REVIEW OF
MULTILINGUAL CORPORA IN TEACHING AND RESEARCH

Multilingual Corpora in Teaching and Research (From the series Language and Computers: Studies in Practical Linguistics, No 22) Simon P. Botley, Anthony M. McEnery, and Andrew Wilson, Eds. 2000 ISBN: 90-420-0541-6 US $19.00 (Paperback) 208 + vi Editions Rodopi B.V. Amsterdam (Netherlands) and Atlanta, GA (USA)	

Reviewed by John M. Lawler, University of Michigan.

Multilingual corpora are those consisting of texts in more than one language, often a monolingual original and a translation. These translations vary greatly in their faithfulness, accuracy, style, and order of presentation, as well as in granularity of translation, that is, the size of the chunks being translated (e.g., word-to-word, sentence-to-sentence, paragraph-to-paragraph, or idea-to-idea). Since the reasons for constructing multilingual corpora include being able to correlate individual pieces of one text with corresponding parts of another, their use immediately raises the problem of *text alignment*, or computing which chunk of a text in one language corresponds to a given chunk of the parallel text in another language.

This is the major focus of *Multilingual Corpora in Teaching and Research.* Indeed, this book could more accurately have been titled *Text Alignment in Multilingual Corpora: Overview and Case Studies.* Text alignment, it quickly becomes clear, is the outstanding problem in research on multilingual corpora, and thus -- to the extent that progress has been made in its solution -- its outstanding success story. The problems that arise in alignment research reprise practically every issue in Natural Language Processing (NLP) and Automatic Translation, (e.g., sentence division, anaphor tracking, ambiguity resolution), and the peculiar limitations of the alignment task make the application of alignment strategies to these broader problems surprisingly productive, as is discussed in detail in this volume.

Multilingual Corpora consists of two introductory chapters, covering theoretical and methodological issues, the literature, and the state of the art (up to early 1998), as well as 10 individual case studies, each describing an existing corpus project, 2 in the US and the rest in Europe. All the case studies except the last (on problems aligning English and Chinese texts) deal strictly with Indo-European languages (Danish, English, French, German, Greek, Italian, Norwegian, Spanish, and Swedish) and most of the corpora discussed contain texts in just two languages.

Chapter 1, "Bilingual Text Alignment -- An Overview," by Michael Oakes and Tony McEnery (one of the editors) of Lancaster University, is typical of recent work in CL/NLP in that it distinguishes sharply between statistical and linguistic methods of text alignment. As these authors put it (p. 4) "Statistical methods tend to work better for large corpora, since they are relatively rapid, while linguistic methods can be better for small corpora." The vast majority of the article is a survey of the statistical methods used in various alignment projects, including formulae and discussion of results, although three varieties of

linguistic techniques are also covered. This disparity reflects the simple fact that statistically-based NLP has been far more successful overall than linguistically-based approaches, especially in tasks involving corpora (see Bayer, Aberdeen, Burger, Hirschman, Palmer, and Vilain [1998] and Hoard [1998] for discussion.).

Chapter 2, "Bilingual Text Alignment: Where Do We Draw the Line?" by Michel Simard, George Foster, Marie-Loise Hannan, Elliott Macklovitch, and Pierre Plamondon of Canada's Centre d'Innovation en Technologies de l'Information, takes up the question of granularity in the context of Isabelle's (1993) concept of *Translation Analysis (TA),* that is, "the reconstruction of the correspondences between segments of a source text and segments of its translation" (p. 39), a principled approach to alignment. Before concluding on a generally sanguine note, they discuss three alignment programs at different granularity levels: *JACAL* (Just Another Cognate ALignment program), a character-level program; *Salign,* a sentence-level program that can be used in conjunction with JACAL (though it need not be); and *TMAlign,* a lexical-level alignment program.

Chapter 3, "Corpus and Terminology: Software for the Translation Program at Göteborgs Universitet, or Getting Students to Do the Work," by Pernilla Daniellson and Daniel Ridings, deals with a suite of programs developed for training translators. This is one of the most obvious educational uses of multilingual corpora; the software described here is designed to be used by future translators to pick out "terminology" (i.e., technical terms that may be unfamiliar outside a particular specialty) in context, and create their own personal terminology bank for future use, in the process learning a great deal about translation. It is built from more or less off-the-shelf software (i.e., *Microsoft Access)* and is seen to be robust, simple, and easy to use, as well as meeting the needs of students.

Chapter 4, "Parallel and Comparable Bilingual Corpora in Language Teaching and Learning," by Carol Peters, Eugenio Picchi, and Lisa Biagini of Istituto di Linguistica Computazionale in Pisa, discusses the interesting distinction between *parallel corpora,* or "translationally equivalent texts," and *comparable corpora,* for which they adopt Laffling's (1992) description: "texts which, though composed independently in their respective language communities, have the same communicative function." *PiSystem DBT,* an Italian/English bilingual text query program implemented for language learners, is used to highlight these issues in this chapter. A demo version is available on the Web at http://www.ilc.pi.cnr.it/pisystem/demo/demo_dbt/demo_bilingui/index.htm (this is a different URL from the one given in the book, which now returns an error message). As expected, analyses of comparable corpora are more difficult and pose unique problems. Thus, the implementation discussed is still experimental.

In chapter 5, "Using Authentic Corpora and Language Tools for Adult-Centred Learning," Renée Meyer, Mary Ellen Okurowski, and Thérèse Hand of New Mexico State University explore an application, *OLEADA* (not an acronym, but rather the Spanish word for "tidal wave"), developed at NMSU. *OLEADA* is a complete learning environment, integrating "three language technologies: on-line text corpora, information retrieval, and language analysis tools. A single user interface allows seamless access to the texts and tools in ten languages" (p. 87). This short chapter doesn't go into design or performance specifics, but rather concentrates on the varying uses of *OLEADA's* three customer groups: language training developers, classroom developers, and independent students.

Chapter 6, "Teaching Terminology Using Electronic Resources," by Jennifer Pearson of Dublin City University, is concerned, like Chapter 3, with an application designed to help future translators experience and learn to handle real use of technical jargon and phrases of art in a realistic context. This is an extremely interesting chapter, with many examples of terminological variation, and especially of culture-specific terms for which there are usually no good equivalents.

Chapter 7, "Parallel Texts in Language Teaching," by Michael Barlow of Rice University, shows how even a simple concordance program (*ParaConc*, a simple parallel version of Barlow's *MonoConc*, reviewed this issue and by Lawler, 2000) can be of great use to teachers and students for exploring the wide variety of ways in which a single word or phrase gets translated, especially as part of an idiomatic or metaphoric expression. The result, as anyone who's spent enough time with a good bilingual dictionary can attest, can be eye-opening.

David Woolls of Birmingham University, extends this concept in a different direction in Chapter 8, "From Purity to Pragmatism; User-Driven Development of a Multilingual Parallel Concordancer." The software involved, part of the European Union's LINGUA project, produces various types of concordances over parallel texts in Danish, English, French, German, Greek, and Italian. Rather than focusing on its usage and applications, the chapter is a developmental history of the program, from initial specifications through iterative cycles of construction, testing, and revision of the corpus and the various software tools associated with it, and the inevitable problems that arose at each stage, and how they were handled -- generally by downsizing expectations. This is an article that can be read with sympathy and profit by anyone involved in large-scale distributed development schemes.

Chapter 9, "The English-Norwegian Parallel Corpus: Current Work and New Directions," by Stig Johansson and Knut Hofland of the University of Oslo, is a progress report on an ongoing project, with sections on its uses and recent multilingual extensions to French and German parallel corpora. Of particular linguistic interest are the extensive discussions, with examples, of the occurrence of the Norwegian modals *skal* (p. 135) and *nok* (p. 137); modals are often problematic, but examples like this can help understand something of their vagaries. The section on multilingual extensions is highlighted by an equally extensive and equally interesting discussion of cleft sentences ("That's what I meant," and its ilk) and other clausal anaphora, and their translated equivalents; any syntactician reading this section would yearn for such a tool. This is a good example of how corpus linguistics can inform theoretical linguistics, as well as language learning.

Chapter 10, "Unlocking the power of the SMEMUC," by Raphael Salkie, of the University of Brighton, coins what the author admits is an "ugly acronym" for Small and MEdium-sized MUltilingual Corpus. He argues that such corpora are "a good way forward for those of us who want to take corpora out of the computer laboratory and into the hands of teachers, students, and language researchers," (p. 148) and goes on to describe the step-by-step development and subsequent pedagogic uses of INTERSECT, a French-English parallel corpus massaged to fit the needs of ParaConc (discussed in Chapter 7). His conclusion is one that is easy to agree with: "Sometime in the future, when today's computers seem like little toys and the Internet is fast and freely available, large multilingual corpora will be available for everyone. For now, it is corpora like INTERSECT which can take a lead in convincing linguists, language teachers and translators that multilingual corpora have a lot to offer them" (p. 156).

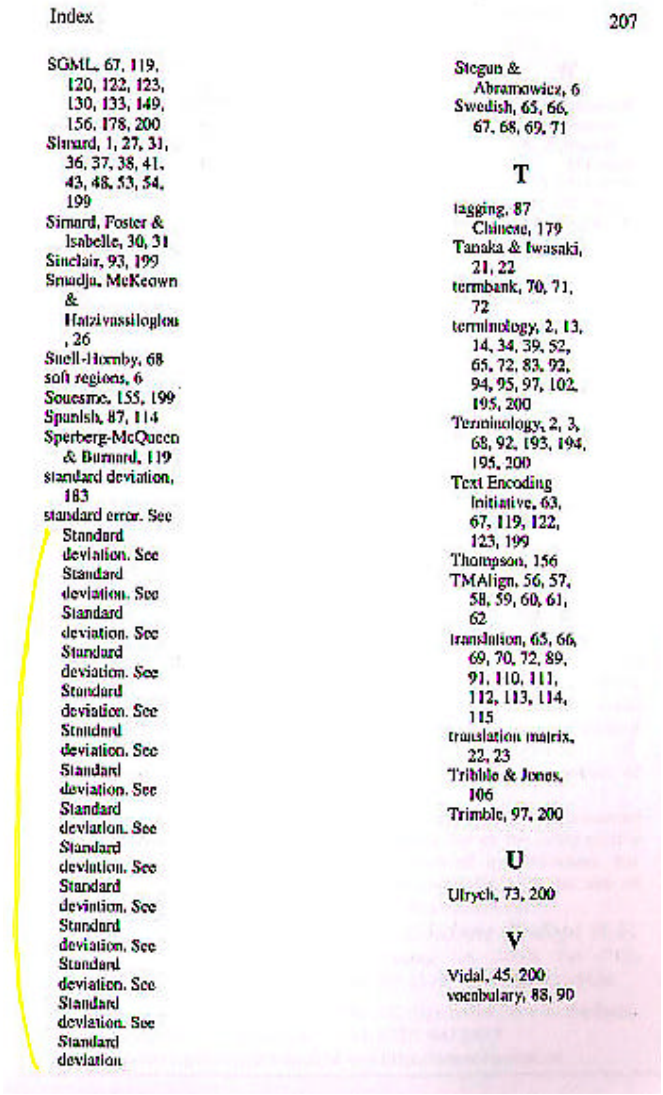
Chapter 11, "Corpus-Based Contrastive Lexicography: The Case of English *with* and its German Translation Equivalents," by Josef Schmied and Barbara Fink of the University of Chemnitz, focuses on the use of a bilingual parallel corpus to research the syntax and semantics of the preposition *with*, in all its uses and collocations. The lexicographic results are the stars here, while the software plays a supporting role; this is a good example of the kind of research that would have been impossible even to conceive of, let alone carry out, before the advent of aligned multilingual corpora. It will be of interest not only to computational linguists, but also to translators, semanticists, lexicographers, and language teachers.

Finally, Chapter 12, "Parallel Alignment in English and Chinese," by Tony McEnery, Scott Piao, and Xu Xin of the University of Lancaster, addresses the challenges for multilingual parallel corpus research posed by non-European and non-Indo-European languages. Many new methods are still needed, and so far the work is largely experimental and the results rather sketchy. Nevertheless, the authors produce a useful discussion of the problems they encountered and report on one alignment method, based on bi-

variate distribution, that they tried out on a sample corpus. They conclude, "Aligning languages which are not genetically related is a challenge for computational linguists, and may well stretch the 'language independence' claim of some current alignment algorithms to the breaking point." The chapter includes an appendix containing a short set of tags that were used in the alignment task.

Overall, this is a really interesting book for a linguist to read. All the articles are well-written and accessible at any level of knowledge about corpora (although readers of chapters 1 and 12 might benefit from familiarity with Oakes, 1998), and the problems encountered are diverse and challenging enough to engage anyone with an interest in language. This would serve nicely as a source of additional readings for courses in corpus linguistics, translation theory, or software design, as well as being a good source of good ideas and potential pitfalls for corpus and software designers themselves.

For such a useful book, though, it is a shame that the index is so sparse, consisting of only seven pages, each of which is mostly white space, with one 12-character-wide column on either side. The index could have been printed in three pages with more appropriate use of space, especially when one considers that the entry for "standard error," a cross-reference to the immediately preceding entry for "standard deviation" on page 207, takes up an entire quarter-page (see Figure 1).



Index 207

SGML, 67, 119, 120, 122, 123, 130, 133, 149, 156, 178, 200
Simard, 1, 27, 31, 36, 37, 38, 41, 43, 48, 53, 54, 199
Simard, Foster & Isabelle, 30, 31
Sinclair, 93, 199
Smadja, McKeown & Hatzivassiloglou, 26
Snell-Hornby, 68
soft regions, 6
Souesme, 155, 199
Spanish, 87, 114
Sperberg-McQueen & Burnard, 119
standard deviation, 183
standard error. See Standard deviation. See Standard deviation. See Standard deviation. See Standard deviation. See Standard deviation. See Standard deviation. See Standard deviation. See Standard deviation. See Standard deviation. See Standard deviation. See Standard deviation. See Standard deviation. See Standard deviation. See Standard deviation. See Standard deviation. See Standard deviation. See Standard deviation. See Standard deviation
Stegun & Abramowicz, 6
Swedish, 65, 66, 67, 68, 69, 71

T

tagging, 87
Chinese, 179
Tanaka & Iwasaki, 21, 22
termbank, 70, 71, 72
terminology, 2, 13, 14, 34, 39, 52, 65, 72, 83, 92, 94, 95, 97, 102, 195, 200
Terminology, 2, 3, 68, 92, 193, 194, 195, 200
Text Encoding Initiative, 63, 67, 119, 122, 123, 199
Thompson, 156
TMAlign, 56, 57, 58, 59, 60, 61, 62
translation, 65, 66, 69, 70, 72, 89, 91, 110, 111, 112, 113, 114, 115
translation matrix, 22, 23
Tribble & Jones, 106
Trimble, 97, 200

U

Ulrych, 73, 200

V

Vidal, 45, 200
vocabulary, 88, 90

Figure 1. Index entries

Indexes are hard to make, and good quality control is often outside the reach even of editors, but a well-made index repays an editor's labor in the form of usefulness for readers. There are a few other infelicities; in addition to the ones remarked on in Dash (2001), such as the absence of Section 3.1.1 mentioned on page 179, I might add the running head for chapter 7, which renames the chapter to "Parallel texts in English teaching."

But all these are very minor matters; this is a **really** good book, worth its price and bound to be useful for a long time to come.

ABOUT THE REVIEWER

John Lawler, Associate Professor of Linguistics at the University of Michigan, Ann Arbor, former chair of the LSA Computer Committee, and software author (MONOSYL, A World of Words, The Chomskybot), has published on topics including metaphor, Acehnese syntax, generic reference, second-language learning, English syntax and semantics, negation and logic, sound symbolism, UNIX, and popular English usage, and has consulted on software development for industry and academia.

E-mail: jlawler@umich.edu

REFERENCES

Bayer, S., Aberdeen, J., Burger, J., Hirschman, L., Palmer, D., & Vilain, M. (1998). Theoretical and computational linguistics: Toward a mutual understanding. In J. Lawler & H. Dry (Eds.), *Using Computers in Linguistics* (pp. 231-255). New York: Routledge. A chapter overview is availableon the Web: http://www.routledge.com/linguistics/introduction.html#chapter.8.

Dash, N. S. (2001). Review of Botley, McEnery, & Wilson (2000), *Multilingual Corpora in Teaching and Research*. *LINGUIST, 11*(2537). Retreived June 1, 2001 from the World Wide Web: http://linguistlist.org/issues/11/11-2537.html.

Hoard, J. E. (1998). Language Understanding and the Emerging Alignment of Linguistics and Natural Language Processing. In J. Lawler & H. Dry (Eds.), *Using Computers in Linguistics* (pp. 197-230). New York: Routledge. A chapter overview is available on the Web: http://www.routledge.com/linguistics/introduction.html#chapter.7.

Laffling, J. (1992). On Constructing a Transfer Dictionary for Man and Machine. *Target 4*(1), 17-31.

Lawler, J. M. (2000). Review of *MonoConc Pro 2.0* Concordancing Software. *LINGUIST*, 11(1411). Retrieved June 1, 2001 from the World Wide Web: http://linguistlist.org/issues/11/11-1411.html.

Oakes, M. (1998). *Statistics for corpus linguistics.* Edinburgh: Edinburgh University Press.

