Nor Jnl Ling 37.2, 135–139 © Nordic Association of Linguists 2014 doi:10.1017/S0332586514000158

Andersen, Gisle & Daniel Hardt. 2014. Introduction: Corpus linguistics and the Nordic languages. *Nordic Journal of Linguistics* 37(2), 135–139.

Introduction: Corpus linguistics and the Nordic languages

Gisle Andersen & Daniel Hardt

Gisle Andersen, NHH Norwegian School of Economics, Helleveien 30, NO-5045 Bergen, Norway. gisle.andersen@nhh.no

Daniel Hardt, Department of IT Management, Copenhagen Business School, Howitzvej 60, DK-2000 Frederiksberg, Denmark. dh.itm@cbs.dk

Recent decades of research in linguistics have seen a shift towards empirical methods and an increased use of data from corpora as a basis for making claims about language (Sampson 2005). This trend has made its mark on research on the Nordic languages also, and the current special issue aims to show some of the breadth of research in this field. The issue is in its entirety devoted to contributions that use the methodology of corpus linguistics on Nordic language data. This includes research that investigates both historical and contemporary aspects of the languages of the Nordic region.

Since the advent of corpus linguistics in the early 1960s, the English language has had a privileged position, and it was not until much later that corpora were developed for the Nordic languages. However, scholars from the Nordic region were engaged in corpus linguistics from a very early stage. This is seen, for instance, from their involvement in technological, methodological and theoretical developments in the field of corpus linguistics through participation in ground-breaking projects such as the Lancaster–Oslo–Bergen corpus (Johansson, Leech & Goodluck 1978), the London–Lund Corpus (Svartvik & Quirk 1980) and the English–Norwegian parallel corpus (Johansson & Hofland 1994). But it was not until the 1990s that the compilation of corpora representing the Nordic languages really took off, and subsequently a number of language corpora have been made accessible under the auspices of key organisations such as The Text Laboratory and Uni Research in Norway, the Swedish Language Bank (*Språkbanken*) in Sweden, Center for Language Technology (CST), and Society for Danish Language and Literature (DSL) in Denmark, the University of Helsinki in Finland and the University of Iceland.

During the same period, there have also been major advances in the computational tools and statistical techniques available to researchers. The corpus linguistic landscape has been extended to include not only hand-crafted and manually

edited corpora but also large web-based corpora that are especially fit for the study of lexical and other innovation (Renouf 2007, Andersen & Hofland 2012). In variationist studies, statistical methods such as cluster analysis and regression analysis are used to describe more reliably the correlation between the variables of time and frequency with other sociolinguistic variables (Gabrielatos et al. 2012). Association measures are used to account for word co-occurrence features, and measures for keyness and n-gram frequency convey variation in and between corpora. Further, more sophisticated schemes for the annotation of linguistic categories have been developed; these include the annotation of deep-level syntactic structure, such as the system for treebanking currently being developed in the Infrastructure for the Exploration of Syntax and Semantics (INESS) project (Rosén 2012), and an approach to annotating multiple levels of linguistic description in parallel corpora in the Copenhagen Dependency Treebank project (Buch-Kromann & Korzen 2010). To facilitate variationist studies, advanced dialect corpora now enable the study of variation between speakers and speaker groups, with links between transcribed speech and audio/video files and geodata, as in the Nordic Dialect Corpus and Syntax Database, developed in the ScanDiaSyn project (Johannessen et al. 2009). Jointly, the recent development of new corpus technology has contributed to more refined annotation methods and more sophisticated analytical tools for the benefit of users of corpora.

Another significant technological development is that an increasing amount of corpora and other language resources are being made accessible through federated technical infrastructures, most notably the pan-European Common Language Resources and Technology Infrastructure (CLARIN), in which all of the Nordic countries play an active role through their respective national nodes (http://clarin.eu/). This initiative is likely to stimulate an increased effort in Nordic corpus linguistics and makes for fruitful exploitation of language resources and cross-institutional cooperation in future research.

The six papers in this volume display and exemplify many of the recent advances and developments in Nordic corpus linguistics. What they have in common is the use of corpus linguistic procedures for the empirical study of language, with a special focus on Nordic languages.

In an account of historical language resources from the Swedish Language Bank, PETER ANDERSSON looks into the Swedish lexeme *fast* and explores its diachronic development. The data show how this form has undergone grammaticalization from an adjective meaning 'steady', 'robust' into a concessive construction *fast*(*än*). Andersson's functional analysis reveals the rise of a concessive subordinator through the conventionalisation of a concessive inference which occurs in the critical context of a construction called the universal concessive conditional clause (roughly the equivalent of the English 'however much').

The paper by TAM BLAXTER looks at the speech of male and female characters in the Old Norse *Íslendingasögur* (Icelandic sagas), which is a series of narrative prose

texts produced in Iceland in the thirteenth and fourteenth centuries. It uses keyword analysis to compare the speech of the male and female characters in an attempt to shed light on the social construction of gender in the society of the Icelandic Sagas. The analysis points to evidence for systematic power imbalances between male and female, as well as differences in forms of address. While Blaxter emphasizes that conclusions about gender differences based on the analysis must be made with caution, the research provides a fascinating window on gender differences in this society.

The paper by SIGNE OKSEFJELL EBELING is a study within the fledging field of cross-linguistic phraseology. Looking into the English/Norwegian cognate pair *eye/øye*, her study illustrates that a phraseological approach is needed in order to fully account for the meaning of units that contain these nouns. On the basis of data from bilingual translation corpora and monolingual corpora, she performs an analysis within Sinclair's (1991) extended-units-of-meaning model that conveys the wide-ranging semantic potential of *eye/øye* expressions. The cognates are seen to share a common phraseology in some of their most frequent uses; on the other hand, there are also notable differences in the metaphorical extensions of the words in the two languages.

ANTON GRANVIK and SUSANNA TAIMITARHA look into topic-marking expressions in Swedish in a corpus-based analysis of prepositional synonymy. Their contribution is a quantitative study of four complex near-synonymous Swedish forms that express the 'aboutness' of the discourse, namely *angående*, *beträffande*, *rörande* and *gällande*. Originally participial expressions, these are used in contemporary Swedish as topic-marking prepositions, on a par with English *concerning*, *regarding*, etc. The data for their study are drawn from a large Swedish newspaper corpus, in relatively formal written language, and it is subjected to two types of statistical analysis, a collexeme analysis (Stefanowitsch & Gries 2003) and a logistic regression analysis. Observing that the four items behave quite similarly in terms of their collocational features, the authors show that the forms studied are characterised by a high degree of interchangeability.

The paper by LUDOVIC DE CUYPERE, KRISTOF BATEN & GUDRUN RAWOENS offers a corpus-based analysis of the passive alternation in Swedish, that is, the alternation between a morphological realisation of the passive with the clitic *-s* (*s*-passive), and a periphrastic realisation with the auxiliary *bli* 'become' (*bli*-passive). The paper uses corpus linguistics techniques to account for the relative impact of a variety of lexico-grammatical factors on the use of either the morphological or periphrastic passive. Restricting themselves to a selection of the three lexical verbs, *acceptera* 'accept', *behandla* 'treat' and *välja* 'choose', the authors complement previous research by applying a multivariate analysis (logistic regression) that evaluates the simultaneous effect of the factors of subject animacy and number, aspect, modal verb and presence/absence of a 'by'-phrase.

Finally, KELLY SMITH, BEATA MEGYESI, SUMITHRA VELUPILLAI & MARIA KVIST investigate Swedish clinical texts from electronic health records and compare their linguistic characteristics to those of standard Swedish texts and biomedical journal texts. A number of interesting differences emerge: the clinical text contains more technical terms and in general a lower level of lexical variance. Grammatical differences were also found: clinical texts tend more frequently to omit subjects, verbs and function words, are more likely to use passive voice, and in general are shorter and more telegraphic than the other text types. The study is of interest as a contribution to genre and domain analysis. The authors also suggest that their study might provide a basis for development of automatic methods for simplifying or otherwise processing electronic health records.

REFERENCES

Andersen, Gisle (ed.). 2012. *Exploring Newspaper Language: Using the Web to Create and Investigate a Large Corpus of Modern Norwegian*. Amsterdam: John Benjamins.

Andersen, Gisle & Knut Hofland. 2012. Building a large monitor corpus based on newspapers on the web. In Andersen (ed.), 1–28.

Buch-Kromann, Matthias & Iørn Korzen. 2010. The unified annotation of syntax and discourse in the Copenhagen Dependency Treebanks. *Proceedings of the Fourth Linguistic Annotation Workshop* (Association for Computational Linguistics), 127–131.

Gabrielatos, Costas, Tony McEnery, Peter J. Diggle & Paul Baker. 2012. The peaks and troughs of corpus-based contextual analysis. *International Journal of Corpus Linguistics* 17(2), 151–175.

Johannessen, Janne Bondi, Joel Priestley, Kristin Hagen, Tor Anders Åfarli & Øystein Alexander Vangsnes. 2009. The Nordic Dialect Corpus: An advanced research tool. In Kristiina Jokinen & Eckhard Bick (eds.), *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009* (Northern European Association for Language Technology (NEALT) Proceedings Series), Tartu, 73–80.

Johansson, Stig & Knut Hofland. 1994. Towards an English–Norwegian parallel corpus. In Udo Fries, Gunnel Tottie & Peter Schneider (eds.), *Creating and Using English Language Corpora*, 25–37. Amsterdam: Rodopi.

Johansson, Stig, Geoffrey Leech & Helen Goodluck. 1978. Manual of information to sccompany the Lancaster–Oslo/Bergen Corpus of British English, for use with digital computers. Oslo: Department of English, University of Oslo. [http://clu.uni.no/icame/manuals/LOB/INDEX.HTM]

Renouf, Antoinette. 2007. Corpus development 25 years on: From super-corpus to cyber-corpus? In Roberta Facchinetti (ed.), *Corpus Linguistics 25 Years On*, 27–49. Amsterdam & New York: Rodopi.

Rosén, Victoria. 2012. Exploring corpora through syntactic annotation. In Andersen (ed.), 67–78.

Sampson, Geoffrey. 2005. Quantifying the shift towards empirical methods. *International Journal of Corpus Linguistics* 10(1), 15–36.

Sinclair, John. 1991. *Corpus, Concordance, Collocation* (Describing English Language). Oxford: Oxford University Press.
Stefanowitsch, Anatol & Stefan Th. Gries. 2003. Collostructions: Investing the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2), 209–243.
Svartvik, Jan & Randolph Quirk (eds.). 1980. *A Corpus of English Conversation*. Lund: Gleerup.