The Nordic Dialect Corpus – an advanced research tool

Janne Bondi Johannessen
University of Oslo
Oslo, Norway
jannebj@iln.uio.no

Joel Priestley
University of Oslo
Oslo, Norway
joeljp@gmail.com

Kristin Hagen
University of Oslo
Oslo, Norway
kristiha@iln.uio.no

Tor Anders Åfarli
Norwegian Univ. of Science & Tech.
Trondheim, Norway
tor.aafarli@hf.ntnu.no

Øystein Alexander Vangsnes
University of Tromsø
Tromsø, Norway
oystein.vangsnes@hum.uit.no

Abstract

The paper describes the first part of the Nordic Dialect Corpus. This 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 search. The corpus is web-based and features full audio-visual representation linked to transcripts.

1 Credits

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.

2 Introduction

In this paper, we describe the first, completed part of the Nordic Dialect Corpus. The corpus has a variety of features that combined makes it a very 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 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.

3 Why the Nordic Dialect Corpus was developed

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

Kristiina Jokinen and Eckhard Bick (Eds.)
NODALIDA 2009 Conference Proceedings, pp. 73–80

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.

4 Description of the Corpus

4.1 Linguistic contents and numbers

The corpus contains dialect data from the national languages Danish, Faroese, Icelandic, Norwegian, and Swedish. It is steadily growing, since there are still new recordings that are being done, or planned, while other recordings are in various stages of finishing. At the moment, it contains speech data from approximately 170 informants with 466 000 words, unevenly spread between the five countries. Eventually, this will rise to around 600 informants and the number of words will likely be more than doubled. The numbers for the corpus as of today are given below.

Country	**No of informants**	**No of words**
Denmark	7	19 088
Faroe Islands	3	16 794
Finland	0	0
Iceland	4	10 287
Norway	45	132 417
Sweden	125	287 639
Sum	**184**	**466 225**

Table 1: Corpus contents by 9. January 2009.

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.)

Some recordings, such as those from Norway, the Swedish dialect of Oevdalian 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.

4.2 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 par 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.[1] For the Norwegian dialects and the

[1] The Norwegian phonetic transcription follows that of Papazian and Helleland (2005). The transcription of the Oevdalian dialect follows the Oevdalian orthography (stan-

Oevdalian 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 Oevdalian:

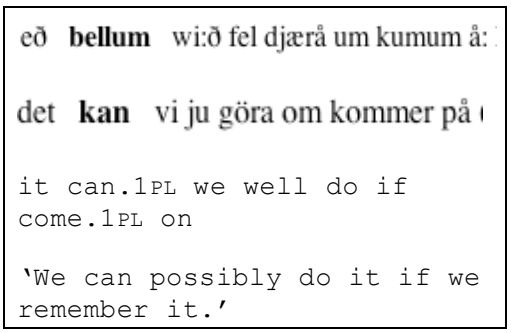

Figure 1. Two transcriptions for Oevdalian.

The Text Laboratory at the University of Oslo has the responsibility for the further technical devopment, 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 Swedish data will be tagged soon. 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 boundary 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 repeatedly 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 is being 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). The new data will be used as training data for a new Swedish speech TreeTagger.

4.3 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. Below, the user has searched for the suffix *–ig*, which can be found in Norwegian, Swedish, and Danish.

dardised in 2005 by the *Råðdjärum* (The Oevdalian Language Council).

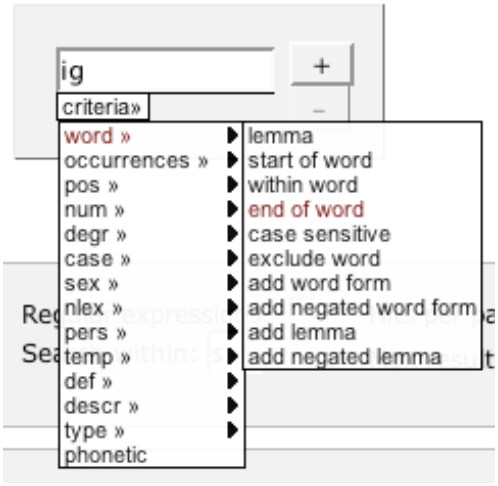

Figure 2: Search for suffix *-ig*

Notice that since nothing else was specified, this search would querying the whole corpus, i.e. amongst all the languages. Below 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 found	Translation	Language
7	særlig	especially	No, Da
7	farlig	dangerous	No, Sw, Da
7	þannig	thus	Ice
7	kjedelig	boring	No
6	väldig	very	Sw
5	rigtig	right	Da
5	otrolig	unbelievable	Sw
4	konstig	strange	Sw
1	sjómannaslig	sailorlike	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 words in between:

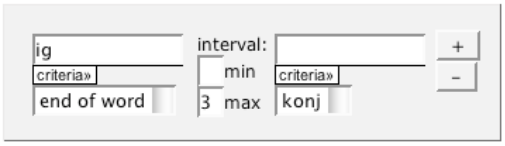

Figure 3. Searching for two words.

The illustration shows a search for a word ening 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 the figure above, 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 exemplifies many of the search options that are available for a word.

Phonetic querying: The user can choose to query the corpus by specifying 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 Oevdalian, we can simply write for example the 1pl suffix as it is:

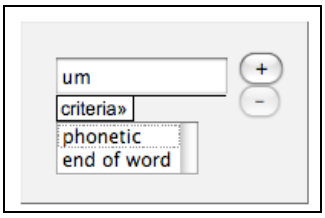

Figure 4. Searching in phonetic mode.

This will give results that would have been impossible to get from the orthographic text only. We refer to Figure 1 for an illustration, 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. Below, we show how we can choose individual places from the com-

plete list, to be able to query only the informants from these places, which happen to be the area of Älvdalen in Sweden.

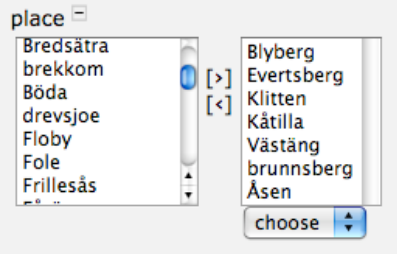

Figure 5. Delimiting the corpus by choosing some places from the full list.

4.4 Displaying 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 trancribed speech from five countries and spans four decades. Some of the speech was naturally recorded using a tape recorder and later mp3 recorder, and some was recorded by videocamera. The result is accompanied by a clickable symbol to show the audio and video of that particular speech sequence. This is illustrated below.

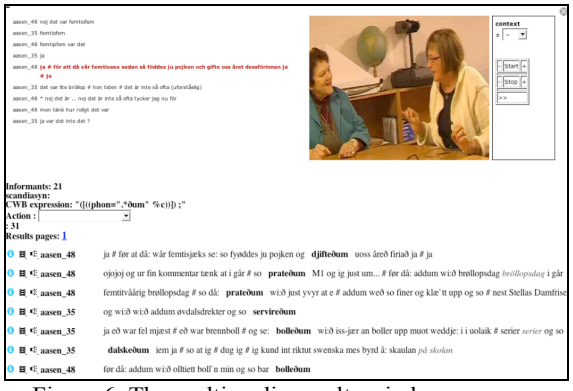

Figure 6. The multimedia results window.

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 window when navigating the mouse over the text:

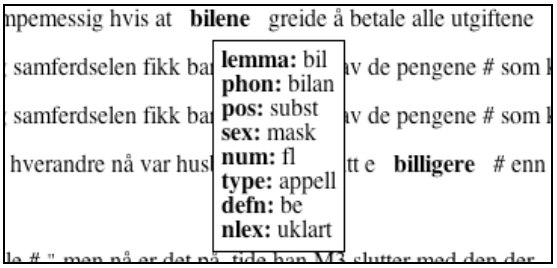

Figure 7. A window shows all information for each word that is moused over.

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 4.6). The functionalities that follow in this subsection are choices in this menu.

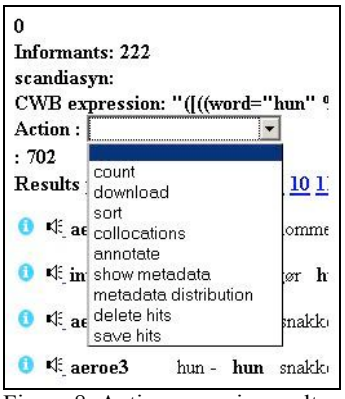

Figure 8. Action menu in results window.

Count: Choosing the Count option gives the search results as a list of all the hits sorted by frequency. Below, a bit of a list is shown as a result of the search for nouns starting with *bil-* 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 9. Some nouns beginning with *bil-* (‘car’).

The count results can be shown in a number of ways, such as histograms and pie charts. The same result as above is shown below as a pie chart:

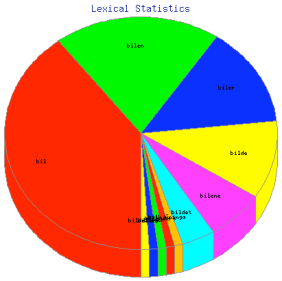

Figure 10. The same information as in figure 9.

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 coeffiency, log-likelihood ration etc., with a choice between neighbouring bigrams and trigrams. The example below 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; the frequencies of the collocations are the same independently of language.

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
kjørte **	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'.

4.5 Displaying information on informants

There are two ways of finding information on the informants.

Via results page: Each concordance line has an "i" symbol on its very left. Clicking on this symbol reveals the following information on the informant in question: informant code, sex, age group, country, place, number of words, recording year.

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., like above, 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. 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.

4.6 Further processing of results

Deleting or choosing some results: In a corpus search it is often the case that the user get 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. Let us use a simple example: We want to find all and only the occurrences of the 3sgF pronoun ('she') used as a determiner with something between and then a noun. This search will give a lot of unwanted hits that we want to remove. We choose the Delete option from the Action menu and get the figure below:

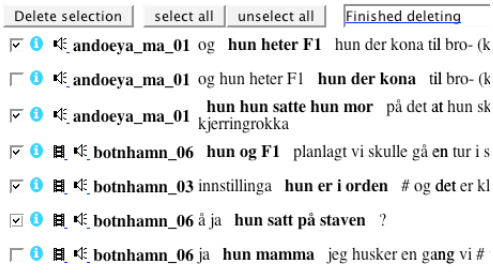

Figure 12. Results window with Delete option.

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 in-

stead have chosen the Choose option, which functions in the same way.

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. Below, we have chosen to annotate the examples by two categories: Demonstrative or Other:

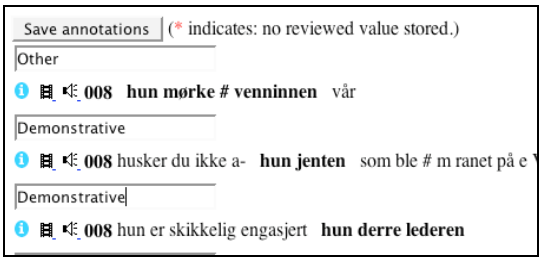

Figure 13. Results window with Annotate option.

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 imported as for instance tab-separated text.

5 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 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. However, 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.

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.

6 Conclusion

We have presented the first version of the Nordic Dialect Corpus. It contains nearly half a million words of Nordic dialects. Most of them have been collected recently, but we have also included some old speech data. The Nordic Dia-

lect 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 alle the languages, and all present and future texts will be tagged.

Acknowledgements

In addition to participants in the ScanDiaSyn and NORMS networks, we would like to thank three anonymous NODALIDA-09 reviewers for valuable comments.

References

Allwood, Jens, Joakim Nivre, Elisabeth Ahlsén. 1989. Speech Management - On the Non-Written Life of Speech. *Gothenburg Papers in Theoretical Linguistics*. University of Gothenburg.

Christ, Oli. 1994. A modular and flexible architecture for an integrated corpus query system. COMPLEX'94, Budapest.

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

Hagen, Kristin, Janne Bondi Johannessen and Anders Nøklestad. 2000. A Constraint-based Tagger for Norwegian. I Lindberg, Carl-Erik og Steffen Nordahl Lund (red.): *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, Joel Priestley and Lars Nygaard. 2007. An Advanced Speech Corpus for Norwegian. In: *NODALIDA Proceedings*. Tartu: University of Tartu, p. 29-36

Johannessen, Janne Bondi and Kristin Hagen. 2008. *Språk i Oslo. Ny forskning omkring talespråk.* Novus forlag, Oslo.

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 and Fredrik Jørgensen. 2006. Annotating and Parsing Spoken Language. In Peter Juel Henrichsen, Peter Rossen Skadhauge (eds.): *Treebanking for Discourse and Speech*. København: Samfundslitteratur. s. 83-103

Johansson, Sofie Kokkinakis. 2003. *En studie över påverkande faktorer i ordklasstaggning. Baserad på taggning av svensk text med EPOS.* (PhD dissertation). Gothenburg University.

Jørgensen, Fredrik. 2008. Automatisk gjenkjenning av ytringsgrenser i talespråk. In Johannessen and Hagen (eds.).

Nivre, Joakim and Leif Grönqvist. 2001. Tagging a Corpus of Spoken Swedish. *International Journal of Corpus Linguistics 6(1*), 47-78.

Nygaard, Lars. 2007. The Glossa Manual. The Text Laboratory. www.hf.uio.no/tekstlab/glossa.html

Papazian, Eric and Botolv Helleland. 2005. *Norsk talemål*. Høyskoleforlaget, Kristiansand.

Rosén, Victoria. 2008. Mot en trebank for talespråk. In Johannessen and Hagen (eds).

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*.

Søfteland, Åshild and Anders Nøklestad. 2008. Manuell morfologisk tagging av NoTa-materialet med støtte fra en statistisk tagger. Johannessen and Hagen, 226-234.

Thompson, Paul. 2008. Summary on Info of audio-visual corpora. *Corpora List*, 15 December 2008.

Corpora and web resources

Barbiers, S. et al (2006). Dynamic Syntactic Atlas of the Dutch dialects (DynaSAND). Amsterdam, Meertens Institute. URL: http://www.meertens.knaw.nl/sand/

British National Corpus: http://www.natcorp.ox.ac.uk/

Corpus Gesprochen Nederlands. http://lands.let.kun.nl/cgn/ehome.htm

Nordic Dialect Corpus: 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/

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

ISSN 1736-6305 Vol. 4
http://hdl.handle.net/10062/9206