A multilingual speech corpus of North-Germanic languages

Janne BONDI JOHANNESSEN (University of Oslo)
Kristin HAGEN (University of Oslo)
Joel PRIESTLEY (University of Oslo)

In this paper we describe challenges regarding the development and presentation of the Nordic Dialect Corpus (Johannessen et al. 2009, Johannessen 2011). The corpus was initiated by a group of syntacticians from the Scandinavian Dialect Syntax Network (ScanDiaSyn) from six countries in Northern Europe, representing the North Germanic languages – also called the Nordic or Scandinavian languages, i.e. Danish, Faroese, Icelandic, Norwegian, Swedish (in Sweden and Finland).

A problem for a lot of linguistic research is that there is a data problem. In order to have controllable data, linguists like to turn to corpora, but these are most often only based on written texts, and do not reflect how language is used in much broader contexts, and they totally miss out on the kind of language one gets in dialogue and on dialectal variation. The ScanDiaSyn researchers thus needed a corpus representing the speech of their countries. Since the five languages are very close to each other, and often have features that overlap in their dialects, if not in their standard written varieties, it was felt that one corpus covering spontaneous speech of all their languages and as many dialects as possible would be the best option. This presented challenges with regard to data collection, and especially with regard to transcription, grammatical annotation and search options in the corpus. Also, a best possible availability of the corpus was deemed important, including the possibility of a good web-interface and a presentation of both audio and video.

We have not found much literature on speech corpora that are available on the web, that represent several languages or speech varieties, and that have available audio and video. The most famous speech corpus for English is the speech part of the British National Corpus (10 million words). However, it says in their own search interface distribution that its dialect categorisation is unreliable. Further, as a dialect corpus, the BNC has limited value, since it is not represented with audio, and the speech is transcribed only orthographically. The Scottish Corpus of Text and Speech is available on the web, but of its 4 million words, only 20 texts have spoken texts, provided with orthographic transcription, synchronised with the audio or video. It is not grammatically annotated and does not cover the full geographical area. The DynaSand web-based dialect database consists of information on various syntactic features and their distribution geographically in the Netherlands and Belgium. It contains recorded material from the project's questionnaire sessions, but the conversations contain to a large extent read sentences and meta-linguistic discussions, and less spontaneous speech (Barbiers et al. 2006). The Corpus of French Phonology (La phonologie du français contemporain: usages, variétés et structure – PFC) is a web-based corpus of spoken French from across the Francophone world. It is searchable both phonologically and w.r.t. informant characteristics, and has transcriptions linked to sound. Since it is developed for phonological research purposes it is not grammatically annotated, but is otherwise close to the ideas of the Nordic Dialect Corpus.

In the talk we will say something about the methodology of getting spontaneous speech data for all the languages, which involved quite a lot of new recordings, but also in some cases re-use of previous recordings. We will then go on to tlk about the transcription practices and the various decisions that were made to accomodate the characteristics of the languages involved and the financial limitations for each language. All of the languages are represented with orthographic transcriptions, in order to facilitate corpus search and grammatical annotations. Some also have phonetic transcriptions to be able to get a written representation of the speech

involved. All the languages have been grammatically tagged mostly using existing written taggers with individual tagsets that have in some cases been modified for spoken language, and in the talk we will discuss how we have solved the problem of different tagsets for the common corpus. The taggers used are a constraint grammar (CG) tagger for Faroese (Trosterud 2009), a CG tagger for Danish (Bick 2003), a Hunpos tagger for Swedish based on a written-language tagger (Johanson-Kokkinakis 2003), and an Icelandic tagger (Loftsson 2008).

Finally we will argue for the significance of having a user-friendly interface that is at the same time simple and advanced with regard to search options. All information in and about the corpus should be searchable, whether it is the transcriptons (both kinds) and the tags, or metadata about the informants (language, location, age, sex). We show how we have solved this for the corpus, and how we present the results so that he researchers can immediately get the audio and video linked to the transcriptions, and even maps for geographical locations of the search results and translations of the individual transcriptions to English (using Google technology for the latter two).

The corpus is still under development. By July 2011 there is a total of 2,472,401 words representing 693 from 191 places in 5 countries.

References

Barbiers, S. et al (2006). Dynamic Syntactic Atlas of the Dutch dialects (DynaSAND). Amsterdam, Meertens Institute. http://www.meertens.knaw.nl/sand/

Bick, Eckhard (2003), PaNoLa - The Danish Connection, In: Henrik Holmboe (red.) Nordic Language Technology, Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000-2004 (Yearbook 2002). pp. 75-88. Copenhagen: Museum Tusculanum.

British National Corpus: http://www.natcorp.ox.ac.uk/

Johannessen, Janne Bondi, Kristin Hagen og Anders Nøklestad. 2000. A Constraint-based Tagger for Norwegian. I Lindberg, Carl-Erik og Steffen Nordahl Lund (red.): 17th Scandinavian Conference of Linguistics. Odense Working Papers in Language and Communication 19, 31-48, University of Southern Denmark, Odense.

Johannessen, Janne Bondi, Joel Priestley, Kristin Hagen, Tor Anders Åfarli, and Øystein Alexander Vangsnes. 2009. The Nordic Dialect Corpus - an Advanced Research Tool. In Jokinen, Kristiina and Eckhard Bick (eds.): Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009. NEALT Proceedings Series Volume 4.

Johannessen, Janne Bondi. 2011. The Nordic Dialect Corpus – a joint research infrastructure. In J. B. Johannessen (ed.), Language Variation Infrastructure, Oslo Studies in Language 3(2), 45–62.

Johansson-Kokkinakis, Sofie. 2003. En studie över påverkande faktorer i ordklasstaggning. Baserad på taggning av svensk text med EPOS. Göteborg University.

La phonologie du français contemporain : usages, variétés et structure (PFC): http://www.projet-pfc.net/pfc-recherche

Loftsson, Hrafn. 2008. Tagging Icelandic text: A linguistic rule-based approach. Nordic Journal of Linguistics 31.1.

Nordic Dialect Corpus: http://www.tekstlab.uio.no/nota/scandiasyn/

Scottish Corpus of Text and Speech: http://www.scottishcorpus.ac.uk/

Trosterud, Trond. 2009. A constraint grammar for Faroese. NEALT Proceedings Series.