Review of: Martin Weisser (2016), Practical Corpus Linguistics

Naomi Truan

LINGUIST List 27.5091

Tue Dec 13 2016

Review: Computational Ling; Text/Corpus Ling: Weisser (2016)

Editor for this issue: Clare Harshey <clare@linguistlist.org>

Date: 30-Jun-2016
From: Naomi Truan <truan.naomi@gmail.com>
Subject: Practical Corpus Linguistics


E-mail this message to a friend


Discuss this message

Book announced at http://linguistlist.org/issues/27/27-1427.html

AUTHOR: Martin Weisser
TITLE: Practical Corpus Linguistics
SUBTITLE: An Introduction to Corpus-Based Language Analysis
PUBLISHER: Wiley-Blackwell
YEAR: 2016

REVIEWER: Naomi Truan, Université Paris Sorbonne - Paris IV

Reviews Editor: Robert A. Cote

SUMMARY

In “Practical Corpus Linguistics. An Introduction to Corpus-Based Analysis”, Martin Weisser offers an overview of methods and techniques to practice Corpus Linguistics as a student, researcher or teacher. It aims at raising awareness of how corpus evidence can be used for linguistic purposes. “Practical Corpus Linguistics” is conceived as a textbook and, therefore, does not engage into theoretical discussions on Corpus Linguistics. Rather, it puts the emphasis on how to collect, prepare and archive your data to make it suitable for linguistic analysis. The book consists

of twelve chapters, including the introduction and the conclusion. Every chapter relies on many practical exercises commented in detail at the end of each chapter, so that the reader can test his/her knowledge and understanding throughout the book.

The first four chapters begin by defining what a corpus is and asking how it should be collected, to which purpose, and with which implications. Chapter 1, titled “Introduction”, presents the topic of linguistic data analysis. It clearly explains what data is and how it relates to one’s research question. Already relying on corpus examples, it shows the practical and technical difficulties a corpus linguist might be confronted with, especially when the data is not ‘clean’ yet, which makes any automatic analysis of the data very unreliable. Weisser explicitly specifies that theoretical implications of Corpus Linguistics will not be thoroughly analysed, but that book practical exercises will enable the corpus linguist to directly experience Corpus Linguistics ‘at work’.

In Chapter 2 entitled “What’s Out There? A General Introduction to Corpora”, Weisser defines a corpus as “any collection of texts that has been systematically assembled in order to investigate one or more linguistic phenomena” (p. 13, double emphasis from the author). The distinctions between various types of corpora (synchronic vs. diachronic, general vs. specific, static vs. dynamic) are presented in relation with the main corpora available online (summarised in tables).

Chapter 3, “Understanding Corpus Design”, presents the main difficulties regarding the construction of a corpus: sampling, size, legal issues. It then gives an overview of text structures (text body, headers, footers and meta-data) with some practical exercises to look for theses pieces of information in HTML documents, for instance.

Chapter 4 is devoted to “Finding and Preparing Your Data”. Both existing corpora available for analysis (Project Gutenberg and Oxford Text Archive, among others) and corpora that need to be collected by the researcher are extensively presented. Very practical advice is given including which extension should be used for files (.txt) and which encoding format should be the default one (UTF-8). In order to avoid potential errors when running statistics on a corpus, Weisser insists on ‘cleaning up’ manually data and also on documenting and preparing it for further distribution or archiving.

As Weisser recalls in the conclusion, Chapters 5 to 10 deal with “various techniques for analysing language data using established methods of corpus linguistics” (p. 255). “Concordancing” as “an analysis technique that allows linguists to investigate the occurrences and behaviour of different word forms in real-life contexts” (p. 67) is the topic of Chapter 5. It focuses on the free program AntConc. The chapter then deals with the installation on Windows, Mac and Linux, the selection of the files, and the sorting and saving of the results, showing “how useful it is to be able to create concordances like this within a few seconds” (p. 74).

Chapters 6, “Regular Expressions”, states that regular expressions are “an important and very powerful means of specifying […] complex search terms for concordances or computer program for language processing”. The specific options for quantification of regular expressions are listed, showing how qualitative and quantitative analysis of the results might be combined.

Chapter 7, “Understanding Part-of-Speech Tagging and Its Uses” addresses morpho-syntactic annotation, more commonly referred to as “Part-of-Speech (or PoS) tagging”, which the author considers to be “one of the main breakthroughs in corpus linguistics” (p. 101). The Penn Treebank Tagset, a relatively simple one with only 48 tags, is introduced to provide a first insight into possible grammatical categories, followed by the CLAWS (Constituent Likelihood Automatic Word-tagging System), a “far more detailed” (p. 105) one. Both tagsets are displayed in tables, so that an overview is easily accessible. The exercices for this chapter include raising awareness of tagging errors, in order to make the reader more confident in post-editing.

Modern mega corpora such as the BNC, ANC or COCA are addressed in Chapter 8 “Using Online Interfaces to Query Mega Corpora”. The web-based interfaces BNCweb and BYU Web-Interfaces are comparatively presented with an emphasis on the BNCweb. The chapter not only deals with basic standard queries, but also with the navigation online and with headword and lemma queries.

Chapter 9 is entitled “Basic Frequency Analysis – or What Can (Single) Words Tell Us About Texts?”. Nevertheless, “what exactly we should treat as a word” (p. 147) remains a delicate topic, since compounds (all three variants icecream, ice-cream and ice cream are to be found in the BNCweb), multi-words units such as phrasal (prepositional) verbs (e.g. give in/up) or contractions (e.g. can’t, she’s) “are often largely neglected in the analysis of corpus data,

especially in more automated and quantitatively oriented types of corpus analysis" (p. 149). As a result, Weisser introduces the type/token distinction, 'type' being "a representative instance/word form in a frequency list", whereas 'token' refers to "each individual occurrence of a particular type" (p. 149). The chapter further explores tools the reader is already acquainted with at this point, such as AntConc and BNCweb, distinguishing between word (frequency) lists and keyword lists.

"Exploring Words in Context" is the aim of Chapter 10, which proposes to extend the queries to bigger units such as n-grams, word clusters, lexical phrases, or lexical bundles, and colligations ("the co-occurrence […] of specific word classes or lexical items with particular parts of speech, p. 200)" in BNCweb, COCA, and AntConc.

Finally, chapters 11 and 12 show further perspectives for more advanced linguists. In Chapter 11, "Understanding Markup and Annotation", the principles of linguistic annotation, and more specifically of XML files, are introduced. After a brief history of SGML (Standard Generalized Markup Language) and HTML (Hypertext Markup Language), this chapter is a crash course into XML annotation and Text Encoding Initiative (TEI): tags, attributes, and style sheets will no longer be a mystery for you.
Chapter 12 offers a "Conclusion and Further Perspectives", recapitulating what one should have learnt through the book. In order to avoid "specific issues and errors in later analysis stages" (p. 254-255), preparing the data carefully and consciously is crucial. Weisser concludes with a link to a corpora mailing list (http://www.hit.uib.no/corpora/) which might be of interest for many readers.

EVALUATION

This textbook makes Practical Corpus Linguistics accessible to everyone. The focus on methodological and technical aspects and the instructive dimension of the book – nothing is considered obvious or already known – make it very useful to any corpus linguist aiming at a better understanding of his/her data, even though one can regret that the emphasis is clearly on English language.

The prior concern is to show how Corpus Linguistics actually functions and to raise awareness, and this goal is more than fulfilled. Through the various exercises, it is very easy to test one's comprehension and the reader gradually gains confidence. The educational, sometimes entertaining tone as well as the glossary also contribute to gradually enhance the reader's learning capacities in a field in which many feel insecure, and Weisser repetitively encourages the reader: "don't be alarmed if you see a lot of coding […]. We'll learn more about this a little later" (p. 35) or "'[d]on't despair if you end up with lots of errors in the beginning" (p. 237).

Moreover, Weisser's book shows how much data is never 'given' (as the Latin etymology misleadingly says), but is a construction from the researcher, who is responsible for every step on the way to a corpus: collection, preparation, analysis, and archive. The documentation should be part of the process, and this is why Weisser tackles it from the very beginning.

At the same time, this "book being (only) an introductory textbook" (p. 256), it might be too elementary for advanced linguist students or researchers. It is also a pity that the book mainly – if not only – deals with lexical and morpho-syntactic issues, fully neglecting how pragmatic markers can also be explored through Corpus Linguistics techniques. Weisser recognizes that one should not restrict to a kind of lexical or semantic analysis "that largely ignores the fact that words really only gain their 'true meanings' in context" (p. 193), but still addresses principally lexical research questions, even though terms of address or discourse markers are briefly discussed in XML linguistic annotation (p. 238).

It is all the more unfortunate that Aijmer & Rühlemann (2014) recently showed how fruitful Corpus Pragmatics combination of pragmatics and corpus linguistics can be. Discursive, textual and enunciative aspects are also not tackled. In this respect, the free, open source program TXM (http://textometrie.ens-lyon.fr/?lang=en), which is very broadly used (mainly by French speaking scholars since it was developed in France) could also have been mentioned.

Despite this limitation, the book is very readable and well structured. It should accompany scholars at the beginning of any research to raise awareness about technical issues that are too often overlooked, although they play a crucial part in linguistic analysis.

REFERENCES

Aijmer, Karin & Christoph Rühlemann. 2014. Corpus Pragmatics. A Handbook. Cambridge: Cambridge University Press.

ABOUT THE REVIEWER

Naomi Truan is a PhD Student in Contrastive Linguistics at the Université Paris-Sorbonne (Paris IV) and the Freie Universität Berlin (“cotutelle de these”). Her research interests include Pragmatics, Discourse Analysis, Corpus Linguistics and Cognitive Linguistics. Her current work focuses on the category of person, pronouns, terms of address and reported speech in political discourse in France, Germany and Great-Britain.