

The Nordic Dialect Corpus and the Nordic Syntactic Judgments Database anno 2011

Janne Bondi Johannessen, Kristin Hagen, Anders Nøklestad, and Joel Priestley

(The Text Laboratory, ILN, University of Oslo)

1. Introduction

Creating corpora and databases for linguistic research is an ongoing effort with two almost conflicting goals. On the one hand, the users (linguists and philologists) want as much data as possible, i.e. the more words and the more meta-variables in the corpus, the better. The users want as many different search options and combinations as possible. On the other hand, such resources are hard to combine with another wish by the same users: maximum user-friendliness¹.

In this paper we present two dialect research tools, which, we believe, fulfill both requirements: The Nordic Dialect Corpus (Johannessen et al. 2009a) and The Nordic Syntactic Judgments Database (Lindstad et al. 2009). In particular, we will focus on how geographical maps can be used to enhance the amount of information available for the users while simultaneously making the resources easier to use. The map functionality in these applications is provided through the integration of Google Maps².

2. Nordic Dialect Corpus

The Nordic Dialect Corpus (Johannessen et al. 2009a) is the result of the collaboration in the research networks Scandinavian Dialect Syntax and Nordic Centre of Excellence in

¹ This is an updated and extended version of a paper published in the proceedings of the seventh international conference on Language Resources and Evaluation (LREC 2010) (Johannessen et al. 2010). The present paper deals with the 2011 versions of the corpus and database, which contain more data from more informants than previous versions. We have also included a section on results management.

² <http://maps.google.no/>

Microcomparative Syntax. The technical development is carried out at the Text Laboratory at the University of Oslo.

The corpus is currently under development, in the sense that the number of words is still growing and new functionalities are still being added, but it is already fully usable. At the moment it contains 2.1 million words.

The corpus contains recordings of dialects from the more or less mutually intelligible languages of the five countries Denmark, Iceland, Faroe Islands, Norway and Sweden. Recordings have mostly been done on a national basis, which means that there is some variation between them. For example, the Swedish material was mainly recorded during a different project a decade ago, while the Danish and Norwegian material was mainly recorded by national projects with this particular corpus in mind. The Faroese recordings were done as part of the dialect project, while the Icelandic recordings have been done partly in the project and partly before. For each language, there are recordings from the last decade. In addition, there are some older recordings from Norway. All the recordings in the corpus contain spontaneous speech, but while for some of the languages there are conversations between informants, for others, the conversations are between one informant and one project assistant, and for others still both types of conversations are available. The number of dialects recorded in each country varies due to differences in financing and of course in the linguistic situation. For example, there are around ten in Denmark, but around 100 in Norway and Sweden.

The recordings are presented in audio and, for some, in video format. All the dialects have been transcribed orthographically, and some phonetically. Each place is ideally represented with informants of both sexes, and sometimes also of different age groups. Part of the corpus is POS tagged, but all of it will be thus tagged in the end. Like with other corpora, an important use is thought to be linguistic searches for words, parts of words or strings of words, and in combination with POS tags. In addition, meta-linguistic variables can be used as search filters. For the benefit of users who do not understand Nordic languages, search results can be automatically translated using Google Translate in order to give the users a rough impression of what the utterance is about.

The corpus is used with the web-based Glossa corpus system (Nygaard et al. 2008), and is user-friendly with pull-down menus and clickable boxes and no need for regular expressions at the interface level.³

³ The Glossa corpus system was developed at the Text Laboratory at the UiO, and is used for a wide range of corpora: monolingual and parallel translation corpora, written and spoken language corpora. A number of universities use Glossa, due to its user-friendliness even as more and more search variables are

However, when a corpus includes hundreds of geographical locations from many countries, the typical user will need help to find where the places are actually located. Furthermore, for certain searches, the distribution of hits will in actual fact represent an isogloss for a particular phenomenon. This will not be visible unless the results are projected onto a map.

2.1. Maps in the Nordic Dialect Corpus

In order to enhance the visualisation of search hits in the corpus, we have implemented the use of maps. We have chosen to use the Google Maps API in our implementation, for several reasons: its flexibility w.r.t. the projection of information onto a map, their open licensing terms and the fact that they cover the whole area. We were looking at other solutions, such as the excellent free online services of the Norwegian Mapping Authority, but since they do not cover the whole Nordic area, they were not an option.

In the corpus, maps are used in three different ways: 1) for each concordance line, a clickable information button gives geographical information about that hit, 2) the geographical distribution of all the hits are represented in one map, and 3) geographical filters can be specified on the map instead of via the list of place names. The first two have been implemented, while the third option is currently being developed.

Option 1 gives the name of the place that particular speaker is from (providing further details such as age-group, sex, recording year). Importantly, since the place-name may not be known to the researcher, a map is also presented alongside. Figure 1 shows an extract of the search results, with the information button on the left.

negation adverb in Norway in the phonetic transcription.⁴ We can illustrate by searching in the phonetic transcription for the clitic variants *n̄te*, *k̄je* and *k̄ke*; these are all dialectal variants of the equivalent written standards *ikk̄je* and *ikke* (both meaning ‘not’). It is obvious that a long list of hits presented as a concordance would not give the linguist a nice overview of where each form is used, almost no matter how the list is presented or sorted. A distribution indicated on a map, on the other hand, would immediately give us a nice overall picture. We click on a map button that covers the whole resulting concordance.



Figure 3: Concordance list with map button

Starting with *n̄te*, we find that there are 102 hits in the concordance, found in five places; Aremark, Fredrikstad, Rømskog, Råde, and Trysil, all of them close to the Swedish border.

⁴ The Norwegian part of the corpus is transcribed both phonetically ó following the transcription standard in Papazian and Helleland 2005 ó and orthographically.

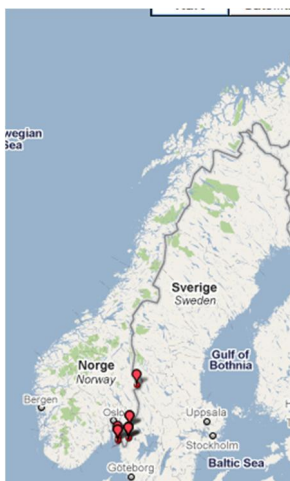


Figure 4: Five places have negator *nte*.

It is not surprising that this is the area where we find it, since *inte/nte* is the main negation word in Swedish.

Turning now to *kje*, we find that this negator covers a big part of Norway, as shown in Figure 5.

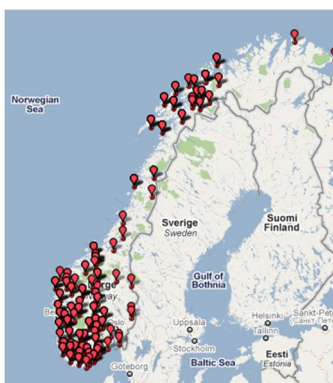


Figure 5: The negator *kje* is almost everywhere

This result is probably somewhat surprising for many people. The negation *kje* (with its full form *ikkje*) is the one that is used in the written standard *Nynorsk*, which is generally considered to be close to the dialects of West Norway, but we see here a much wider geographical distribution. The corpus concordance gives us 5594 occurrences, but it is the map that really shows how widespread it is.



Figure 6: The negator *kke* is quite rare compared to *kje*.

The corpus concordance shows 1590 occurrences of the negator *kke*. Given that this is the negation (with its full form *ikke*) that is considered the standard in Norway, it is quite interesting that it is rarer both in number and in geographical distribution than *ikkje* (*kje*).

At the moment, only the Norwegian data and certain parts of the Swedish data are grammatically tagged. Until we have grammatically tagged all the five languages it will be difficult to search in languages across the whole area, since the orthographies (and lexicons to some extent) differ. We can illustrate a cross-linguistic search, however, by doing a string search: all words starting with *hop*. The result is 497 hits distributed as shown in fig. 7.

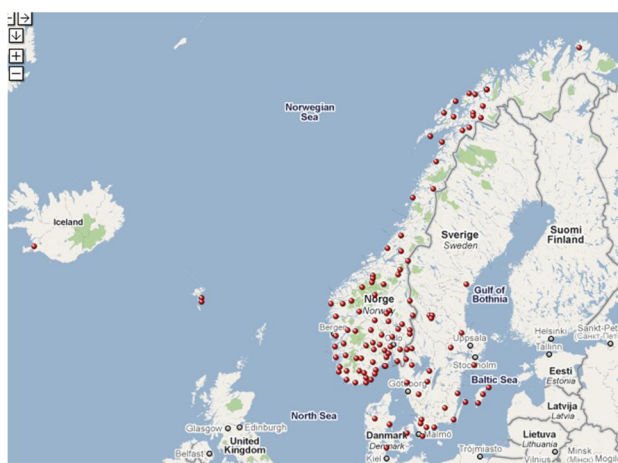


Figure 7: The distribution of the string *hop-* shown by the red dots.

This time we have results for (counted from west) Iceland, Faroe Islands, Norway, Denmark and Sweden. Checking the concordance, we find that some actual words are

hoppede (DK), *hoppa* (NO), *hoppade* (SW), all meaning *jumped* as well as some country specific ones, like *hoppas* (SW *hopes*) and *hopa* (FA). The equivalents in the other languages will not be found, due to differences in spelling and lexicon: *håper* (NO), *håber* (DK), *vona* (ICE).

We think these examples illustrate how useful the illustration of corpus search results on a map can be.

2.2. Results management in the Nordic Dialect Corpus

In addition to map visualisations, automatic translations, and audio and video playback, Glossa allows the user to perform a wide range of other functions on the search results:

- The distribution of matching words, lemmas or grammatical tags can be shown as lists, as a pie chart, or as a horizontal or vertical histogram, or it may be downloaded in a tab-separated or comma-separated format or as an Excel spreadsheet.
- The search results themselves may be downloaded as HTML, a tab-separated or comma-separated file, or an Excel spreadsheet.
- The search results may be saved to the server and can be retrieved on subsequent logins by the same user.
- Optionally, selected search results may be deleted before the results are saved. This is useful if some of the results do not actually represent what the user was looking for, perhaps because of erroneous grammatical tagging or simply because the user was not able to specify their query in sufficient detail.
- The results may be sorted according to the matching expression, its left or right context, or its sentence ID, or they may be shown in random order. Both a primary and a secondary search criterion can be specified. The sorting can be based on word form, lemma, or grammatical tag.
- Collocation lists (lists of words typically co-occurring with the search expression) may be generated, again based on word form, lemma or grammatical tag. The size of the context that is considered can be either a bigram (one word before or after the search expression) or a trigram (two words before the search expression, two words after it, or one word on each side of the search expression). In order to find words that tend to co-occur with

the search expression, it is not always sufficient just to look at the frequency of the words. It will usually be the case that function words, for example, will be the most frequent with most other words. However, sometimes we want to identify words that (whether frequently or infrequently) typically occur along with the search expression we have specified. Such a task requires more sophisticated statistical measures, and Glossa provides a wide range of measures (through the use of Banerjee and Pedersen's Ngram Statistics Package [Banerjee and Pedersen 2003]):

- Dice coefficient
 - Fisher's exact test ó left-sided
 - Fisher's exact test ó right-sided
 - Log-likelihood ratio
 - Mutual information
 - Pointwise mutual information
 - Odds ratio
 - Phi coefficient
 - T-score
 - Pearson's chi-squared test
- Users can create their own categories and annotate the search results with these. Such categories can represent grammatical distinctions that are not already annotated by the grammatical tagger, phonetic or phonological distinctions that are not included in the semi-phonetic transcriptions, information structure, or any other type of information that the user may find valuable. A user may create any number of annotation sets in order to annotate their search results in different ways. These annotations can be saved on the server along with the search results themselves, or they may be downloaded in any of the file formats mentioned above.

3. Nordic Syntactic Judgments Database

The Nordic Syntactic Judgments Database is the other research tool developed under the ScanDiaSyn umbrella. It contains speakers' intuitions, i.e. speakers' evaluation of test

sentences (many of which are grammatical only in some dialects) presented to them in a questionnaire. The ScanDiaSyn project has gathered data at 270 measure points in Scandinavia, of which 133 places have been put into the database so far.

A common Nordic pool of around 1400 sentences has been created, and national subprojects have selected a subset of them: in Norway, 140 sentences are tested, in Denmark 240. The informant judges each sentence on a scale from 1 (ungrammatical) to 5 (grammatical). Where possible, the database informants are the same as those in the corpus, making it possible to test whether what people claim about their language is in accordance with their actual language use. The linguistic literature has shown that there is often divergence between the two kinds of data, and it turns out that our data are no different, as shown with respect to the use of dative case in Norwegian by Johannessen et al. (2009b). In spite of some methodological challenges, judgments questionnaires are indispensable for syntactic research, where some constructions are rare and unlikely to be found in abundance in a corpus, and also because some of the research will be to test which constructions are actually ungrammatical. The maps below will show that the information from the judgments database is actually consistent in each area, and therefore prove their usefulness. More information on the database can be found in Lindstad et al. (2009).

Each test sentence has been appended with a number of linguistic categories describing in as much detail as possible the linguistic property that is tested by that particular sentence. An illustration is given with *wh*-questions differing in the placement of the finite verb, as V3 or V2 (verb in the third or the second sentential position), (1) and (2); with the linguistic description for both given in (3):

(1) Hva du heter?
what you is.called
=What is your name?∅

(2) Hva heter du?
what is.called you
=What is your name?∅

(3) word order, interrog., question, constituent question, simple *wh*-word V3/V2

Querying the database will typically be done in order to find how many have accepted a certain construction and where they are located. The results can be seen as a list of all informants with their judgment for that sentence. We think the database can be used for illustrations at this point, even though at the moment there are only 133 measuring points included. We illustrate with two sentences tested on the Norwegian informants. One is sentence no. 988 (the same as (1) above), and the other is an exclamation, sentence no. 311, here (4):

- (4) Hva biler det var her
 what cars it was here
 =What a lot of cars there are here!ø

The result is presented as in figure 8. The judgments are represented both by number and colour, where red colour and low number mark means that a particular sentence is considered ungrammatical by a particular speaker in this particular dialect, while green colour and high number mark signal the opposite.

311	kå bilar det var her	left periphery, exclamative, wh-exclamative, fronting, wh-phrase, fronted wh-phrase	Evje	evje01um	1
311	kå bilar det var her	left periphery, exclamative, wh-exclamative, fronting, wh-phrase, fronted wh-phrase	Evje	evje02uk	1
311	kå bilar det var her	left periphery, exclamative, wh-exclamative, fronting, wh-phrase, fronted wh-phrase	Evje	evje03gm	1
311	kå bilar det var her	left periphery, exclamative, wh-exclamative, fronting, wh-phrase, fronted wh-phrase	Evje	evje04gk	1
988	ke du heite?	word order, interrogative, question, constituent question, simple wh-word, V3	Gausdal	gausdal01um	5
988	ke du heite?	word order, interrogative, question, constituent question, simple wh-word, V3	Gausdal	gausdal02um	5
988	ke du heite?	word order, interrogative, question, constituent question, simple wh-word, V3	Gausdal	gausdal03gk	5
988	ke du heite?	word order, interrogative, question, constituent question, simple wh-word, V3	Gausdal	gausdal04gk	5
311	ke bile det var her	left periphery, exclamative, wh-exclamative, fronting, wh-phrase, fronted wh-phrase	Gausdal	gausdal01um	1
311	ke bile det var her	left periphery, exclamative, wh-exclamative, fronting, wh-phrase, fronted wh-phrase	Gausdal	gausdal02um	1

Page 2 of 12 Displaying 21 - 40 of 232

Figure 8. Search result for two sentences.

In our map application, we can choose a number of options. We illustrate with the top-most choice in figure 9. Here we get, for the sentences or categories we have chosen, all the locations where the informants have valued the sentence at 4-5 on average, i.e. accepted the sentence as fully grammatical. Seeing a long list of place names with individual judgments, like in figure 8, would not be as revealing as a map. Sentence (4) is represented by blue (dark) colour, and sentence type (1) by grey (light).

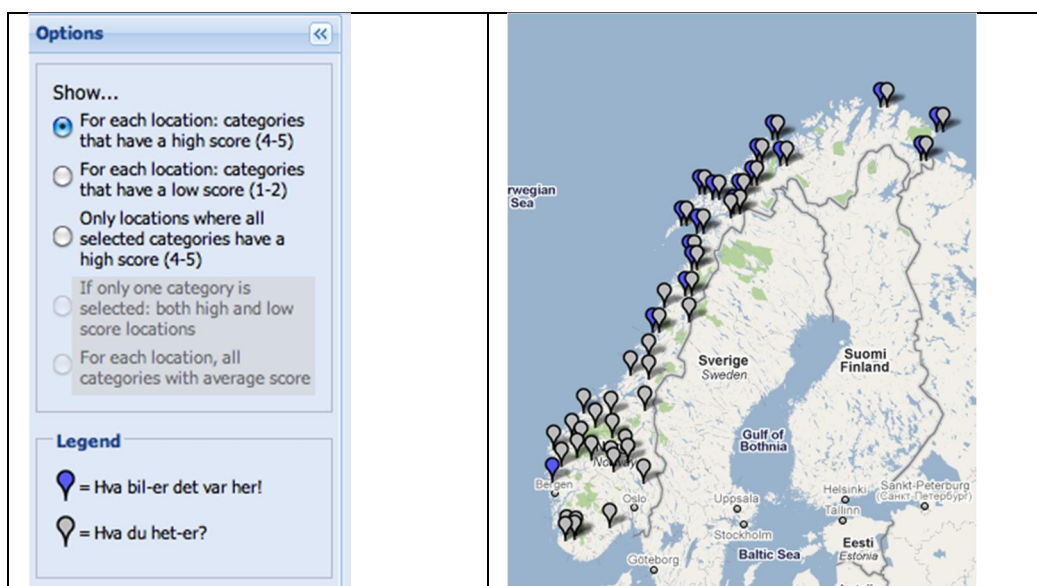


Figure 9: Positive informant judgments show wide geographical distribution of the non-standard V3 question type in Norway. The exclamation type is mostly found in the north.

We see that the sentence type exemplified in (1), which has the non-standard word-order V3, is actually accepted as grammatical, perhaps surprisingly, by informants in much of Norway.

We also illustrate option two, in figure 10, which shows where each selected sentence type has got a low score (1-2) in the questionnaire.

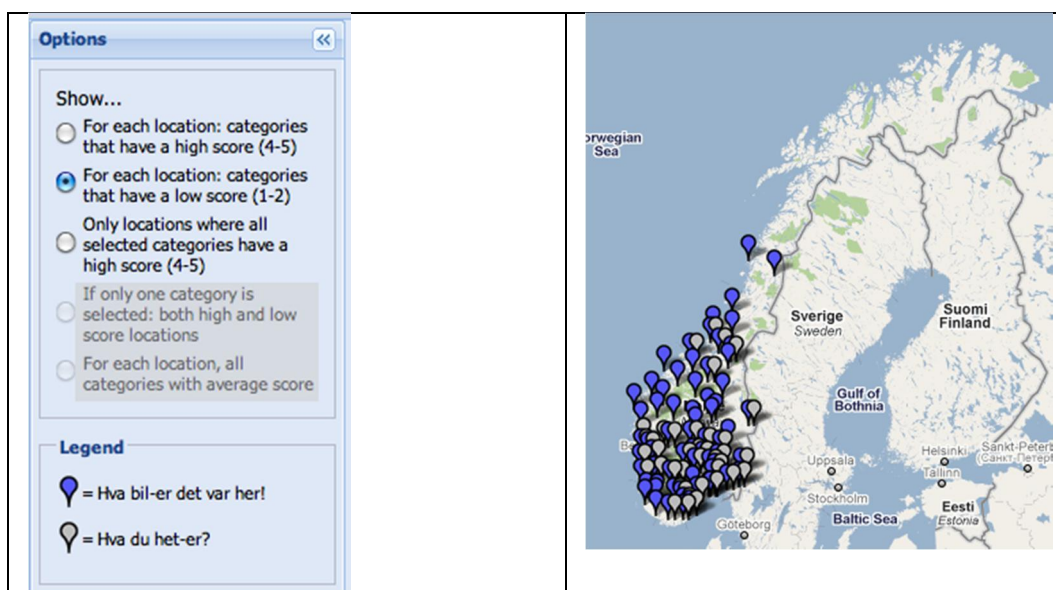


Figure 10: Negative informant judgments show that both the V3 question type and the exclamation type are rejected in parts of Southern Norway, with no rejections in the north.

It is obvious that using maps to illustrate the geographical distribution of features based on answers by informants is a very useful tool that can reveal surprising and hence new results.

4. Comparison with other map resources

Our language resources are not the first ones to use maps. We would like to mention *The World Atlas of Language Structures Online* (WALS) (Haspelmath et al. 2008) and the *Dynamische Syntactische Atlas van de Nederlandse Dialecten* (DynaSAND) (Barbiers et al. 2006). Both are advanced research tools, but different from ours.

WALS presents a vast array of maps with information from languages across the world. The maps are ready-marked with codes denoting several grammatical features in a particular category, but can be combined with information on other maps. Since the information encoded in the maps come from different authors and different investigations, there is huge variation as to how many and which languages are presented on a map with a particular feature. For example, the WALS map showing word order types includes information on as many as 1228 languages, while the map containing information on nominal plurality contains only 291 languages (and lack most of the Nordic ones!). The original information on the linguistic features comes from a variety of written linguistic literature.

The DynaSAND also presents many possibilities for features to be combined and shown in maps. Here the information is on Dutch dialects in Netherlands and Belgium. The Dutch language data all come from the same research project, so here all places have been tested w.r.t. the same features.

The maps we have shown from the Nordic Syntactic Judgments Database can, like those of WALS and DynaSAND, be generated from all the data in the database, and any combination is possible. In addition, the Nordic Syntactic Judgments database can show negative results as well. This is a major difference between our maps and the WALS ones. There, if a feature is not marked on a map, one does not know whether the reason is that the feature does not exist in that language, or whether it has not been investigated there. The DynaSAND maps can show negative results for some test sentences, however, not all test sentences were given in all regions, which limits this option somewhat (thanks to Jan Pieter

Kunst, p.c. for this information). The possibility of showing where negative values are given, together with the fact that all features have been tested everywhere (or will be when the database is finished) gives a very good picture of the distribution of a certain phenomenon.

The maps that are presented for hits from spontaneous speech in the Nordic Dialect Corpus do not have a counterpart in the other two. Here we generate maps on whatever the user wants to test, whether it is a word, a part of a word, a grammatical category, a string of letters written in a certain way or pronounced in a certain way, a combination of any of these, and even filtered with information on particular variables like sex, age or country.

Finally, we can mention that there have been attempts to combine geographical information with linguistic information in order to test hypotheses about what determines linguistic distance. One such attempt is reported in De Vriend et al. (2008), but they report their results as less interesting than expected.

5. Conclusions

Having developed two advanced language resources, it was soon clear that illustrating results with maps would enhance their usability. We have shown how maps can give the users immediate and valuable information that lists of concordances in the corpus or lists of table results in the database could not give. The visualisation makes it clear whether some phenomenon is distributed evenly across the whole area, or can only be found in more restricted areas. It will also show whether there is a clear geographical dividing line for a certain phenomenon, suggesting an isogloss. It furthermore makes it easy to see correlations between phenomena; whether two or more phenomena have the same geographical distribution, making possible new research questions as to what this kind of correlation might mean or entail in a specific case.

At the moment we are still in the process of putting more data into both the corpus and the database, but both resources are fully usable and can be used for many types of research in their current stage.

Acknowledgements

The work reported in this paper owes a lot to a number of research networks, funding bodies and individuals. First we should mention the Scandinavian Dialect Syntax research network (ScanDiaSyn), and the Nordic Centre of Excellence in Microcomparative Syntax (NORMS), under whose umbrellas the present work has been conceived, planned and developed. We would further like to thank those who have given us non-Norwegian material for the corpus and database, viz. Swedia 2000 (Anders Eriksson) for Swedish, Háskoli Islands (Ásta Svavarsdóttir) for Icelandic, DanDiaSyn (Henrik Jørgensen) for Danish, and who helped us to get recordings from Faroese (Zakaris Svabo Hansen). Also, we would like to thank our former colleague Arne Martinus Lindstad for the work he has done towards the linguistic categorisation of the Nordic Syntactic Judgments Database. One central person in the overall project to be mentioned especially is Øystein Alexander Vangsnes. Numerous other people have contributed to data collection, recording, transcription, tagging and preparation of data in different ways. We refer to the Nordic Dialect Homepage for more details on these. Finally, a number of funding bodies have contributed directly to the development of the corpus and database: The Research Council of Norway, The University of Oslo and the Nordic Research Councils NOS-HS and Nordforsk.

References

- Banerjee, Satanjeev and Ted Pedersen (2003). The Design, Implementation, and Use of the Ngram Statistics Package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, February 17-21, 2003, Mexico City.
- Barbiers, Sjef et al. (2006). *Dynamische Syntactische Atlas van de Nederlandse Dialecten (DynaSAND)*. Amsterdam, Meertens Instituut.
- De Vriend, Folkert, Jan Pieter Kunst, Louis Ten Bosch, Charlotte Giesbers and Roeland Van Hout (2008). Evaluating the Relationship between Linguistic and Geographic Distances using a 3D Visualization. In *Proceedings from LREC Marrakech*.
- Haspelmath, M., M.S. Dryer, D. Gil, and B. Comrie (eds) (2008) *WALS online*. Max Planck Digital Library, Munich.
- Johannessen, Janne Bondi, Kristin Hagen, Anders Nøklestad, and Joel Priestley (2010). Enhancing Language Resources with Maps. In *Proceedings of the Seventh conference*

on International Language Resources and Evaluation (LREC'10). European Language Resources Association 2010 ISBN 2-9517408-6-7, pp. 1081-1088.

Johannessen, Janne Bondi, Joel Priestley, Kristin Hagen, Tor Anders Åfarli, and Øystein Alexander Vangsnes (2009a). 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, Arne Martinus Lindstad and Signe Laake (2009b). Dative in Norwegian: Evidence from four dialects. The Maling Seminar; 30. November 2009, University of Iceland, Reykjavik.

Lindstad, Arne Martinus, Anders Nøklestad, Janne Bondi Johannessen, and Øystein Alexander 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*.

Nygaard, Lars, Joel Priestley, Anders Nøklestad, and Janne Bondi Johannessen (2008). Glossa: A multilingual, multimodal, configurable user interface. *Proceedings from LREC Marrakech*.

DynaSAND:

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

Glossa corpus system:

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

Google Maps API:

<http://code.google.com/intl/nb/apis/maps/index.html>

Nordic Dialect Corpus:

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

Nordic Syntactic Judgments Database:

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

NORMS:

<http://norms.uit.no/>

ScanDiaSyn:

<http://uit.no/scandiasyn>

WALS online:

<http://wals.info/>