DE GRUYTER OPEN

Review in ICAME Journal, Volume 38, 2014, DOI: 10.2478/icame-2014-0014

Lidun Hareide, Christer Johansson and **Michael Oakes** (eds.). *The many facets of corpus linguistics in Bergen – in honour of Knut Hofland* (Bergen Language and Linguistics Studies 3). 2013. ISBN: 978-82-998587-3-1. Reviewed by **Sebastian Hoffmann**, University of Trier.

As most readers of the *ICAME Journal* will be well aware, the work conducted in Bergen has been instrumental in developing and expanding corpus linguistics as a research methodology from its very early days onwards. Much of this success story is in some way or other related to Knut Hofland and his work at what is today known as the *Computational Language Unit* (CLU) – a research and development centre for language technologies that has its roots in the early 1970s when the *Norwegian Computing Centre for the Humanities* was founded. The selection of papers under review celebrates Knut's pivotal role in corpus linguistics over the past four decades by presenting an overview of the breadth of research that is conducted in Bergen (and some of its associated institutions) in the 21st century. All papers are freely available via an open-access online publication at https://bells.uib.no/index.php/bells/issue/view/56/showToc, although a print-on-demand version can be obtained by readers who prefer to have a physical copy of the book.

The publication consists of 18 chapters, including a brief introduction by the editors and, as the last paper of the book – which will not be further discussed here – a speech held by Øystein Reigem (in Norwegian) on the occasion of Knut's 60th birthday. Although this is not reflected in the table of contents, the collection of papers is – according to the editors' comments in the introduction – divided into four main sections, covering the general topic areas of 'parallel corpora' (5 papers), 'domain specific corpora' (5 papers), 'language development' (2 papers) and 'statistical analysis of corpora' (3 papers). These sections are preceded by **Geoffrey Leech**'s outline of the development of ICAME and the Brown Family of Corpora and the crucial role played by Knut Hofland at Bergen – and Stig Johansson at Oslo – in the development of corpus linguistics from the "derided fringe" (p. 3) of linguistics towards its current status as a mainstream methodology. Few people outside the group of core ICAME confer-

ence participants are likely to know about these early stages of corpus linguistics, and Leech's first-hand account provides interesting reading. One important focus is on the creation of the Lancaster-Oslo/Bergen corpus (LOB), the British counterpart to the first electronic corpus, the Brown corpus, and how this joint British-Norwegian project led to the establishment of ICAME in 1977. The paper also briefly touches on the historical importance of ICAME as a distributor of corpora (nowadays a minor role), its yearly conferences and the CORPORA discussion list.

The section on parallel corpora is opened with a contribution by an unlikely first author, **Knut Hofland**, and his co-authors **Paul Meurer** and **Andrew Salway**. According to the editors, Knut was 'tricked' into writing a contribution for his own festschrift. Again, this is an overview paper, but in this case the focus is more specifically on the research and corpus compilation projects conducted in Bergen over the last 40 years, covering aspects such as the Ibsen Corpus/Concordance, LOB, the Bergen Corpus of London Teenage Language (COLT), the English-Norwegian Parallel Corpus (ENPC) and the Translation Corpus Aligner (TCA) tool that was developed to facilitate the alignment of electronic parallel corpora. The authors also describe more recent corpus projects that make use of web-derived data – e.g. for the Norwegian Newspaper Corpus, a continuously updated monitor corpus of now more than one billion words, with contents dating back to 1998, or a Norwegian Twitter corpus. Further sections of the paper focus on tools such as *Corpuscle* (Section 3), a fast web-based search interface whose functionality goes beyond that of other corpus tools in allowing researchers to query parallel corpora (e.g. for words in one corpus that are not translated by a particular word or expression in another language), on corpus annotation in Treebanks (Section 4) and on the automated information extraction from corpora and the visualisation of such data (Section 5). Given the explicit overview nature of this contribution, covering a wide range of research resources and tools, its inclusion in the 'parallel corpora' section of the book (as indicated by the editors in the introduction) is perhaps a little odd. In my view, it would have been more appropriate to group it together with Geoffrey Leech's historical overview contribution.

The remaining four papers of the section do, however, exclusively deal with parallel corpora. First, **Signe Oksefjell Ebeling** and **Jarle Ebeling** use the English Norwegian Parallel Corpus (ENPC) to investigate mismatches in sentence alignment between Norwegian originals and their translations into seven target languages, i.e. cases where a Norwegian sentence is either split up into two or more target language sentences or where it is merged with another sentence into a single target language sentence. A number of factors such as indi-

vidual author/translator styles, target language constraints/preferences and country/language-specific guidelines for translation are considered, but findings unfortunately remain fairly inconclusive. This is followed by a contribution by **Rosa Rabadán** and **Marlén Izquierdo**, who study the English approximate negators *scarcely, rarely, barely, hardly* and *seldom* and their translations into Spanish. One of their findings is that Spanish translations cover a large range of realisations and tend to over-specify the approximation expressed in the English original. In the next paper, **Carla Parra Escartín** provides a work-in-progress report that recounts the steps involved compiling a parallel corpus of Technical Regulations Information System (TRIS corpus). The author's focus is on the discussion of various standards for corpus annotation – e.g. by comparing TEI and (X)CES – and her motivations for choosing a particular standard for the markup of her own corpus over other options, thus potentially providing helpful guidance for comparable future projects. The final paper of the section by **Pedro Patiño** introduces the Corpus of English and Spanish Free Trade Agreements (FTA corpus), a parallel corpus of 233 files amounting to approx. 1.4 million words each, which is complemented by a much smaller section of Norwegian–Spanish/English translations of trade agreements. His research aims at the detection of specialised collocations in this particular text type, but no conclusive results are as yet available.

The section on 'domain specific corpora' is opened by **Ingrid Simonnæs** and **Sunniva Whittaker** in their paper on the Bergen translation corpus TK-NHH, which is in fact also a parallel corpus in that it contains translations of domain-specific texts into various languages produced by candidates taking the National Translator Accreditation Exam. The authors discuss possible uses of the corpus in teaching and research, supporting their views with a sample analysis involving the translation of culture-bound legal concepts for which no straightforward translation equivalents exist in the target languages (e.g. *medmor,* literally 'co-mother', a concept that was introduced in Norway when same-sex and hetero-sex marriages were given the same status). In his paper on recent developments in Norwegian corpus lexicography, **Gisle Andersen** provides a description of the Norwegian Newspaper Corpus – a web-based monitor corpus of more than one billion words dating back to 1998 – and shows how even methodologically fairly simple search and retrieval strategies lead to results that can significantly support the work of Norwegian dictionary-makers. Lexicography is also at the core of the paper by **Marita Kristiansen** who, however, uses a much smaller dataset – a corpus of researchers' blogs relating to economic-administrative domains specifically compiled by Knut Hofland for this project – to retrieve specialised neologisms that might not be detected by researchers in

other types of data sources and that would therefore escape the attention of lexicographers compiling specialised dictionaries. **Kjersti Fløttum**, **Trine Dahl**, **Anders Alvsåker Didriksen** and **Anje Müller Gjesdal** report on previously conducted research on KIAP, a corpus of academic publications in three languages (Norwegian, English and French) across three disciplines. In particular, they focus on self and other representations (e.g. via personal/indefinite pronouns, adversative conjunctions and metatextual and metadiscursive expressions) as indications of culture/country-specific style. Their most important finding is that discipline trumps language in that for example Norwegian and French medical papers exhibit more similarities than Norwegian papers on medical and linguistic topics. The final paper of the section is by **Annette Myre Jørgensen**, who relates her findings about the language used by Spanish teenagers that she has retrieved from the Madrid subcorpus of the Corpus Oral de Lenguaje Adolescente (COLAm). Her focus is on a variety of features – such as discourse markers, taboo words and hyperbolic intensification – that have been isolated as indicative of adolescent speech in previous research.

The third section of this festschrift is devoted to the topic of 'language development'. The first of the two papers is by **Martha Thunes** who presents her work on the encoding of inalienability in English and Norwegian, which formed part of her PhD thesis (submitted in 2011). English requires overt inalienability marking via a possessive determiner while Norwegian realises the same semantic content via the marking of definiteness. Thunes observes that this translational correspondence results in linguistically predictable translations between the two languages, and can therefore potentially be handled via machine-translation. The theme of 'language development' is taken up in **Benedikte Vardøy** and **Margje Post**'s paper by looking at recent English loanwords ending in *-ing* in Russian. They make use of the Russian National Corpus and complement this data source with Integrum, which is a much larger commercial database containing text of more than 1,200 central and local Russian newspapers dating back to 1996. Their study reveals, perhaps not surprisingly, that loanwords in *-ing* have become more frequent in Russian since perestroika. On the basis of a closer analysis of five frequent items, *rejting* ('rating'), *kasting* ('casting'), *trening* ('training'), *bodibilding* ('body-building') and *lifting* ('lifting'), the authors also provide a more fine-grained analysis of the semantic properties – and changes – exhibited by this type of loanwords in Russian.

The final three research papers of the collection relate to the 'statistical analysis of corpora'. The section opens with a study by **Gard B. Jenset** and **Lidun Hareide** on a topic that was also at the heart of the contribution by Ebeling and Ebeling, viz. which factors can be seen to influence the use of different

sentence alignment patterns in parallel corpora. Using hierarchical clustering techniques on a set of aligned Norwegian-Spanish parallel data, the authors convincingly show that both the translator and the genre can be seen as having an effect on sentence alignment, but that neither factor is sufficient to explain the variation observed in the data on its own. The second paper of the section is by **Christer Johansson**, who discusses both statistical significance and effect size and shows how their combined application can help linguists in evaluating and interpreting quantitative differences observed in corpora. Finally, **Michael Oakes** and **Alois Pichler** demonstrate – based on comparisons between various texts by Wittgenstein and two of his amanuenses – how computational stylometry can be employed to answer questions of authorship. Their findings suggest that Wittgenstein's 'Diktat fur Schlick' – a text made available by the Wittgenstein Archives at the University of Bergen and whose authorship is debated by scholars – is indeed much closer to other texts by Wittgenstein than to texts by the two other possible authors.

Evaluation

As indicated above, the purpose of this volume of papers is to celebrate Knut Hofland and his great contributions to corpus linguistics both in Bergen and beyond. To do so by producing a showcase of the range of current research carried out in Knut's immediate surroundings certainly strikes me as a great idea, and the variety of topics covered by the papers in this collection certainly seems to suggest that Knut's light shines brightly and his influence still continues in Bergen. Indeed, there are a number of papers that fully live up to the expectations one would have of such an endeavour. Just to name one example of a successful research-based paper, the contribution by Jenset and Hareide is not only expertly written but also provides very interesting methodological insights into the use and interpretation of one particular set of statistical techniques. Furthermore – and not unexpectedly – the overview papers by Leech and Hofland *et al*. also competently deliver relevant information to readers wishing to learn more about the history of corpus linguistics in general and research at Bergen in particular.

Unfortunately, however, this publication contains some of the worst editing I have ever seen, and the editors' oversights affect the quality of this festschrift on various levels. First of all, it is quite obvious that very little peer reviewing can have taken place in the preparation of this publication. A number of papers should simply not have been published in their current state, and it is hard to understand that the editors have not caught some of the more obvious problems themselves.

Also, many papers of this collection would have greatly profited from a language check by a native speaker of English. Moreover, in one case, it is obvious that the authors have in fact partially misunderstood the meaning of some of the English examples they have worked with. Finally, there is a plethora of formal errors and inconsistencies that further contribute to the overall impression that the editors have not done their work properly. For example, Rabadán and Izquierdo's conclusion refers to sections of the paper that do not – or no longer? – exist (e.g. "see 4.2.2.1.c", p. 58), and there are many obvious typos (e.g. "Together these three projects embrace the breath of research at NHH", p. 127; *translitteration*, p. 180*; ialienability*, p. 185). Figures and tables are separated from their headings by page breaks or are as a whole misplaced in the paper (e.g. figure 3 on pp. 96–97, which should be found after the last sentence on p. 97), and the alignment of examples is broken (e.g. examples (4)–(6) on p. 129). The list of such formal problems could be vastly expanded. It is difficult to imagine that the collection of papers was proofread before it was published.

I have no doubt that Knut Hofland felt honoured and happy when he was presented with this festschrift, and the great impact he has had on the field certainly deserves this kind of attention. But it is unfair to Knut to offer him a product that is so obviously flawed.

One possible explanation for the obvious lack of quality control could have been that this is an online publication. Perhaps this – still relatively new – mode of publication encouraged less rigorous editing and lower overall quality thresholds. The ease with which online data can be amended or deleted has no doubt led to a more relaxed attitude to exactitude and consistency in the case of websites or blogs. However, this should not extend to the text-type discussed here. A work like the one reviewed here will clearly not improve the reputation of online publications.