Corpus Linguistics

HSK 29.1
Handbücher zur Sprach- und Kommunikationswissenschaft

Handbooks of Linguistics and Communication Science

Manuels de linguistique et des sciences de communication

Mitbegründet von Gerold Ungeheuer (†)
Mitherausgegeben 1985–2001 von Hugo Steger

Herausgegeben von / Edited by / Edités par
Herbert Ernst Wiegand

Band 29.1

Walter de Gruyter · Berlin · New York
Corpus Linguistics

An International Handbook

Edited by
Anke Lüdeling and Merja Kytö

Volume 1

Walter de Gruyter · Berlin · New York
Introduction

1. Why a handbook on corpus linguistics?

Corpus linguistics today is often understood as being a relatively new approach in linguistics that has to do with the empirical study of “real life” language use with the help of computers and electronic corpora. In the first instance, a “corpus” is simply any collection of written or spoken texts. However, when the term is employed with reference to modern linguistics, it tends to bear a number of connotations, among them machine-readable form, sampling and representativeness, finite size, and the idea that a corpus constitutes a standard reference for the language variety it represents. While linguistics divides up into many research areas depending on complexes of research questions, corpus linguistics in essence behaves diametrically: it offers a set of methods that can be used in the investigation of a large number of different research questions.

For a number of reasons, we think that the time is right for a handbook on this approach: we now have access to large corpora and rather sophisticated tools to retrieve data from them. Over the past few decades, corpus linguists have gained a great deal of experience in dealing with both theoretical and practical problems in their research. In other words, we are now much wiser about the ways in which legitimate claims can be made about language use on the basis of corpora. There is also a new focus on empirical data in theoretical linguistics, with growing interest in the techniques and procedures practised within the corpus linguistic approach.

Our handbook is intended to sketch the history of corpus linguistics, and describe various methods of collecting, annotating and searching corpora as well as processing corpus data. It also reports on a number of case studies that illustrate the wide range of linguistic research questions discussed within the framework.

In this Introduction, we will survey the main areas covered in the 61 articles included in the handbook. In the next section, we first give a brief overview of the “roots” of corpus linguistics and then discuss the role played by corpus linguistics in a number of central fields of linguistics. Our aim is to highlight the various ways in which techniques deriving from these fields have contributed to modern corpus linguistics, and vice versa, the ways in which corpus linguistics has been able to contribute to the advances made in these fields. In section 3, we look at the kinds of data that corpora can give us and the kinds of research questions these data can be used for. In section 4, we look into the issues relevant to corpus design, and in section 5, we turn to the links between corpus linguistics and computational linguistics. Finally, in section 6, we introduce the structure of the handbook and the way in which the articles have been grouped under the main section headings.

2. Origins and history of corpus linguistics

As a methodology, the rise of modern corpus linguistics is closely related to the history of linguistics as an empirical science. Many techniques that are in use in corpus linguis-
tics are much older than computers: many of them are rooted in the tradition of the late eighteenth and nineteenth century when linguistics was for the first time claimed to be a “real”, or empirical, science.

2.1. Historical linguistics: Language change and reconstruction

One of the main contributors to modern corpus linguistics is the area of comparative and historical linguistics, the reason being, of course, that historical linguists have always made use of texts or text collections as their material. Many techniques developed in the nineteenth century for reconstructing older languages or recognizing relationships between languages are still in use today. In the Indo-European tradition, the study of language change and the reconstruction effort were dependent on early texts or corpora (cf. Sprachdenkmäler, or “language monuments”). Jacob Grimm and later the Neogrammarians supported their claims about the history and grammars of languages by citing textual data. The Neogrammarians declared that it was the study of present-day language use evidenced in dialects (and not only the study of early texts) that was of utmost importance.

Many ideas and techniques deriving from nineteenth-century scholars were adopted and further developed in modern corpus linguistics. There is currently a great interest in the compilation of historical corpora; historical corpora were also among the first electronically available corpora (e.g. Roberto Busa’s pioneering work on the writings of St Thomas Aquinas and Louis Milič’s Augustan Prose Sample).

The advent of electronically available texts made it possible to collect vast amounts of data relatively rapidly. This in turn enabled scholars to profit from statistical methods in linguistic analysis and to design and develop new sophisticated tools and models for their research purposes. Today, mathematically complex models of language change can be computed by using data drawn from electronic corpora. Both long-term developments and recent change have been found to be of interest.

2.2. Grammar writing, lexicography and language teaching

Nineteenth-century grammarians illustrated the statements they made by examples taken from the writings of recognized authors, e.g. Hermann Paul in his Prinzipien der Sprachgeschichte (Halle: Max Niemeyer, 1st ed. 1880) used German “classics” to exemplify virtually every claim, be it in phonology, in morphology or in syntax. Today, grammar writers may also adopt a corpus-driven approach, but the corpora they now use include not only classics but all kinds of texts. To remedy the so far very much written-biased view, there is also a growing interest in the grammar of spoken language. In grammatical descriptions of a language, corpora can be exploited to arrive at frequency information on usage characteristic of different varieties, registers and so forth.

In lexicography, to take some early examples, the Oxford English Dictionary, and many dictionaries of dead languages give citations from texts containing the word under scrutiny in a context. In modern corpus linguistics, this method is materialized in the form of KWIC concordances. Even though computers make it easier to find and classify
examples and retrieve multi-word entries (modern lexicographic tools use sophisticated
statistics to extract collocations and interesting patterns for each word), the underlying
ideas of exploiting corpus texts are still very similar to those used by early lexicographers
and philologists who had no access to computer technology.

Traditional school grammars and textbooks tend to contain constructed or edited
eamples of language use. In the long run, these can provide only limited support for
students who are sooner or later faced with authentic language data in their assignments.
In this respect, corpora as sources of empirical data play an important role in language
pedagogy. They can also be used to assess the validity of the more traditional teaching
materials based on constructed examples of language use. In language teaching, corpora
provide a source for activating students and engaging them in independent study of
authentic usage. An important recent application is computer-assisted language learning
(CALL), where corpus-based software is used to support the interactive learning activi-
ties students carry out with the help of the computer.

2.3. Sociolinguistics: Language varieties

Another root of modern corpus linguistics is the research on language varieties. It all
began with the compilation of dialect maps and collections of dialect expressions in the
last third of the nineteenth century. The methods were similar to the methods used in
historical linguistics at that time, with one big difference: dialect corpora were systemati-
cally compiled according to given criteria. This might perhaps be seen as a precursor of
the still ongoing discussion of what to include in a corpus (although, of course, today
the issues involved can be very different).

Currently electronic corpora are often used in the research on language varieties (e. g.
dialects, sociolects, registers), and mathematical methods (e. g. multifactorial analyses)
crucially rely on computationally available data.

2.4. Psycholinguistics and language acquisition

Research in psycholinguistics has traditionally meant experiments carried out in the lab-
oratory environment. However, in the study of many important phenomena, for instance
the familiarity of words to a language user in word recognition experiments, corpora
have been profitably used for basic frequency data. Similarly, a speaker’s overall output
can be compared with speech error data drawn from corpora. The analysis of language
pathologies also has a great deal to gain from the study of corpus data: with reference
to data from corpora, it is possible to build hypotheses on the factors underlying the
possible abnormalities detected in a speaker’s output.

In the late 1800s and early 1900s, corpora of child language were composed on the
basis of parental diaries containing records of children’s language. Interest in corpus
collection continued even after the diary studies period. Up until the 1960s, large samples
of data were collected and analysed in order to arrive at norms of development in child
language acquisition. From the 1960s onwards, longitudinal studies have replaced large
sample studies; in longitudinal studies data is collected from individual children over
time.
In a similar fashion, second or foreign language acquisition can now be studied quantitatively and qualitatively much more easily than ever before by examining learner corpora.

2.5. Structuralism

The Indo-European descriptivist tradition continued in a way in American Structuralism (the focus being of course on the synchronic rather than on the diachronic perspective). The goal was to acquire data from many different languages and to develop systematic ways of describing them. In their work American Structuralists contributed to corpus linguistics in at least two ways. First, they raised the issue of data collection. Second, they developed ways of transcribing oral production in languages that had not been recorded in written form earlier on.

2.6. The philosophical tradition

Language philosophy and rationalist theories (Friedrich Frege, Rudolf Carnap, later Noam Chomsky) are not concerned with the empirical study of language. Instead, statements on language use are based on constructed examples and conscious introspective reflections. The aim is to make claims about the ways in which human beings process language and to arrive at a cognitively plausible model of language. The criticism raised by Chomsky against the empirical corpus linguistics approach is justified in many respects. Among other things, it must be recognized that a corpus is inherently a collection of performance data, mostly limited and constrained by external circumstances. How could such data be the appropriate guide for a linguist aiming at modelling linguistic competence? Yet related criticism can be raised against rationalist methodology. How can any serious model of human behaviour be based on artificial examples?

The debate carried out on such fundamental issues as the aims of linguistic analysis and description, and the nature of the data and the analytical tools used has helped corpus linguists to define and justify their position. Today generative linguists are becoming more and more interested in empirical questions and the use of corpus data. Nor can corpus linguists ignore the need to reflect on the issues relevant to language theory and modelling when working on their data (cf. Charles J. Fillmore’s visionary contribution on “‘Corpus Linguistics’ or ‘Computer-aided Armchair Linguistics’” to Directions in Corpus Linguistics edited by Jan Svartvik (Berlin and New York: Mouton de Gruyter, 1992, pp. 35–60)).

3. What can corpus data contribute?

What can corpus data contribute, or in other words, what kinds of data can be used for answering which linguistic research questions? Corpus data cannot replace introspection or provide grammaticality judgments. Nor can corpus data replace experimental data or
fieldwork. The present section deals with corpus data in theoretical linguistics; we come back to the use of corpus data in computational linguistics, in section 5 below.

Corpora can in principle give us three different kinds of data: (1) empirical support, (2) frequency information, and (3) meta-information.

(1) Many linguists use a corpus as an “example bank”, that is, they try to find empirical support for whatever hypothesis, principle or rule they are working on. Examples can, of course, also be made up or simply found by chance, but the corpus linguistics approach provides a search tool which usually enables a good recall of relevant examples in a given corpus. It is often the case that one just “doesn’t come to think of” a relevant example. Many long-standing “truths” have been refuted by corpus data (one example is the often repeated claim that particles in German particle verbs do not topicalize; there are many perfectly grammatical examples of particle topicalization in any corpus). Corpus evidence can be found for verifying hypotheses on each linguistic level from speech sounds to entire conversations or texts. Within the framework, it is possible to replicate the analysis and thus reproduce the results, something which is not possible (and not even intended to be possible) in introspection.

(2) While (1) is a qualitative method of corpus exploitation, corpora provide frequency information for words, phrases or constructions that can be used for quantitative studies. Quantitative studies (which are, of course, often based on a qualitative analysis) are used in many areas of theoretical linguistics as well as in computational linguistics. They show similarities and differences between different groups of speakers or different kinds of texts, provide frequency data for psycholinguistic studies, and so forth. This area is especially interesting because of the different mathematical models that can be used in the analysis.

(3) In addition to the linguistic context, a corpus may provide extra-linguistic information (or meta-data) on such factors as the age or gender of the speaker/writer, text genre, temporal and spatial information about the origin of the text, etc. This extra-linguistic information allows comparisons between different kinds of text or different groups of speakers.

To profit from information that can be retrieved from corpora, the end user needs to know how a given corpus was collected, how it is annotated, how it can be searched, and which statistical techniques can be used. It is necessary to understand the potential (and problems) of corpus data. Corpus linguistics provides a methodology that can be rigidly defined (in the same sense that methodology is defined and standardized in psychology or sociology, and in contrast to introspection or consulting language users about e.g. grammaticality judgments).

4. Corpus design

Many agree that the decisive factor in corpus design is the purpose a corpus is intended to be used for. This purpose largely decides whether a corpus will contain written or spoken language, or both, and what registers and varieties will be represented in it.
Multi-purpose corpora typically contain texts representative of various genres whilst specialized corpora can be limited to highlight only one genre, or a family of genres. In balanced corpora the various components are represented in a stratified way, which enables the scholar to map occurrences of the linguistic phenomena investigated against the extra-linguistic variables involved. In unbalanced corpora, other guidelines than representativity have been followed. Understandably enough, it is not always possible to collect data in similar (or even sufficient) quantities for each text category represented in the corpus; this is often the case with historical corpora.

Though a corpus is traditionally understood as a closed set of textual material, some so-called monitor corpora have also been developed, with the aim of storing a continuously growing (non-finite) body of texts in the memory of the computer. These monitor corpora are useful in lexicographical work. Electronic large-scale dictionaries can also be used for data retrieval in ways similar to corpora (e.g. the *Oxford English Dictionary*).

In multilingual corpora two or several languages are brought together to enable research on translation and contrastive studies, language engineering applications, language teaching, etc. Parallel corpora, on the other hand, bring together the same text in its original language and in translation(s) (cf. the “polyglot” bibles in the Middle Ages). Not only translations but also texts from one and the same topic area written independently in two or several languages can be included in a parallel corpus. A relatively recent introduction to the variety of corpora is learner corpora, where samples of learner production appear and enable systematic study of L2 properties.

5. Corpus linguistics and computational linguistics

Ever since computers were introduced in linguistic analysis, computational linguistics and corpus linguistics have been linked in three ways. In computational linguistics and corpus linguistics, techniques have been developed for structuring, annotating and searching large amounts of text. Techniques have also been designed for qualitative and quantitative study of corpus data. In computational linguistics, corpus data are exploited to develop Natural Language Processing (NLP) applications.

5.1. Preprocessing of corpora

Many corpora that are used by linguists are preprocessed in one or another way: for instance, corpora can be tokenized, split into sentences, and annotated with part-of-speech information. In many cases, information about the (inflectional) morphology of a token has also been added, together with the lemma form. Higher-level annotation such as information on syntactic structure (tree banks), or semantics (anaphoric relations, dependency structures) is also available for certain corpora.

Several competing techniques (e.g. rule-based, statistical, constraint-based, hybrid) have been developed to automate these preprocessing steps, enabling the annotation of large amounts of data within a reasonable amount of time and research effort. Two points are important. Firstly, each preprocessing step forces the corpus compiler to make linguistic decisions that will influence further preprocessing steps and the evaluation of
the corpus. The end user must be aware of these decisions in order to find what he or she is looking for. For instance, the tokenizer has to decide whether to treat *New York* or *Weil der Stadt* (the name of a German town) as one token or several; it will also need to deal with open word classes such as English compound nouns. Similarly, a lemmatizer has to decide what to do with for example German particle verbs. Secondly, the end user must be aware of the quality of the work done over the preprocessing stage and of the error rates involved, as any coding errors, especially systematic errors, may influence the results.

Once end users are aware of the preprocessing principles, corpus texts can be exploited for solving linguistic research questions and for developing NLP products. In our view it is important for our readers to learn about such issues as the techniques that can be used to preprocess a corpus, the (linguistic) decisions that have been made by corpus compilers, and the errors that can be encountered when exploiting corpora. This is why a number of articles on preprocessing have been included in the handbook.

## 5.2. Qualitative and quantitative study of corpus data

Developing techniques for the preprocessing of text is not the only computational linguistic contribution to corpus linguistics. In addition, many methods and standards for the qualitative and quantitative study of text have been developed within the two frameworks. Let us mention only two issues in this context, the work done on grammar development and on morphological productivity.

## 5.3. Corpus data in NLP applications

Many NLP applications rely on the availability of large amounts of textual data. Today, many applications use statistical algorithms that are trained on electronic corpora.

Machine translation is a case in point: with the arrival of fast computers and large amounts of text in the 1970s, it was possible to start using computational techniques for translation purposes. Today, an increasing use is made of parallel corpora and various alignment techniques. While many problems are still waiting for solutions, promising advances have been made within e.g. statistical translation and example-based machine translation.

## 6. Structure of the handbook

The handbook consists of 61 articles organized in five sections.

*Section I: Origins and history of corpus linguistics – corpus linguistics vis-à-vis other disciplines*

In section I, articles 1–3 relate the origins of corpus linguistics to the disciplines in linguistics in the nineteenth and twentieth century. Articles 4–8 show how corpus linguistics and corpus study are positioned with respect to certain central branches of linguistics.
Section II: Corpus compilation and corpus types
Section II explains how corpora are collected, discusses the different corpus types, and introduces some existing corpora. In article 9, different strategies for collecting text and designing a corpus are dealt with. Articles 10–19 describe different types of corpora and data collections.

Section III: Existing corpora
In section III, article 20 provides brief introductions to over a hundred influential corpora (different corpus types, different languages), whilst article 21 introduces some examples of corpora of less studied languages and discusses the special problems related to the collection of these corpora.

Section IV: Preprocessing corpora
In section IV, the different preprocessing steps for corpus data are introduced in articles 22–32. Here our focus is not on the detailed description of algorithms (which are dealt with in the HSK volume on Computational Linguistics) but rather on the decisions that have to be made, on the potential and limitations of the procedure at different stages, and on possible sources of errors. Articles 33–34 introduce search strategies for linear and non-linear material, and article 35 addresses three important aspects of linguistically annotated corpora, i.e. their quality assurance, reusability and sustainability.

Section V: Use and exploitation of corpora
Section V is devoted to the use that can be or has been made of corpora. Article 36 lays the groundwork for articles 37–38: here the different statistical distributions found in corpora are described (normal distribution, LNRE distribution etc.). Articles 39–60 then show how corpus data can be exploited statistically or made other use of. Finally, article 61 surveys the major research designs used in statistical corpus-based analysis.

Anke Lüdeling (Berlin, Germany) and Merja Kytö (Uppsala, Sweden)
Acknowledgments

This project has been an exciting exploration into the world of words and computers! We wish to express our warmest thanks to all those who contributed to the completion of this handbook.

We are indebted to the reviewers of the handbook articles, who generously gave their time to help our authors improve their texts in so many ways. Among the reviewers were our authors, who all reviewed one or more articles, and the following external reviewers: Jan Aarts, Gisle Andersen, Jörg Asmussen, Angelika Becker, Bettina Berendt, Silvia Bernardini, Lars Borin, Peter Bosch, Christian Chiarcos, Massimiliano Ciaramita, Bengt Dahlqvist, Katrin Dohlus, Adrienne Dwyer, Jürg Fleischer, Felix Golcher, Angela Grimm, Peter Grund, Jörg Hakenberg, Patrick Hanks, Andrew Hardie, Sebastian Hoffmann, Graham Katz, Graeme Kennedy, Göran Kjellmer, Peter Kolb, Olga Krasavina, Brigitte Krenn, Emil Kroymann, Jonas Kuhn, Andreas Lücking, Claudia Maienborn, Anna Mauranen, Paola Merlo, Britta Mondorf, Joybrato Mukherjee, Frank Müller, Victoria Oketch, Ralf Plate, Ines Rehbein, Randi Reppen, Beatrice Santorini, Michael Schiehlen, Thomas Schmidt, Bettina Schrader, Hinrich Schütze, Serge Sharoff, Anna-Britta Stenström, Martin Volk, and Benjamin Weiss. Heartfelt thanks to you all.

We were also fortunate to have Professor Tony McEnery (Lancaster University) join us for the initial stages of the editorial project. We take the opportunity of acknowledging the inspiration and insights with which he contributed to our work.

Further, we would have been lost without our assistants at Institut für deutsche Sprache und Linguistik at Humboldt-Universität zu Berlin. Sabine Krämer processed the first batch of articles, and Amir Zeldes drilled through the rest. It was very much thanks to the unfailing stamina and exemplary organisational skills of Amir that we were able to deal with the stream of contributions flooding in at the various stages of the editorial process.

Finally, we wish to express our gratitude to the Series Editor, Professor Herbert Ernst Wiegand, and Dr Anke Beck at Mouton de Gruyter for trusting us with this book. We are also indebted to Ms Barbara Karlson and Ms Monika Wendland at the publisher’s, who always had answers to our questions.

The Editors