

## A multilingual speech resource: The Nordic Dialect Corpus \*

Janne Bondi Johannessen, Joel Priestley, Anders Nøklestad

The Text Laboratory, Department of Linguistics and Nordic Studies,  
University of Oslo, P.O.Box 1102, Blindern, N-0317 Oslo, Norway  
{jannebj, joel.priestley, noklesta}@iln.uio.no

**Abstract.** This paper describes the Nordic Dialect Corpus, a corpus that consists of transcribed spoken dialects, with sound and video, from five North European languages (Danish, Faroese, Finnish, Icelandic, Norwegian and Swedish). The paper focuses on recent developments that have been added as a result of wishes expressed by the linguist users. These include map views of various selections of search results, English translations of every dialect concordance, and search possibilities and presentation of both orthographic and phonetic transcriptions.

**Keywords:** Nordic languages, corpus, multilingual, multimodal, speech

### 1 Introduction

We present a corpus that is one of the major outcomes of the big collaborative dialect project Scandinavian Dialect Syntax (ScanDiaSyn) and the Nordic Centre of Excellence in Microcomparative Syntax (NORMS). In this paper we will focus specifically on features that the linguist users have asked for. The project has members from universities representing dialects from five languages in six countries in Northern Europe (Denmark, Faroe Islands, Finland, Iceland, Norway and Sweden). The majority of the project members (as well as present and future users) are linguists and dialectologists, and have little interest in and knowledge of technical matters; their main focus is dialectological research and data collection. The development of the Nordic Dialect Corpus (Johannessen et al. 2009; Johannessen et al. 2010), which we focus on in this paper, and the Nordic Syntactic Judgments Database (Lindstad et al. 2009) are the technical responsibility of the Text Laboratory, University of Oslo. When planning and constructing the corpus, we had to pay special attention to user-friendliness, but also to the requirements of the users w.r.t. search options and results handling.

There are a number of factors that are challenging with respect to constructing search options and results handling in this corpus:

- there are five different standard orthographies corresponding to the five standard languages (Danish, Faroese, Icelandic, Norwegian and Swedish)
- the corpus contents consists of transcribed speech
- some of the recordings have a double set of transcriptions – orthographic and phonetic
- transcriptions should be linked to audio and video and presented nicely
- the corpus should be tagged, needing five spoken language taggers, but the tagsets cannot be the same due to linguistic differences
- where possible, the same tags should refer to the same types of entity in all the languages
- metadata on informants should be usable as filters in search (age, sex, place)

\* The technical development of the corpus and the Norwegian recordings have been generously funded by NordForsk, NOS-HS, The University of Oslo and the Norwegian Research Council. In addition, the national research councils of Denmark, Iceland and Sweden have funded the national projects DanDiaSyn, IceDiaSyn, Swedia 2000 and SweDiaSyn, responsible for recordings that are now part of the corpus.

- different levels of geographical belonging should be specifiable (country, area, place)
- search results should be possible to handle in a number of different ways, including exporting of different formats.
- all text from all languages should be searchable at the same time
- the users want the search results to come with a translation into English
- the users want maps to see where the informants are from

Because of the prospective non-technical users of the corpus (mainly linguists and dialectologists), all solutions should be user-friendly and searches with regular expressions should be avoided. We use the corpus system Glossa (Johannessen et al. 2008), which has been developed at the Text Laboratory and which uses the corpus search system Corpus Workbench (Evert 2005). The latter is based on a query language of regular expressions. Glossa has a front end of clickable boxes and menus, whose information is translated to regular expressions inside the system. To our knowledge no other corpus system exists with this variety of options.

## 2 Languages and dialects contained in the corpus

The idea behind the Nordic Dialect Corpus is that the Northern Germanic languages (the Scanavian languages) are so similar that they could be considered dialects of one language. Most of the dialects in the area of Norway, Sweden, Denmark and Swedish Finland are mutually understandable, as are Icelandic and Faroese. Much of the vocabulary and the grammar is the same across all the languages, but with small syntactic differences that are interesting to study comparatively. The research of ScanDiaSyn and NORMS has been particularly aimed at syntax, and some of the development of the Nordic Dialect Corpus is done with this in mind. In addition, the two research networks have developed a syntactic questionnaire whose answers have been put into the Nordic Syntactic Judgements Database, which will not be further described here.

There are recordings from all the countries involved, performed with partly national and partly common Nordic funding. They are therefore somewhat different from each other w.r.t. a number of variables. The recordings from Denmark, Faroe Islands, Norway and partly Iceland were done for this corpus, while most of the Swedish recordings were done by a previous research project: Swedia 2000.

The number of measuring points where recordings were done differs, so that Denmark has currently 14, Faroe Islands 5, Iceland 3, Norway 94, and Sweden 40. There are at least two informants from each place, while the ideal is four (to ensure one old and one young speaker of each sex). At the moment there are 525 speakers producing 1.7 million words in the corpus, but this will rise as new recordings are added over the next year.

## 3 Challenging searches and results presentation, and their solutions

### 3.1 How to deal with so many languages in one corpus?

Although the dialects in the Nordic area can be argued to form a continuum, there are five standard languages, with different orthographies, to which they belong. There was a wish from the corpus users to be able to search for all the languages at the same time in one search window. While we discussed having a possibility of direct translation of each search word, in order to search simultaneously in all the languages, even when the search word would be represented by different lexical items in the languages, we abandoned this idea due to a lack of appropriate dictionaries to and from all the languages, and the fact that online translation at the word level is very arbitrary. Instead we offer the possibility of using one search box for each orthographic form for disjunctive search. The user can try to find translations by following a link to the common Nordic online dictionary Tvärslå, which consists of mostly automatically generated

multilingual wordlists from a number of sources, in the project Nordisk Nätordbok (see reference list).

For example, the multilingual dictionary tells us that the Nordic language variants of the word ‘not’ are *ikke* (Norwegian, Danish), *inte* (Swedish) and *ekki* (Icelandic). We had to find the Faroese version, *ikki*, elsewhere. With this information we can use the disjunctive search shown in the left-hand part of figure 1 below, which gives us 25,935 results, some of them shown in the right-hand part of figure 1. The user-friendly interface with boxes is translated to the Corpus Workbench regular expression in (1).

(1) "`((word="ekki" %c)) | ((word="inte" %c)) | ((word="ikki" %c)) | ((word="ikke" %c))` ;"

The screenshot shows the 'Scandinavian Dialect Corpus' search interface. On the left, there are four search boxes, each with a 'criteria»' button and '+' and '-' buttons. The search terms entered are 'ikki', 'inte', 'ikke', and 'ekki'. On the right, a list of search results is displayed, each starting with a blue information icon and the text ' aarhus6'. The results show various sentences containing the search terms, with '[translate]' links below each sentence.

Figure 1: Disjunctive search in the whole corpus plus some of the search results.

### 3.2 How to deal with several transcriptions?

Most of the dialects in the corpus have been transcribed with one transcription only; that of the standard orthography of that country. However, the dialects of Norway and of the two areas of Älvdalen and Gotland (considered to have dialects very far from Standard Swedish) have also been transcribed with a script that is phonetic-like (the Norwegian version is compliant with the transcription used in Papazian and Helleland (2005), the Övdalian one is the one recommended by the Övdalian Language Council, while the Gotland one has been used by the Swedia 2000 project).

Since there are two transcriptions, it should be possible to search in either transcription, and it should be possible to see both transcriptions aligned. We have made sure that the transcriptions are fully word-aligned with each other. This can be ensured in most cases because the transcriptions are first done manually on a phonetic basis, and then the orthographic transcriptions are developed using a semi-automatic dialect-transliterator. The transliterator takes a phonetic transcription as input and performs a first translation from it to the standard orthography based on word-lists learned from previous transcriptions from the current and other similar dialects. The translation is then corrected by a linguistically trained person.

The way to search phonetically is shown below in figure 2: The text box is the same as for orthographic searches, but the menu underneath the box gives the option of specifying phonetic search.

The screenshot shows a search interface. At the top, there is a text box containing the word 'itte'. Below the text box is a dropdown menu labeled 'criteria' with 'phonetic' selected. To the right of the text box and dropdown menu are two circular buttons, one with a '+' sign and one with a '-' sign.

**Figure 2:** Phonetic search for the phonetic form *itte* ('not').

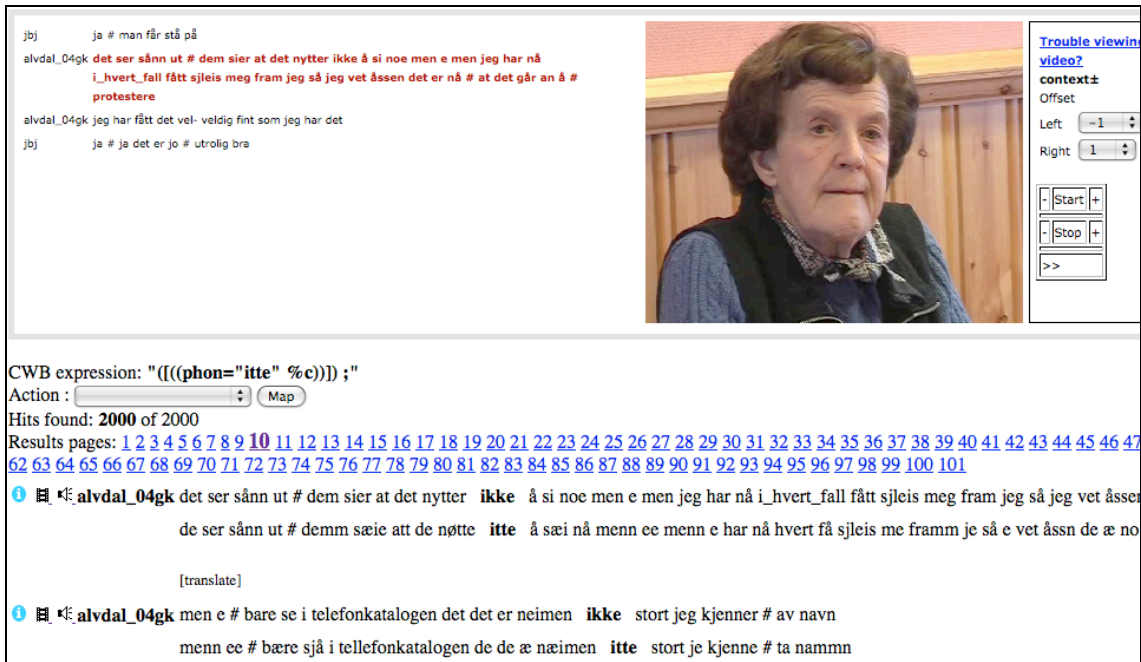
The results can also be specified to show either or both transcriptions, shown in the left-hand part of figure 3, with results on the right-hand side.

<p>Orthographic <input type="radio"/></p> <p>Phonetic <input type="radio"/></p> <p>Both <input checked="" type="radio"/></p>	<p><b>aal_04gk</b> men det får ta det <b>ikke</b> kort det er ikke anna gjøre ved menn dæ får ta d <b>itte</b> korr de e kji anna jera ve [translate]</p> <p><b>aasnes_ma_02</b> men han kom nok <b>ikke</b> att den n- natta # fordi det hadde l- de menn hænn komm nukk <b>itte</b> att dænn n- nætta # færde de hac [translate]</p> <p><b>aasnes_ma_02</b> så bjørnen kom nok <b>ikke</b> att denne e (uforståelig) ei natt eller så bjønn komm nukk <b>itte</b> att denna ee kommentar æi natt æll [translate]</p>
--	--

**Figure 3:** Specification of desired presentation and some results.



### 3.3 How to link transcriptions to audio and video?



All the transcriptions are linked to audio files and some also to video files. The linking has been done in the transcription program Transcriber, and the files are played in QuickTime. It is important for the sake of user-friendliness that these files are available directly from the concordance. We have chosen to do this with a clickable button to the left of each concordance line. This can be seen in figure 4.



jbj ja # man får stå på  
 alvdal\_04gk det ser sånn ut # dem sier at det nytter ikke å si noe men e men jeg har nå  
 i\_hvert\_fall fått sjeis meg fram jeg så jeg vet åssen det er nå # at det går an å #  
 protestere  
 alvdal\_04gk jeg har fått det vel- veldig fint som jeg har det  
 jbj ja # ja det er jo # utrolig bra

CWB expression: "(((phon="itte" %c)))";  
 Action :    
 Hits found: 2000 of 2000  
 Results pages: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#) [11](#) [12](#) [13](#) [14](#) [15](#) [16](#) [17](#) [18](#) [19](#) [20](#) [21](#) [22](#) [23](#) [24](#) [25](#) [26](#) [27](#) [28](#) [29](#) [30](#) [31](#) [32](#) [33](#) [34](#) [35](#) [36](#) [37](#) [38](#) [39](#) [40](#) [41](#) [42](#) [43](#) [44](#) [45](#) [46](#) [47](#) [48](#) [49](#) [50](#) [51](#) [52](#) [53](#) [54](#) [55](#) [56](#) [57](#) [58](#) [59](#) [60](#) [61](#) [62](#) [63](#) [64](#) [65](#) [66](#) [67](#) [68](#) [69](#) [70](#) [71](#) [72](#) [73](#) [74](#) [75](#) [76](#) [77](#) [78](#) [79](#) [80](#) [81](#) [82](#) [83](#) [84](#) [85](#) [86](#) [87](#) [88](#) [89](#) [90](#) [91](#) [92](#) [93](#) [94](#) [95](#) [96](#) [97](#) [98](#) [99](#) [100](#) [101](#)

ⓘ   alvdal\_04gk det ser sånn ut # dem sier at det nytter ikke å si noe men e men jeg har nå i\_hvert\_fall fått sjeis meg fram jeg så jeg vet åssen  
 de ser sånn ut # demm sæie att de nøtte itte å sæi nå menn ee menn e har nå hvert få sjeis me fram je så e vet åssen de æ no  
 [translate]

ⓘ   alvdal\_04gk men e # bare se i telefonkatalogen det det er neimen ikke stort jeg kjenner # av navn  
 menn ee # bære sjå i tellefonkatalogen de de æ næimen itte stort je kjenne # ta namnn

**Figure 4:** By clicking the video button, the appropriate video, belonging to that particular transcribed segment, is displayed.

### 3.4 How to deal with tags from five different languages?

Tagging the corpus with POS and morphological tags has taken more time than expected, because it has been impossible to find spoken language taggers for these languages. For Norwegian we had a spoken language tagger already, but for the other four languages we have had to develop new taggers based on existing written language taggers. We find it important that the tagsets are as similar as possible, so that the parts of speech have the same name. Nevertheless, there will be differences with respect to morphosyntactic categories that do not exist in all the languages. Thus, we expect all the five languages to have nouns and verbs, for example, but not four case distinctions. Mainland Scandinavian languages have no case distinctions on their nouns, while Icelandic and Faroese do. We are in the process of translating all the tags to English, which we use as a meta-language in the corpus, thus unifying the terminology as we go along. At the moment the user interface is catering only for Norwegian tags, but this will be changed before the end of the year. An illustration is given in figure 5.

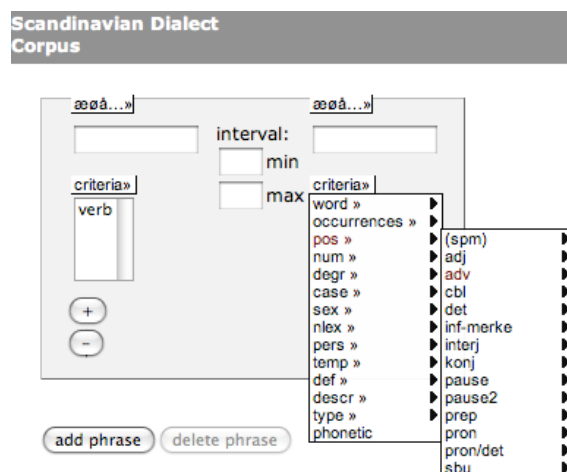


Figure 5: The search box for parts of speech and other tags.

The search in figure 5 will return all strings in the corpus that have a verb followed by an adjective – independently of language.

### 3.5 How to filter metadata?

For each informant there is information about homeplace, age, sex, and country, to mention the most important ones. These can be used to filter a search or pick particular informants with desired characteristics. The search page has its own section where such data can be chosen. It is illustrated in figure 6. It should be pointed out that each category is expandable, revealing a menu of choices. For example, when the sex menu is expanded, a list of two choices is revealed.

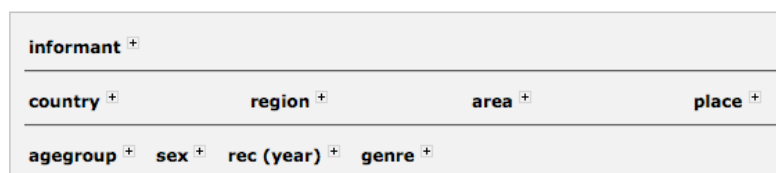


Figure 6: Metadata box, with expandable choices.

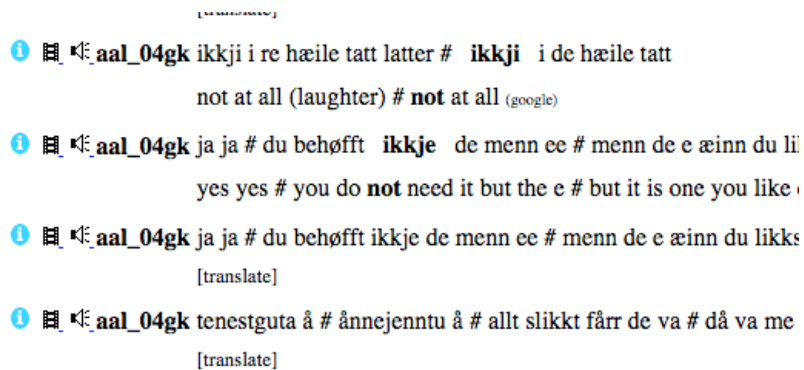
Unless one of these choices is expanded and one alternative is chosen, one's search will be completely general, covering all countries, place, ages, sexes, recordings years etc.

### 3.6 How to translate a corpus of spoken dialects?

Many linguists are interested in dialectology and the possibility of studying languages that differ only marginally from each other. However, they may not be very knowledgeable of the vocabulary that all these dialects and languages contain. So for example, if they make a very general search using parts of speech, they may need translation help to get an idea of what the resulting sentences say. The linguists of the ScanDiaSyn network have therefore expressed a wish to see the results translated.

We found that Google Translate could be used for this purpose. This could be done with no extra expense. Since we have at least one orthographic transcription for each dialect, we send that string to the online Google Translate service, which we can do even in the cases where the user has made a phonetic search and asked only for a phonetic view of the results. Since most users will probably not ask for a translation, we have chosen to have a button in the results view,

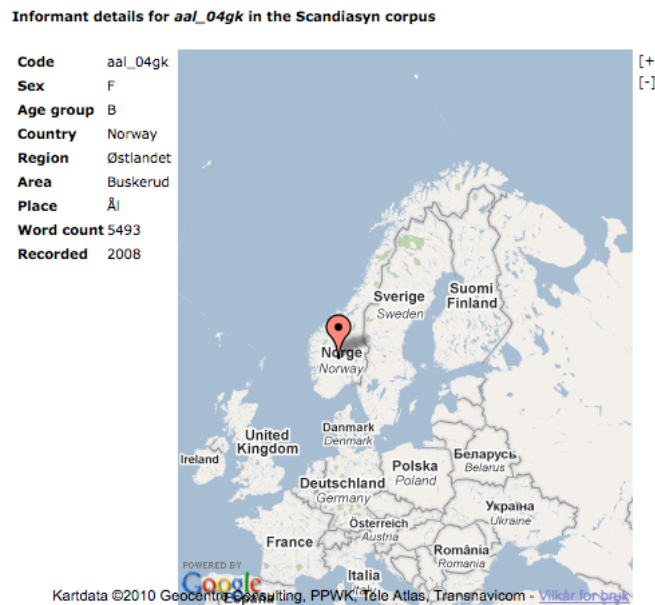
so that individual sentences can be translated by the choice of the user. This is illustrated in figure 7.



**Figure 7:** Results illustrating a phonetic view of a dialect with translation buttons (and two translations).

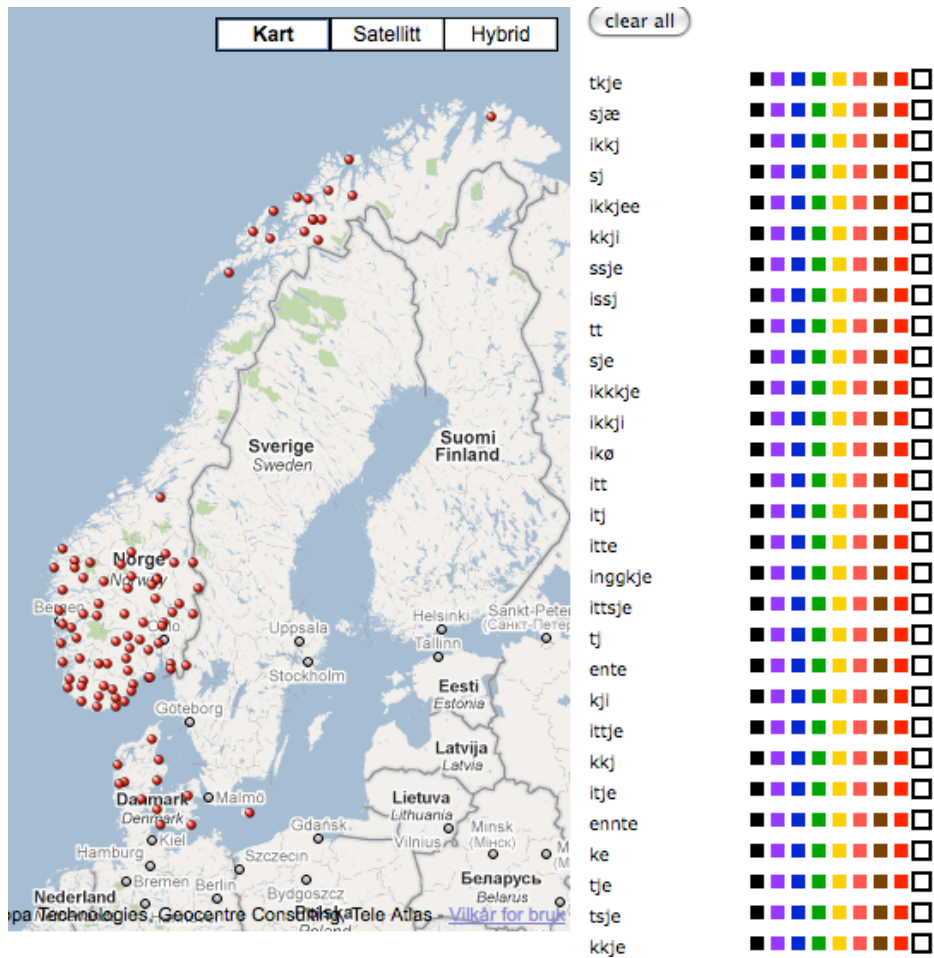
### 3.7 How to enhance the results with maps?

The ScanDiaSyn linguists may be good at the geography of one or two of the countries involved, but very few know where all the recording places are situated. They have expressed a wish to have information on the users as well as the total results illustrated with maps. We have therefore used Google Maps in two ways. First, for each individual result it is possible to click an information button to see an information box about a particular informant. The information button can be seen to the left of each concordance line, for example in figure 7. One such information box is illustrated in figure 8.



**Figure 8:** Information box about one informant.

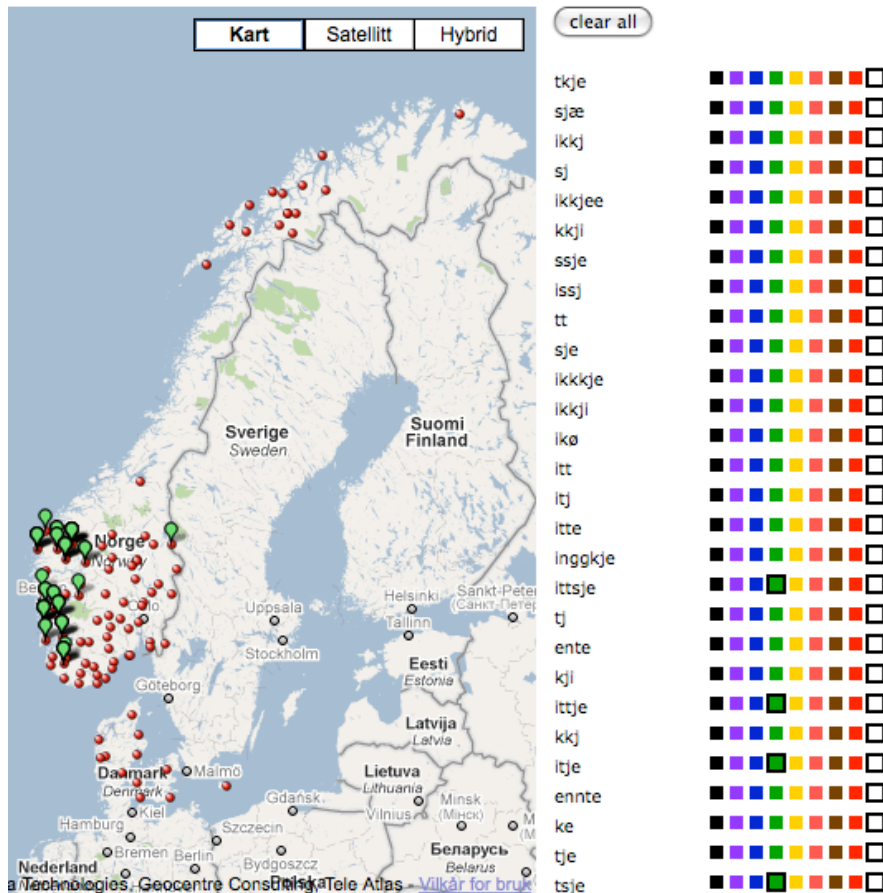
The other way to use maps is to get all the results presented on one map. We have chosen to do this by showing all results first, irrespectively of variation in phonetic realisation, and then present a list to the user, who can choose which results to be displayed on the map, see figure 9.



**Figure 9:** Results for the orthographic search for *ikke* ('not'), shown as red dots. The colour boxes on the right can be used to pick one or more particular pronunciations.

In figure 10 we illustrate how the user can pick one or more pronunciation variants from the phonetic transcription alternatives on the right side of the map. We have chosen to display those transcriptions that are pronounced with affricates.





**Figure 10:** Results for *ikke* ('not'), but with selected affricate pronunciations chosen from the column on the right.

There are still map solutions that have not been implemented yet. The most important is that we want to be able to have searchable maps as well. This way the user should be able to click on places in a map instead of specifying them through lists of place and country names.

#### 4 Conclusion

The Nordic Dialect Corpus has been developed in very close cooperation with dialectologists and linguists in two Nordic research networks. This way we have found solutions that we have seen in no other corpus, including hundreds of dialects, five languages, several transcriptions, maps, translations, as well as multimodal representations with audio and video. Everything is presented in a user-friendly web interface.

#### References

- Evert, Stefan. 2005. The CQP Query Language Tutorial. Institute for Natural Language Processing, University of Stuttgart.
- Johannessen, Janne Bondi, Lars Nygaard, Joel Priestley and Anders Nøklestad. 2008. Glossa: a Multilingual, Multimodal, Configurable User Interface. In: *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. Paris: European Language Resources Association (ELRA).
- Johannessen, Janne Bondi, Joel Priestley, Kristin Hagen, Tor Anders Áfarli, and Øystein Alexander Vangsnes. 2009. The Nordic Dialect Corpus - an Advanced Research Tool. In

- Jokinen, Kristiina and Eckhard Bick (eds.): *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009. NEALT Proceedings Series Volume 4.*
- Johannessen, Janne Bondi, Kristin Hagen, Anders Nøklestad and Joel Priestley. 2010. Enhancing Language Resources with Maps. In Calzolari, Nicoletta , Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, Mike Rosner and Daniel Tapias (eds.) *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA).
- Lindstad, Arne Martinus, Anders Nøklestad, Janne Bondi Johannessen and Øystein A. Vangsnes. 2009. The Nordic Dialect Database: Mapping Microsyntactic Variation in the Scandinavian Languages. In Jokinen, Kristiina and Eckhard Bick (eds.): *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009. NEALT Proceedings Series Volume 4.*
- Papazian, Eric and Botolv Helleland. 2005. *Norsk talemål*. Høyskoleforlaget, Kristiansand.

## URLs

- Corpus Query Language: [http://cogsci.uni-osnabrueck.de/~korpora/ws/CWBdoc/CQP\\_Tutorial/](http://cogsci.uni-osnabrueck.de/~korpora/ws/CWBdoc/CQP_Tutorial/).
- Corpus Work Bench: <http://cwb.sourceforge.net/>
- Google Translate: <http://translate.google.com/>
- Google Maps API: <http://code.google.com/intl/no/apis/maps/>
- Nordic Centre of Excellence in Microcomparative Syntax: <http://norms.uit.no/>
- Nordic Dialect Corpus: <http://www.tekstlab.uio.no/nota/scandiasyn/>
- Nordic Syntactic Judgements Database: <http://www.tekstlab.uio.no/nota/scandiasyn/>
- Nordisk nätordbok - Tvärså och Tvärsök <http://www.csc.kth.se/tcs/projects/netordbog/>
- Scandinavian Dialect Syntax network (ScanDiaSyn): <http://uit.no/scandiasyn/>
- The Text Laboratory: <http://www.hf.uio.no/tekstlab/>