

# THE NORDIC DIALECT CORPUS – A JOINT RESEARCH INFRASTRUCTURE

JANNE BONDI JOHANNESSEN

*Department of Linguistics and Nordic Studies, University of Oslo*

## ABSTRACT

The paper describes the Nordic Dialect Corpus as of June 2010. The corpus (see Johannessen et al. 2009) is steadily growing, and new features are constantly added, so the version we describe is that of June 2010, while the corpus work has funding for another two years. The corpus is a tool that combines a number of useful features that together makes it a unique and very advanced resource for researchers of many fields of language studies. The corpus is web-based and features full audio-visual representation linked to transcriptions and translations.

## [1] INTRODUCTION

In this paper, we describe the Nordic Dialect Corpus<sup>1</sup>. The corpus (see Johannessen et al. 2009) is steadily growing, and new features are constantly added, so the version we describe is that of June 2010, while the corpus work has funding for another two years. 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.

## [2] WHY THE NORDIC DIALECT CORPUS WAS DEVELOPED

The Nordic Dialect Corpus was developed after a need for research material was voiced by members of the NORMS (Nordic Centre of Excellence in Micro-comparative Syntax) and the ScanDiaSyn (Scandinavian Dialect Syntax) networks.

The overarching goal for these researchers is to study the dialects of the North-Germanic languages, i.e., the Nordic languages spoken in the Nordic countries, as dialects of the same language. 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.

Studying the dialects only within the confines of each national language was therefore considered to be misguided from a theoretical and principled point of view. Second, doing research across dialects over such a big area, covering six countries (Denmark, Faroe Islands, Finland, Iceland, Norway, and Sweden), would be almost impossible if each researcher should get hold of relevant data on their own.

Third, the research in NORMS and ScanDiaSyn focusses on syntax – in which case data of many different kinds were necessary. Questionnaires for specific phenomena were needed (but will not be discussed in this paper), and recordings of spontaneous speech as it is used in ordinary conversations were very important. The latter need is satisfied by the Nordic Dialect Corpus.

## [3] DESCRIPTION OF THE CORPUS

### [3.1] *Linguistic contents and numbers*

The corpus contains dialect data from the national languages Danish, Faroese, Icelandic, Norwegian, and Swedish. It is steadily growing, since new recordings are still being done, or planned, while other recordings are in various stages of finishing. At the moment, it contains speech data from approximately 525 informants with 1.8 million words, unevenly spread between the five countries. Eventually, this will rise to around 600 informants. The numbers for the corpus as of today are given in Table 1.

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 record-

---

[1] The Nordic Dialect Corpus is the result of close collaboration between the partners in the research networks Scandinavian Dialect Syntax and Nordic Centre of Excellence in Microcomparative Syntax. The researchers in the network have contributed in everything from decisions to actual work ranging from methodology to recordings, transcription, and annotation. Some of the corpus (in particular, recordings of informants) has been financed by the national research councils in the individual countries, while the technical development has been financed by the University of Oslo and the Norwegian Research Council, plus the Nordic research funds NOS-HS and NordForsk.

|                | Informants | Places | Words     |
|----------------|------------|--------|-----------|
| Denmark        | 75         | 14     | 229 909   |
| Faraoe Islands | 19         | 5      | 48 427    |
| Iceland        | 4          | 1      | 10 287    |
| Norway         | 301        | 94     | 1 200 120 |
| Sweden         | 126        | 40     | 299 86    |
| Total          | 525        | 154    | 1 788 609 |

TABLE 1: Corpus contents by June 2010

ings 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.)

Some recordings, such as those from Norway, the Swedish dialect of Övdalian and the Danish dialect of Western Jutlandic, have two kinds of recordings per informant: one semi-formal interview (informant and project assistant), and one informal conversation between two informants. Some dialects have recordings of both young and old informants, while others are only represented by old 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.

### [3.2] *Annotation: transcription and tagging*

All the dialect data have been transcribed by at least one transcription standard, and this work has been done for the most part in the individual countries: Each dialect has been transcribed by the standard official orthography of that country. (For Norwegian, which has two standard orthographies, Bokmål was chosen since there exist important computational tools for this variant.) In addition, all the Norwegian dialects and some Swedish ones have also been transcribed phonetically<sup>2</sup>. For the Norwegian dialects and the Övdalian Swedish ones that have two transcriptions, the first transcription to be done was in each case the phonetic one, and then the phonetic transcription was translated to an orthographic transcription via a semi-automatic dialect transliterator developed for the project. The fact that there are two transcriptions for dialects that are very different from the standard national orthography makes it possible to search with both transcriptions in the corpus, and present search results in both, as illustrated below for the Swedish dialect of Övdalian in Figure 1. This figure also shows the translation by Google, which is provided as a service in the corpus results presentation.

icke behöver vi (uförståelig) någon lista # om de- det **kan** vi ju göra om kommer på (uförståelig)  
 itjä byövum wi:ð kommentar nån lista # um e- eð **bellum** wi:ð fel djærå um kumum å: kommentar  
 we need not (uförståelig) # no list of **it-surely** we **can** come on in (uförståelig) (google)

FIGURE 1: Two transcriptions for Övdalian and a Google translation.

The Text Laboratory at the University of Oslo has the responsibility for the further technical development, including tagging. The whole corpus will be grammatically tagged with POS and selected morpho-syntactic features language by language. So far, the Norwegian data have been tagged, while the transcribed texts from the other languages are in the process of being tagged now. Tagging speech data is different from tagging written data. Speech contains disfluencies, interruptions and repetitions, and there are rarely clear clause boundaries (Allwood, Nivre and Ahlsén 1989, Johannessen and Jørgensen 2006). This is usually reflected in the transcription of speech, which generally does not contain clause boundaries or sentential markers such as full stops and exclamation marks (Jørgensen 2008, Rosén 2008). Any tagger developed for written language will therefore be difficult to use directly for spoken language. (Though Nivre and Grönqvist 2001 did this, on a material different from ours).

The Norwegian speech tagger was developed for the NoTa Corpus (Norwegian speech corpus – Oslo part). Søfteland and Nøklestad (2008) describe how the corpus was first tagged with the Oslo-Bergen tagger for written Norwegian (Hagen et al. 2000), and then trained with a TreeTagger (Schmid 1994) on the resulting, manually corrected file. The TreeTagger gained an accuracy of 96.9%. This tagger has then been used unchanged for the dialect corpus, under the assumption that the speech as represented in the dialects and in Oslo are sufficiently similar once they are all transcribed by the same transcription standard. The Swedish tagger has been trained in the same way. A written language TnT tagger developed by Sofie Johansson Kokkinakis (2003) has been applied to the Swedish dialect transcriptions (their standard orthographic version). After having been manually corrected and retrained, a spoken language Swedish statistical HunPos tagger has been developed at the Text Laboratory<sup>3</sup>. For Faroese, we have used a Faroese constraint grammar tagger developed for written language (Trosterud 2009), and manually corrected the results<sup>4</sup>.

- 
- [2] The Norwegian phonetic transcription follows that of Papazian and Helleland (2005). The transcription of the Övdalian dialect follows the Övdalian orthography standardised in 2005 by the *Rådjärum* (The Övdalian Language Council).
- [3] The manual corrections of the Swedish tagger were done by Piotr Garbacz, and the tagger was developed by André Lynum, both at the Text Laboratory, UiO.
- [4] The manual corrections of the Faroese tagger were done by Remco Knooihuizen for the Text Laboratory, UiO.

### [3.3] 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.

Several of the features in the search interface and the results display follow suggestions by participants in ScanDiaSyn and NORMS.

**Searching for lemmas and part of words:** For those parts of the corpus that are tagged and lemmatised, it is possible to search for the lemma only. 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, 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. In Figure 2, the user has searched for the suffix *-ig*, which can be found in Norwegian, Swedish, and Danish.

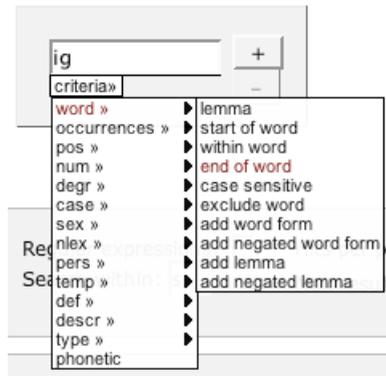


FIGURE 2: Search for suffix *-ig*

Notice that since nothing else was specified, this search would query the whole corpus, all the languages. In Table 2 we can see some of the many hits for the frequent adjectival suffixes *-ig* and *-lig* in the mainland Nordic languages, and a couple of occurrences of words containing the same sequence of letters in the insular Nordic languages (not representing these suffixes, however).

| Freq. | Word         | Translation  | Language   |
|-------|--------------|--------------|------------|
| 7     | særlig       | especially   | No, Da     |
| 7     | farlig       | dangerous    | No, Sw, Da |
| 7     | þannig       | thus         | Ice        |
| 7     | kjedelig     | boring       | No         |
| 6     | vældig       | very         | Sw         |
| 5     | rigtig       | right        | Da         |
| 5     | otrolig      | unbelievable | Sw         |
| 4     | konstig      | strange      | Sw         |
| 1     | sjómannaslig | sailor-like  | Fa         |

TABLE 2: Some results from the *-ig* search

**Searching for more than one word:** In order to specify a search for more than one word, the user clicks on the plus sign in the first box, which gives one more box, with the possibility of specifying a number of words in between (Figure 3).

The screenshot shows a search interface with two main search boxes. The first box contains the text 'ig' and has a pull-down menu with 'criteria»' and a radio button labeled 'end of word'. To its right is an 'interval:' section with a text input field containing '3', a 'min' label, and a 'max' label. Further right is a second search box containing 'konj' with a pull-down menu showing 'criteria»'. To the right of the second search box are two buttons: a '+' button above a '-' button.

FIGURE 3: Searching for two words

The illustration shows a search for a word ending in *-ig* separated by at most three words from a conjunction to the right.

**Searching for part of speech:** The tagged part of the corpus can also be queried directly by part-of-speech tags. This is exemplified in Figure 3, where the second word is specified to be a conjunction. The user can choose whether a search word is specified by a word form (or part of one) and a part of speech or both. The pull-down menus in Figure 2 exemplify many of the search options that are available for a word.

**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 4.2). 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 4):

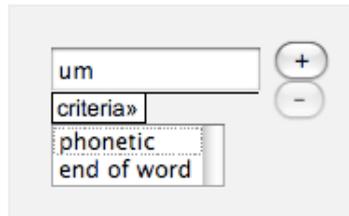


FIGURE 4: Searching in phonetic mode

This will give results that would have been impossible to get from the orthographic transcriptions. We refer to Figure 1, where the dialectal *bellum* ('can' 1pl) is represented by the standard *kan* ('can').

**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. In Figure 5, we show how we can choose individual places from the complete list, to be able to query only the informants from these places, which happen to be the area of Älvdalen in Sweden.

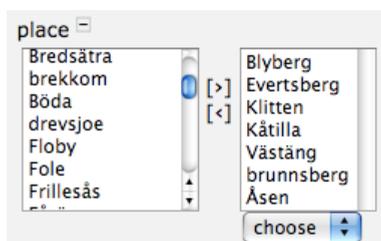


FIGURE 5: Delimiting the corpus by choosing some places from the full list

### [3.4] 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 corpus includes transcribed speech from five countries and spans four decades. Some of the speech was recorded using a tape recorder and later mp3 recorder, and some was recorded by videocamera. 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 6 below.

The screenshot shows a search results window with a list of results on the left and a video player on the right. The video player shows two women sitting at a table, with a 'context' control panel on the right side of the video frame. The control panel includes a search bar, a dropdown menu, and buttons for 'Start', 'Stop', and '>>'. Below the video player, there is a section for 'Informants: 21', 'scandiasyn:', 'CWB expression: (((phon="øðum" %c))) ;', and 'Action:'. Below this, there is a section for 'Results pages: 1' and a list of results. Each result includes a speaker ID, a transcription, and a phonetic transcription. The results are as follows:

- 1 aasen\_48 ja # för at då: vår femtisjæks se: so fyoddes ju pojken og **djifteðum** uoss æreð firiað ja # ja
- 1 aasen\_48 ojojoj og ur fin kommentar tænki at i går # so **prateðum** M1 og ig just um... # för då: addum wi:ð bröllopsdag *bröllopsdag* i går
- 1 aasen\_48 femtitvåårig bröllopsdag # so då: **prateðum** wi:ð just yvyr at e # addum weð so finer og kla:tt upp og so # nest Stellas Damfrisei
- 1 aasen\_35 og wi:ð wi:ð addum øvdalsdreker og so **servireðum**
- 1 aasen\_35 ja eð war fel mjæst # eð war brennboll # og se: **bolleðum** wi:ð iss-jær an bollar upp muot weddje: i i uolaik # serier *serier* og so
- 1 aasen\_35 **dalskeðum** iem ja # so at ig # dug ig # ig kund int riktut swenska mes byrd å: skalan *på skolan*
- 1 aasen\_48 för då: addum wi:ð olltiet boll'n min og so bar **bolleðum**

FIGURE 6: The multimedia results window

**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 (Figure 7).

The screenshot shows a window with three search results. The first result is for the word 'billig'. A tooltip window is open over the word 'billig', showing the following information:

- lemma: billig
- phon: billikt
- pos: adj
- sex: nøyt
- num: ent
- defn: ub
- degr: pos

The search results are as follows:

- 1 aal\_04gk så han må nå få gjort noe med det taket da men det er ikke **billig**  
[translate]
- 1 alvdal\_03gm men så byn- ja ja en gikk den men hun var aldeles sjuk på han ve  
[translate]
- 1 bengtsfors\_om1 hon blev **billig** bara för det att hon såg ut så  
[translate]

FIGURE 7: A window shows all information for each word that is moused over

**Action menu:** On the results page there is an Action menu with a selection of choices for further displaying of results and results handling (the latter of which will be presented in section [3.6]). The functionalities that follow in this subsection are choices in this menu (Figure 8).

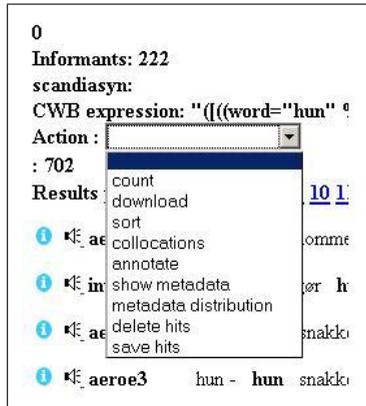


FIGURE 8: Action menu in results window

**Count:** Choosing the Count option gives the search results as a list of all the hits sorted by frequency. In Figure 9, a bit of a list is shown as a result of the search for nouns starting with *bil-* in Norwegian.

| occurrences | match              |
|-------------|--------------------|
| 40          | <b>bil</b>         |
| 20          | <b>bilen</b>       |
| 14          | <b>biler</b>       |
| 11          | <b>bilde</b>       |
| 7           | <b>bilene</b>      |
| 4           | <b>bildet</b>      |
| 1           | <b>bilkjøringa</b> |
| 1           | <b>bilbasert</b>   |
| 1           | <b>bilder</b>      |
| 1           | <b>bilveg</b>      |
| 1           | <b>bildeler</b>    |

FIGURE 9: Some nouns beginning with *bil-* ('car')

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

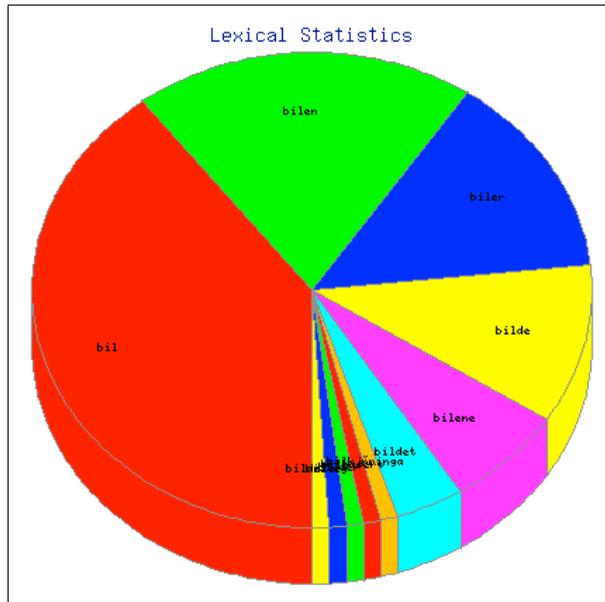


FIGURE 10: The same information as in Figure 9.

**Sort:** The results are by default sorted according to the geographical residence of the informants. However, they can be displayed in many other ways as well. 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 coefficient, log-likelihood ratio etc., with a choice between neighbouring bigrams and trigrams. The example in Figure 11 illustrates the collocations for the word *bil* 'car', used in the three mainland Nordic countries. The value of this choice is clearly illustrated in the example in Figure 11; the frequencies of the collocations are the same independently of language.

**Maps:** Recently an option of displaying 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 12 the different phonetic varieties of the negation are displayed in the right-hand column, giving the user the choice 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.

| Left context |      |           | Right context    |      |          |
|--------------|------|-----------|------------------|------|----------|
| ngram        | rank | AM occ    | ngram            | rank | AM occ   |
| en **        | 1    | 0.3304 19 | ** og            | 2    | 0.1628 7 |
| ha **        | 5    | 0.0800 4  | ** och           | 3    | 0.1412 6 |
| har **       | 5    | 0.0800 4  | ** #             | 3    | 0.1412 6 |
| åker **      | 5    | 0.0800 4  | ** då            | 4    | 0.0964 4 |
| åka **       | 5    | 0.0800 4  | ** ?             | 6    | 0.0732 3 |
| med **       | 7    | 0.0606 3  | ** eller         | 6    | 0.0732 3 |
| köra **      | 7    | 0.0606 3  | ** som           | 6    | 0.0732 3 |
| æ **         | 7    | 0.0606 3  | ** på            | 6    | 0.0732 3 |
| kjøre **     | 7    | 0.0606 3  | ** för           | 8    | 0.0494 2 |
| ikke **      | 7    | 0.0606 3  | ** ##            | 8    | 0.0494 2 |
| egen **      | 7    | 0.0606 3  | ** här           | 8    | 0.0494 2 |
| ingen **     | 9    | 0.0408 2  | ** (uforståelig) | 8    | 0.0494 2 |
| vi **        | 9    | 0.0408 2  | ** nå            | 10   | 0.0250 1 |
| kjøtte **    | 9    | 0.0408 2  | ** stående       | 10   | 0.0250 1 |
| ja **        | 9    | 0.0408 2  | ** dit           | 10   | 0.0250 1 |
| någon **     | 9    | 0.0408 2  | ** ner           | 10   | 0.0250 1 |
| kjører **    | 9    | 0.0408 2  | ** hver          | 10   | 0.0250 1 |
| kör **       | 9    | 0.0408 2  | ** kommer        | 10   | 0.0250 1 |
| * **         | 11   | 0.0206 1  | ** hemma,        | 10   | 0.0250 1 |

FIGURE 11: Some collocations for *bil* ‘car’.

### [3.5] *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 recently we have also included a map for his/her home place, see Figure 13.

**Via search page:** There is a button called “Show Texts”, 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 a descending order. This way, we can see how different the informants are in this respect. For example, one old man from Skreia, Norway, utters 1,300 words during his session, while another old man, from nearby Stange, utters more than 6,400 words.

### [3.6] *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

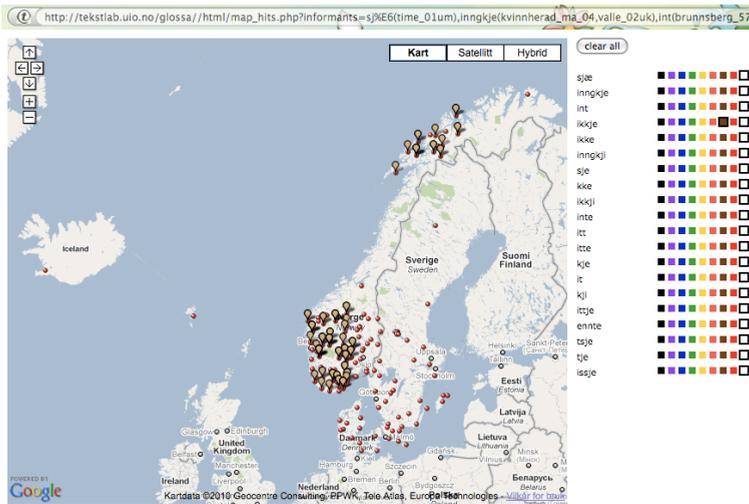


FIGURE 12: A map showing all the places that have hits (all the dots) for the orthographic forms of the negation ‘not’. The column on the right can be specified for a phonetic variant. Here the phonetic form *ikkje* has been chosen. It should be noted that parts of North Norway have not yet been included in the corpus.

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. Let us use a simple example: We want to find all and only the occurrences of the 3sgF pronoun (‘she’) used as a determiner, followed by any word, and then a noun. This search will give a lot of unwanted hits that we want to remove. We can then choose the Delete option from the Action menu and get Figure 14.

Notice in the figure that by having chosen the Delete option, the results come with a little box on the left hand side. In this box we tick the examples that we want to remove. If we suspected that there would only be a few examples that were appropriate for our research, we could instead have used the Choose option, which functions in the same way, but where ticking a box would mean to keep that result and delete the unticked ones.

**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. In Figure 15, we have chosen to annotate the examples by two categories: Demonstrative or Other.

The annotations can be edited and saved as annotation sets, for later reuse

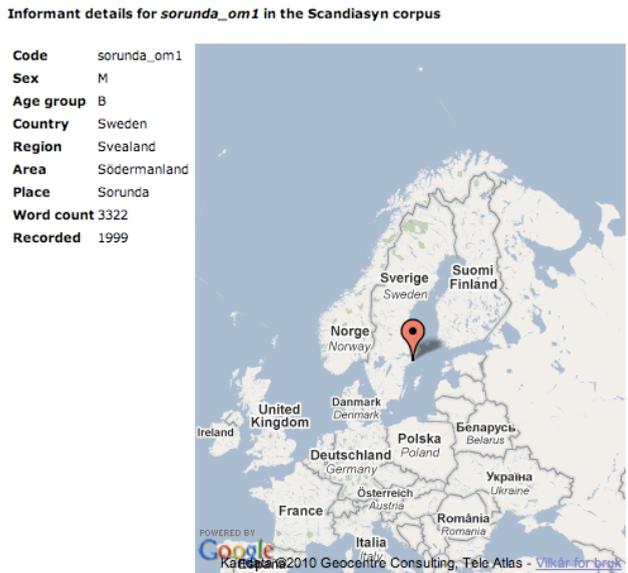


FIGURE 13: Information that appears in the search results window

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 imported as for instance tab-separated text.

#### [4] COMPARISON WITH OTHER DIALECT CORPORA

There are some other dialect resources on the web, but there are to our knowledge few or no available web-based dialect multimedia corpora for other languages. One 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, it is pedagogical, and not aimed at researchers. For example, there is no search option in the transcripts and no grammatical annotation.

The Scottish Corpus of Text and Speech contains 4 million words, 20% of which is spoken texts, provided with orthographic transcription, synchronised with the audio or video. It is not grammatically annotated and is not representative. How-

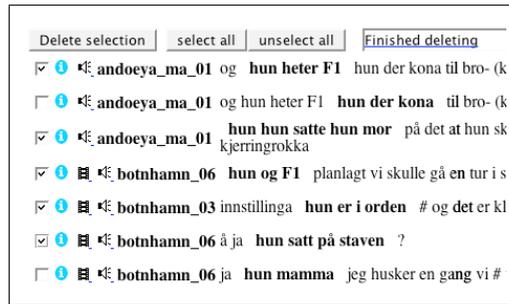


FIGURE 14: Results window with Delete option

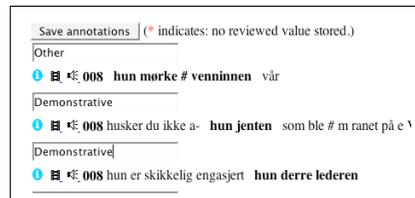


FIGURE 15: Results window with Annotate option

ever, it has a nice search interface.

The British National Corpus contains 10 million words of spoken English, which have been categorised into 28 different dialects. However, it says in their own search interface distribution that this categorisation is unreliable. Further, as a dialect corpus, the BNC has limited value, since it is not represented with audio, and the speech is transcribed orthographically.

The DynaSand web-based dialect database consists of information on various syntactic features and their distribution geographically in the Netherlands and Belgium. It contains recorded material from the project's questionnaire sessions, but the conversations contain to a large extent read sentences and meta-linguistic discussions, and less spontaneous speech.

The Spoken Dutch Corpus is transcribed orthographically, some of it also phonetically, and it is morphologically tagged. It contains spoken standard Dutch, not dialect data, and is not available by a web-interface.

The Corpus of French Phonology (La phonologie du français contemporain: usages, variétés et structure – PFC) is a web-based corpus of spoken French from across the Francophone world. It is searchable both phonologically and w.r.t. informant characteristics, and has transcriptions linked to sound.

There might be web-based dialect corpora for other languages, but information about these is hard to find, and they do not seem to be available on the web.

One such corpus under development is Corpus of Estonian Dialects. Another is Spoken Japanese Dialect Corpus (GSR-JD), available on DVD. Finally we should mention a small dialect corpus of Norwegian (Talesøk). It contains audio and transcriptions, and is available on the web.

There are some general web-based speech corpora that do not focus on dialect classification. For an overview of some Northern European ones, and their state of art w.r.t. topics like technical solutions and audio-visual availability, we refer to Johannessen et al. (2007).

Finally, we would like to mention that Paul Thompson at the University of Reading had a posting at Corpora List on November 30 2008 asking for information on corpus projects in which the developers have linked digital audio and/or video files to the transcripts, to allow access to the precise segment(s) of the audiovisual files that relates to a part of the transcript. In his summary of 15 responses there was only one dialect corpus – our own Nordic Dialect Corpus.

#### [5] CONCLUSION

We have presented the first version of the Nordic Dialect Corpus. It contains nearly 1.8 million words of Nordic dialects as spontaneous, not manuscripted, conversations. Most of them have been collected recently, but we have also included some old speech data. The Nordic Dialect Corpus has an advanced interface for searching and results handling. It is already a great resource for dialect researchers and linguists interested in the Nordic languages. The next version of the corpus will contain more dialect data. Part-of-speech taggers adapted for speech will be developed for all the languages, and all present and future texts will be tagged.

#### [6] ACKNOWLEDGEMENTS

In addition to participants in the ScanDiaSyn and NORMS networks, I would like to thank three anonymous NODALIDA-09 reviewers for valuable comments on an earlier version on this paper. I would also like to thank various funding bodies for funding the technical part of this project: the University of Oslo (both the Humanities Faculty and the central Research Department), The Norwegian Research Council, NordForsk and NOS-HS. In addition the national research councils in Norway, Sweden, Denmark and Iceland have contributed to the projects NorDiaSyn, SweDiaSyn, DanDiaSyn and IceDiaSyn, as has the project Swedia 2000, which has contributed a lot of Swedish recordings, and Norsk Ordbok 2014, which has contributed many months worth of recordings for the old Målførearkiv material that is included in the corpus.

Finally, I would like to emphasise that developing a corpus of the kind that I have described here is team work. I want to mention the cooperation of Øystein

Alexander Vangsnes (University of Tromsø), Tor Anders Åfarli (the Norwegian University of Science and Technology, Trondheim) and the staff at the Text Laboratory, especially Kristin Hagen, Signe Laake, Anders Nøklestad and Joel Priestley.

## REFERENCES

- Allwood, Jens, Joakim Nivre & Elisabeth Ahlsén. 1989. Speech management - On the nonwritten life of speech. In *Gothenburg Papers in Theoretical Linguistics*. University of Gothenburg.
- 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. URL [www.ims.unistuttgart.de/projekte/CorpusWorkbench/CQPTutorial](http://www.ims.unistuttgart.de/projekte/CorpusWorkbench/CQPTutorial).
- Hagen, Kristin, Janne Bondi Johannessen & Anders Nøklestad. 2000. A Constraint-based Tagger for Norwegian. In Carl-Erik Lindberg & Steffen Nordahl Lund (eds.), *17th Scandinavian Conference of Linguistics*, Odense Working Papers in Language and Communication 19, 31–48. University of Southern Denmark, Odense.
- Johannessen, Janne Bondi & Kristin Hagen. 2008. *Språk i Oslo. Ny forskning omkring talespråk*. Novus Forlag, Oslo.
- Johannessen, Janne Bondi, Kristin Hagen, Joel Priestley & Lars Nygaard. 2007. An Advanced Speech Corpus for Norwegian. In *NODALIDA Proceedings*, 29–36. Tartu: University of Tartu.
- Johannessen, Janne Bondi & Fredrik Jørgensen. 2006. Annotating and Parsing Spoken Language. In Peter Juel Henriksen & Peter Rossen Skadhauge (eds.), *Treebanking for Discourse and Speech*, 83–103. København: Samfundslitteratur.
- Johannessen, Janne Bondi, Lars Nygaard, Joel Priestley & 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).
- Johannessen, Janne Bondi, Joel Priestley, Kristin Hagen, Tor Anders Åfarli & Øystein Alexander Vangsnes. 2009. The Nordic Dialect Corpus - an Advanced Research Tool. In Kristiina Jokinen & Eckhard Bick (eds.), *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009. NEALT Proceedings Series Volume 4*.

- Johansson, Sofie Kokkinakis. 2003. *En studie över påverkande faktorer i ordklasstagning. Baserad på taggning av svensk text med EPOS*. Ph.D. thesis, Gothenburg University.
- Jørgensen, Fredrik. 2008. Automatisk gjenkjenning av ytringsgrenser i talespråk. In Janne Bondi Johannessen og Kristin Hagen (ed.), *Språk i Oslo*. Novus Forlag.
- Nivre, Joakim & Leif Grönqvist. 2001. Tagging a Corpus of Spoken Swedish. *International Journal of Corpus Linguistics* 6(1). 47–48.
- Nygaard, Lars. 2007. *The glossa manual*. The Text Laboratory. URL [www.hf.uio.no/tekstlab/glossa.html](http://www.hf.uio.no/tekstlab/glossa.html).
- Papazian, Eric & Botolv Helleland. 2005. *Norsk talemål*. Høyskoleforlaget, Kristiansand.
- Rosén, Victoria. 2008. Mot en trebank for talespråk. In Janne Bondi Johannessen & Kristin Hagen (eds.), *Språk i Oslo*. Novus Forlag.
- Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*.
- Søfteland, Åshild & Anders Nøklestad. 2008. Manuell morfologisk tagging av NoTa-materialet med støtte fra en statistisk tagger. In Janne Bondi Johannessen og Kristin Hagen (ed.), *Språk i Oslo*, 226–234. Novus Forlag.
- Thompson, Paul. 2008. Summary on Info of audio-visual corpora. *Corpora List*.
- Trosterud, Trond. 2009. A constraint grammar for faroese. In Eckhard Bick, Kristin Hagen, Kaili Müürisep & Trond Trosterud (eds.), *Proceedings of the NODALIDA 2009 workshop Constraint Grammar and robust parsing*. NEALT Proceedings Series, vol. 8, 1–7.

## CORPORA AND WEB RESOURCES

Barbiers, S. et al (2006). Dynamic Syntactic Atlas of the Dutch dialects (DynaSAND). Amsterdam, Meertens Institute.

<http://www.meertens.knaw.nl/sand/>

British National Corpus:

<http://www.natcorp.ox.ac.uk/>

Corpus Gesproken Nederlands.

<http://lands.let.kun.nl/cgn/ehome.htm>

Nordic Dialect Corpus:

[http://omilia.uio.no/glossa/html/index\\_dev.php?corpus=scandiasyn](http://omilia.uio.no/glossa/html/index_dev.php?corpus=scandiasyn)

NoTa Corpus (Norwegian speech corpus – Oslo part)

<http://www.tekstlab.uio.no/nota/oslo/>

La phonologie du français contemporain : usages, variétés et structure (PFC)

<http://www.projet-pfc.net/pfc-recherche>

Sounds familiar?

<http://www.bl.uk/learning/langlit/sounds/index.html>

Scottish Corpus of Text and Speech.

<http://www.scottishcorpus.ac.uk/>

Spoken Japanese Dialect Corpus (GSR-JD)

<http://research.nii.ac.jp/src/eng/list/detail.html#GSR-JD>

Swedia 2000.

<http://swedia.ling.gu.se/>

Talesøk.

<http://helmer.aksis.uib.no/talekorpus/Hovedside.htm>

Text Laboratory, UiO:

<http://www.hf.uio.no/tekstlab/English/index.html>

#### AUTHOR CONTACT INFORMATION

Janne Bondi Johannessen

Department of Linguistics and Nordic Studies

University of Oslo

P.O. Box 1102 Blindern

N-0317 Oslo

Norway

[jannebj@iln.uio.no](mailto:jannebj@iln.uio.no)