
Reviewed by Richard Xiao, Lancaster University

Corpus annotation refers to the process of introducing interpretive linguistic information into a corpus. In spite of earlier criticisms against annotation (e.g. Sinclair 2004a,b; see McEnery, Xiao, & Tono 2006:31–32, McEnery & Hardie 2012:153–57 for further discussion), there has been an increasing consensus that annotation enriches a corpus (McEnery & Hardie 2012:31) and represents ‘added values’ (McEnery et al. 2006:30), substantially broadening the range of research questions that corpora can help to address. Given the time and cost of manual annotation, a range of computational tools (many freely available) have been developed to automate or assist in semi-automatic annotation at the morphological, lexical, syntactic, semantic, and discourse levels. Xiaofei Lu’s book provides an up-to-date, hands-on, practical guide to using such computational tools in automatic corpus annotation and analysis.

The book comprises eight chapters. Ch. 1 sets the scene for the book by discussing its objectives and rationale, explaining the importance and necessity of corpus annotation in linguistic research, and outlining the book’s organization, with an overview of each chapter. L justifies here his decision to focus on tools that are accessed through command-line interfaces, the reason being that corpus annotation and analysis tools with a graphic user interface (GUI) or a web-based interface are usually quite intuitive or are accompanied by detailed user manuals.

Given the book’s focus on tools with command-line interfaces and the author’s assumption—correct, in my view—about the reader’s lack of prior experience with such things, Ch. 2 introduces the command-line interface and illustrates the basic commands for file system management (e.g. creating, renaming, moving, deleting files/directories), as well as some commonly used commands and tools for text processing (e.g. pattern matching and regular expressions).

Chs. 3 and 4 focus respectively on corpus annotation and corpus analysis at the lexical level. Ch. 3 introduces part-of-speech (POS) tagging (including tokenization and segmentation) and lemmatization, and provides step-by-step instructions for downloading, installing, and running the Stanford POS tagger and the TreeTagger via the command-line interface. Publicly available GUI versions of the two tools for use on Windows systems—as well as some other tools, such as the Stanford tokenizer for English, Chinese, and Arabic, the web-based CLAWS POS tagger, and the Morpha lemmatizer for English—are also briefly introduced.

Ch. 4 exemplifies how lexically annotated corpora can be analyzed, in the form of various types of frequency lists and n-grams (e.g. word form, POS, lemma, and their combinations), via the command-line interface. It also illustrates lexical richness analysis, that is, lexical density (in terms of the proportion of content words), lexical variation (in terms of the type-token ratio and its variants), and lexical sophistication (in terms of the proportions of words of different frequency bands), using various command-line interface tools and other Windows- or web-based tools that are publicly available.

Chs. 5 and 6 move the discussion from the lexical to the syntactic level by focusing on syntactic parsing and syntactic analysis, respectively. Ch. 5 discusses two grammar formalisms, namely phrase structure grammar and dependency grammar, and introduces two syntactic parsers based on these theories: the Stanford parser and Collin’s parser.

Ch. 6 consists of two parts. The first introduces some key concepts of tree relationships and provides a tutorial on downloading, installing, and running Tregex, a search engine that effectively queries parsed corpora to retrieve parse trees on the basis of tree relationships and regular expressions. The second reviews a range of metrics that measure syntactic complexity in first and second language acquisition research, and introduces a number of software tools that can be used to automate syntactic-complexity analysis based on such metrics.

Ch. 7 discusses the analysis of semantic fields and propositions at the semantic level, conversational acts at the pragmatic level, and coherence-cohesion and text structure at the discourse level. These types of corpus analysis are presently all automated less reliably than the lexical and syntactic analyses discussed in earlier chapters and are thus usually based on selected linguistic
features, or involve substantial human-machine interaction. Unsurprisingly, the discussion in this chapter is largely more of a review of relevant tools than a tutorial on how to use them.

Finally, Ch. 8 concludes by providing a brief summary of the book and exploring three future directions in computational corpus analysis, namely language meaning and use (e.g. word-sense disambiguation and dialogue-act analysis), learner language (e.g. error analysis), and corpus analysis based on specific linguistic theories (e.g. systemic functional linguistics and cognitive linguistics).

The book is pedagogically geared and clearly targeted at a specific readership. Based on a set of lecture notes for a graduate course, the book is designed as a textbook to provide ‘novice language and linguistics researchers’ (vii) without sophisticated programming skills with ‘a systematic and accessible introduction to the state-of-the-art computational systems and software programs that can be used to automate or semi-automate the annotation and analysis of text corpora at diverse levels’ (1). The chapters outlined above demonstrate that this pedagogical goal set for the book has been achieved successfully. While it is undoubtedly beneficial for researchers interested in corpus analysis—and in corpus creation in particular—to acquire some programming skills, it is nevertheless not always necessary or realistic for everyone to learn programming. In such cases, readers will find it more practical and efficient to take advantage of the existing tools introduced in this book in their corpus annotation and analysis. Hence the book is strongly recommended to anyone who is interested in annotating and analyzing corpus data but who cannot or does not want to learn programming, especially those users of operating systems based on command-line interfaces (e.g. UNIX and similar systems such as Mac OS X and Linux). Readers will find that the book is written in an accessible style using simple and clear language and provides adequate (but not too much) technical information and sufficient manipulation details for them to understand and follow the tutorials.

The book would be greatly improved, however, if the following points were taken into account. First, its pedagogical value could be further enhanced if each chapter included a set of tasks or mini-projects providing learners with an opportunity to practice the skills learned, as well as a list of readings for interested readers to further explore particular aspects of corpus annotation and analysis. Second, there are a variety of errors and typos that may confuse the unsophisticated reader.1 There are also numerous language errors and stylistic inconsistencies in the book that are not enumerated here, but these, together with the errors and typos, require a list of corrections to be published; they should also be corrected in any future revised edition of the book. A final possible improvement relates to the book’s organizational structure: it seems odd that a single-authored book has references listed at the end of each individual chapter.

Despite these imperfections, I highly commend this book and believe readers will find it informative, instructive, and enlightening.

---

1 For example, the pattern matching and substitution exemplified by the ‘sed’ command on p. 30 is ‘s/This/THIS/g’, but is incorrectly written as ‘s/This/That/g’ in the text below the example. In the ‘awk’ command on p. 32, the assignment operator ‘=’ in the condition $1 = ‘may’ should be replaced with the comparison operator ‘==’. The name of the output file in the java command on p. 49 (i.e. ‘sample-input.tag’) does not match the file name given in the explanation on p. 50 (i.e. ‘sample-tagged.txt’). In the discussion of example 5.7b on p. 103, the explanation of the triplet root (‘ROOT-0, bought-7’) is not provided for the example sentence, which has the verb *likes* instead of *bought* as the root. The phrase structure tree that illustrates the analysis of infinitive verb phrases is Fig. 2 rather than Fig. 3, as referenced in the text on p. 119. The tree diagrams and the captions in Fig. 6.4 on p. 121 and Fig. 6.5 on p. 122 do not match; the tree diagrams and the example sentences in the captions should be swapped with each other. The example sentence *John and Mary both left early* (p. 123) should actually be *My bother [sic] and his friends*. In Table 6.1 on p. 125, the pattern ‘A -> B’ is repeated and has contradictory descriptions in rows 20–21 of the table. In Table 6.6 on p. 138, the formula ‘number of words / number of clauses’ for the first instance of the code MLS should be revised to ‘number of words / number of sentences’, while the second occurrence of MLS in the table should be corrected to MLC. In example 7.6 on p. 155, *(CUTE, (ALL, BABIES))’ should be written as ‘(CUTE, (ALL, BABIES))’, with the needed closing parenthesis.

Reviewed by Susan Fischer, Graduate Center, City University of New York

Deaf people cannot hear, so they need a visual means of linguistic expression; sign languages provide that means of expression. Deaf people have been using some form of signing for millennia: Plato observed signing; so did Diderot. Psychologists and linguists in the nineteenth and early twentieth centuries dismissed signing as derivative or primitive; now sign languages are recognized as true languages with expressive capabilities and grammatical complexities equal to those of spoken languages. Modern linguists began studying the structures of sign languages about sixty years ago, first in only a few places, but more recently in many research institutions all over the world; increasing numbers of deaf and hearing linguists have contributed measurably to deeper insights into the structure and use of sign languages.

Pfau, Steinbach, and Woll’s edited handbook of Sign language (SL) constitutes a massive and largely successful effort to bring together a reference on state-of-the-art thinking from a broad range of perspectives on sign languages, and the book’s very existence demonstrates that the study of sign languages has come of age as a legitimate field of endeavor. The first four sections, comprising twenty-one chapters, address the traditional core linguistic categories of phonetics and phonology, morphology, syntax, and semantics/pragmatics. The remaining twenty-two chapters are divided into five sections: communication in the visual modality, psycholinguistics and neurolinguistics, variation and change, applied issues, and handling sign language data. The editors have assembled a group of authors who are among the best qualified to treat the subject matter they have been asked to address: here I would single out chapters by Onno Crasborn on phonetics, Gaurav Mathur and Christian Rathmann on verb agreement, Inge Zwitserlood on classifiers, Diane Lillo-Martin on utterance reports and constructed action, Sarah Taub on iconicity and metaphors, Ronnie Wilbur on information structure, Victoria Nyst on shared (village) sign languages, Susan Goldin-Meadow on home sign, Deborah Chen Pichler on acquisition, Bencie Woll on atypical signing, Dany Adone on language emergence and creolization, Susan McBurney on history, and Christopher Stone on interpreting. Also of particular interest is the integration in several chapters of data with up-to-date linguistic theory. Unfortunately, space does not permit detailed discussion of these interesting efforts.

The earliest and most extensive research has been on American Sign Language (ASL), various western European sign languages, and sign languages in places settled by Europeans, such as Australia and South America. The content in SL is naturally dominated largely by European and North American researchers. Recent work on Asian sign languages, however, has led us to rethink some of our assumptions about how sign languages work and the range of grammatical pos-

Printed with the permission of Susan Fischer. © 2014.