Integrating heterogeneous digital text resources to analyse complex neologisms in Norwegian

Anje Müller Gjesdal, Gunn Inger Lyse

Recent years have seen an increase in research on dynamic language phenomena within the disciplines of linguistics and discourse analysis, especially related to societal issues, such as climate change (see e.g. Tvinnereim & Fløttum 2015). At the same time, research infrastructures such as META-NET (www.meta-net.eu) and CLARIN (www.clarin.eu) work to document language resources and tools (LRT) and to streamline data formats to facilitate the combination of resources in research. The Norwegian contribution to the European research infrastructure for digital linguistic data, CLARIN, is emerging through the national project CLARINO. Research infrastructures such as CLARINO may enable researchers to combine the use of different resources efficiently, allowing for new perspectives on established as well as emerging research methods and questions, opening up, for instance, for a fruitful collaboration between linguistics and discourse analysis on the one hand, and data sciences on the other. However, an effort is needed to provide researchers with a toolbox of data sets and tools that are user-friendly enough to take full advantage of the potential offered by recent technological developments.

In tandem with academic infrastructure initiatives, we observe the emergence of linguistic data resources that were not originally developed for language research purposes, yet reveal an interesting potential for linguistic research, particularly with respect to hypothesis generation. Relevant examples in a Norwegian context are databases such as Atekst and Holder de ord? [‘Do they keep their word?’]. Holder de ord? is based on the proceedings of parliamentary debates, which has been made publicly available as structured data in XML format, while Atekst is a privately owned digital news archive.

This paper presents a use case where separate digital resources, each developed for different purposes, are used for a linguistic analysis of an emerging environmental neologism, det grønne skiftet. By studying this expression across independent resources we measure both its frequency as well as its spread across corpora and domains, which we hypothesize to be an indication of its neology status (see Gjesdal & Lyse 2016). Observations of the expression det grønne skiftet in non-academic text sources (Atekst and Holder de ord) confirmed our initial hypothesis that the use of this expression is indeed increasing in Norwegian. Unfortunately, the query possibilities in these resources are extremely limited for scientific purposes. For instance, Holder de ord? does allow queries using wildcards (a search such as grøn* skift* comprises both the word form grønt and grønne, and skifte|skiftet), but it is not possible to express that the wildcard should stop at word spaces. Thus, a search for grøn* skift* will return any any match where the string grøn* precedes skift*, even if there are five sentences between the two words. A better-suited corpus is therefore arguably the Norwegian Newspaper Corpus (NNC) (Andersen & Hofland 2012), a monitor corpus of Norwegian news text, developed by Uni Research Computing. The NNC is available via the CLARINO infrastructure through the Corpuscle corpus management and analysis system, which has been developed for annotated corpora (see http://clarino.uib.no/korpuskel/). It is an interesting corpus because it is updated with new material every day, and it covers a range of content domains, spanning from political discussions to sports, from 1998 until today. NNC seems to confirm the observation from Atekst and Holder de ord? that the expression has seen a conspicuous growth in the news media.

In conclusion, the paper shows an increase in the expression, but importantly the material analysed provides insufficient for a conclusion at this stage. We therefore argue that improved access to research corpora, as well as corpus analysis platforms with adapted, easily
accessible and user-friendly tools to analyse such data, would significantly advance the analyses of dynamic language phenomena that require access to heterogeneous types of digital language data. We also argue that infrastructures such as CLARINO have a key role to play in this development.

**References**


