Towards historical e-linguistics for Swedish

While corpus linguistics has a long tradition of extensive empirical studies, such work is more scarce in historical linguistics. During the last couple of years there has been an increased interest in the potential of language technology tools for other disciplines than modern linguistic research, for instance social, educational and historical studies (Borin et al 2010). Recent work (see, e.g., Pumfrey et al 2012) has shown that large-scale corpus-based historical linguistics is a fruitful approach.

An obvious area of collaboration between historical linguistics and language technology research is language change, such as grammaticalization. A concrete example concerns modal auxiliaries in the Germanic languages. The development of modals has received a lot of interest in the linguistic literature, and there is evidence of a diachronic relation between root and epistemic modal meanings (Bybee et al 1994). However, there is less certainty regarding the details and potential causes of this development. Earlier studies on epistemic auxiliaries (i.e. the meaning of speaker judgment) focus mainly on theoretical explanations and are usually limited to manually based case studies (Andersson 2007). To reach a more sufficient picture on the mechanisms involved and ultimately automatic sense disambiguation, large-scaled corpus studies are needed.

A step towards historical e-linguistics for Swedish are the Språkbanken tools – Korp for querying concordances and Karp for searching in lexica (Borin et al 2012a/2012b) – which provide a link to the past by diachronically connecting corpora and lexical resources.

Here, we exemplify the symbiosis between historical linguistics and language technology tools with the development of the Swedish verb tör/torde (Old Swedish þora). In capturing the rise and distribution of the epistemic meaning – ‘probably’ in relation to the earlier meaning ‘dare’ – we need to identify a large spectrum of historical forms (tor, pura, thor, thör, thort). This could be done through links between lexical units from different time periods, which help us explore variant forms and meanings in the corpora. We also need to identify the syntactic distribution of the verb. A morphosyntactic analysis of the texts would be helpful to find common constructional features such as subject type, complement type and clause type. Creating this automatically not only for Contemporary Swedish, but for all historical forms of the language, requires developing new and adapting existing tools.

We discuss various extensions of existing tools and resources, such as annotation of text, a morphology for Old Swedish, and the linking between lexical resources, as well as how to create these extensions semi-automatically from digitized historical resources. Ultimately, we develop a platform for exploring the diachronic development of the Swedish epistemic modal auxiliary verb, as well as other research questions in an emerging field we would like to call historical e-linguistics. The benefits of this approach are iterative since the corpus instances can improve the lexical knowledge, which will in turn enhance the quality of the corpus analysis, increasing the quality of the search results.
References


