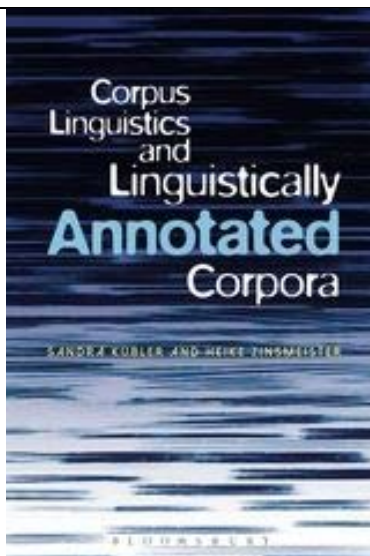


## REVIEW OF *CORPUS LINGUISTICS AND LINGUISTICALLY ANNOTATED CORPORA*

<p><b>Corpus Linguistics and Linguistically Annotated Corpora</b></p> <p>Sandra Kübler &amp; Heike Zinsmeister</p> <p>2015</p> <p>ISBN: 978-1-4411-6447-6</p> <p>US: \$38.61</p> <p>312 pp.</p> <p>Bloomsbury</p> <p>London</p>	
---	--

### Review by **Rodrigo A. Rodríguez-Fuentes, Purdue University**

A corpus without annotations can exist, but its usefulness might be so limited that, for most contemporary linguistic studies, it would not be worth creating. Linguistic Annotation, which signals what research questions can be answered and what hypotheses may or may not be provided with evidentiary support, is a key factor in the way the corpora can be used.

The authors of *Corpus Linguistics and Linguistically Annotated Corpora* (CLLAC) took advantage of being part of two different contexts with overlapping needs in order to write this book. Dr. Sandra Kübler has experience in research in computational linguistics. She is Associate Professor for Computational Linguistics at Indiana University. Heike Zinsmeister has a similar background in computer linguistics annotation and is Professor of German Linguistics and Corpus Linguistics at University of Hamburg, Germany. The book is the outcome of the dialogue established between the expertise of Dr. Kübler in word-level and syntactic annotation and the experience of Dr. Zinsmeister in semantic and dialogue annotation. Together, both authors have nearly 20 years of experience in the field. The authors argue that the book seeks to fill the gap between linguists and computational linguists by establishing a bridge that reflects the confluence of these fields.

This book includes, but it is not limited to, publicly available corpus query and visualization tools. In fact, it is a complete guide to understanding the structure and annotation of different types of linguistically annotated corpora in different languages. The intended audience is not restricted to language researchers. Despite the comprehensive approach of the book, there are many technical considerations that are related to readers trained in computational linguistics or computer programming.

The authors outline the structure of the book based on four questions: what corpus is and why linguistic annotation may provide useful information (Part I), what linguistic annotation is and how it appears (Part II), how to use linguistic annotation as empirical evidence (Part III), and how to retrieve linguistic information from corpus annotation in practice (Part IV).

Each part is subdivided into chapters. Each of the 13 chapters begins with an outline of the content of the respective chapter. This orientation is useful for individualizing the specific aspects of annotation. It is particularly practical in cases when the linguistic features being explored may seem to fit in different

types of annotations, and the terminology might be confusing or technical. The outline also helps the reader to engage in aspects of interest and to establish limits and priorities. At the end of each chapter, there is a section called *Further Reading* with detailed information about each of the corpora cited and references to other books and papers that may be useful for developing more knowledge or doing research in the aspects addressed. CLLAC also has a companion webpage with a list of linguistically annotated corpora and query tools.

Part I (Chapters 1 and 2) is a contextualization of the field in general terms with explanations of basic terms that need to be clearly defined in order for the book to be most advantageous to readers. This section is the shortest and has considerable overlap with other corpus linguistics books, such as Biber, Conrad, and Reppen (1998); McEnery, Xiao, and Tono (2006); and McEnery and Hardy (2012). Nevertheless, this is just an introduction to more in-depth explorations. It includes the historic background (not limited to electronic corpora) that constitutes the principles and caveats that were transferred to electronic corpora and that practically defined the basic architecture of any contemporary corpus. As the reader advances through the book, more background information is required in the field of linguistic annotation. Part I includes a section that provides background about annotation standards for linguistically annotated corpora, which could be helpful in the event of designing a tagger for specific purposes. Fundamental principles related to segmentation are cited but not discussed deeply. Likewise, terms such as *tokenization*, *operationalization*, *representativeness*, *sampling criteria*, *sampling frame*, and *metadata* are defined, and the importance of consistence across approaches and annotation is stressed.

Part II (Chapters 3–6) walks readers through the actual praxis of the principles of linguistic annotation and the basic (but not simple) types of annotation, i.e. Word Level, Syntactic, Semantic, and Discourse. This part also provides views on the importance of understanding annotation for corpus linguists. In the word level annotation, the authors show how to establish a productive relation with the linguistic paradigm of interest and the type of annotation required, as well as how to adapt the linguistic annotation software or tagger to each research interest. In fact, as the focus is on annotation, it makes the information presented relevant across languages and different notions of linguistic elements to provide a more vast understanding and a more complete picture of the field as a whole. The syntactic annotation chapter depends mainly on the concepts of *constituent* and *dependency structure*, which are studied using English, German, and multilingual annotated corpora. Moreover, the authors explain the variables of annotation using *grammatical functions* and give us a detailed introduction to the concept of *treebanks*. Semantic annotation is studied from lexical semantics to semantic relations. These concepts, which seem somewhat simple, could be complex as well. For example, a single lemma might need to be associated with multiple meanings, and annotations need to account for these types of complexities. In Chapter 5, there are detailed explanations of semantically annotated corpora and current projects that are interested in exploring this feature. Following with this stream of thought, Part II ends with *discourse annotation*, a type of tagging that requires semantic features beyond words and sentences and focuses on context and use. Semantic annotation and discourse annotation do not have an opposing relationship as one might think, but a complementary one, which is why pragmatics relies on semantics. In this section, the key concepts are *anaphora* and *coherence* because they are the base upon which the different styles and specificity of discourse annotation occur. In Part II, more specialized knowledge of linguistics is required due to the variables involved in the presentation of subjects, and although it may not be overwhelming for a general audience, the use of certain key terms might not allow an entirely informed reading if the reader is not acquainted with such terms and concepts.

Part III is comprised of only two chapters (7–8), and it discusses how linguists or corpus linguistic researchers can use the corpus tools adequately—once they understand the architecture of the different types of corpora—by taking advantage of the characteristics and singularities that each collection of texts and annotations provide. Chapter 7 establishes an explicit account of uses and limitations of linguistically annotated corpora. These elements are determining factors to take into account for doing research. From

here, key elements in planning, such as the amount of work (automatic annotation and manual checking) required and resources to be used can be foreseen. Likewise, the limitations of each type of corpus could be an important addition to the limitations of a study carried out using certain types of corpus. For the researcher, finding ways to approach these variables could determine the methodology of the study and also could help to set realistic expectations. Chapter 7 also explores and sorts the types of (common) systematic errors that can be expected from automatic annotation software at different linguistic levels. The chapter starts with examples of complex issues with temporal NPs (Noun Phrases) in different types of Parts of Speech (POS) annotation software and advances to more intuitively expected errors in computer-based programming software. Limitations of corpus linguistics in general are addressed carefully, as these are not to be confused with methodological errors.

Chapter 8 holds what might be called a “metadisciplinary” status. This chapter shows the ultimate goal of linguistically annotated software: using it for a research purpose, which requires a constant reflection on the type of annotation used. At this stage of the reading, and especially with the large amount of information already provided, the authors zero in on specific examples of questions to be addressed. The situations presented “show by example how a linguistic question or hypothesis can be translated into a form that allows us to find evidence and counter-evidence in a corpus” (p. 169). The studies cited include detailed and outlined explanations of the linguistic features explored and the type of corpus used, including the Corpus of Contemporary American English (COCA), the British National Corpus (BNC), the Penn Treebank, and the OntoNotes Corpus. The insights in this chapter are extremely detailed with type of study conducted, chronological exposition of procedures, decisions (and their reasons) made by the researchers, type of searches carried out, problems encountered, and print screens of the procedures followed by the results, keeping as the main focus the query tools and their flexibility in favor of the studies.

The last section of the book, Part IV is comprised of five chapters (9–13). This part is indeed an extension of the previous one, as it considers different ways of querying linguistically annotated corpora. Each chapter concentrates on a different type of search, which are related to the type of annotation used in the corpora; however, before the book gets into specific querying, it addresses concordance, and the concordancers AntConc and ParaConc are acknowledged as important query tools that allow users to find concordances in collections of texts without annotation. Chapter 10 examines diverse ways to search for regular expressions of syntax by different classes (e.g. sequences of literal characters, disjunctions, negation, wildcard, counters, classes, ranges, etc.) Chapter 11 has to do with word level queries, supported by COCA, BNCweb, and Corpus Query Processor (CQP). This chapter goes beyond individual words and addresses search for collocations and the association measures used by different corpora to find them. Chapter 12 focuses on syntactic query, based in syntactically annotated corpora. Syntactic features, as Chapter 4 reveals, are usually complex in annotation and relationships. Thus, different types of precedence and dominance are crucial to establish accurate query tools for a corpus linguistic study. Similarly, the architecture and type of annotations used in the corpus are important. In this chapter, Tgrep and Tgrep2 are briefly described as basic query tools. More content is devoted to Tigersearch and Fangorn for constituent annotation, and Netgraph for dependency annotation. Lastly, Chapter 13 focuses on step-by-step query for semantically and discourse annotated corpora. The obligatory connection to Chapters 5 and 6 reminds the reader of the fact that this type of annotation is still scarce, and it is a developing area of linguistic annotation. Still, this chapter presents the process of building a query in FrameNet and its subtools, based on semantic frames, and TimeBank, based on temporal information. For discourse relations, the Penn Discourse TreeBank (PDTBrowser) and the Discourse Relations Reference Corpus (RSTTool), as well as ANNIS3 (Annotation of Information Structure, version 3) are succinctly explored. Part IV relies more than any other part on references to previously presented information in the book due to a particular character that requires some understanding of the information related to linguistic annotation and different uses of corpora tools. It might be extremely technical in both

linguistic and annotation vocabulary, but this technical vocabulary is necessary for this to be a useful reference book that might be re-read by individual parts depending on the features of interest to readers.

CLLAC is a necessary publication to establish a connection between two areas of study (Computer Programming and Linguistics) that are now dependent upon one another in corpus linguistics. This volume is a recommended reference for annotating corpora or undertaking various enterprises with linguistic corpora, whether those endeavors are small or ambitious. It is also a great tool that will help researchers plan and foresee limitations and convey research expectations. One of the interesting features of CLLAC is that it gets the reader acquainted with tools and available software through representative examples of the software's characteristics. All the examples cited in the book correspond to real-life examples from linguistic corpora, annotation software, or codes. Names and trademarks are provided directly. This, besides the illustrative purposes, gives a glimpse into the way annotation software actually works and the different features each brand/type has. For this book, linear reading is suggested due to the number of concepts and ideas that gain complexity as the reader advances through the chapters. At any given point, even for the linguist or the computer programmer, it could be necessary to go back to study and to understand ideas previously explicated.

CLLAC could be defined as a technical book that presents complex information in accessible language. I would certainly recommend it for researchers of language studies, from undergraduates in linguistics and applied linguistics to senior researchers who rely on corpus linguistics and could use it as a resourceful reference tool. Corpus annotation software and tagger initiatives are growing due to the demand for specific features in the annotation and corpus architecture. Despite the fact that the book does not include all available software (indeed, that is not its goal) and would need adaptation for pedagogical use, it absolutely allows readers to get an accurate panoramic view of corpus annotation. It also allows readers to get as specific as desired concerning the available tools for doing linguistic research through the use of corpora and the structure of these. Furthermore, it could be an excellent reference for starting to create a tagger.

All in all, CLLAC has a group of characteristics that very few books in the market can claim: it explores the dimensions of building and annotating corpora in detail, it works as a manual to aid in the understanding of factors for choosing a certain type of annotation tool or annotated corpus, and it shows the actual use of corpora in research with hands-on examples from actual publications that use linguistically annotated corpora.

---

## ABOUT THE AUTHOR

Rodrigo A. Rodríguez-Fuentes is a doctoral candidate in ESL/Second Language Studies at Purdue University. He has taught English as a Foreign Language at all academic levels and has coached English teachers of public schools in Colombia. He has developed on-line materials for the Purdue OWL and has published articles about the relationship between technology and literacy. His current research interests include validation in assessment and testing procedures applied to EFL, much of which integrates technology.

**E-mail:** [rodri246@purdue.edu](mailto:rodri246@purdue.edu)

---

## REFERENCES

McEnery, T., & Hardy, A. (2012). *Corpus linguistics*. Cambridge University Press.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. New York: Routledge.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.