the compilers and the users vastly outnumber the computer-science contingent in this volume, as questions of tagging and parsing, that is the automatic grammatical analysis of natural language data, are touched on in only one paper (Voutilainen/ Haikkilä).

3. The Major Resources for the Corpus-Linguist Working on the English Language

In what follows I will discuss the most important English-language corpora, focussing chiefly on those that can be installed on personal-computers and are thus available for desk-top research. For a fuller picture, the reader is referred to Taylor/ Leech/ Fligelstone 1991, who list the resources more completely, and Altenberg 1991, a bibliography which documents the major research done using them. Annual updates are provided in the *ICAME Journal* (Bergen, Norway).[2]

The corpora to be discussed fall into two groups: (a) those that are in the range of roughly one million words, carefully sampled and proofread, inevitably aging, and - owing to their limited size - chiefly of use for the study of the most frequent words and grammatical structures in the language, and (b) large and sometimes open-ended collections of text in which proofreading and principled sampling techniques have to some extent been sacrificed on the altar of size.

The prototype of the type (a) corpus is the Standard Sample of Edited American English, named the Brown Corpus after the institution the project was based at. It contains 500 samples of about 2,000 words each, spanning 15 textual genres from press reportage to various types of lowbrow fiction. The British LOB (for Lancaster-Oslo/Bergen) corpus followed, with the new opportunity for systematic research into British-American differences soon yielding an impressive body of research literature. Matching corpora documenting second-language Indian English (Kolhapur), Australian English (Macquarie) and New Zealand English (Wellington) followed suit. The problem was that while the first two corpora, namely LOB and Brown, contained only texts first published in 1961, the later clones included material from the nineteen-seventies and eighties, thus introduc-

ing a most unwelcome distorting factor in the shape of possible linguistic change over this period of time. In order to remedy this difficulty, the present writer decided to compile two new British and American corpora which are to match the originals as closely as possible in size and composition except that the texts included were published in 1991 and 1992. The projects await completion in 1996 and are known in the linguistic community by their somewhat facetious working titles FLOB and Frown (for Freiburg-LOB and Freiburg-Brown). On completion of the new corpora, it will be possible to study systematically not only regional variation between written British and American English but also linguistic change in progress. A question which it will be possible to ask, for example, will be to what extent the grammar of British English has been influenced by American usage over the past thirty years.

The Survey of English Usage corpus (initiated in the pre-computer era by Sir Randolph Quirk, University College London) is a one-million word corpus which contains a sizable amount of surreptitiously recorded spontaneous speech, part of which was later made available in machine-readable form and published in prosodic transcription (Svartvik/ Quirk, eds. 1980). This corpus of English conversation has not been surpassed as a databasa for research on spoken English so far.

The latest venture in the small-corpus field is the International Corpus of English (Sidney Greenbaum, London), which aims to document spoken and written English of the nineties from all major native-speaking and second-language communities. The British component of ICE will be available to the research community shortly.

A pioneering effort in the development of vast and open corpora was the COBUILD corpus (John Sinclair, Birmingham), which from the beginning was planned as a joint venture between academia and the publishing business and has so far resulted in a number of dictionaries and other reference and teaching materials. It has sprouted a number of successor projects, which - like the original - are accessible to the general public with some difficulty only. Most of these ventures are biased towards written language, and it might be argued that they have become the victim of extremely rapid progress as meanwhile vast quantities of machine-readable English, continue to pile up every year "by themselves", as it were, because practically all the major British and American newspapers and journals offer machine-readable editions on CD-ROM.

The crowning achievement in this tradition is undoubtedly the recently published British National corpus, which contains over 100,000,000 words of British English from spoken and written texts and in which - and this is a trailblazing innovation - every word has been automatically tagged for part-of-speech membership. It is the product of a consortium bringing together major dictionary publishers, academic institutions (chiefly the

University of Lancaster) and the British Library (Research and Development Division). It probably offers the best of both the "small" and the "large" corpus worlds, because it is distributed at a very low price and yet offers a hitherto unattained amount of clean and orderly data. Also, it goes some way towards redressing the major drawback in most previous projects, namely the under-representation of spontaneous speech.[3]

It is impossible for me to give a similarly detailed survey of corpora available for the study of other languages. However, one resource which needs to be mentioned is the Multilingual Corpus published in 1994 as a compact-disc by the European Corpus Initiative of the Association for Computational Linguistics. People interested in computer-based work on German language and literature will find annual updates in the first yearly issues of the journal *Germanistik*. The major centre of corpus-based descriptive work on modern German is, of course, the Institut für Deutsche Sprache (IDS) in Mannheim.

4. Conclusion: Corpus - Linguistics and Neighbouring Fields

The greater part of corpus-based research in linguistics concerns questions and problems that are specific to the discipline and of little interest to outsiders. However, the corpora I have described are available to researchers from other fields for their own purposes. In some areas a dialogue between corpus-linguists and other scholars using the same resources is bound to yield interesting results. In conclusion, I should like to mention three areas in which such interdisciplinary dialogue has in my opinion been long overdue.

In philosophy, Ludwig Wittgenstein started a tradition in which philosophical inquiry was to be preceded by a close scrutiny of the natural-language uses which crucial terms of the analysis were put to. Philosophers of the natural-language school have usually relied on their introspection to establish unreflective natural uses of words and expressions. Without being polemical, however, one might ask whether a trained philosopher's intuition is the closest we can get to current communicative practices in a community of speakers. An analysis of thousands of actual uses of a critical expression as could be culled from suitable corpora is certainly a better indicator of community consensus in usage.

Lexicographers working on large and continually updated corpora are in a perfect position to record the emergence, spread and establishment of new words. Developments in the vocabulary of a language, however - be they the introduction of new words or subtle changes in the meanings of existing ones - frequently mirror changes in society and speakers' attitudes. Take, for example, the increasingly frequent use of the adjective *aggressive* as a positive evaluation of behaviour (as in "aggressive negotiaton tactics", which presumably means "successful tactics," or "an aggressive and ambitious business student"). To the linguist, this is one more example of semantic change of the ameliorative type, in which the meaning of a word loses some of the negative connotations originally associated with it; to the social scientist it may be an indicator to value changes in the community.

To end with a self-evident example, one might point out that the quality of a computer-scientist's or an artificial intelligence researcher's work on natural-language processing is in direct proportion to the corpora that he or she can use as a testing ground for tagging and parsing programs.

Notes

1) Jespersen is here quoted as an illustrative example of corpus-use before the advent of computers. His grammar is otherwise largely based on the author's inexhaustible collection of citations, which he amassed over a lifetime of diligent scholarship, and can thus serve as a perfect example of the third of the four data-gathering strategies mentioned above.

2) ICAME, P.O. Box 53, N-5027, Bergen, Norway, the Association for Computational Linguistics, c/o D. E. Walker, Bellcore, MRE 2A 3-79, 445 South Street Box 1910, Morriston, NJ 07960, USA, and Oxford University Computing Services, 13 Banbury Rd., Oxford OX 2 6 NN, England, are the three major clearing-houses for up-to-date information on resources available and other logistical matters in the field of English corpus-linguistics.

3) To mention one of the more ingenious ways of doing so, for example, certain demographically representative individuals were wired with recording equipment during a set period of time in order to document and obtain a sample of their most authentic speech.

Literature

Aijmer, K. & Altenberg, B. (eds.) (1991): *English corpus linguistics*. London: Longman.

Altenberg, B. (1991): A bibliography of publications relating to English computer corpora. In: Johansson, St. & Stenström, A.-B. (eds.): *English computer corpora: Selected papers and research guide* (pp. 355-396). Berlin: Mouton de Gruyter.

Atkins, B.T.S., Levin, B. & Zampolli, A. (1994): Computational approaches to the lexicon: An overview. In Atkins, B.T.S. & Zampolli, A. (eds.): *Computational approaches to the lexicon (*pp. 17-45). Oxford: OUP.

Butler, C. (ed.) (1992): *Computers and written texts*. Oxford: Blackwell.

Fries, Ch. (1940): *American English grammar*. New York.

Fries, U., Tottie, G. & P. Schneider (eds.) (1994): *Creating and using English language corpora*. Amsterdam: Rodopi.

Jespersen, O. (1909-49): *An English grammar on historical principles*. 5 vols. Copenhagen: Munksgaard.

Sinclair, J. (1991): *Corpus, concordance, collocation*. Oxford: OUP.

Smith, G.W. (1991): *Computers and human language*. Oxford: OUP.

Svartvik, J. (ed.) (1992): *Directions in corpus linguistics*. Berlin: Mouton de Gruyter.

Svartvik, J. & Quirk, R. (eds.) (1989*): A corpus of English conversation*. Lund: Lund University Press.

Taylor, L., Leech, G. & Fligelstone, St. (1991): A survey of machine-readable corpora. In: Johansson, St. & Stenström, A.-B. (eds.): *English computer corpora: Selected papers and research guide* (pp. 319-354). Berlin: Mouton de Gruyter.

Address:

Professor Dr. Christian Mair, Albert-Ludwigs-Universität Freiburg, Englisches Seminar I, Institut für Englische Sprache und Literatur, Kollegiengebäude IV, D-78095 Freiburg, Germany, Tel: +49-761/203-3336, Fax: +49-761/203-3340