From Fieldwork to Speech Corpus: The American Norwegian Heritage Language and CANS

Janne Bondi Johannessen

1. Introduction

The emigration from Norway to America lasted for ca. 100 years, from 1825 to the 1920s. During this period, 800,000 Norwegians emigrated, mostly to the American Midwest, but also to Canada (Johannessen & Salmons 2015: 1-20). In the period of 1875-1905, more than 30% - 40% of the population left from the rural communities in Southern Norway (Myhre 2015).

Many of those who came to America chose to settle together in the countryside, where they could perform farming activities. From a linguistic perspective, this was fortunate, as the Norwegian language continued to be used in the stable Norwegian communities that were founded in the new homeland. Their Norwegian language became a heritage language ("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", Rothman 2009: 156). Even now, ca. 200 years after the first emigrants left Norway for "the promised land," the Norwegian language is still used by individuals and groups in rural pockets in the USA and Canada. This language has been an intriguing research object for linguists and laymen for 150 years (Rynning 1838: 26; Duus 1855-1858; Flaten 1900-1904; Flom 1902, 1926; see Hjelde & Johannessen 2017 for more).

Studying a heritage language that is the result of migration is a fascinating enterprise for linguists. The way it has developed away from the language of the homeland can be studied in any of the linguistic disciplines, such as phonetics and phonology, morphology, syntax, semantics, and lexicography. It is often necessary to relate empirical findings to characteristics of the speakers (e.g., their age at the time of recording, age at the time of acquisition of both the heritage and dominant language, their literacy in the heritage language) and the society they live in (e.g., how big the heritage community is, if and when instruction at school was in the heritage language, the language of the church, newspapers, and other institutions). These days, heritage language studies are popular, and it is possible to compare research of one heritage language with that of other heritage languages. It is also interesting to relate the study of a heritage language to studies of immigrant languages in the homeland. The study of a heritage language is one window to understand the nature of human language generally, together with studies of first and second language acquisition and language attrition (for general literature on these topics, see Benmamoun, Montrul & Polinsky 2013a, 2013b; Johannessen & Salmons 2015; Johannessen 2018; Schmid & Köpke 2019; Lohndal et al. 2019; Montrul & Polinsky, forthcoming; Johannessen & Putnam 2020).

[*] I would like to thank my collaborators and friends at WILA 10, Østfold University College, for inviting me to give a keynote talk and giving me the chance to talk about the importance of data and corpora. I would also like to thank my colleagues at the Text Laboratory, UiO, Kristin Hagen, Anders Nøklestad and Joel Priestley, for the great work they have done and continue to do in implementing our ideas into concrete reality, and to Kristin for helping to keep the projects going and taking care of the millions of things that must be seen to. I am also grateful to earlier and present fieldworkers for their efforts, and to all the people who let their language be available for research. The work described in this paper has been funded by many sources, but the most important ones are: The Research Council of Norway (RCN) projects NorAmDiaSyn and Centres of Excellence funding scheme, MultiLing for fieldwork, and the RCN infrastructure projects CLARINO for developing the corpus system Glossa, and LIA for transcription of much of the material.

© 2021 Janne Bondi Johannessen. Selected Proceedings of the 10th Workshop on Immigrant Languages in the Americas (WILA 10), ed. Arnstein Hjelde and Åshild Søfteland, 1-10. Somerville, MA: Cascadilla Proceedings Project.

2. Collection of Heritage Norwegian Data

2.1. Until 1990

Ernst W. Selmer and Didrik Arup Seip were professors at the University of Oslo and went to the American Midwest in 1931 to document and study the language there. Seip (1933: 257-259) says that their main questions were: *Have the peculiarities of Norwegian dialects been maintained in the Norwegian communities in America?* and *How have the Norwegian dialects over there influenced each other?* However, these questions were difficult to answer because the dialects had already been in contact with each other, with Swedish, and of course English. Seip and Selmer made recordings, which they kept at the Phonetics Department at the University "as a testimonial to the language our countrymen used in the new world" [transl. JBJ], but did no linguistic research based on them. The recordings were done on wax cylinders; however, the cylinders were not well preserved, and many have been broken or lost. There were 292 informants on 354 rolls, but very few were from the known Norwegian American rural areas. Many instead were educated from urban areas and spoke a Norwegian close to the written standard ("book Norwegian"). Some were even born in Norway, according to Haugen (1992).

What is left of the recordings is now available online at the Text Laboratory, UiO (see URLs at the end of this paper). From Seip and Selmer, there are original lists with 101 names and six tapes with several recordings on them; many of them destroyed. The recordings were handed over to Prof. Hallvard Dørum, UiO, by Einar Haugen. Later, the waxed cylinders were copied onto audio tapes by the Swedish national radio. In 2009, Dørum handed the tapes to me, and they were copied onto a DVD and put on a server at the Text Laboratory, ILN, UiO.

Einar Haugen recorded the speech of Norwegian Americans together with Magne Oftedal in the period of 1935–1948. They visited Wisconsin, Minnesota, Iowa, and Illinois. Haugen and Oftedal used a meticulous system for annotating information about the participants and even wrote assessments of each, along with a wealth of other information. They recorded 207 informants, most of whom were heritage speakers, but some were also born in Norway. The recordings were distributed on 105 tapes which were later divided into eight DVDs digitized by the National Library (after having been made into cassette copies in 1969 at Harvard Audio Laboratory for Prof. Inger Moen, UiO). The speakers on the recordings of the first two DVDs have been identified by Arnstein Hjelde (Østfold University College). The recordings are available online at the Text Laboratory, together with images of the assessments and transcriptions for some of the recordings.

Arnstein Hjelde made recordings in Minnesota, South Dakota, and North Dakota, with heritage speakers who spoke the Norwegian Trønder dialect. The Text Laboratory has 79 online recordings from 1987 available, as well as metadata about the speakers. A selection of recordings from all these three sets of sources has been transcribed, annotated, and included in the CANS corpus (see Section 3).

2.2. From 2010

2.2.1. The Start

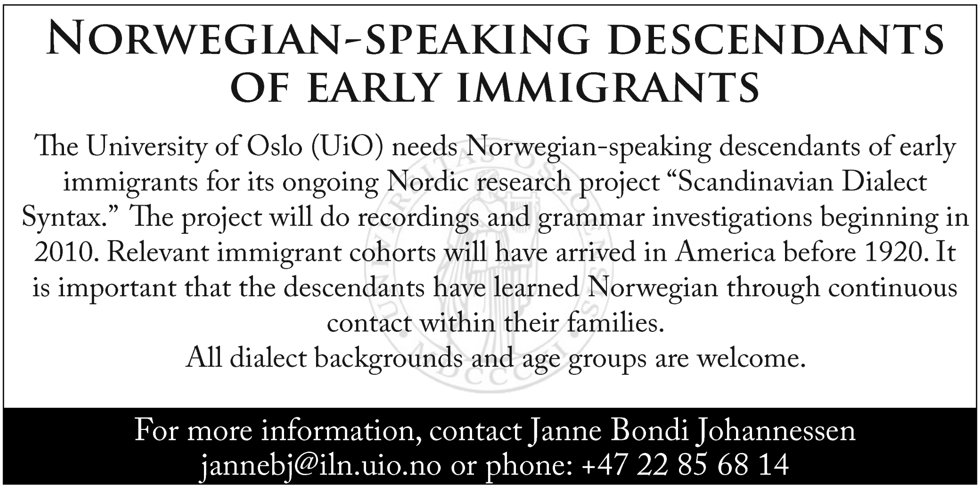
In 2010, the present author, Janne Bondi Johannessen, was the project leader of a major research project called Norwegian Dialect Syntax – NorDiaSyn (which again was connected to a series of other Nordic dialect projects under the umbrella Scandinavian Dialect Syntax), financed by the RCN. I was contacted by the research council, which could offer extra money if the project could be expanded to cooperation with researchers in a selection of countries, including the United States of America. This seemed like a golden opportunity to study the remains of the American Norwegian language, though we were doubtful there was anything left of it. Through contacts, I was given the names of Joseph Salmons (University of Wisconsin, Madison) and Arnstein Hjelde (Østfold University College, Halden); the latter had worked on this language in the past (see Section 2.1).[1] While Hjelde could assure us that there must still be Norwegian-speaking Americans, Salmons enthusiastically agreed to arrange a seminar with his

[1] I am grateful to Bert Vaux, University of Cambridge, and Ingeborg Kongslien, UiO for these contacts, and also to Ingeborg for sharing her knowledge on the Norwegian culture in America, including magazines and newspapers which could receive our advertisements.

colleagues and Norwegians in the summer of 2010. This was the start of the Norwegian American Dialect Syntax Project (NorAmDiaSyn), which resulted in the fieldwork and corpus presented in this paper, and also in the international WILA (Workshop of the Immigrant Languages in the Americas) series, which has been organized every year since 2010.

2.2.2. Field Trips

In the meantime, I had put advertisements in various magazines asking for people who were Norwegian speakers, who had learned the language in their families, and whose ancestors had immigrated before 1920 (see Figure 1). We received around 30 answers, mostly from neighbors and relatives of elderly speakers. This resulted in the first field trip by the present author in Norwegian America in March 2010, followed by nine more until 2016. On the first trip, I was accompanied by research assistant Signe Laake. On the second, we arranged fieldwork for many of our Norwegian and American colleagues in connection with the first workshop. On subsequent trips, Arnstein Hjelde was a faithful companion, who could also offer some of his old acquaintances, as well as various other research assistants and students (for more information on these trips, see the field trip web site, URL at the back, and Figure 2).



NORWEGIAN-SPEAKING DESCENDANTS
OF EARLY IMMIGRANTS

The University of Oslo (UiO) needs Norwegian-speaking descendants of early immigrants for its ongoing Nordic research project "Scandinavian Dialect Syntax." The project will do recordings and grammar investigations beginning in 2010. Relevant immigrant cohorts will have arrived in America before 1920. It is important that the descendants have learned Norwegian through continuous contact within their families.
All dialect backgrounds and age groups are welcome.

For more information, contact Janne Bondi Johannessen
jannebj@iln.uio.no or phone: +47 22 85 68 14

Figure 1: Advertisement in an American magazine.

To date, several hundred speakers of the American Norwegian heritage language have been interviewed and recorded. This paper presents version 3 of the CANS corpus, which contains 205 speakers of Norwegian (version 3.1 has since been published with 268 speakers). Being aware that this heritage language is moribund, it has been central to collect as much material as possible by as many speakers as possible. Furthermore, for some research it may be crucial to know something about the English language that heritage Norwegian speakers speak. Therefore, in 2016, the English speech of some of the participants was also recorded. This way, it will be possible to study whether their English language is different from that of people who are not Norwegian heritage speakers.

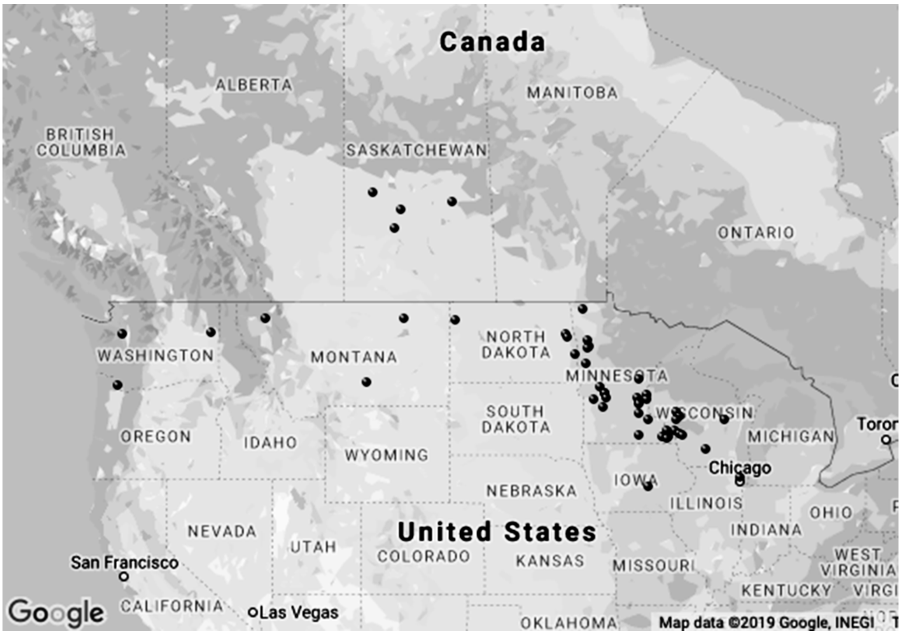


Figure 2: Places where recordings were done 2010-2016.

2.2.3. Method

When given the opportunity, it was preferable to record two speakers engaged in conversation together, but also to have individual interviews. This method was also the one used in Norway and the other Scandinavian countries when dialect speakers were recorded (cf. 2.2.1). It was assumed that the Norwegian Americans would be more relaxed when speaking to each other than when speaking to the researchers. However, the researchers also wanted the more formal interview because then they had a chance to ask the same questions to all (mainly background questions on their lives), and also to perhaps see whether they had more than one linguistic register. In order for the participants to get into a Norwegian language mode, they were encouraged to speak about the olden days, which would typically include the family and farm life of yesteryear, as these were experiences that would have been accompanied by the Norwegian language.

We also made sure to ask about factors that might be relevant in understanding the participants' language performance, such as age, birth year, immigrant generation, when they learned the heritage language, what language they spoke at school and confirmation, whether they were literate in Norwegian, etc. Each speaker was recorded for 30-60 minutes, sometimes on more than one trip. As the speakers were elderly, they got tired quite quickly. They often had appointments with the doctor, where they had to take each other or go out on the field with their big combines (the latter happened mainly in Canada). Here it should be mentioned that Einar Haugen actually interviewed his own informants for 12 hours. From today’s perspective where people are always busy, this is almost unimaginable.

The fieldwork in the NorDiaSyn project in Norway included grammaticality judgement tasks, but this was in a country where speakers had Norwegian as their dominant language, were literate, and usually also younger. We were skeptical about using these kinds of tasks with the elderly, less fluent heritage speakers in America, as we expected them to accept too much (i.e., that they were not so certain about their own judgments). In addition, our heritage speakers were also illiterate in Norwegian and had little faith in their knowledge. Such tasks would also remind them of school, something of which they only had a faint memory, and maybe not so nice either. Our participants were usually not highly educated, and many had struggled at school during their early years due to language difficulties (i.e., school was in English, while they mostly knew Norwegian). However, we did try a few tests on the most recent trips, and they worked reasonably well as long as the tests were supported with pictures.

Finally, heritage speakers feel vulnerable about their language skills. Therefore, the study's method used the same type of introductory speech with these speakers that was used with dialect speakers in the Scandinavian countries. We first explained that the documentation of their speech was part of a large project in Europe, where we tried to document dialects across all the Nordic countries, and that the Norwegian dialects in America were no less interesting. We also explained that there was a general interest in heritage languages in linguistics at the moment, and that the special qualities their languages have are intriguing to us. Finally, we told them we knew their language would be different from the language used in Norway (e.g., there would be many words from English, and there could be words they could not remember). We also mentioned that we were interested in their language precisely because it was not ordinary Norwegian; if we had wanted that sort of speech, we would not have come to America to study theirs. Providing this background information had the desired effect. The participants became more relaxed, as they understood we really meant it when we said we were interested in their particular way of speaking.

3. A Searchable Speech Corpus: Steps from Data to Corpus

Within 10 field trips, we had collected many hours of recordings, which ought to be shared with present and future researchers. As the director of the Text Laboratory, UiO, I knew that the obvious way of doing this was to make a searchable speech corpus with direct access to sound and video. We had already done this for other speech corpora containing dialects in Norway and Nordic countries, and we knew how this could generate a lot of new research.

The corpus is transcribed both phonetically (using Elan software) and orthographically (using the semi-automatic Oslo transliterator). The latter orthography is useful for searching for words or forms across pronunciations (i.e., generalized searches). This way, one can see all phonetic variants of a certain orthographic word, which makes it possible to find (new) patterns and to get an overview of the vocabulary. It is also necessary for applying other tools of language technology, such as morphological taggers and parsers. The phonetic transcription can be used to search directly for certain pronunciations, to see the pronounced forms in writing, and the different pronunciations displayed on a map.

The main cost associated with building this kind of corpus is related to transcription. Many researchers were interested in this corpus, and some even helped pay for transcriptions (Marit Westergaard and Merete Anderssen, UiT The Arctic University of Norway, Arnstein Hjelde, Østfold University College, and Ida Larsson and Kari Kinn, UiO), which were otherwise financed by the Text Laboratory, UiO, and the Department of Linguistics and Scandinavian Studies. Later, the large project entitled Language Infrastructure made Accessible (LIA), financed by the Research Council of Norway 2014–2019, would pay for many more transcriptions. In the meantime, Prof. Ida Larsson (UiO) got the idea of adding Swedish data to the corpus. The name was changed accordingly, from Corpus of American Norwegian Speech to Corpus of American Nordic Speech, using the same acronym of CANS.

CANS is also morphologically tagged. For Norwegian, a TreeTagger trained on corrected output from the Oslo-Bergen-tagger was used (Nøklestad & Søfteland 2007; Johannessen et al. 2012). For Swedish, a TnT tagger trained on the Swedish PAROLE corpus (Kokkinakis 2003) was used (see also Johannessen 2015).

In addition to recordings, transcription, and tagging, a corpus needs good metadata. In CANS, there is a multitude of information on each speaker and situation: age, gender, heritage, Norwegian background, American background, country, place, area (state), number of visits to Scandinavia, confirmation language, school language, emigration year, generation, recording year, and genre. All of the above components are then inserted into a corpus system. The CANS corpus uses Glossa, developed at the Text Laboratory (Nøklestad et al. 2017) over several years, in close consultation with users.

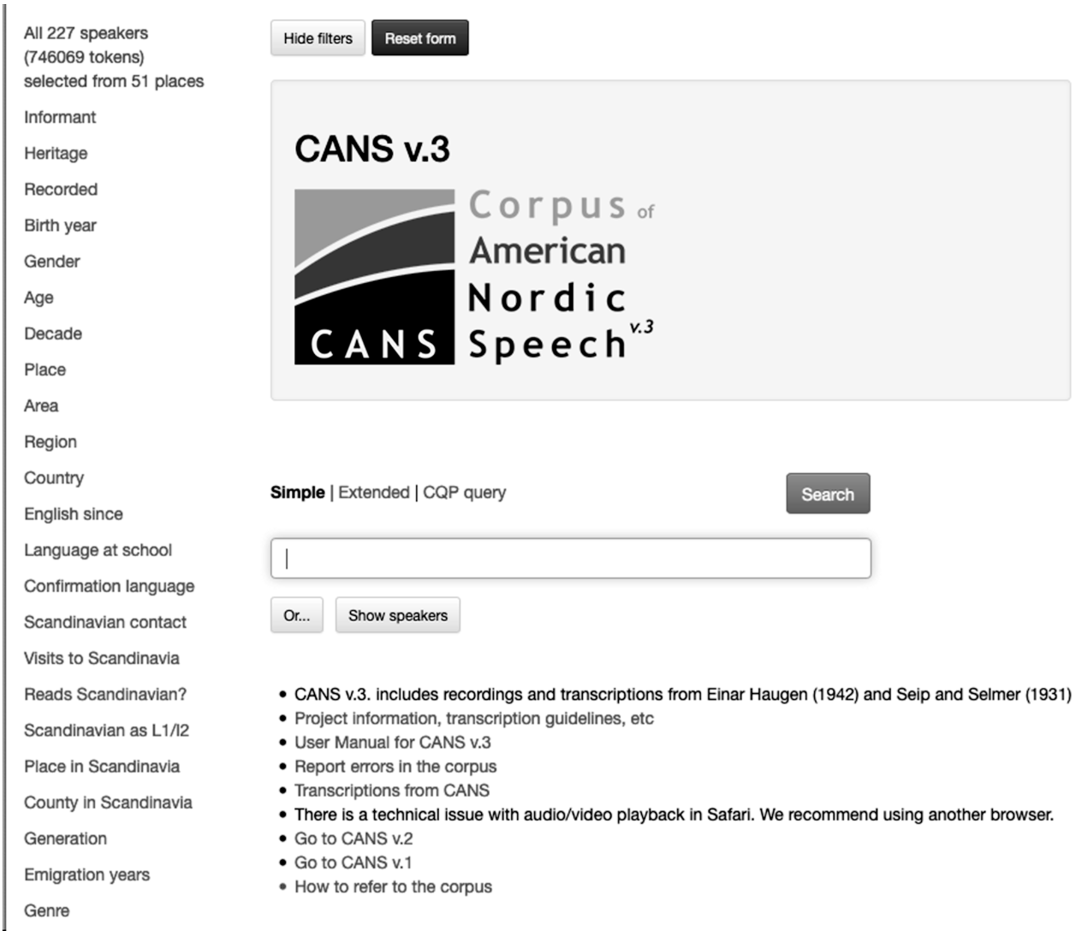
The version of the corpus presented here contains data from recordings done by Seip and Selmer, Haugen, Hjelde, and those conducted by myself and colleagues. This covers a period of 85 years. The oldest speaker was born in 1850 and the youngest in 1999, representing a 150-year span. There are recordings from two countries and two closely related heritage languages (see Table 1). The corpus will continue to grow as there are still recordings that are not transcribed. If more data are collected, this will also be added to the corpus.

Table 1: Some basic facts of the corpus of American Nordic Speech v.3

No. of speakers	227
Heritage Swedish speakers	22
Heritage Norwegian speakers	205
No. of tokens	746,069
Places in Canada	4 (Archerwill, North Battleford, Outlook, Saskatoon)
Places in USA	47 (in Illinois, Iowa, Minnesota, Montana, North Dakota, South Dakota, Washington, Wisconsin)
Oldest recordings	1931 and 1942 (47 speakers)
Most recent recordings	2016 (2 speakers)

4. Use of the CANS Corpus of American Nordic Speech

A searchable electronic corpus has the potential to be used for many different types of research questions. If it is available on the web and easy to use, like the CANS corpus, the chances of it being used for research are high. Below I give illustrations of some of the ways that the CANS corpus can be used, and a discussion of its actual use by researchers follows.



All 227 speakers (746069 tokens) selected from 51 places

- CANS v.3. includes recordings and transcriptions from Einar Haugen (1942) and Seip and Selmer (1931)
- Project information, transcription guidelines, etc
- User Manual for CANS v.3
- Report errors in the corpus
- Transcriptions from CANS
- There is a technical issue with audio/video playback in Safari. We recommend using another browser.
- Go to CANS v.2
- Go to CANS v.1
- How to refer to the corpus

Figure 3: The CANS corpus search interface.

The long list on the left in Figure 3 contains metadata that can be used as a filter for the linguistic search. The ‘Show Speakers’ button in the middle of the figure can be used to see a list of the speakers that are left after metadata filtering. The input field in Figure 3 shows the simplest one of three different alternative interfaces. However, it is often useful to use the extended interface, which makes it possible to extend the search words with different properties. In Figure 4, this is illustrated by a search for any determiner in the singular form, followed by a word that is not in the Norwegian dictionary (categorized as ‘x’), which is likely to be a loanword. This is possible since the corpus is morphosyntactically tagged. This kind of search makes finding the gender of loanwords quite easy, for instance. Figure 5 shows some of the results from this search in the form of a concordance. As expected, the search renders loanwords: *business college, beans, spring break*.

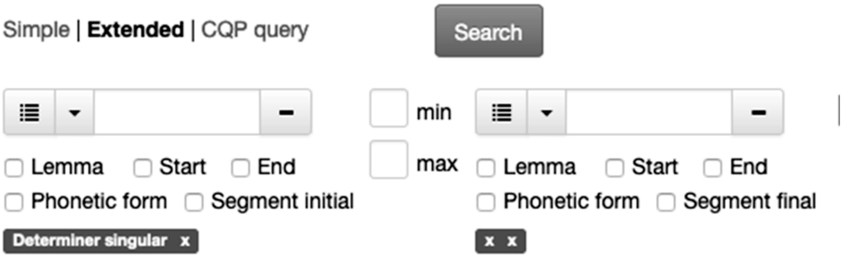
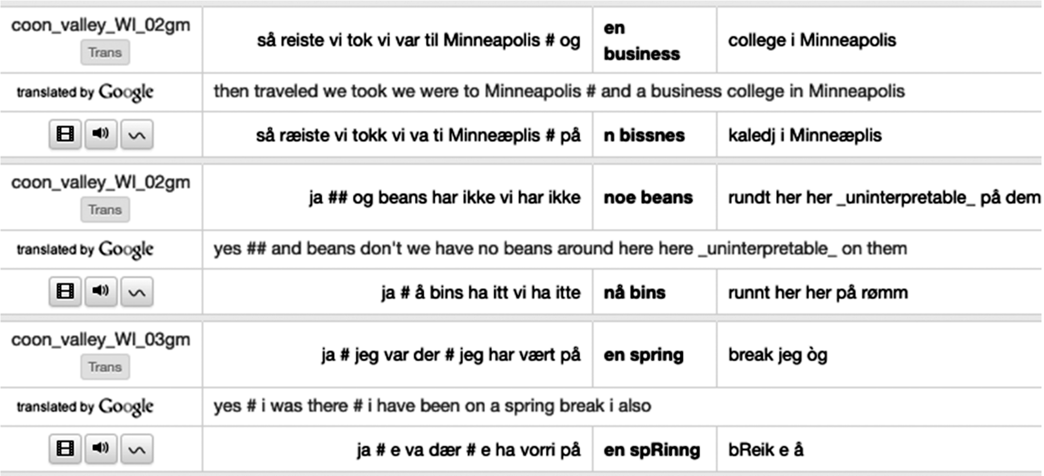

Simple | Extended | CQP query
Search
min
max
☐ Lemma ☐ Start ☐ End
☐ Phonetic form ☐ Segment initial
Determiner singular x
☐ Lemma ☐ Start ☐ End
☐ Phonetic form ☐ Segment final
x x

Figure 4: The extended search option: Two words, where the first should be a singular determiner and the second a word not in the dictionary.



coon_valley_WI_02gm Trans	så reiste vi tok vi var til Minneapolis # og	**en business**	college i Minneapolis
translated by Google	then traveled we took we were to Minneapolis # and a business college in Minneapolis		
	så ræiste vi tokk vi va ti Minneæplis # på	**n bissnes**	kaledj i Minneæplis
coon_valley_WI_02gm Trans	ja ## og beans har ikke vi har ikke	**noe beans**	rundt her her _uninterpretable_ på dem
translated by Google	yes ## and beans don't we have no beans around here here _uninterpretable_ on them		
	ja # å bins ha itt vi ha itte	**nå bins**	runnt her her på rømm
coon_valley_WI_03gm Trans	ja # jeg var der # jeg har vært på	**en spring**	break jeg òg
translated by Google	yes # i was there # i have been on a spring break i also		
	ja # e va dær # e ha vorri på	**en spRinng**	bReik e å

Figure 5: Some results from the search for a singular determiner plus a word not in the dictionary. (Notice that the concordance shows both the orthographic and the phonetic transcription, as well as the built-in translation from Google.)

The metadata menu that was shown in Figure 3 is clickable, and the values within each category depend on the type of data. For the birth year, all known years for all the participants are given (see Figure 6).

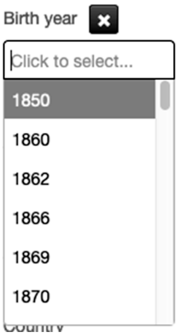


Figure 6: An item in the metadata menu, Birth year, is chosen and the values that are associated with participants in the corpus pop up.

In addition to viewing results as a concordance, it is possible to see the hits distributed on a map. This is illustrated in Figure 7, where the search item was *travle*, which seems to be a new word in American heritage Norwegian. It is a verb meaning ‘to walk,’ replacing the homeland verb *gå*, which has taken on the same meaning as the English *go*, discussed in Haugen 1953, Hjelde & Johannessen (2017) and Johannessen & Laake (2017). The map view can show how widespread a certain item is, and in this case, it shows clearly that the new meaning stretches over an area that covers two countries and five states.

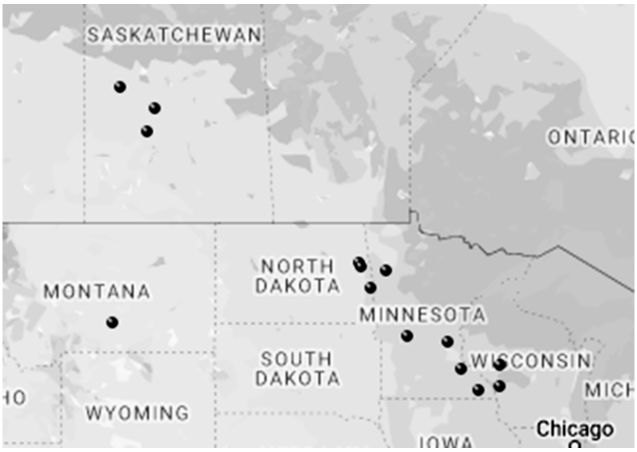


Figure 7: The result of a search for *travle*, meaning ‘to walk,’ where the results are given on a map rather than as a concordance (only the places where this word has been used are marked on the map).

For further examples and information on how to use the corpus, I refer to the User Manual for CANS (see References). There is no doubt that researchers have embraced the opportunity to get empirical data from the CANS corpus. A search on Google Scholar using “corpus of American Norwegian speech” (74 hits), “corpus of American Nordic speech” (11 hits), and “the cans corpus” (19 hits) yields 104 scientific publications where this corpus is mentioned. In addition, there will be studies that have not been picked up by Google Scholar because researchers have not mentioned the corpus explicitly. Furthermore, the recordings on the Text Laboratory website are popular with some researchers.

The CANS corpus has been used by many different scholars since it was launched in 2014. Thirty-five are mentioned here (from Google Scholar, 2019): A Alexiadou, G Andersen, M Anderssen, M Andréasson, L Annear, Y van Baal, P Bartásková, A Bjerkstig, JR Brown, J Bousquette, R Eik, C Furiassi, MB Grimstad, GF Hansen, A Hjelde, JB Johannessen, S Khayitova, K Kühl, AA Kåsen, I Larsson, KG Leskinen, T Lohndal, B Lundquist, AK Lykke, S Laake, E Olsen, JH Petersen, B Riksem, LIS Rødvand, K Speth, MS Strand, E Tengesdal, AM Tjugum, M Westergaard, and TA Åfarli. The topics are varied, including grammatical gender, word order in subordinate clauses, word order in main clauses, verb inflection, complex definiteness in the noun phrase, subject shift and object shift, cross-linguistic comparisons, attrition, frequency effects in the heritage language, language variation, stability, and change. Many of the studies are PhD and MA theses.

In the future, the CANS corpus should be used for more morphology and syntax, but also typology (cross-linguistic studies), comparison (across speakers and areas), sociolinguistics (such as identity issues), language mixing in syntax and discourse, discourse analysis, language processing (structure, input frequency), and lexical and textual contents. The recordings that were collected before 1990 and described above in Section 2.1 were previously not available to the general research community, mostly for technological reasons. The opportunities of modern technology, especially the web-based search system, give possibilities that were not imaginable before. It is a joy to see that Norwegian and Swedish heritage languages have been studied with enthusiasm by so many scholars. The corpus will not be used up and will continue to be there for new researchers and ideas in the years to come.

5. Conclusion: The Advantages of a Corpus Such as CANS

A brief tour of the early and current fieldwork and data collection has been given. What is needed to build a searchable corpus, and how the CANS corpus can be used for research has also been discussed. A corpus is in many ways the ultimate research tool for gathering empirical data. The data—spontaneous dialogues—are very general and can be used for many different studies (unlike specific elicitations). A modern corpus tool is efficient and easy to use, with buttons and menus, compared to old-style corpora where search expressions must be written in code. It is also efficient when compared with doing fieldwork yourself to obtain data, which in this case will be impossible in the future anyway as the relevant informants will no longer be present. A corpus saves the speech of informants for eternity. It is cost-efficient, as collected data can be shared by many researchers anywhere and anytime. Moreover, a corpus makes it possible to replicate studies by others, so that results can be verified.

Fieldwork is a necessary, often fun, but slow and expensive way of getting speech data of a certain type. It is essential that researchers who undertake this kind of effort make the results and data material (e.g., recordings and metadata) available for other researchers. Since this work has been done with public funds, it is reasonable that the public receives the material outcome. For many types of research, a corpus is the ultimate way of making data available.

References

Benmamoun, Elabbas, Silvina Montrul & Maria Polinsky. 2013a. Heritage languages and their speakers: Opportunities and challenges for linguistics. *Theoretical Linguistics* 39(3-4). 129-181.

Benmamoun, Elabbas, Silvina Montrul & Maria Polinsky. 2013b. Defining an "ideal" heritage speaker: Theoretical and methodological challenges. Reply to peer commentaries. *Theoretical Linguistics* 39(3-4). 259-294.

Duus, Olaus Fredrik. 1855-1858 [1947]. *Frontier parsonage: The letters of Olaus Fredrik Duus, Norwegian pastor in Wisconsin.* Northfield: The Norwegian American Historical Association.

Flaten, Nils. 1900-04. Notes on the American-Norwegian with vocabulary. *Dialect Notes* 2. 115-126.

Flom, George T. 1902. English elements in Norse dialects of Utica. *Dialect Notes* 2. 115-26.

Flom, George T. 1926. English loanwords in American Norwegian. *American Speech* 1. 541-58.

Haugen, Einar. 1953. *The Norwegian language in America: A study in bilingual behavior*. Philadelphia: University of Pennsylvania Press.

Haugen, Einar. 1992. A language survey that failed: Seip and Selmer. *American Speech* 67(3)(autumn). 330-336. Duke University Press.

Hjelde, Arnstein & Janne Bondi Johannessen. 2017. Amerikanorsk: Orda vitner om kontakt mellom folk. In Terje Mikael Hasle Joranger (ed.), *Norwegian-American Essays 2017*, 257-282. Oslo: Novus forlag.

Johannessen, Janne Bondi. 2015. The Corpus of American Norwegian Speech (CANS). In Béata Megyesi (ed.), *Proceedings of the 20th Nordic Conference of Computational Linguistics, NODALIDA 2015, May 11-13, 2015, Vilnius, Lithuania*. NEALT Proceedings Series 23.

Johannessen, Janne Bondi. 2018. Factors of variation, maintenance and change in Scandinavian heritage languages. In Jonathan Richard Kasstan, Anita Auer & Joe Salmons (eds.), Special issue on ‘Heritage-language speakers: Theoretical and empirical challenges on sociolinguistic attitudes and prestige’, *International Journal of Bilingualism* 22(4). 447-465. DOI: https://doi.org/10.1177/1367006918762161

Johannessen, Janne Bondi, Kristin Hagen, André Lynum & Anders Nøklestad. 2012. OBT+stat. A combined rule-based and statistical tagger. In Gisle Andersen (ed.), *Exploring Newspaper Language. Corpus compilation and research based on the Norwegian Newspaper Corpus*, 51-65. Amsterdam: John Benjamins Publishing Company.

Johannessen, Janne Bondi & Signe Laake. 2017. Norwegian in the American Midwest: A common dialect? In *Journal of Language Contact* 10(1), 5-21.

Johannessen, Janne Bondi & Michael T. Putnam. (2020). Heritage Germanic languages in North America. In Richard Page & Michael T. Putnam (eds.), *The Cambridge Handbook of Germanic linguistics*, 783-806. Cambridge: Cambridge University Press.

Johannessen, Janne Bondi & Joseph C. Salmons. 2015. The study of Germanic heritage languages in the Americas. In Janne B. Johannessen & Joseph C. Salmons (eds.), *Germanic heritage languages in North America: Acquisition, attrition and change*, 1-20. Amsterdam: John Benjamins Publishing Company.

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

Lohndal, Terje, Jason Rothman, Tanja Kupisch & Marit Westergaard. 2019. Heritage language acquisition: What it reveals and why it is important for formal linguistic theories. *Language and Linguistics Compass* 13(12).

Montrul, Silvina & Maria Polinsky (eds.). (Forthcoming). *The Cambridge Handbook of Heritage Languages and Linguistics.* Cambridge: Cambridge University Press.

Myhre, Jan Eivind. 2015. Utvandring fra Norge. https://www.norgeshistorie.no/industrialisering-og-demokrati/artikler/1537-utvandring-fra-norge.html

Nøklestad, Anders, Kristin Hagen, Janne Bondi Johannessen, Michal Kosek & Joel Priestley. 2017. A modernised version of the Glossa corpus search system. In Jörg Tiedemann (ed.), *Proceedings of the 21st Nordic Conference on Computational Linguistics (NoDaLiDa) 2017*, 251-254.

Nøklestad, Anders & Åshild Søfteland. 2007. Tagging a Norwegian Speech Corpus. *NODALIDA 2007 Conference Proceedings*.

Rynning, Ole. 1839. *Sandfærdig Beretning om Amerika, til Oplysning og Nytte for Bonde og Menigmand. Forfattet af en Norsk, som kom derover i juni Maaned 1837*. Christiania: Guldberg & Dzwonkowski.

Schmid, Monika S. & Barbara Köpke (eds.). 2019. *The Oxford Handbook of Language Attrition*. Oxford: Oxford University Press.

Seip, Didrik Arup. 1933. Nordmenn og norsk språk i Amerika. *Ord och Bild* 1933. 253-59. Rpt. “with some additions” in Seip, Didrik Arup. 1931. *Studier i Norsk Språkhistorie*, 280-296. Oslo: Aschehoug.

Web Sites

All recordings: http://www.tekstlab.uio.no/norskiamerika/english/recordings.html
Fieldwork 2010-2016: http://tekstlab.uio.no/norskiamerika/english/field-work.html
Haugen’s recordings: http://www.tekstlab.uio.no/norskiamerika/english/recordings/haugen.html
Hjelde’s recordings: http://www.tekstlab.uio.no/norskiamerika/english/recordings/hjelde.html
Seip and Selmer’s recordings: http://www.tekstlab.uio.no/norskiamerika/english/recordings/seip-selmer.html
User Manual for CANS: http://tekstlab.uio.no/brukerveiledninger/CANS/index_eng_v3.html

Selected Proceedings of the 10th Workshop on Immigrant Languages in the Americas (WILA 10)

edited by Arnstein Hjelde
and Åshild Søfteland

Cascadilla Proceedings Project   Somerville, MA   2021

Copyright information

Selected Proceedings of the 10th Workshop on Immigrant Languages in the Americas (WILA 10)
© 2021 Cascadilla Proceedings Project, Somerville, MA. All rights reserved

ISBN 978-1-57473-478-2 hardback

A copyright notice for each paper is located at the bottom of the first page of the paper.
Reprints for course packs can be authorized by Cascadilla Proceedings Project.

Ordering information

Orders for the printed edition are handled by Cascadilla Press.
To place an order, go to www.lingref.com or contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, fax: 1-617-776-2271, sales@cascadilla.com

Web access and citation information

This entire proceedings can also be viewed on the web at www.lingref.com. Each paper has a unique document # which can be added to citations to facilitate access. The document # should not replace the full citation.

This paper can be cited as:

Johannessen, Janne Bondi. 2021. From Fieldwork to Speech Corpus: The American Norwegian Heritage Language and CANS. In *Selected Proceedings of the 10th Workshop on Immigrant Languages in the Americas (WILA 10)*, ed. Arnstein Hjelde and Åshild Søfteland, 1-10. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #3535.