given in corpus linguistics. The title of the book, *Corpus Linguistics 25 Years on*, however may be a bit misleading, as only Part I really provides the historical account many readers might be looking for. Parts II and III really more offer a snapshot of the current state-of-the-art, not unlike conference proceedings. However, editor Roberta Facchinetti has skilfully managed to maintain a well-rounded topical spread over the contributions. In doing so, the book provides an interesting showcase of the added scientific value corpus-based methods bring to the field of linguistics.

Guy De Pauw, CNTS-Language Technology Group, Department of Linguistics, University of Antwerp, Belgium

doi:10.1093/llc/fqm033
Advance Access Published on 12 November 2007

Corpus Linguistics Beyond the Word–Corpus Research from Phrase to Discourse.

Eileen Fitzpatrick (ed.). *Amsterdam/New York: Rodopi, 2007. Language and Computers: Studies in Practical Linguistics, nr. 60. 277 pp. ISBN: 90-420-2135-7. $81/€58 (hardback)*

This volume contains 15 selected papers originally presented at the 5th North American Symposium on Corpus Linguistics at Montclair State University, New Jersey, in 2004. Focusing on the use of corpora to study domains, 'beyond the word', the symposium covered a wide range of topics, from corpus creation, discourse and register variation to applications in language or medical education, most of them involving tools, approaches, and statistical techniques new to corpus linguistics. The editor uses the phrase 'beyond the word' to indicate linguistic productions longer than the word, from phrases to pragmatics. The papers are arranged in two sections: one on syntactic analysis tools and corpus annotation, and a second on applications in pedagogy and linguistic analysis.

In the first chapter, Barrett, Greenberg, and Schwartz report on exploratory research in automatically selecting documents from distinct domains for machine translation corpora in order to improve results in machine translation. Their method relies on the assumption that texts belonging to different domains have a different syntactic profile. A comparison of part-of-speech tag densities in seven hand-selected documents in four different domains (medical, financial, military, and narrative) suggests that there is a direct correlation between texts from the same domain. This method could also be applied to genre, register or authorship analysis, and in the second chapter, Grieve-Smith investigates whether it is actually possible to exclude grammatical sources of covariation from a list of markers of register, genre or style variation. Using Douglas Biber's (e.g. Biber, 1988) notion of the 'envelope of variation', where grammatical features are counted 'as a proportion of the opportunities for these features to be produced' (p. 21), Grieve-Smith analyses two features that correlate but do not co–vary according to Biber in the Michigan Corpus of Academic Spoken English (MICASE) corpus.

Whereas the first two papers zoom in on words, the focus shifts to phrases in the third paper. Deane and Higgins use singular value decomposition (SVD), a dimensionality reduction technique which compresses a matrix and extracts significant features from it, in order to derive a 'measure of constructional similarity' (p. 43) from Local word contexts extracted from the Lexile corpus. This method allows generalisation over classes and testing for synonyms. Part-of-speech tagging and shallow or partial parsing are essential for all methods described above, although there are quite a few syntactic patterns that may lead to errors, according to van Delden (Chapter 4). He distinguishes between two types of problems: those due to incorrect tags and those occurring in spite of correct tags. The suggested solutions involve adding extra arcs to the finite state automata (FSA), semantic rules or verb sub-categorization. In the fifth paper, Davies reports on a large-scale investigation of register-based variation in a corpus of modern Spanish covering various registers. The corpus was tagged according to 150 syntactic features that might be interesting for a register study. A web-based interface was developed which allows the user to check the relative frequency of all features in each register.

Contrary to words and phrases, discourse phenomena like animacy or linguistic weight involve knowledge that is too complex to be automatically coded. According to Garretson and O'Connor (Chapter 6), these phenomena require a 'combined manual-and-automatic analytical approach' (p. 89) depending on the task. Using the Boston University Noun Phrase Corpus, they illustrate the approach on a case study concerning the possessive alternation in English. Maynard and Leicher (Chapter 7) provide the MICASE corpus with pragmatic annotation in order to expose interesting linguistic phenomena. Whereas teachers usually rely on their intuition, adding this extra annotation layer will allow them easy access to authentic examples. The study of politeness in speech is central in the last paper on tools and annotation. García Vizcaíno discusses the implications of using the British National Corpus (BNC) and the Corpus Oral de Referencia del Español Contemporáneo (COREC) for the study of intonation.

The second part of the book focuses on applications in pedagogy and linguistic analysis, of tools and annotations similar to the ones described in the first part. Davis and Russell-Pinson (Chapter 9) use the Charlotte Narrative and Conversation Collection (CNCC) for two educational purposes: teacher training and medical education. The corpus contains speech data from 'a range of ages, ethnicities, cultures and native languages' (p. 143) which helps teachers to understand their students' backgrounds and conversations with people diagnosed with dementia, allowing professionals to build strategies for communicating with them. Another educational application is the GRIMMATIK method presented by Zinggeler (Chapter 10) which provides a research-based German grammar for students based on fairy tales from the brothers Grimm. The electronic corpus supplies students with grammatical information, meaning, frequencies and usage patterns of queried words.

The next two papers discuss methods for assessing students' foreign language writing skills. De Haan and van Esch (Chapter 11) suggest using syntactic and lexical features in order to indicate progress after one and two years of study of English or Spanish as a Foreign Language (respectively, EFL and SFL), but find that this approach does not lead to unambiguous results. Neff *et al.* present a corpus-based study on errors made by Spanish students in EFL in the twelfth paper. Tagging every error for error type (e.g. form, lexico-grammatical aspects, register, style) leads to the conclusion that two-thirds of the errors can be accounted for, by grammar and lexis.

The last three papers give an idea of what influence and implications corpus linguistics research has for planning the structure of science articles (Chapter 13), writing a grammar of Albanian (Chapter 14) or the study of 19th-century written English (Chapter 15). Shehzad's corpus consists of articles from journals from the Institute of Electrical and Electronics Engineers (IEEE) Computer Society. There is only limited structural variation in these articles, which implies that it can easily be taught. Murzaku offers a quantitative analysis of the third person personal pronoun in Albanian using a corpus of website content and scanned material. Where no diachronic or synchronic study so far has been able to state the presence or absence of this pronoun, corpus research is the answer. In the final paper, Johansson investigates the role of the relativizer *that* in the Corpus of Nineteenth-Century English (CONCE). Generally, *that* is seen as less formal than the *wh*-forms, but a comparison between trials, drama, and letters shows differences in usage.

This book offers a glimpse on the computational tools and statistical techniques for corpus study, corpus annotation schemes and applications of corpus linguistics. The editor provided a thorough introduction and a logical structure. By broadening the scope from words to discourse phenomena, the reader gets an idea of the gradual increase in complexity of research on linguistic productions 'beyond the word'. Because of the diversity of perspectives, any corpus or computational linguist will be activated and brought to new ideas.

A few critical observations can be made concerning the first paper, in which the authors list some limitations of word-based methods for text classification. They looked into literature on

authorship attribution and spam detection, only to find that 'little attention has been paid in these efforts to parts of speech' (p. 2). They state that the work of Brainerd (1973) is one of the exceptions. Nevertheless, there are numerous articles in the field of computational linguistics that investigate the use of syntactic features for text classification (e.g. Argamon *et al.*, 2003; Baayen *et al.*, 1996; Gamon, 2004; Kukushkina *et al.*, 2001; Stamatatos *et al.*, 2001). In the 13th chapter, Shehzad concludes that there is only little structural variation in science articles, without considering that this may be due to corpus structure and journal editing. All science articles are taken from journals issued by the same society (in this case IEEE), whose editors supply authors with style sheets and clear instructions concerning structure. This also implies that authors who deviate too much from these recommendations are asked to restructure or rewrite the article. The reliability of Shehzad's research would definitely have benefited from selecting articles from a few different journals or societies.

In summary, this book should become a resource for anyone interested in corpus linguistics. It broadens the reader's perspective by introducing new tools, approaches and applications. Because the focus is on linguistic productions longer than the word, the first part of the book will be valuable to readers with backgrounds in linguistics, corpus linguistics or computational linguistics. Teachers in foreign language learning will be particularly interested in the second part, where the stress is on pedagogical applications.

References

Argamon, S., Koppel, M., Fine, J., and Shimoni, A. (2003). Gender, genre, and writing style in formal written texts. *Text*, **23**(3): 321–46.

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Brainerd, B. (1973). The computer in statistical studies of William Shakespeare. *Computer Studies in the Humanities and Verbal Behavior*, **4**(1): 9–15.

Baayen, H., Van Halteren, H., and Tweedie, F. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, **11**(3): 121–31.

Gamon, M. (2004). Linguistic correlates of style: Authorship classification with deep linguistic analysis features. *Proceedings of COLING 2004*611–7.

Kukushkina, O., Polikarpov, A., and Khmelev, D. (2001). Using literal and grammatical statistics for authorship attribution. *Problemy Peredachi Informatsii*, **37**(2): 96–108 Translated as 'Problems of Information Transmission'.

Stamatatos, E., Fakotakis, N., and Kokkinakis, G. (2001). Automatic text categorization in terms of genre and author. *Computational Linguistics*, **26**(4): 471–95.

Kim Luyckx, CNTS-Language Technology Group, Department of Linguistics, University of Antwerp, Belgium

doi:10.1093/llc/fqm034
Advance Access Published on 12 November 2007

Deep Time of the Media: Toward an Archaeology of Hearing and Seeing by Technical Means.

Siegfried Zielinski. *Translated by Gloria Custance. Cambridge, MA: MIT Press, 2006. 375 pp. ISBN 0-262-24049-1*. £25.95 (hardback). (Originally *Archäologie der Medien: Zur Tiefenzeit der technischen Hörens und Sehens*. Reinbeck bei Hamburg: Rowohlt Taschenbuch Verlag, 2002.)

'Our culture', physicist and philosopher Ernst Mach (1838–1916) declared at the close of the 19th Century, 'has gradually acquired full independence, soaring far above that of antiquity. It is following an entirely new trend. It centres around mathematical and scientific enlightenment. The traces of ancient ideas, still lingering in philosophy, jurisprudence, art and science constitute impediments rather than assets, and will come to be untenable in the long run in face of the development of our own views.' Such robust triumphalism may seem rather quaint now, but it remains a triumph of the imagination actually to write a history of technology and the sciences that does not in one way or another share in Mach's error. In *Nature and the Greeks*, fellow physicist Erwin Schrödinger quotes that passage from Mach to mark his own attempt to avoid it in