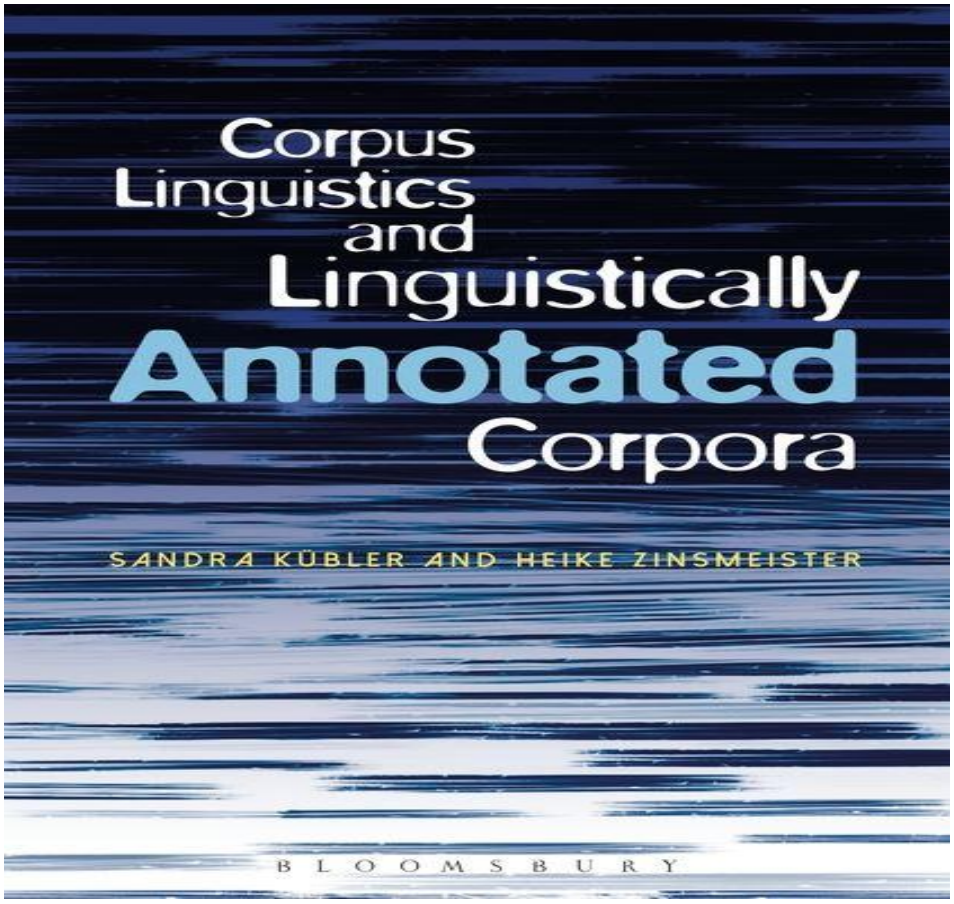
Arab World English Journal
INTERNATIONAL PEER REVIEWED JOURNAL ISSN: 2229-9327
مجلة اللغة الانكليزية في العالم العربي

Arab World English Journal (AWEJ) Volume.7 Number.1 March, 2016

Pp.521-524

Book Review

Corpus Linguistics and Linguistically Annotated Corpora



Authors: Sandra Kubler & Heike Zinsmeister

Title of the Book: Corpus Linguistics *and* Linguistically Annotated Corpora

Publication Date: February 2015

Publisher: Bloomsbury – London & New York

Pages: 320

Reviewer: Lee McCallum

Corpus Linguistics and the creation of corpora center on the creation of a collection of naturally occurring authentic texts that can be electronically stored and their language patterns studied through the use of corpus software. Corpus Linguistics has seen rapid growth in the last two decades especially with the use of learner and multilingual corpora covering both European and non-European languages. Corpus Linguistics and Linguistically Annotated Corpora by Sandra Kubler and Heike Zinsmeister aims to introduce a range of corpus tools and the uses of annotated corpora to traditional linguistics and linguistics students who are not familiar with corpora or the linguistic value they offer. The concept of annotating a corpus involves 'tagging' the corpus' texts to highlight word classes and semantically and syntactically relevant words or longer phrases known as language 'chunks'.

The 320 page book is supplemented by a companion webpage that lists annotated corpora and existing query tools, a further reading list and an exercise section at the end of each chapter. These additions allow readers to build on their initial reading in this book and also practice using the tools the book promotes. In a similar manner, the appendices provide the basic annotation schemes of two well-known corpora: the Penn Treebank corpus and the International Corpus of English (ICE) while the bibliography provides readers with a reference dense reading and resource list of both historic and modern articles, books and tools.

Part 1, from pages 1-21, gives an introduction to Corpus Linguistics and the annotation of corpora. The introduction covers important terminology including what a corpus is and the long-standing history of manual paper-based corpora and now electronically stored corpora. This introductory chapter also details the sampling and collecting of texts when designing a corpus. The authors also stress that the issues of balance and representativeness depend on the purpose of the corpus. The chapter then moves onto the language choice and time frames that the chosen texts will cover i.e. whether or not the corpus will be a static corpus of texts or if it will be a monitor corpus that is regularly updated and added to. Similarly, the reader is encouraged to consider if the texts will be all from one source/genre or if they will be taken from a range of genres/sources. These selection decisions depend on the purpose of the corpus and its intended uses.

Part 1, Chapter 2 covers the different levels of annotation a corpus may take. It introduces word-level and discourse-level annotation involving whole texts. The authors note that the surrounding word company is important for allocating a label to the word under study. The chapter details how English nouns and verbs may be annotated according to mood and number whereas other European languages can annotate adjectives according to gender and number. Nouns can also be annotated according to word class by the use of a POS (Part-Of-Speech) tagger. Words can also be syntactically and semantically annotated at sentence or larger word chunk level to determine structural or meaning patterns. Finally, the authors illustrate discourse annotation to gauge the coherence of a text. The chapter concludes by outlining the argument for and against annotating

a text, searching an annotated corpus and where to find further guidelines on annotation. Part 1 succeeds in raising awareness of the breadth and depth of corpora use and the value of annotated corpora. The strengths of the book become apparent with corpus output introduced and reference is made to other languages besides English.

Part 2, from pages 44-156, expands on Part 1 by devoting an individual chapter to annotation at word-level, syntactic annotation, semantic annotation and discourse annotation. Each chapter goes into a lot of detail and the output from the actual corpora followed by an explanation of how to read and analyze it means the average linguist is walked through the generation of output and its subsequent analysis. As a novice corpus creator and user myself, I feel this is a really valuable addition to the book because first time corpora users can find being faced with seemingly complex output daunting and overwhelming.

Part 3, from pages 157-194, starts by outlining the advantages and limitations of using annotated corpora and serves to remind us that computer annotation, while efficient, is never 100% accurate and manual correction will still in many cases be required. Chapter 8 then uses output from well-known corpora such as the British National Corpus (BNC), the Brigham Young University (BYU) search interface and the Corpus of Contemporary American English (COCA) to demonstrate what scholars have used annotation for. The selection of output here is useful as it exposes the reader to (1) more than one corpus and (2) different varieties of English however, readers who want to see other languages in action may be disappointed that they do not feature in this particular chapter.

Part 4, from pages 196-273, defines concordance lines as lines of text taken from a corpus that show the occurrence of a search word in its text position. Part 4 outlines what concordance lines are used for and where to download and access commercial concordancers that allow texts to be queried. Once again, concordance output is vital for the reader to understand how to read the lines and how they might be used for research and pedagogic purposes. Chapter 10 deals with searching for common/frequent expressions or expressions that interest the linguist. This chapter contains useful shortcuts and search tips to generate different corpus output. The chapter outlines how to search for instances of more than one word or class. For example, the reader is shown how to search for the colours red, blue, yellow or green and this is shown as: "red|blue|yellow|green" where | = or. The chapter also outlines how to search for number and letter ranges as well as different word forms like: begins, begun and beginner. The remainder of Part 4 continues in a similar manner by using symbols and shortcuts to search for word level items (Chapter 11); syntactic structures (Chapter 12) and semantic and discourse phenomena (Chapter 13). Part 4 is resource dense with many corpus tools demonstrated which opens up endless research and pedagogic possibilities to readers.

Overall this book is useful for novice corpus linguists or those who have extensive linguistics experience and are interested in discovering how corpora can assist and deepen their

understanding. The book offers extensive training in annotation and adequately demonstrates how annotation can be linguistically exploited. While the book covers other languages beside English, there is a lack of non-European corpora referenced and this is something readers should be aware of. Despite this, the book remains a solid contribution to Corpus Linguistics and will greatly help those needing guidance in annotation through its clear explanations, extensive demonstrations and example packed chapters.

Reviewer: Ms. Lee McCallum, Prince Sultan University.