A Multimodal Dialect Corpus

Janne Bondi Johannessen
University of Oslo
Øystein Alexander Vangsnes
UiT The arctic University of Norway
Signe Laake
University of Oslo
Tor Åfarli
Norwegian University of Science and Technology

Using language corpora in linguistic research helps ensure that the empirical basis of a linguistic argument or discussion reflect the speakers of the language. It also means that claims about linguistic practice can be tested and repeated, thus satisfying a need in modern science.

In this paper we will present a spoken language corpus of dialects of the North Germanic languages. The corpus consists of ca. 2 million words (transcribed) from five languages, whose dialects are so close that many of them can be considered variants of the same language.

The technology behind the corpus, the Glossa search system (Johannessen et al. 2008), is partly newly developed and partly relies on existing resources, like Corpus Work Bench. Glossa is available for other corpora and freely downloadable from GitHub.

In the paper we will demonstrate some very nice features of the corpus, such as: aligned transcriptions, available sound and video directly from the result concordances, online illustrations of formants, waveforms and spectograms., and importantly: maps. A lot of effort has been put into a user-friendly interface, making it easy to use for non-technical dialectologists and linguists in general.

The Nordic Dialect Corpus was developed in a project that included linguists from five countries, and who were involved in collecting dialects as well as in transcription. The linguists also have used the corpus afterwards and discovered many new isoglosses. This has resulted in publications and a dialect atlas available online. This will also be presented in the talk.