The Nordic Dialect Corpus

Janne Bondi Johannessen*, Joel Priestley*, Kristin Hagen*, Anders Nøklestad* and André Lynum*+

* Text Lab, ILN, University of Oslo,Norway,

+ IDI, Norwegian University of Science and Technology – NTNU

E-mail: jannebj@iln.uio.no, joeljp@ iln.uio.no, kristiha@ iln.uio.no, noklestad@ iln.uio.no, andrely@idi.ntnu.no

Abstract

In this paper, we describe the Nordic Dialect Corpus, which has recently been completed. The corpus has a variety of features that combined makes it an advanced tool for language researchers. These features include: Linguistic contents (dialects from five closely related languages), annotation (tagging and two types of transcription), search interface (advanced possibilities for combining a large array of search criteria and results presentation in an intuitive and simple interface), many search variables (linguistics-based, informant-based, time-based), multimedia display (linking of sound and video to transcriptions), display of results in maps, display of informant details (number of words and other information on informants), advanced results handling (concordances, collocations, counts and statistics shown in a variety of graphical modes, plus further processing). Finally, and importantly, the corpus is freely available for research on the web. We give examples of both various kinds of searches, of displays of results and of results handling.
Keywords: Scandinavian languages, speech, maps, audio, video

1 Introduction

The Nordic Dialect Corpus is a speech corpus with natural speech from informants from five Nordic countries. The main purpose of the corpus is to facilitate the study of the dialects of the North Germanic languages, i.e., the Nordic languages spoken in the Nordic countries. The languages are closely related to each other, and three of them are mutually intelligible (Norwegian, Swedish and Danish), as are two others (Faroese and Icelandic). All of them have some mutual intelligibility with each other if we consider written forms.

In this paper we will present Nordic Dialect Corpus: the linguistic contents, annotation, the search interface, the display of results and informant information and how to further process the results. In the last chapter we will compare the Nordic Dialect Corpus with some other dialect corpora on the web.

2 Description of the corpus

2.1 Linguistic contents and numbers

The corpus contains dialect data from the national languages Danish, Faroese, Icelandic, Norwegian, and Swedish. Some numbers for the corpus are given in Table 1.

	Informants	Places	Words
Denmark	81	15	211,266
Faraoe Islands	20	5	62,411
Iceland	10	2	23,626
Norway	557	162	2,137,368
Sweden	126	39	307,861
Total	794	223	2,742,532

Table 1. Corpus contents.

Due to differences in the financing of the data collection in the different countries, the data are less uniform than one might have wanted ideally. (Some recordings and transcriptions were done for this corpus, while others were already done, such as most of the Swedish ones, which were generously given us by the earlier project Swedia 2000, see reference list.)

Some dialects have recordings of both young and senior informants, while others are only represented by senior ones. Some dialects are represented by both old and new recordings, where old ones are generally around fifty years old. Some dialects have been recorded by audio only, while others have been recorded by both audio and video. All the dialects have recordings of informants belonging to both genders. Most importantly, however, all the recordings represent spontaneous speech.

2.2 Annotation: transcription and tagging

Each dialect has been transcribed by the standard official orthography of that country. In addition, all the Norwegian dialects and some Swedish ones have also been transcribed phonetically. For these dialects, the phonetic transcription was translated to an orthographic transcription via a semi-automatic dialect transliterator developed for the project. The whole corpus is grammatically tagged with word class and selected morpho-syntactic features language by language.

2.3 Search interface

The corpus uses an advanced search interface and results handling system, Glossa (Nygaard, 2007; Johannessen et al, 2008). The system allows for a large variety of search combinations making it possible to do very advanced and complex searches, even though the interface is very simple, with pull-down menus, and boxes that expand only when prompted by the user. The corpus search system Corpus Work Bench (Christ, 1994; Evert, 2005) is used, so that the simple corpus queries are translated to regular expressions before querying – something that is invisible to the user.

We chose to use Glossa because it is a complete system

allowing both advanced searches and result display, including multimedia viewing and map presentation, together with advanced result handling with possibilities for deleting, annotating and saving of results. To our knowledge, no other system can offer all these features.

Searching for lemmas and part of words: In addition to searching for word forms, it is also possible to search for lemmas. This way we get all inflected forms of one lexeme. This feature is very useful when there is suppletion in the stem of the word. For example, a search for the Norwegian lemma *gås* ('goose') will give the results *gås*, *gåsa*, *gjess*, *gjessene* (various combinations of number and definiteness). The same box where the user can write a full search word or a lemma can also be used to write part of a search word. This way the user can, for example, search for a particular suffix.

Searching for more than one word: An arbitrary number of search boxes may be opened in order to search for more than one word, with the possibility of specifying a number of words in between (Figure 1).

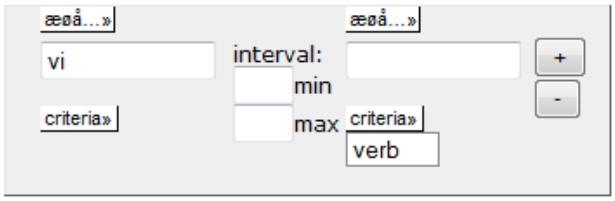

Figure 1: Searching for two words.

Searching for part of speech: The corpus can also be queried directly by part-of-speech tags that include word class and morphosyntactic features.

Phonetic querying: The user can choose to query the corpus by giving a phonetically specified string. This works only for the dialects that have two transcriptions (cf. section 2.2). An example of a situation in which this is useful will be where we want to query person-number inflection on verbs. Here, tagging will not help, since each tagger is trained on the standard orthographic version of the texts, and person-number inflection is only a dialect feature. Searching for this feature in Övdalian, we can simply write for example the 1pl suffix as it is (Figure 2):

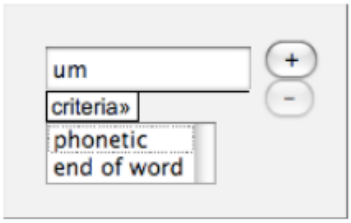

Figure 2: Searching in phonetic mode.

This will give results that would have been impossible to get from the orthographic transcriptions.

Informant-based querying: There are a number of ways to query the corpus in addition to the linguistics-based ones that we have seen above. All the details that are known about each informant are also searchable in the search interface. Thus, it is possible to specify as search criteria: age, sex, recording year, place of residence, country, region and area.

2.4 Display of search results

Each search in the corpus gives a standardised view of the results in the form of a classical KWIC concordance. The results can be viewed in a number of additional ways which we will present below.

Multimedia display: The search result is accompanied by a clickable symbol to show the audio and video of that particular speech sequence. This is illustrated in Figure 3 below.

Display of transcriptions and tagging: For those linguistic variants that have two transcriptions, either transcription can be chosen for displaying the result. The grammatical tags and the phonetic transcription of each standard orthographic word are visible in a box when mousing over the text.

Translations: We have tested the possibility of translating search results automatically into English using Google Translate. While the quality of the translation may vary, especially considering the fact that we are translating spoken language, the translations may at least give non-native speakers an idea of what the utterances are about. Unfortunately, this has become a commercial service, and we are currently reviewing what to do.

Result actions: On the results page there is a menu with a selection of choices for further displaying of results and results handling (the latter of which will be presented in section 2.6). The functionalities that follow in this subsection are choices in this menu.

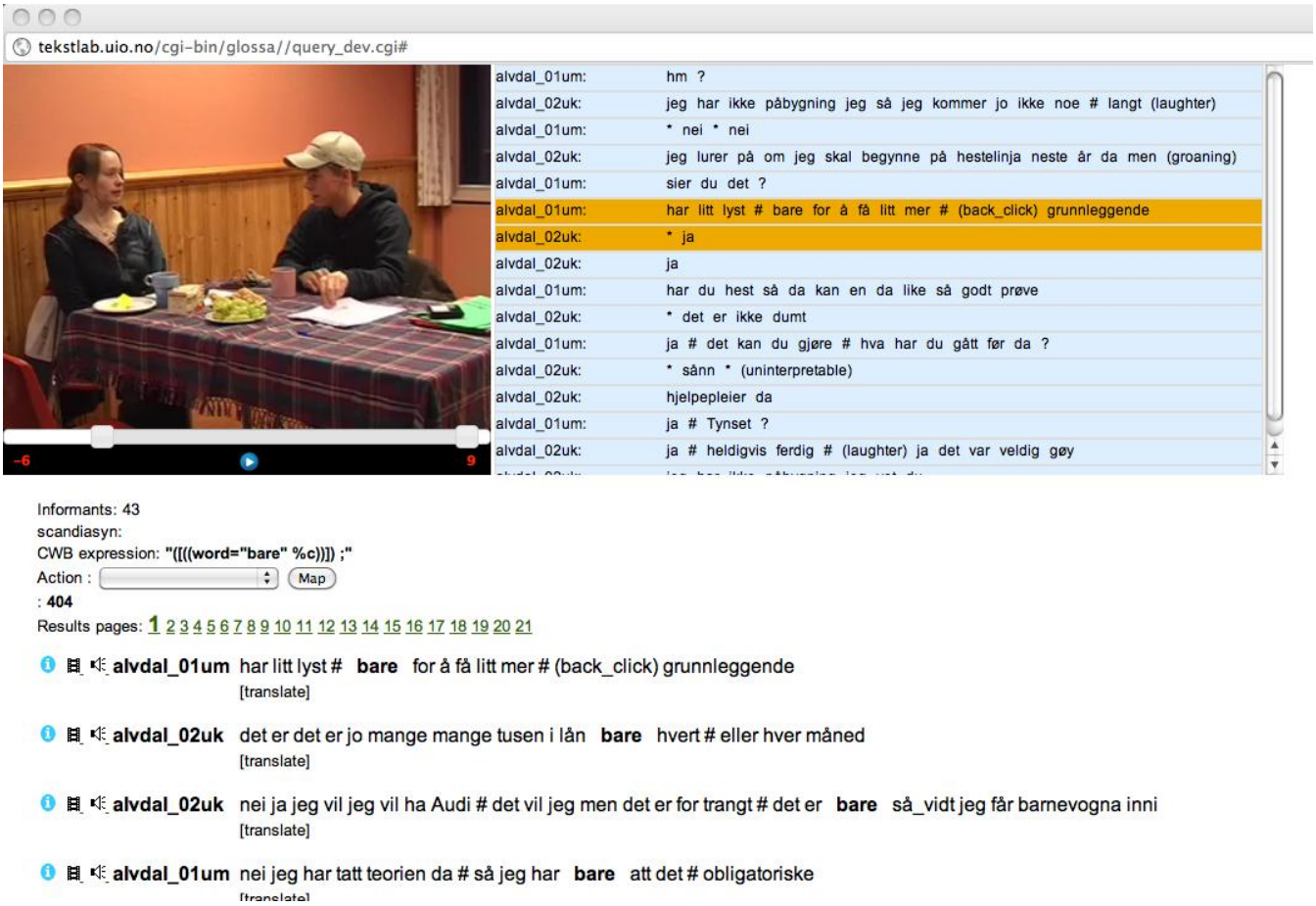

Figure 3. The multimedia result window.

Count: Choosing the Count option gives the search results as a list of all the hits sorted by frequency. In Figure 4, a bit of a list is shown as a result of the search for nouns starting with *bil-* ('car') in Norwegian.

occurences	match
40	bil
20	bilen
14	biler
11	bilde
7	bilene
4	bildet
1	bilkjøringa
1	bilbasert
1	bilder
1	bilveg
1	bildeler

Figure 4. Some nouns beginning with bil- ('car').

The count results can be shown in a number of ways, such as histograms and pie charts:

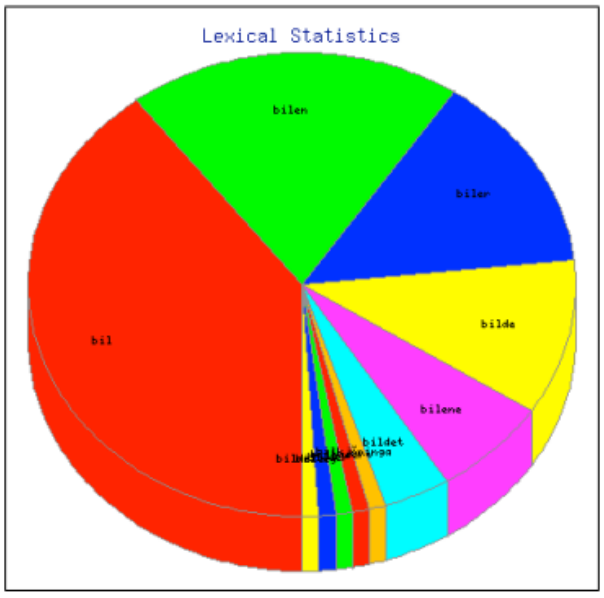

Figure 5. The same information as in Figure 4.

Sort: The results can be sorted in many ways; the most useful ones are perhaps those that sort the matches by the next word to the right or left.

Collocations: The results can be shown as collocations

according to many different statistical measurements such as dice coeffiency, log-likelihood ratio etc., with a choice between neighboring bigrams and trigrams.

Maps: Recently an option to display the search results on maps (using Google Maps technology) has been added. Since one search can cover a variety of results, for example when one orthographic word covers many different phonetic varieties, an additional option has been added in which each variety can be selected independently. In the map in Figure 6 the different phonetic varieties of the pronoun *hun* ('she') are displayed in the right-hand column, giving the user the option to choose one or more and have them independently shown on the map. The orthographic variety has been displayed by a neutral dot covering all pronunciations.

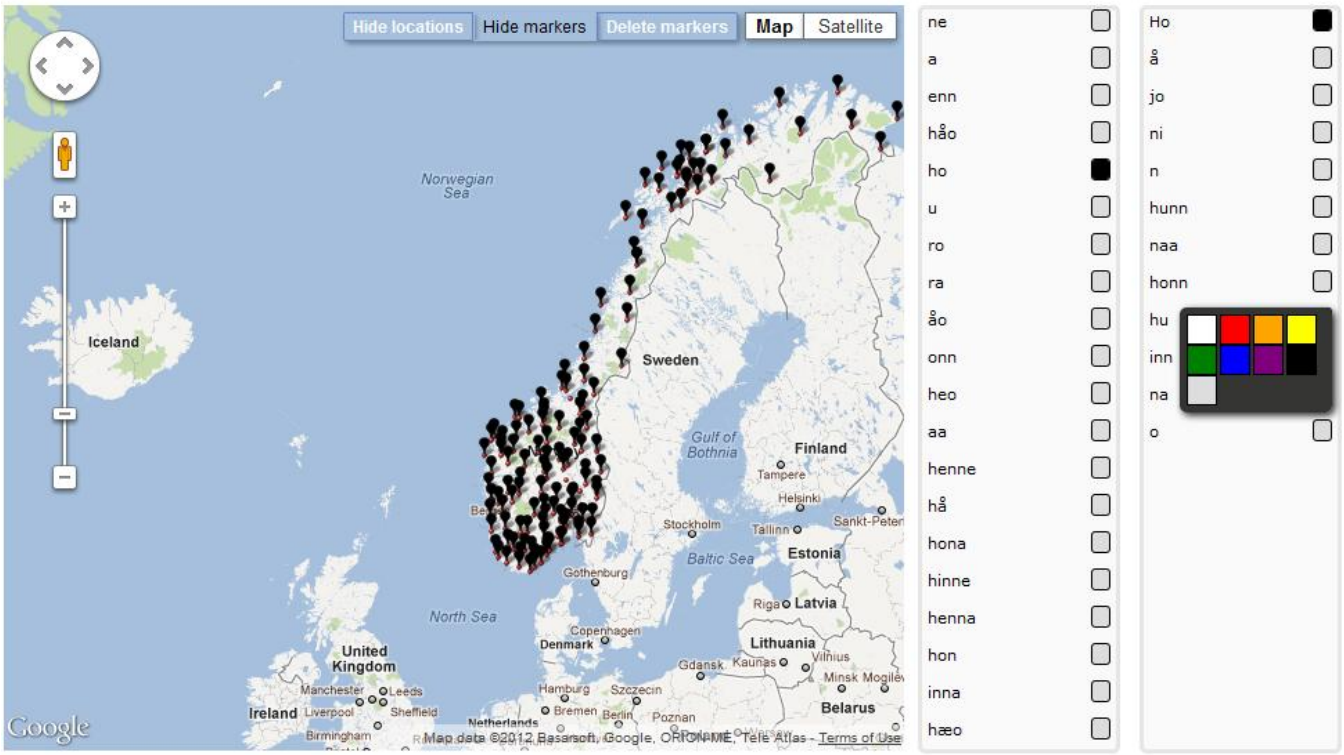

Figure 6. A map showing all the places that have hits (all the black dots) for the phonetic form *ho* of the pronoun *hun* ('she').

2.5 Displaying information on informants

There are two ways of finding information on the informants.

Via results page: Each concordance line has an information symbol on its very left. Clicking on this symbol reveals information on the informant in question: informant code, sex, age group, country, place, number of words, recording year, and a map for his/her home place.

Via search page: There is a button called "Show informants", which shows information on which informants are included in a particular query. For example, if the user wants to query the corpus on Swedish data only, (s)he can press this button and immediately see how many informants are represented in the selection, how many words each informant has uttered etc., and this information can also be sorted by category to present for example number of words in descending order. This way, we can see how different the informants are in this respect.

2.6 Further processing of results

Deleting or choosing some results: In a corpus search it is often the case that the user gets more results than intended. Sometimes the search expression just was not good enough, which can best be corrected by a new and more precise search.

However, sometimes it is impossible to formulate better search criteria, whether it is because there is too much homonymy in the corpus, or because it just is not annotated for all imaginable research features. In these cases, the user can delete individual hits, or alternatively choose to keep only selected hits, and then save the modified search results.

Annotating results: The individual researcher often needs to further annotate the results, for example according to pronunciation of certain sounds or words, or specific syntactic patterns. The annotations can be edited and saved as annotation sets, for later reuse with other results.

Saving and downloading results: All results can be saved and/or downloaded, whether we choose the raw results or those that we have further processed by deletion, choice or annotation. By saving we get the opportunity to look at the results later, and with exactly the same possibilities for further processing and displaying of results in the corpus interface. Downloaded results, on the other hand, are not thus available in the corpus system, but can be stored as for instance tab-separated text and imported into other applications.

3 Other dialect corpora on the web

To our knowledge, there are not many dialect corpora on the web. The perhaps most well known, *The British National Corpus*, contains 10 million words of spoken English, categorised into 28 different dialects. Unfortunately, the speech is only transcribed orthographically and so far there are no audio available.

The Scottish Corpus of Text and Speech contains 800,000 spoken words, orthographically transcribed with links to audio or video. The search interface is nice with a possibility to search through maps. The corpus is not grammatically annotated and there is no result handling.

Another interesting resource is *Sounds familiar? Accents and Dialects of the UK*. It contains information on British dialects, and recordings of the dialects with transcripts, all presented via a web map. However, the aim is pedagogical, and the web site is not developed for searching the transcriptions.

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, and has little spontaneous speech.

Finally we will mention a Norwegian speech corpus from the University of Bergen (*Norsk dialektkorpus*). It contains audio and transcriptions from different projects, and is planned to be revitalized in a new search interface.

For a more thorough listing of dialect corpora on the web, see Johannessen et al (2009).

4 Conclusions

The Nordic Dialect Corpus is a valuable resource for researchers working on Nordic languages and dialects. While there are some other dialect resources on the web, to our knowledge there are few available web-based dialect multimedia corpora for other languages. Among those that do exist, none seem to offer the full combination of features offered by the Nordic Dialect Corpus including grammatical annotation, searches in both orthographical and phonetic transcriptions, display of audio and video linked to the search results, map displays for result distributions, and extensive results management options.

5 References

Christ, Oliver (1994). A Modular and Flexible Architecture for an Integrated Corpus Query System. In *COMPLEX'94*. Budapest.

Evert, Stefan (2005). *The CQP Query Language Tutorial*. Institute for Natural Language Processing, University of Stuttgart. http://www.ims.unistutgart.de/projekte/CorpusWorkbench/CQPTutorial.

Johannessen, Janne Bondi; Nygaard, Lars; Priestley, Joel and Nøklestad, Anders (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).

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

Nygaard, Lars. (2007). *The glossa manual*. The Text Laboratory. http://www.hf.uio.no/tekstlab/glossa.html.

Swedia 2000. http://swedia.ling.gu.se/