Large-scale corpus-based analysis of regional varieties of Danish

As long as it has been physically possible, the Danish dialect research centers in Aarhus and Copenhagen have been recording dialect speakers throughout the country in order to preserve their language and culture for future generations (Andersen 1981). During a particularly active period of recording between 1971-1976 alone, more than 500 hours of recordings were made covering more than 500 different parishes. In the time since, the research centers have focused mostly on developing dictionaries of dialect-specific vocabulary (*Jysk Ordbog* and *Ømålsordbogen*), for which the recordings have only been used very sparingly. Instead, these have relied on older written sources and questionnaires.

As part of two related research projects on regional phonetic and phonological variation in Denmark, we have been gathering the recordings of Danish dialects and updating the existing metadata into a modern standardized format. Taken together, the corpus is unique in its sheer scope and breadth of coverage, but in spite of that, it has never been used for any kind of systematic research. However, it is crucial that such systematic research be done, as the bulk of Danish dialects are either extinct or moribund (Kristiansen 1998, but see also Monka & Hovmark 2016).

In this poster, we discuss challenges and opportunities of working with a corpus of this size. Aside from the sheer scope of the corpus, challenges are posed by the varying and generally subpar quality of the recordings, as well as the lack of systematic transcriptions. That being said, the scope of the corpus also provides unique opportunities to identify and map patterns of variation at a very high level of detail, e.g. by using modern computational methods in dialectometry (e.g. Prokic & Nerbonne 2008, 2013). Using the data, we will be exploring variation in syllable structure, consonant realization, intonation and tonal phenomena in regional variants of Danish. This poster showcases opportunities provided by the corpus and methodology by presenting preliminary research on microvariation in prosodic features and in the realization of laryngeal contrasts in stop consonants (expanding on Puggaard 2018). This will show how large-scale corpus-based studies using fine-grained acoustic analysis and statistics can be used to illuminate aspects of Danish regional variation which have previously rarely been subject to empirical study.

References

Andersen, T.A. (1981): Dialektbånd og databehandling. *Ord & Sag 1*: 11-18.

Kristiansen, T. (1998): The role of standard ideology in the disappearance of the traditional Danish dialects. *Folia Linguistica 32 1-2*: 115-129.

Monka, M. & H. Hovmark (2016): Sprogbrug blandt unge i Bylderup anno 2015. *Danske Talesprog 16*: 73-112.

Prokić, J. & J. Nerbonne (2008): Recognising groups among dialects. *International Journal of Humanities and Arts Computing 2 1-2*: 153-172.

Prokić, J. & J. Nerbonne (2013): Analyzing dialects biologically. In Fangerau, H. et al. (eds.): *Classification and evolution in biology, linguistics and the history of science. Concepts, methods, visualization*: 147-161. Stuttgart: Franz Steiner.

Puggaard, R. (2018): Realizations of /t/ in Jutlandic dialects of Danish. *Linguistica Lettica 26*: 368-395.