CHAPTER 2

A multilingual speech corpus of North-Germanic languages

Janne Bondi Johannessen*, Øystein Alexander Vangsnes**, Joel Priestley* and Kristin Hagen*
*University of Oslo / **University of Tromsø

The Nordic Dialect Corpus project was initiated by the Scandinavian Dialect Syntax Network (ScanDiaSyn). In order to be able to study the North Germanic (i.e., Nordic) dialects, proper documentation of the dialects was needed. A corpus consisting of natural speech by dialect speakers was developed in order to systematically map and study syntactic variation across the Scandinavian dialect continuum. The corpus was to be comprised of transcribed and tagged speech material linked to audio and video recordings. Further, it was decided that a user-friendly interface should be developed for the corpus, and that it should be available on-line. The corpus is now ready for use, and is described here.

1. Introduction

The Nordic Dialect Corpus1 project was initiated by the Scandinavian Dialect Syntax Network (ScanDiaSyn). Documentation of the dialects was required, and the researchers agreed that, in addition to questionnaire-based sampling of syntax data, a corpus of natural, spontaneous speech would be desirable in order to systematically map and study syntactic variation across the Scandinavian dialect continuum. This corpus should comprise transcribed and tagged speech material linked to audio and video recordings. Further, it was decided that a user-friendly interface should be developed for the corpus, and that it should be available on-line. The corpus is now ready for use and described in this paper.

The ScanDiaSyn network is a project umbrella where ten Scandinavian research groups collaborate. The groups are spread across all of the five Nordic countries (Iceland, Denmark, Norway, Sweden, Finland) and one self-governed area (the Faroe Islands). Three non-Nordic groups and a group working on Finnish dialect syntax were liaised with the project through a NordForsk network in the period 2005–2010.

1. <http://www.tekstlab.uio.no/nota/scandiasyn/>
In addition to several national project grants, the ScanDiaSyn project umbrella has also included the Nordic Center of Excellence in Microcomparative Syntax (NORMS) which was generously funded by two Nordic research funding bodies.2

Table 1. The Nordic Dialect Corpus in numbers

<table>
<thead>
<tr>
<th>Country</th>
<th>Total population</th>
<th>Informants</th>
<th>Places</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Denmark</td>
<td>5.5 million</td>
<td>81</td>
<td>15</td>
<td>211,266</td>
</tr>
<tr>
<td>Faroe Is.</td>
<td>50 000</td>
<td>20</td>
<td>5</td>
<td>62,411</td>
</tr>
<tr>
<td>Iceland</td>
<td>322 000</td>
<td>10</td>
<td>2</td>
<td>23,626</td>
</tr>
<tr>
<td>Norway</td>
<td>5 million</td>
<td>564</td>
<td>163</td>
<td>2,186,318</td>
</tr>
<tr>
<td>Sweden</td>
<td>9.5 million</td>
<td>126</td>
<td>39</td>
<td>307,861</td>
</tr>
<tr>
<td>Total</td>
<td>801</td>
<td>224</td>
<td>2,791,482</td>
<td></td>
</tr>
</tbody>
</table>

The corpus is installed in the Glossa corpus system for user-friendly search and results handling (Johannessen et al., 2008; Johannessen, 2012). There were a number of challenges that needed to be addressed, that we shall focus on in this paper. Some of the challenges were due to the fact that some recordings and transcriptions had been done in other projects long before the present project started:

– data collection had to be carried out in several different countries
– the corpus, consisting of different languages, was to be grammatically tagged
– the recordings were to be transcribed, but with different transcription standards and types for the individual languages (the funding for transcription was different for each country, and hence not all languages could be transcribed by both a coarse and a fine-grained transcription, due to the high cost)
– different tags had to refer to the same entities for uniform search possibilities (there are different tags to begin with, since the project used existing, different taggers for each language)
– informant metadata (gender, age, sex, place etc.) were to be used as filters for search
– different geographical divisions were to be specifiable (e.g. country, county, town)
– all text from all languages had to be accessible in the same search
– transcriptions had to be linked to audio and video
– results were to be available in a number of different ways, including different export formats
– informant data were to be plotted on map

In the paper we will discuss these points and explain how we chose to solve them. We will also briefly show how the corpus has already been put to use for linguistic research.

2. The Joint Committee for Nordic Research Councils in the Humanities and Social Sciences (NOS-HS) and the Nordic Research Board (NordForsk).
The project has also used information from grammaticality judgments presented to informants in questionnaires. This information has been collected into a database, the Nordic Syntax Database, and is now available to researchers in addition to the corpus.

2. Challenges in the corpus design and development

2.1 Methodology for collecting speech

The corpus comprises recordings made in the five constituent countries of the Nordic, North Germanic language area. From each country a number of sample points were selected specifically to capture dialect variation. Norway is considered to have a wide variety of dialects, and this country has the largest number of measure points by far. Denmark is considered to have very little dialectal variation left, and the Danish research group therefore decided that a small number of places would be enough. The Faroese and Icelandic measure points are even fewer, but this is due to an unfortunate financing situation rather than deliberate choice.

There is some variation as to the combination of gender, age and number of speakers in the corpus, given that the recordings were mostly done on national research funding and national research management. In Norway, the Norwegian Dialect Syntax Project (NorDiaSyn) was funded by the Norwegian Research Council, a regional bank (Sparebank1 Nord-Norge), and the University of Oslo. This ensured full funding of the recordings in a way that satisfied the criteria given by the researchers. From each point, four informants were identified, two men and two women, old and young. The informants were paired and asked to converse freely for approximately 30 minutes. Care was taken to create comfortable, informal surroundings, in order to encourage spontaneous, unaffected speech. Video equipment was set up, but the informants were left to themselves. Due to privacy legislation, a list of topics deemed off-limits was provided. This included subjects such as trade/labor union and political party membership, as well as the naming of third parties, with the exception of public figures. Each informant also partook in a more formal interview, answering a standard set of questions. The Norwegian part of the corpus has later on been enriched with a number of old recordings from 1950–1980, provided by the dialect archive (Målformearkivet) at the University of Oslo, and with transcriptions funded by the Norwegian Dictionary 2014 project. The main idea of the corpus has been to find examples of the traditional dialects, which caused some limitations as to the selection of possible informants. It was seen as an advantage if the speakers had only very little formal education, since this would mean that they had had less chance of being influenced by other people in a college or university situation away from their home place. There was also a requirement that they should not have lived away for any period of time longer than seven years in total, and it was also desirable that their family should have lived in the same place for generations. Although it was not always easy to find informants who satisfied all the criteria, by and large this part of the data collection has been successful.
The majority of the Swedish recordings (including Finland Swedish) were generously provided for use in the Nordic Dialect Corpus by the SweDia 2000 project. This project was originally aimed at collecting data for phonological research, but the data are by and large fully usable for our corpus, since the recordings contain free speech. Since the SweDia 2000 project had somewhat different goals than the Nordic Dialect Corpus project, some of the data collection was done differently, for instance in that not all the recordings contain a conversation between informants. In retrospect it would have been a good idea to supplement the existing recordings with new ones. But the corpus is not set in stone, and we welcome additions at any time.

The Danish recordings were done by the Danish Syntax Project, funded by the Danish Research Council, and contains six recordings from each place, but with no young people, and sometimes like the Swedia 2000 recordings, with only an assistant talking with an informant rather than two informants talking to each other. Additional recordings in Denmark have been carried out in the Western Jutland area during fieldwork organized by the NORMS project, and these include both children and conversations between dialect speakers.

The Faroese recordings were also sampled during NORMS-funded fieldwork, and these contain both young and old speakers. For Icelandic, the recordings have been less systematic, given a combination of funding and unfortunate chronological synchronization with the rest of the project. Some recordings have been generously provided by the University of Iceland, and some have been done by the ScanDiaSyn network.

In spite of the diverse ways the recordings have been collected, the corpus has become a unique source of spontaneous speech well suited for dialect research in syntax, but also for other linguistic disciplines.

2.2 Transcription and tagging

It was decided early on that all recordings should be transcribed with standard orthography rather than just phonetic transcription. This was vital in order for the corpus to be grammatically tagged by existing software, but also to be easily searchable. It is assumed that all users of the corpus know how their desired search words are written in standard orthography. (And if they do not know that of the other languages, they can look them up in the cross-Scandinavian wordlist Tvärslå, which is provided at the search page.) There is an almost infinite way in which pronunciations can differ across a whole country, so opting only for a phonetic type transcription would make searching virtually impossible.

However, a visualization of the dialectal variation is of course very useful, as it is quicker to be able to find interesting things by reading than by listening through recordings, which is the alternative if there is no phonetic transcription available. Therefore, all the Norwegian recordings and some of the Swedish ones (those of the Övdalian dialect) have been transcribed in a more phonetic way, following, for Norwegian, the method described in Papazian and Helleland (2005) and, for Övdalian, the orthography standardised by the Övdalian language council Rådjärum.
For each language, transcription software was used that inserts time codes directly into the transcribed text at suitable intervals, enabling the transcription later to be presented with its corresponding audio and video. The transcriptions were partly done within each national subproject, and partly in Oslo also for non-Norwegian recordings. Different software were used, but all transcriptions were adapted to the Transcriber format, which is the interchange format used in this project.

For the Norwegian and Swedish recordings that have also been phonetically transcribed, the process started with the phonetic transcription. The transcriptions were then translated to standard orthography using a program developed at the Text Laboratory, University of Oslo: an automatic dialect transliterator. The program takes as input a phonetically transcribed text and an optional dialect setting. Sets of text manually transliterated to orthography provide a good basis for training the program, enabling it to accurately guess the transliteration in subsequent bodies of text. The training process can be repeated, and the trained version can be used for similar dialects. Transcribing each recording twice, phonetically and orthographically, therefore does not take as much as twice the time of transcribing only once.

It is important that all words from the original phonetic transcription have an equivalent in the orthographic transcription. The two must be totally aligned for the results to be used in the corpus search system. Figures 1–3 show how the phonetic transcription can be used in search and results presentation.

**Figure 1.** Searching for two words in sequence. The first is transcribed phonetically: *itte* for the orthographic word *ikke* ‘not’

**Figure 2.** The Both button is ticked, in order to have both kinds of transcription presented in the search results

**Figure 3.** Part of the search result for the query in Figures 1 and 2
The languages are tagged individually with taggers for the respective languages. This means that each language has an individual tag-set decided by those who developed the taggers originally, in most instances long before the Nordic Dialect Corpus was planned.

The Danish transcriptions are lemmatized and POS tagged by a Danish Constraint Grammar Tagger developed for written Danish, see Bick (2003).

The Faroese transcriptions were first tagged with a Constraint Grammar Tagger for written Faroese, see Trosterud (2009). Since spoken Faroese has a lot of words that are not approved in written standard Faroese, about half of the material was manually corrected after the Constraint Grammar tagging. Finally a TreeTagger was trained on the corrected material, and the rest of the transcriptions were tagged again.

The Icelandic transcriptions were first tagged with a tagger for written Icelandic, see Loftsson (2008), and manually corrected afterwards.

The orthographic version of the Norwegian corpus was lemmatized and POS tagged by a TreeTagger originally developed for Oslo speech. The Oslo speech tagger was trained on manually corrected output from the the written language Oslo-Bergen tagger, see Nøklestad & Søfteland (2008). The tagger was then further adapted to the dialect corpus.

The Swedish subcorpus was tagged by a modified version of the TnT tagger developed by Kokkinakis (2003). The tagger was trained on the Swedish PAROLE corpus before it was applied to the corpus material. The Övdalian material was also tagged with this tagger. Recall that Övdalian was transcribed both by the official Övdalian orthography and by a word by word translation into Swedish. The tagger was applied to the latter version.

Each language subcorpus has its own tag-set, but the tags have been standardized in the search system, making it possible to search for the same category across all the corpora. The linguist can choose for example all adjectives to be shown, irrespective of language. This is illustrated in Figure 4.

Figure 4. Querying for adjectives across languages in the corpus
Chapter 2. A multilingual speech corpus of North-Germanic languages

2.3 Metadata

The corpus has metadata relating to each informant and recording. There is information on the sex, age group, and place of origin, where the latter is divided into country, region, area, and place. Also, there is information on the year of recording, which is crucial for the Norwegian subcorpus, which contains both modern and old recordings, with 30–60 years between them. Finally, some recordings are distinguished according to genre: either interview or conversation.

The metadata can be used to create search filters for search in the corpus interface, as depicted in Figure 5.

![Figure 5. Metadata filter in the corpus interface](image)

The metadata is simply represented in a MySQL database, from which the corpus interface system Glossa picks the correct data according to the user’s needs.

Informant metadata can alternatively be found by clicking on the blue i-button (i for information) on the left of each concordance line in the results view, as in Figure 3, yielding the information displayed in Figure 6.

2.4 Multilingual search

Users in the ScanDiaSyn network originally wanted the possibility for multilingual search. They imagined that if they wanted, say, all occurrences of the negation equivalent to ‘not’ in English, a full results list would appear for all languages. However, this would have required a full multilingual dictionary, which does not exist either in paper or digital format for the North Germanic languages.

Instead, we put a link on the search interface to a multilingual word-list (Tvärslå) compiled by several previous language technology projects, including ScanLex, in which two of the authors of the present paper (Johannessen and Vangsnes) were also in charge. This way the user can look up the equivalents of particular words in the other languages. The multilingual list is far from comprehensive, and also contains some faulty language equivalents, since it is partly developed using automatic methods.

The search system Glossa allows for disjunctive searches, making it possible for several strings to be looked up at the same time. This is illustrated in Figure 7, for the orthographic versions of ‘not’ for Faroese, **ikki**, Swedish, **inte**, Danish and Norwegian, **ikke**, and Icelandic, **ekki**.
Figure 6. Metadata on each informant is available via a clickable button

Figure 7. Disjunctive search for the word for ‘not’ in several languages
2.5 Links to audio and video

Using a corpus for searching for particular words, categories or constructions is good, but for a speech corpus it is crucial that the sound and video are also available as part of the search results. In the Nordic Dialect Corpus the user can click on the film or sound symbol to get the desired multimedia display. Figure 8 depicts the display.

![Figure 8. Results with selected video presentation](image)

The transcriptions have time codes, implemented as XML tags, at regular intervals, inserted at the time of transcription. This way there is a direct link between text and audio and video files, to be used by the corpus search system. These files are made available in Flash and Quicktime (the user can choose which).

2.6 Results presented on maps

For a corpus aimed at dialect research, getting results on a map view is very useful. Each geographic measure point is located by GIS coordinates and the Google Maps API is used. Since every item in the corpus is connected to an informant, it means that for each word, string, piece of word or syntactic construction, there is a geographical location.

We have incorporated two ways of displaying results via maps. One way is that all hits are simply marked on the map. Figure 9 shows a search that asks for all hits where, in a subordinate clause, the negation *ikke* or *inte* (Norwegian, Danish, Swedish) precedes the subject. The geographical distribution is shown in Figure 10.

There is a debate in the literature as to whether this word order is allowed or not (see Section 3 below, and Johannessen & Garbacz, 2011). The red dots on the map in Figure 10 show where the hits are. Even if there are more recording places in Norway
than in Sweden and Denmark, cf. Table 1, we see immediately that there are many more places where this construction is found in Norway than in Sweden. Since stress patterns also interfere with the generalisations, it is necessary for the researcher to listen to selected results, but the first picture given by the map is a very useful start.

The other way to use maps is only possible for those search results that belong to a set of two transcriptions. All the phonetic varieties are presented on a chart with the option of coloring each according to a classification of one's own choice.
The chart in Figure 11 shows all the phonetic versions of the pronoun *vi* ‘we’ in Norwegian. We have chosen to color those variants that are pronounced with an initial bilabial /m/ sound with a dark color, while the initial /v/ sounds are colored light. For this illustrative example we ignore the other pronunciations. The result is shown in Figure 12.

![Figure 11. Chart for coloring in the phonetic variants of the pronoun vi ‘we’ in Norwegian](image)

It should be quite clear from the map example that the opportunity of using a corpus combined with maps is an excellent way of finding isoglosses. The geographical limits for a phenomenon are readily apparent on the map. It should be noted in this respect that dialect maps are not a new thing. However, in the past, researchers rarely had the chance to cover many places, so the present corpus may contain data that has never been known before. Secondly, the old maps were rarely the result of spontaneous speech, but rather of words and lists given by the researcher to the informants. The present solution, with a corpus of spontaneous speech as a direct basis for maps, gives good opportunities for both a comprehensive and a correct view of the geographical language variation, and many new isoglosses.
Figure 12. Map of two phonetic variants of the pronoun *vi* ‘we’ in Norwegian: /m/ variants are colored dark, while /v/ variants are colored light

3. Results from research on the Nordic Dialect Corpus

Here we will present some examples of linguistic research that has used the Nordic Dialect Corpus already.

Figures 9 and 10 in Section 2.6 showed the search for clauses in which negation follows a subjunction but precedes a subject. The Norwegian Reference Grammar (Faarlund et al., 1997) says that this word order is only possible if the subject is heavy, i.e. carries some stress. Otherwise the negation should follow the subject.

Johannessen and Garbacz (2011) studied the search results and listened to all the video recordings for each hit to determine whether the pronoun was stressed or unstressed. It turned out that a substantial number of the results had an unstressed subject. Studying the data more carefully, taking full advantage of all the linguistic and metadata information available, the authors found that the results were consistent for many different kinds of subjunctions (the equivalents of ‘since’, ‘even if’, ‘if’, ‘that’ etc.). The results were spread across the whole of Norway and across a long time period, with several hits from the old language archives as well as the newest recordings. The authors of the Norwegian Reference Grammar did not have access to a speech corpus at the time it was written, and it is obvious that claims they made on this topic were not based on empirical facts.

Øystein A. Vangsnes, one of the authors of the present papers, has already used the corpus in investigations of a variety of issues such as exclamatives, wh-questions

We expect much more research to be carried out using the Nordic Dialect Corpus in the years to come, as it becomes more well-known and researchers discover what a great source of data this is, whether one's interests are in phonology, morphology, syntax, semantics or variation across time and space.

4. Conclusion

We have presented the Nordic Dialect Corpus, a dialect corpus consisting of transcribed speech from five countries, containing nearly 3 million words. We have shown how challenges posed by researchers in this project initiated by linguists have been met and solved. The corpus user interface provides access to audio and video, as well as transcriptions – many of which are both phonetic and orthographic. All transcriptions are tagged. Everything is accessible in the Glossa search system, with monolingual or multilingual search options, specified linguistically with possible filtering via metadata.

We have shown how the map options work, and how combining a corpus with a map solution provides advanced possibilities for identifying and representing iso-glosses in a simple way. Finally we have described some recent linguistic research that could hardly have been done without the corpus, and which, in one case, showed that previous claims in a renowned work of the Norwegian language, about the distribution of a particular grammatical phenomenon, were not in accordance with the facts as they appear in the corpus. It is to be hoped that Nordic Dialect Corpus will be used as a major source of data for linguists working on the Nordic languages in the years to come.

Acknowledgements

We are grateful to all the people who have taken part in the corpus data collection and in the many discussions about methods and design leading up to that. We are also grateful to our old and new, permanent and temporary, colleagues at the Text Laboratory, UiO, who have helped at various points in the process, from transcription via transliteration of transcriptions to tagging. This work has been funded by Nordic and national research councils in the Nordic countries, and by universities and smaller research funds including a bank.
References


Chapter 2. A multilingual speech corpus of North-Germanic languages

Links

DanDiaSyn: <www.hum.au.dk/dandiasyn/>
Glossa: <www.hf.uio.no/tekstlab/glossa.html>
Nordic Dialect Corpus: <www.tekstlab.uio.no/nota/scandiasyn/index.html>
NorDiaSyn: www.tekstlab.uio.no/nota/NorDiaSyn/index.html
Norsk Ordbok 2014: http://no2014.uio.no/perl/ordbok/no2014_resultatliste.cgi
SWEDIA 2000: http://swedia.ling.gu.se/
Text Laboratory: <http://www.hf.uio.no/tekstlab/>
Tvärslå: <http://ordbok.nada.kth.se:8070/ordbog_module/ordbog>