Reviews

Lars Borin (ed.). *Parallel corpora, parallel worlds. Selected papers from a Symposium on Parallel and Comparable Corpora at Uppsala University, Sweden, 22–23 April, 1999.* Amsterdam – New York, NY: Rodopi, 2002. i + 220 pp. ISBN 90-420-1530-6. Reviewed by **Martha Thunes**, University of Bergen.

In April 1999 Lars Borin with colleagues hosted a two-day symposium on parallel and comparable corpora at Uppsala University. Contributions from Swedish research institutions form a clear majority in Borin's collection of selected papers from the symposium. However, the book is international in scope and also contains articles by researchers working in Norway, Germany, the United Kingdom, and the United States.

The book opens with a 40-page state-of-the-art article by the editor, arguing that within corpus linguistics there are two separate research traditions, emerging from, respectively, general and computational linguistics. He presents *Parallel corpora, parallel worlds* as a meeting place for the two, and we shall try to see to what degree there is contact between the traditions in the various articles.

Borin's introductory article presents "an overview of the state of the art of parallel corpus research, particularly the situation in Scandinavia" (p. 2). He names the field *parallel corpus linguistics*, and describes its place within the larger field of corpus linguistics. The division between the two directions within corpus linguistics is explained as being caused by the existence of different research traditions before the field of corpus linguistics evolved. Within *plain (parallel) corpus linguistics,* corpora are used as sources of empirical data in the investigation of linguistic phenomena, whereas in *computational (parallel) corpus linguistics*, corpora function as a test-bed for algorithms developed within the theory of linguistic computation. Borin argues that both traditions would benefit from contact with the other, and stresses how useful the tools of the computational camp can be for researchers in the other tradition. He is, however, not specific about what aspects of plain corpus linguistics would be particularly valuable for the computational direction. Altogether, Borin's opening article is a

valuable, enlightening introduction to research on parallel corpora, and, with its numerous references to related works, a good starting point for further study.

The rest of the book is organised into four sections of articles. In Part I, "Parallel and comparable corpus projects", Stig Johansson reports on the Oslo Multilingual Corpus (OMC), where the English-Norwegian Parallel Corpus (ENPC) is extended by the addition of parallel texts in more European languages. The ENPC and OMC illustrate how the two traditions can meet in a fruitful way. These corpora are excellent sources of empirical data for linguistic research, and their value is enhanced by the application methods developed within the computational camp. Johansson's paper falls within the linguistic tradition: the OMC is used to investigate translations of the English verb *spend* into German and Norwegian. The results reveal "how time may be construed differently in different languages" (p. 57).

Anna Sågvall Hein presents aims and achievements of the PLUG project, which belongs to the computational tradition. 'PLUG' is an acronym for 'Parallel corpora in Linköping, Uppsala, Göteborg'. Three university research groups have worked together under the leadership of Sågvall Hein, and there are substantial achievements to report on. Key words are quadrilingual corpus building, search tools, sentence and word alignment, extraction of contrastive lexical data, and evaluation of the latter. One important aim is to improve existing machine translation systems by providing them with bilingual data. But the achievements of the PLUG project may also enhance translation tools used in human and computer-assisted translation.

Margareta Westergren Axelsson and Ylva Berglund discuss a project in the field of foreign language acquisition studies. The paper describes how a corpus of English essays produced by Swedish learners is compiled. In the field of teaching, a returning question is whether students are performing worse or better than they used to, and in the course of time a resource like Axelsson and Berglund's learner corpus of English must be invaluable to the settling of such debates.

Part II of the book is entitled "Linguistic applications of parallel corpora". Here Raphael Salkie poses the question "How can linguists profit from parallel corpora?". Through a discussion of inventiveness in translation, he positions himself within the tradition of plain corpus linguistics, but also approaches the field of machine translation. Salkie proposes a contrastive database, which would be a corpus-derived, multilingual archive of translational correspondences. The conclusion concerning the fruitfulness of bringing together linguistics and translation theory in parallel corpus research is unproblematic, but the

premise that it is very difficult "to find insights from linguistics which can usefully be applied to translation" (p. 93) is surprising.

Trond Trosterud brings the topics of minority language research and language planning into the context of parallel corpus linguistics. As a linguist, he exploits methods of the computational tradition. His project is to transfer the methods and results of parallel corpus research on majority languages to minority languages, in order to support the investigation, preservation, and development of the latter. Trosterud's contribution is innovative and eye-opening, but its importance is slightly weakened as some of his generalisations, although plausible, are not supported by references to empirical facts or related works.

Christer Geisler's article "Reversing a Swedish-English dictionary for the Internet" deals with computational lexicography, thus representing a merge of the linguistic and computational traditions. In dictionary reversal, an existing bilingual dictionary is transformed into a new dictionary by reversing source and target language. Geisler discusses advantages and drawbacks of reversing dictionaries, especially the problem of maintaining translational equivalence.

Part III "Computational tools for parallel corpus linguistics" opens with Gregory Grefenstette's article on multilingual corpus-based extraction, a fine example of how a topic of computational linguistics can be made accessible to the general linguistics community. Grefenstette explains the notion of text abstraction and shows how it lies at the bottom of computational tools for linguistic analysis. His illustration deals with automatic extraction of translation equivalents from parallel texts, and he presents the possibility of creating automatically a Very Large Lexicon from multilingual texts available on the World Wide Web.

Magnus Merkel, Mikael Andersson, and Lars Ahrenberg present one of the projects participating in the PLUG cooperation. The PLUG Link Annotator, together with the Link Scorer, are computational tools developed at Linköping University for the evaluation of word alignment systems. It seems clear, as Merkel et al. indicate, that the Link Annotator could also be of great value to researchers within contrastive linguistics and translation studies, as a tool for registering translational correspondences. However, their article is not readily accessible to readers unfamiliar with the alignment field.

Peter Stahl discusses technical issues involved in building and processing parallel corpora, dealing especially with the software tool Tuebingen System of Text Processing Programs (TUSTEP). It is a powerful and flexible tool for manipulating text, designed primarily for humanists. Ironically, using the system appears to be a technically complex task, and the high level of detail in parts of Stahl's article does not invite the uninitiated reader.

Jörg Tiedemann presents the Uplug system, another part of the Swedish PLUG project. His article addresses readers interested in computer science and system architecture, and with a substantial background from the field of information technology. Uplug is a platform offering a computational environment where different text processing tools are integrated, such as The Uppsala Word Alignment system. Attractive features of the platform are modularity and the use of different text storage formats, implying flexibility and user-friendliness.

Part IV "Issues in parallel corpus annotation" contains two contributions. Klas Prütz's description of a part-of-speech (POS) tagger for Swedish falls within the computational tradition. A tagger is a program designed to annotate the words in a running text with labels, or tags, indicating word class and grammatical features. In the study two different versions of such tag sets were used, one more limited than the other. It would have been interesting to learn more about the motivation behind applying two tag sets.

Among the papers in *Parallel corpora, parallel worlds* it is perhaps the final contribution in part IV, Lars Borin's "Alignment and tagging", which displays the strongest wish to build a bridge between the two camps of general and computational linguistics. The paper reports on a testing of the hypothesis that "[i]t should be possible to use POS tagging for one language in combination with a word alignment system, in order to obtain a (partial) POS tagging for another language" (p. 207). Borin provides an interesting discussion of the experiment in relation to issues of language typology, and concludes that the method is fruitful in the case of closely related languages, where translationally equivalent words of different languages tend to be of the same category.

As we have seen by now, *Parallel corpora, parallel worlds* is a meeting place for the two traditions of corpus linguistics, but the degree of contact varies between the different contributions. Several of the articles by representatives of the computational direction would have been more interesting to the linguistics camp if greater weight had been put on qualitative evaluations of quantitative data and on discussing linguistic implications of choices made in the design of computational models. On the other hand, corpus linguistics would surely also benefit from seeing more linguists do like Trond Trosterud and venture to adopt, perhaps even adapt, methods developed within the computational tradition.