The Corpus of American Norwegian Speech (CANS)

Janne Bondi Johannessen
The Text Laboratory & MultiLing, University of Oslo / P.O. Box 1102
Blindern, 0317 Oslo, Norway
jannebj@iln.uio.no

Abstract

This paper contains a description of the Corpus of American Norwegian Speech, a new tool for heritage language research. We present the background for its existence, the linguistic contents and its main technical features. The demonstration will show the corpus in use, focussing on problems that are specific to heritage language research, and how the corpus can be searched to provide relevant data.

1 Introduction

The American Norwegian language is a dying language. It is therefore important to record it and make it available for research. The best way to do this, is to transcribe the recordings, morphologically tag the transcriptions and make them available in a searchable corpus, The Corpus of American Norwegian Speech (CANS). The Text Laboratory at the University of Oslo (UiO) has developed many speech features for its corpus system Glossa (Johannessen et al. 2008), which is already used for many other speech corpora, so it has been a relatively easy task to put one more corpus into the same architecture. The Glossa architecture has proved to be a valuable corpus search system, and is used for many corpus projects outside the UiO, most recently the Finland Swedish speech corpus Talko. Glossa is currently undergoing further development under the CLARIN umbrella and the Norwegian project Clarino.

The paper is structured as follows: Section 2 gives some background on the American Norwegian heritage language, while Section 3 describes the main features of the corpus, first with a presentation of the informants in the corpus in numbers, then with a list of the special metadata of these heritage informants, and finally with the way the corpus has been transcribed and annotated. Section 4 presents some examples of search possibilities, and Section 5 sums up the paper.

2 The Norwegian Heritage language in America

Norwegian emigration to America started in 1825, and by 1930, 850 000 people had left for the new world, i.e. the USA and Canada, approximately the same as the whole population of Norway in 1800. Most settled in the Midwest. In these rural areas, whole communities had a Norwegian population, and for a long time they had their own schools, churches and newspapers, thereby keeping the language alive. Researchers have been interested in the American Norwegian language at several points in time. Didrik Arup Seip and Ernst W. Selmer did recordings in the 1930s, Einar Haugen in 1940s, Arnstein Hjelde in the 1990s, and the present author with colleagues in the 2010s.

Research into heritage language has recently become a major field of linguistics, Rothman (2009:159) defines this way: "A language qualifies as a heritage language if it is a language spoken at home or otherwise readily available to young children, and crucially this language is not a dominant language of the larger (national) society." The last years an annual workshop series has been established (the last was held in 2014 at UCLA), and in 2012 a special issue of *Norsk Lingvistisk Tidsskrift* was devoted to Norwegian heritage language (for both, see References).

Heritage languages usually differ from their mother languages (if these are a majority language elsewhere), in that they have a rather large number of loanwords, and that their phonology, morphology and syntax may have features from the neighbouring language or at least are different from those of their mother language.

For these reasons, studying heritage language may bring new knowledge on the human language capacity. It is interesting to see which linguistic features that change, and how, and even to compare the changes with the language of first and second acquisition of the mother language. It is obvious that a good corpus of heritage language is a valuable resource for such research.

3 The CANS corpus

3.1 The contents of the corpus: informants and numbers

The corpus will be growing as there are many recordings that are in the process of being transcribed and annotated for corpus adaption, but at the moment the corpus consists of 131 000 words based on the speech of 36 informants from 13 different speakers from Illinois, Iowa, Minnesota, South Dakota and Wisconsin. The speakers range in age from 67 to 96, but the majority are in their 80s. Since none of them have transferred the language to the next generation, this language is dying. The recordings are from the fieldwork conducted by the present author, from the 2010s.

3.2 Metadata on informants

For each informant, a variety of metadata is available, and searchable. This includes place and state of informant, age, year of birth, language of instruction at school, how much contact they have with Norway, how many times they have visited Norway, whether they read Norwegian, whether they have Norwegian as their mother tongue (L1), which generation immigrant they are, area in Norway where ancestors came from, number of words in the interview, and year of recording. The heritage corpus-specific metadata has been selected by consulting researchers who were using the corpus from the start (see Acknowledgments).

3.3 Transcriptions

All the recordings have been transcribed in two ways: a phonetic-like one and a standard orthographic one. The phonetic transcription was done first via the free software Transcriber and later by Elan. The result of this manual transcription has then been translated to orthographic transcription using a semi-automatic Dialect Transliterator, developed at the Text Laboratory. This transliterator uses a "bi-lingual" word list consisting of dialectal, phonetically written word forms and their standard orthographic equivalents, and which is a result of previous transliterations. It translates each phonetically written form to an orthographic word. The result is then inspected manually, checking the two transcription equivalents and comparing the transcriptions to the original audio versions.

After manual inspection and correction, the new transliterated set of phonetic and orthographic text is fed back into the transliterator, improving the word list for further use for that particular dialect or language variety. Whenever a word in its phonetic transcription is not found in the word list, the same word is used for transliteration, to be given an orthographic form later in the manual inspection. At this stage, certain other annotations are also added, see for example Section 3.5.

The two transcriptions are strictly aligned word by word, and linked together in Glossa, and the user can choose to search in only one of them, or in both simultaneously. (See list of web sites at the end of this paper.)

3.4 Tagging

The CANS corpus has been tagged with a TreeTagger (Schmid 1994, 1995) trained on a speech version of the Oslo-Bergen tagger, developed for a speech corpus of the Oslo dialect, and then measured by 10-fold cross validation at an accuracy of 96, 9 % (Søfteland and Nøklestad 2008). The accuracy has not been measured for the CANS corpus, but with its high number of English loan words and dialectal word order, the result is likely not as good.

3.5 Other annotation

Since CANS contains heritage language it has many loanwords. These have been annotated manually in the transliteration process with the tag x. Even if the corpus contains only about 130 000 words, the number of words tagged with x is nearly 4000.

4 Searching the CANS corpus

The corpus can be searched using words, parts of words or word combinations and by annotations. The metadata can also be used as search filters, see Johannessen et al. (2012) and (2014) for a general introduction to the corpus search features.

In this paper the focus is on what is special for the CANS corpus.

Figure 1 illustrates a search for the x-tagging described in Section 4.5; the x is chosen from the Criteria menu that filters any search given in the word-box above it. (The box can be empty, too, as it is here.)

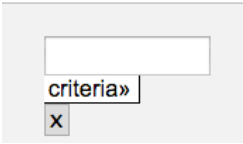

Figure 1: Searching for x-tagged words

This search gives 3857 hits. One is illustrated below, in Figure 2. (The orthographic transcription is on the first line, followed by the phonetic one on the second.)

å jeg driver og **raiser** noe hester

å e driv å **reiser** no hæsster

'Oh, I raise some horses.' (blair_WI_01gm)

Figure 2: One of the results from the search for x-tagged words (*raise* is an English loanword; the Norwegian equivalent is *avle*.)

Other typical words tagged with x are interjections like *oh, huh, right*, and words that have replaced Norwegian ones, like *cousin* (instead of fetter), *back* (tilbake), *figure* (think), *telle* (fortelle).

An example of filtering a search by informant metadata is given in Figure 3, where informants are specified to be fourth generation immigrant. This box comes in addition to the general word-box, this time searching for the word *ikke* 'not'.

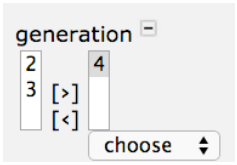

Figure 3: Filtering a search by speaker's generation

The search yielded 467 results divided amongst 7 informants. It can be filtered further, e.g. by the home county of the ancestors, age or any other metadata.

5 Conclusion

This paper has presented the Corpus of American Norwegian Speech, a heritage language corpus. The paper has focussed on heritage languages in general and American Norwegian in particular, and the most central details of the corpus. Some statistics on the informants were offered, the special metadata of these heritage informants were presented, and transcription and annotation were briefly presented. Finally, some illustrations of search possibilities were given. For more general speech corpus features of CANS, the readers are referred to Johannessen et al. (2012, 2014). Expansions of the corpus are expected, as there is presently some funding for more transcriptions. The corpus may also be expanded with other heritage Scandinavian languages.

Acknowledgments

The Corpus of American Norwegian Speech could not have been developed without several people. The most important ones are the American Norwegians who have willingly sat down in front of a camera to talk with their fellow heritage speakers and the researchers. Signe Laake and Arnstein Hjelde have been excellent collaborators and company at fieldwork on several occasions from 2010 to 2014.

The people at the Text Laboratory, UiO, have a major role in the corpus development. Kristin Hagen has taken the responsibility for technical decisions and coordinating work. Joel Priestley has designed the corpus interface and is chief programmer for all the specific speech features. Eirik Olsen, André Kaasen and Eirik Tengesdal have been vital in observations and suggestions related to transcription.

The collection of material, i.e. fieldwork and recordings, has been funded by The Research Council of Norway (RCN) under the project Norwegian Dialect Syntax; the Department of Linguistics and Scandinavian Studies, UiO. Transcriptions have been funded partly by the Text Lab, UiO, by the University of Tromsø (thanks are due to Marit Westergaard and Merete Anderssen, who have also given input on search options and metadata in the corpus), and by the RCN project LIA, no. 225941. The fieldwork was partly supported by the RCN through its Centres of Excellence funding scheme, project no. 223265.

References

Johannessen, Janne Bondi, Lars Nygaard, Joel Priestley, and Anders Nøklestad. 2008. Glossa: a Multilingual, Multimodal, Configurable User Interface. In: *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. Paris: European Language Resources Association (ELRA). http://www.hf.uio.no/iln/tjenester/kunnskap/sprak/glossa/LREC-glossa_2008.pdf

Johannessen, Janne Bondi, Joel Priestley, Kristin Hagen, Anders Nøklestad, and Andre Lynum. 2012. The Nordic Dialect Corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*. European Language Resources Association, p. 3388-3391. http://dblp.uni-trier.de/db/conf/lrec/lrec2012.html

Johannessen, Janne Bondi, Øystein Alexander Vangsnes, Joel Priestley, Kristin Hage,. 2014. A multilingual speech corpus of North-Germanic languages. In Raso, Tommaso; Mello, Heliana (eds.): *Spoken Corpora and Linguistic Studies*. John Benjamins Publishing Company, p. 69-83. https://www.benjamins.com/#catalog/books/scl.61.02joh/fulltext

Norsk Lingvistisk Tidsskrift [Norwegian Linguistics Journal]. 2012. Special issue on the Norwegian Language in America (edited by Janne Bondi Johannessen and Joe Salmons).

Rothman, Jason. 2009. Understanding the Nature and Outcomes of Early Bilingualism: Romance Languages as Heritage Languages. *The International Journal of Bilingualism* 13: 155-163.

Schmid, Helmut. 1995. Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Søfteland, Åshild and Anders Nøklestad. 2008. Manuell morfologisk tagging av NoTa-materialet med støtte fra en statistisk tagger. In Johannessen, Janne Bondi og Kristin Hagen (eds.) *Språk i Oslo. Ny forskning omkring talespråk*. Novus, Oslo.

Web sites

5th Annual Workshop on Immigrant Languages in the Americas, UCLA, October 17-19, 2014. http://tekstlab.uio.no/WILA5/index.html

CLARIN: http://www.clarin.eu/

Clarino: http://clarin.b.uib.no/

Corpus of American Norwegian Speech (CANS): http://tekstlab.uio.no/glossa/html/?corpus=amerikanorsk

DialectTransliterator: http://omilia.uio.no/scandiasyn/translit/

Elan: https://tla.mpi.nl/tools/tla-tools/elan/

Glossa corpus search and processing tool: http://www.hf.uio.no/iln/english/about/organization/text-laboratory/services/glossa.html

Oslo-Bergen Tagger: http://tekstlab.uio.no/obt-ny/english/index.html

Talko: http://www.sls.fi/doc.php?category=2&docid=943

Text Laboratory: http://www.hf.uio.no/iln/english/about/organization/text-laboratory/

TreeTagger: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

Transcriber: http://trans.sourceforge.net/en/presentation.php