Corpus Linguistics and Linguistically Annotated Corpora

SANDRA KÜBLER AND HEIKE ZINSMEISTER

BLOOMSBURY

CORPUS LINGUISTICS AND LINGUISTICALLY ANNOTATED CORPORA

Sandra Kübler and Heike Zinsmeister

BLOOMSBURY
LONDON • NEW DELHI • NEW YORK • SYDNEY

Bloomsbury Academic
An imprint of Bloomsbury Publishing Plc

50 Bedford Square
London
WC1B 3DP
UK

1385 Broadway
New York
NY 10018
USA

www.bloomsbury.com

First published 2015

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

ISBN: HB: 978-1-4411-6447-6
PB: 978-1-4411-1675-8
ePDF: 978-1-4411-1991-9
ePub: 978-1-4411-1980-3

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress.

Typeset by Fakenham Prepress Solutions, Fakenham, Norfolk NR21 8NN
Printed and bound in India

CORPUS LINGUISTICS AND LINGUISTICALLY ANNOTATED CORPORA

Also available from Bloomsbury

AN INTRODUCTION TO CORPUS LINGUISTICS
by David Oakey

COMPUTATIONAL AND QUANTITATIVE STUDIES
by M. A. K Halliday (edited by Jonathan J. Webster)

CONTEMPORARY CORPUS LINGUISTICS
edited by Paul Baker

CORPUS LINGUISTICS: A SHORT INTRODUCTION
by Wolfgang Teubert & Anna Cermáková

CORPUS LINGUISTICS: READINGS IN A WIDENING DISCIPLINE,
by Geoffrey Sampson & Diana McCarthy

LINGUISTICS: AN INTRODUCTION
by William McGregor

WEB AS CORPUS
by Maristella Gatto

WORKING WITH PORTUGUESE CORPORA
edited by Tony Berber Sardinha and Telma de Lurdes São Bento Ferreira

PREFACE

The idea for this textbook emerged when Sandra was teaching corpus linguistics to linguistics and computational linguistics students at Indiana University. One of the goals of this course was to demonstrate to her students how useful annotated corpora and tools established in computational linguistics are. She soon realized the two groups of students differed considerably with regard to prior knowledge. Many concepts that were familiar to her computational linguistics students were new to the linguists. She also found it necessary to introduce students to tools that allow easy access to corpora, especially those that go beyond pure text. Annotated corpora offer two types of challenges: On the one hand, they provide annotations that are often not familiar to linguists. Annotations need to cover complete texts, and thus many phenomena that are not well discussed in linguistic literature. For this reason, they tend to make fewer distinctions than linguistic analyses. On the other hand, the search in annotations requires specialized search tools, which are difficult to figure out on one's own. Additionally, the documentation of annotations and of tools often assumes knowledge that is not readily available to an uninitiated user. The goal of this book is to bridge the knowledge gap between linguistic users and the available documentation of the resources, as well as to promote the use of linguistically annotated corpora to the linguistic community in general.

This book has been a true collaboration between the two authors, with Sandra bringing her expertise in word-level and syntactic annotation to the project and Heike her expertise in semantic and dialogue annotation. But the final form was determined through dialogue. In the end, we both learned a lot about the topics covered in the book, and we also learned that the book is more than the sum of its parts.

We could not have completed this effort without the help of many colleagues in the field. We would like to thank Olga Babko-Malaya, Fabian Barteld, Kathrin Beck, Kelly Harper Berkson, Steven Bird, Donna Byron, Markus Dickinson, Stefanie Dipper, Kerstin Eckart, Johanna Flick, Steven Franks, Ulrich Heid, Erhard Hinrichs, Graeme Hirst, Varada Kolhatkar, Jonas Kuhn, Natalia Modjeska, Anna Nedoluzhko, Stella Neumann, Petya Osenova, Martha Palmer, Massimo Poesio, Arndt Riester, Sabine Schulte im Walde, Heike Telljohann, Yannick Versley, Bonnie Webber, and Tom Zastrow for sending examples, screenshots, and providing general support. Even if we could not integrate all material they provided, they contributed to the content of this book in a very substantial way.

Furthermore, we would like thank the team at Bloomsbury Academic for their immense patience and support during the creation process. We would also like to thank Sandra's and Heike's families, colleagues, and students, who suffered along with us, especially during the final stages. Their patience was greatly appreciated.

Finally, our warmest thanks go to Olga Scrivner for help with proofreading and many useful suggestions that improved the text considerably. All remaining errors are, of course, ours.

Last but not least, we would like to thank all the linguists, annotators, and programmers who were involved in creating the resources and tools that we describe in this book. Without their work, this book would never have been written.

Sandra Kübler, Heike Zinsmeister
April 2014

CONTENTS

PART I
INTRODUCTION

CHAPTER 1
CORPUS LINGUISTICS

1.1 Motivation

Corpus linguistics has a long tradition, especially in subdisciplines of linguistics that work with data for which it is hard or even impossible to gather native speakers' intuitions, such as historical linguistics, language acquisition, or phonetics. But the last two decades have witnessed a turn towards *empiricism* in linguistic subdisciplines, such as formal syntax. These subdisciplines of linguistics used to have a strong intuitionistic bias for many years and were traditionally based on introspective methods. Thus, linguists would use invented examples rather than attested language use. Such examples have the advantage that they concentrate on the phenomenon in question and abstract away from other types of complexities. Thus, if a linguist wants to study fronting, sentences like the ones listed in (1) clearly show which constituents can be fronted and which cannot. The sentence in (2) is an attested example[1] that shows the same type of fronting as the example in (1-a), but the sentence is more complicated and thus more difficult to analyze.

(1) a. In the morning, he read about linguistics.
    b. *The morning, he read about linguistics in.
(2) In the 1990s, spurred by rising labor costs and the strong yen, these companies will increasingly turn themselves into multinationals with plants around the world.

Nowadays, linguists of all schools consult linguistic corpora or use the world wide web as a corpus not only for collecting natural sounding examples, but also for testing their linguistic hypotheses against *quantitative* data of attested language use.

The amount of linguistically analyzed and publicly available corpora has also increased. Many of them had originally been created for computational linguistic purposes, to provide data that could be used for testing or developing automatic tools for analyzing language and other applications. In addition to their original purpose, many of the corpora have been made accessible, for example, in terms of online search interfaces to be readily used and explored by the linguistic community. But even if the resources are available, it is not always straightforward to determine how to use and interpret the available data. We can compare this to arriving in a foreign city. You can wander around on your own. But it is tremendously helpful to have a guide who shows you how to get around and explains how to profit from the characteristics of that

particular city. And if you do not speak the local language, you need a translator or, even better, a guide, who introduces you to it.

This book is intended to guide the reader in a similar way. It guides the reader in how to find their way in the data by using appropriate query and visualization tools. It also introduces the reader to how to interpret annotation by explaining linguistic analyses and their encodings. The first part of the book gives an introduction on the general level, the second part deepens the understanding of these issues by presenting examples of major corpora and their linguistic annotations. The third part covers more practical issues, and the fourth part introduces search tools in more detail. The book has as its goal to make readers truly 'corpus-literate' by providing them with the specific knowledge that one needs to work with annotated corpora in a productive way.

This current chapter will motivate corpus linguistics per se and introduce important terminology. It will discuss introductory questions such as: What is a corpus and what makes a corpus different from an electronic collection of texts (section 1.2)? What kinds of corpora can be distinguished (section 1.3)? Is corpus linguistics a theory or a tool (section 1.4)? How does corpus linguistics differ from an intuitionistic approach to linguistics (section 1.5)? The chapter will end with an explanation of the structure of the book and a short synopsis of the following chapters (section 1.6). Finally, this chapter, like all chapters, will be complemented by a list of further reading (section 1.7).

1.2 Definition of Corpus

A modern *linguistic corpus* is an electronically available collection of texts or transcripts of audio recordings which is sampled to represent a certain language, language variety, or other linguistic domain. It is optionally enriched with levels of linguistic analysis, which we will call *linguistic annotation*. The origin of the text samples and other information regarding the sampling criteria are described in the *metadata* of the corpus.

The remainder of this section will motivate and explain different issues arising from this definition of corpus. For beginners in the field, we want to point out that the term corpus has its origin in Latin, meaning 'body'. For this reason, the plural of corpus is formed according to Latin morphology: one corpus, two *corpora*.

As indicated above, nowadays the term corpus is almost synonymous with *electronically available corpus*, but this is not necessarily so. Some linguistic subdisciplines have a long-standing tradition for working with corpora also in the pre-computer area, in particular historical linguistics, phonetics, and language acquisition. Pre-electronic corpora used in lexicography and grammar development often consisted of samples of short text snippets that illustrate the use of a particular word or grammar construction. But there were also some comprehensive quantitative evaluations of large text bodies. To showcase the characteristic properties of modern corpora, we will look back in time and consider an extreme example of quantitative evaluation in the pre-computer area in which relevant processing steps had to be performed manually.

At the end of the nineteenth century, before the invention of tape-recorders, there had been a strong interest in writing shorthand for documenting spoken language. Shorthand was intended as a system of symbols to represent letters, words, or even phrases, that allows the writer to optimize their speed of writing. The stenographer Friedrich Wilhelm Kaeding saw an opportunity for improving German shorthand by basing the system on solid statistics of word, syllable, and character distributions. In order to create an optimal shorthand system, words and phrases that occur very frequently should be represented by a short, simple symbol while less frequent words can be represented by longer symbols. To achieve such a system, Kaeding carried out a large-scale project in which hundreds of volunteers counted the frequencies of more than 250,000 words and their syllables in a text collection of almost 11 million words. It is obvious that it had been an enormous endeavor which took more than five years to complete.

To make the task of counting words and syllables feasible, it had to be split into different subtasks. The first, preparatory task was performed by 665 volunteers who simply copied all relevant word forms that occurred in the texts on index cards in a systematic way, including information about the source text. Subsequently, all index cards were sorted in alphabetical order for counting the frequencies of re-occurring words. Using one card for each instance made the counting *replicable* in the sense that other persons could also take the stack of cards, count the index cards themselves, and compare their findings with the original results. As we will see later, replicability is an important aspect of corpus linguistics.

The enormous manual effort described above points to a crucial property of modern linguistic corpora that we tend to take for granted as naïve corpus users: A corpus provides texts in form of *linguistically meaningful and retrievable units* in a reusable way.

Kaeding's helpers invested an enormous amount of time in identifying words, sorting, and counting them manually. The great merit of computers is that they perform exactly such tasks for us automatically, much more quickly, and more reliably: They can perform search, retrieval, sorting, calculations, and even visualization of linguistic information in a mechanic way. But it is a necessary prerequisite that relevant units are encoded as identifiable entities in the data representation. In Kaeding's approach, for example, he needed to define what a word is. This is a non-trivial decision, even in English if we consider expressions such as **don't**[2] or **in spite of**. How this is done will be introduced in the following section and, in more detail, in Chapter 3.

1.2.1 Electronic Processing

Making a corpus available electronically goes beyond putting a text file on a web page. At the very least, there are several technical steps involved in the creation of a corpus. The first step concerns making the text accessible in a corpus. If we already have our text in electronic form, this generally means that the file is in PDF format or it is a MS Word document, to name just the most common formats. As a consequence, such files can only be opened by specific, mostly proprietary applications, and searching in

such files is restricted to the search options that the application provides. Thus, we can search for individual words in PDF files, but we cannot go beyond that. When creating a corpus, we need more flexibility. This means that we need to extract the text and only the text from these formatted files. If our original text is not available electronically, we need to use a scanner to create an electronic image of the text and then use an Optical Character Recognition (OCR) software that translates such an image to text. Figure 1.1 shows a scanned image of the beginning of the "Roman de Flamenca," a thirteenth-century novel written in Old Occitan, on the left and the results of using OCR on the image on the right. Generally, the OCR output is not free of errors; see for example the word **di[s]** in the image in Figure 1.1, which is recognized as **dies]**. Thus, if possible, the OCR output should be corrected manually.

We also need to decide how much of the original text formatting we need to represent in the corpus. Some corpus linguists are convinced that it is important to keep information about the number and type of white spaces or which words are written in bold face or in italics. However, if we want to perform linguistic annotations, then such formatting information is often more of a hindrance than an asset. Thus, if we want to keep such information, it would make sense to separate it from the text. Then, we have one file with pure text without any formatting and one file with the additional information.

The next problem that we need to solve is *character encoding*, i.e. the internal representation of characters on the computer. If our texts are in English, this is not a

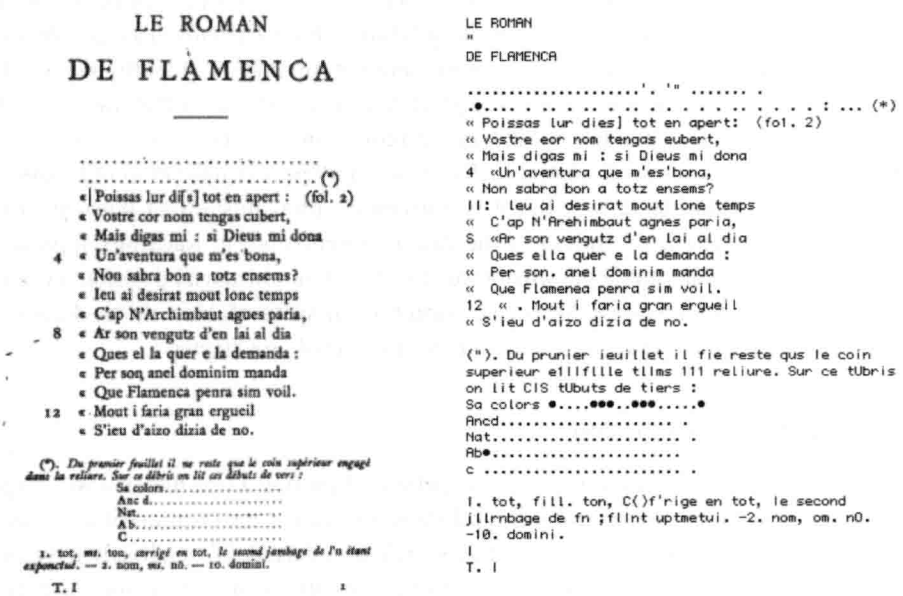

Figure 1.1 An image of the first page of the "Roman de Flamenca" and the OCR output.

problem, but if we have additional characters, such as the accented characters in Figure 1.1, they have to be represented correctly; otherwise, they may be displayed as different characters if the text is opened in a different application. This problem becomes even more challenging for languages such as Arabic or Chinese, which are not based on a character set like English.

Character encodings are not corpus-specific but used for text representation in general. The most basic one is the ASCII format (American Standard Code for Information Interchange), which is a numeric encoding defined to represent the English alphabet, digits from 0 to 9, punctuation symbols, and some formatting-related commands such as space or line break. A more powerful encoding is Unicode, a comprehensive extension of ASCII that allows us to represent all kinds of scripts including Arabic, Cyrillic, or ancient scripts like Egyptian hieroglyphs, and also Chinese characters. A very common specification of Unicode is UTF-8. An important aspect here is that spaces or other empty text positions are also represented as atomic elements in character encoding.

The last problem concerns the *segmentation* of the text, which at this point is a long sequence of characters, into meaningful units, as mentioned above. The units that are generally used are sentences and words. The segmentation into word tokens is called *tokenization*. Non-word symbols can serve as word boundaries but this is not necessarily so. For example, **don't** is often tokenized into two tokens **do** and **not** as in (3-a), where || marks word boundaries. In contrast, multi-word expressions, such as **couch potato** or **in spite of**, are very often tokenized into their components defined by white space irrespective of their non-compositional semantics, cf. (3-b). Punctuation is stripped off the word string to which it is attached in the text and constitutes a token on its own, whereas abbreviations or ordinals are not segmented, as exemplified in (3-c).

(3) a. don't → ||do||not||
b. in spite of → ||in||spite||of||
c. Mr. Spielberg produced "Men in Black". →
||Mr.||Spielberg||produced||"||Men||in||Black||"||.||

The tokenization of corpus text can be done on the fly, for example, based on rules, such as: *Split at white-space and strip off punctuation.* Given the examples in (3), it is clear that there is still room for interpretation, which calls for more detailed rules about how contracted word forms, multi-word expressions, and punctuation are to be handled. In addition, there are scripts like Chinese that generally do not mark word boundaries by spaces. It is clear that tokenization is a much bigger challenge there.

The disambiguation of '.' as period versus other usages such as abbreviation marker plays a central role in automatic sentence segmentation. For human readers, this distinction is easy to make. But for computers, this disambiguation is not straightforward and requires rules or corpus-derived statistical evidence. Other phenomena in sentence segmentation are not as straightforward, even for the human reader. For example, does the colon in (4-a) mark a sentence boundary or not?[3] An argument

against a boundary here is that **What's on the test** functions as the object of **know** hence the first part of the sentence would be an incomplete sentence without the second part after the colon. But should this decision hold even if there is a whole list of dependent sentences, for example, if the students wanted to know several things at the same time as in (4-b), or if the dependent sentence was presented as direct speech as in (4-c)? The way to deal with these cases has to be defined for corpus annotation. There is no predefined correct answer to this question. But it is important that these cases are handled consistently throughout the corpus.

(4) a. ||Since chalk first touched slate, schoolchildren have wanted to know: What's on the test?||
b. ||Since chalk first touched slate, schoolchildren have wanted to know:||What's on the test?||How many correct answers do I need to pass?||...
c. ||Since chalk first touched slate, schoolchildren have wanted to know:||"What's on the test?"||

Hard-wiring tokenization and sentence segmentation by encoding it in the corpus itself is more replicable than doing it on the fly. This can be done implicitly by formatting the text, for example, in the *one-token-per-line format* as in (5-a) for the phrase **in spite of**.[4] Explicit marking involves markup that labels the word units. (5-b) exemplifies a simple markup which encloses each token by an opening tag (<token>) and a closing tag (</token>). For more on the encoding of tokenization, see section 2.2.1.

(5) a. in
spite
of
b. <token>in</token><token>spite</token><token>of</token>

The explicit encoding of tokenization can be seen as an electronic version of Kaeding's index cards with the advantage that the tokens are identified at their text position itself in a re-usable way.

For many corpora, corpus processing does not end with the encoding of tokens or sentences for that matter. These corpora comprise different types of linguistic analyses in terms of annotation. We will introduce the different types of annotation in greater detail in Part II of the book. The take-home message here is the following: A characteristic property of a corpus is that it encodes linguistic meaningful units and potentially other linguistic analyses in a way that they can be searched automatically.

1.2.2 Text Collection

Apart from the more technical issues in representing a corpus, there are also many decisions with regard to the content of a corpus. We will demonstrate these as far as possible using Kaeding's project described above. Kaeding planned to improve

shorthand for German as a language as a whole and not just for a specific variety such as the language spoken in court sessions. To this end, he compiled his corpus from a whole range of genres including topics such as law, economics, theology, medicine, history, mixed newspapers and books, military, private letters, literature, etc. He tried to achieve a database that was *representative* for the use of German in a way that his frequency counts would generalize also to texts not included in his corpus.

This leads to the question of the criteria on which we should sample texts for a corpus. Here, expert opinions differ over whether it is a necessary condition that the collection has been sampled according to explicit sampling criteria or not. For some corpus linguists, there is no corpus without careful sampling. For many computational linguists, on the other hand, any text collection can be dubbed 'corpus' as long as the collection is used to gain some linguistic information from it, i.e. as long as the collection is used as a corpus, it can be called a corpus. Possible sampling criteria include the text type (used by Kaeding), the age of the speakers, or the time when a text was published, to name only a few.

As in Kaeding's early project, the motivation for sampling criteria is that the corpus is built to derive a representative sample for the language or language variety that is the object of investigation. A fundamental problem with this concept is that it is very hard, if not impossible, to define what a particular language or language variety is. For example, what is English? Is it all utterances and written texts produced by English native speakers? Produced today, in the last 20 years, or when? Produced by children and adults alike? Produced by dialect speakers, academics, or also people with little formal education? Produced in private settings or for official publication? It is clear, that we cannot answer the original question properly and, hence, we do not know how a representative sample should look, since we do not know the original object (or *population* in statistical terms). What we can do is *operationalize* the boundaries for the object under investigation.

Such an operationalized sampling strategy has been used, for example, in the BROWN Corpus, for which "English" is operationalized as all English texts published in the year 1961 in the United States—more precisely, as all English-language American publications of 1961 listed in two specific library catalogs. This *sampling frame* provides the population from which the actual corpus texts were randomly sampled, taking further *sampling criteria* into account. One of these criteria is the communicative purpose of a text in terms of its genre. The BROWN Corpus distinguishes fifteen different genres such as press reportage or press editorials. For practical reasons, the BROWN Corpus consists of text samples of only 20,000 tokens each. Given that the limits of computer storage space are decreasing, it is no longer necessary to pose such restrictions on text samples. Another aspect of collecting samples of the same size is that it ensures a *balanced* collection.[5] As a consequence of the size restriction, frequency counts of different subcategories in the corpus can be more easily compared if the relevant reference sets have the same size. However, this consideration is no longer relevant if modern corpus statistical tests are used. Instead of comparing frequency counts, it is advisable to compare relative frequencies complemented with information about confidence