An Advanced Speech Corpus for Norwegian

Janne Bondi Johannessen, Kristin Hagen, Joel Priestley and Lars Nygaard
The Text Lab, Univ. of Oslo,
P.O.Box 1102 Blindern, N-0317 Oslo, Norway
{jannebj, kristiha, joelp, larsnyg}@iln.uio.no

Abstract

This paper describes a new Norwegian speech corpus – The NoTa Corpus – that exhibits a variety of useful and advanced features. It contains 900 000 words of transcribed, lemmatised and POS tagged Oslo speech (carefully selected to cover many speech varieties), which is linked directly to audio and video. It has advanced search interfaces both for searches and results presentations. Since corpora of this kind are aimed at linguists and non-technical users, our guideline has been to keep user-interfaces maximally simple at all levels. The paper describes the contents of the corpus, and focuses on some nice features of its search interface. Some problems and solutions w.r.t. transcription are discussed, and the corpus is compared with five other speech corpora.

1 Introduction

In this paper, we will present the NoTa Corpus – a new speech corpus for Norwegian. It has been developed in order to serve non-technical linguists as well as developers in language technology. This means that it will be a valuable language resource for a wide range of users, for research problems in such diverse disciplines as lexicography, phonology, morphology, syntax, semantics, pragmatics, dialectology, socio-linguistics, psycholinguistics, speech synthesis, grammatical tagging and parsing, and artificial intelligence.

For linguist users, the search and results interfaces are developed in order to ensure a simple human-machine dialogue, a non-trivial task given the complex searches that can arise from the wide range of possible combinations of search variables, including multimodal options. Both the contents and interfaces for search and presentation of results have been planned in order to give maximal value for the user linguist at a minimum of effort and training.

For language and speech technologists we have focused on a high technical standard for various aspects of the contents, especially audio quality and standardised text markup.

With our aim at developing a high standard speech corpus, we have used a variety of off-the-shelf programs as well as tools and resources that we have developed ourselves, many of which will be available for the larger research community.

The corpus consists of 900 000 words that are transcribed, lemmatised and POS tagged. The transcriptions are linked to audio and video. A result concordance with video is illustrated below:

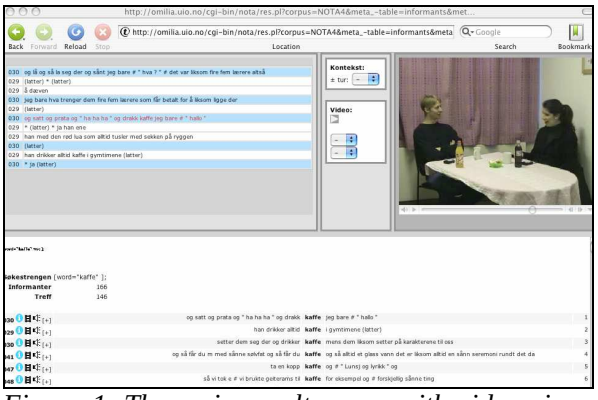

Figure 1. The main results page with video viewing.

Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek and Mare Koit (Eds.)
NODALIDA 2007 Conference Proceedings, pp. 29–36

Section 2 of the paper focuses on the contents of the corpus; annotation, multimedia representation, selection of informants and type of recordings, and transcription. Section 3 describes the search interface with special interest given to criteria regarding linguistic and informant selection. In section 4 we compare the corpus with five other speech corpora along some of the variables that have been highlighted in this paper.

2 Contents of the NoTa Corpus

2.1 Annotation and multimedia representation

The NoTa corpus is transcribed using standard orthography[1]. (The reason for this choice and some discussion about transcription is given in section 2.3.) The corpus is lemmatised and POS tagged by a TreeTagger trained on a manually corrected version of the Oslo-Bergen tagger, which is a written language tagger (for details, see Nøklestad and Søfteland, to appear).

The corpus is represented with video and audio recordings that are linked to the transcriptions[2]. The linking between transcription and audio/video makes it possible for the user to get a direct multimedia representation of any desired fragment of the corpus.

The corpus is searchable via the Internet site using the corpus explorer tool Glossa (Nygaard 2007), a very user-friendly interface built on top of the IMS Corpus Work Bench Query system. The results are shown as concordances linked to the multimedia representations. The Glossa system also allows further processing of the search results by exporting all or a subset of them to external file formats, and by viewing them in a variety of ways, such as frequency counts, collocations, statistical measures, pie charts etc.

All transcriptions of the speech occurring in the corpus are searchable, as are the specially annotated events such as laughter and coughing, plus a variety of interjections and exclamations, extralinguistic noises etc. It is also possible to do searches via grammatical tags. (Some examples will be given in section 3.)

2.2 Informants and recording situation

The corpus consists of the speech of 166 informants from the Oslo area, carefully divided to represent in equal numbers gender, age (three groups; 16–25, 26–50, 51–95), educational background and place of residence. The informants were recruited in a variety of ways, from actively contacting centres for elderly people, schools and work places, to using the press, students and the network of people we knew.

Each informant takes part in a semi-formal ten-minute interview with a project assistant, in which he or she is asked general questions about his or her life. In addition, each informant takes part in an informal 30-minute dialogue with another informant, at which point the informants get served drinks and snacks to add to the informal atmosphere. This way the corpus has two different speech styles from each informant.

Norwegian legislation requires a high level of anonymity and security (to the extent that this is possible when informants appear on audio and video). This has two consequences. First, the topics that are talked about must be "safe": the informants must be instructed not to talk about e.g. politics, religion, illness, criminality, and other people. Second, the informants must not be linked to the data by name or other identification, so the lists of their names and addresses have had to be destroyed.

The second consequence cannot be compensated, but the first consequence turns out not to be a serious problem. The informants get a list of possible topics (such as film, pets, travel, sports) to help them if the conversation goes dead. By comparing the two styles of each informant, it is clear that the limitation on topic is not generally inhibiting.

2.3 Transcription

The NoTa corpus has been transcribed by standard orthography (Norwegian Bokmål), in accord with the practice in other speech corpora, such as the Spoken Dutch Corpus (CGN). The benefits of such a choice over a more phonetic variant are numerous: Transcribers do not need special

[1] All speech is transcribed using the freely downloadable program Transcriber.

[2] We have used Quicktime Pro to convert from .wav-format to AAC in .mov-files, to be played by each user in Quicktime, via a central streamer.

training; inter-annotator agreement in transcriptions is more likely to obtain; fewer options will make transcribing quicker; the resulting transcription can readily be used for searching, reading it will be easier, and tagging and parsing will be easier.

However, speech will always contain linguistic as well as non-linguistic information that standard orthography – as it appears in standard dictionaries[3] – has no remedies for, and which corpus developers want to and sometimes need to cater for, so some concessions will have to be made. We shall mention a few of them here, and otherwise refer to Hagen (2005), Johannessen et al. (2005), Bødal et al. (to appear).

A first challenge is to decide what it means to use an orthographic standard. Should it only count at word-level? How about syntax? Consider the example below. In Norwegian, the standard norm says that 3p pl pronouns are inflected for case, so that nominative is used with subjects, and accusative with objects (example 1 below). However, many people violate that norm in various ways, as in (2).

```
(1a)  De          går
      they-NOM    walk
(1b)  Anne ser    dem.
      Anne sees   them-ACC.

(2a)  Dem         går.
      them-ACC    walk
 (2b) Anne ser de.
      Anne sees   they-NOM.
```

We have chosen to follow the orthographic norms only at word-level, so that it is irrelevant whether a word is used "wrongly"; what is relevant is whether a given spoken word has an orthographic equivalent. Thus, the examples in (2) are acceptable transcriptions in the NoTa corpus. Also, maybe needless to say, "incorrect" word order will never be changed by the transcribers.

A second challenge is words that occur in spoken language only. One difficult type is words that are clearly variants of written ones, but where it is unclear of which particular word. Consider the Norwegian clitic (spoken) pronouns in (3), which are unmarked for case. Choosing an orthographic form for them would be to either force a particular case onto them, something we have already seen can be very difficult due to inter-speaker variation, or to force a choice of animacy even when the context is ambiguous. Some choices of pronouns are given in (4).

```
(3a)  a     3p sg fem
(3b)  n     3p sg masc

(4a)  hun   3p sg fem nom
      henne 3p sg fem acc

(4b)  han   3p sg masc ani-
                       mate nom
      ham   3p sg masc ani-
                       mate acc
      den   3p sg inanimate
```

We have chosen to add these and other words that do not have a clear equivalent in the standard orthography, to a word-list that we ask the transcribers to use.

A second type of words not found in the standard dictionary are typically dialect words or borrowings. We simply use these as they are, and have chosen to tag them in the following way:

```
(5a)  den fisken ser gøllei
            [language=x] ut
      that fish looks "gøllei"
                  (horrible)

(5b)  yes [language=x] det er fint
      "yes" that is good
```

Interjections are a third type of words that are not all found in the dictionary. Like with other non-standard words, we found that we had to add them to our word-list. Distinguishing between interjections and other noises is not necessarily easy, however. Our rule of thumb was to try to fit some constant meaning to the sound sequence. If possible, we treated the candidate as an interjection, and devised a uniform spelling for the word in question. This work was also necessary to distinguish these interjections from similar, but non-identical, ones that already existed in the dictionary. Some of our new interjections together with some old ones (marked with BMO) can be seen in (6):

[3] Standard orthography in the NoTa context is defined as that which can be found in Wangensteen (2005): *Bokmålsordboka*.

(6)

aha (surprised) BMO
e (hesitating – irrespective of the vowel quantity)
eh (indicating distance)
ehe (”I see” – two syllables)
em (hesitation)
heh (impressed)
hm (inquiring, wondering) BMO
hæ (inquiring) BMO
jaha (strengthen ”yes”) BMO
m (hesitation, accepting)
m-m (benektende)
mhm (”I see” – two syllables)
mm (confirming – two syll.)
næ (surprised, wondering)
nja (doubting) BMO
næhei (strengthening ”no”)
ops (something went wrong)
u (impressed)
ææ (confirming – two syll.)
å-å (something went wrong)
å ja (suprised)

In addition to interjections, there are meaningful sounds that many speech corpora annotate, such as laughter. Their meaning is not as conventionalised as that of interjections, and we have chosen to have very coarse-grained categories, (7). They are annotated in the corpus as tags.

(7)

Front clicking sound
Back clicking sound
Sucking noise
Sibilant
Yawning
Laughter
Breathing
Special cough[4]

All transcriptions have been proof-read by other transcribers than those having done the original transcription, and regular transcription meetings were held between the half a dozen transcribers and the project management during the 18 months project period. The correctness and inter-annotator agreement ought therefore to be high, although we have no numbers to show it, and must admit as well that we do still find mis-annotations.

Most of the corpus consists of dialogue. We have taken this genre seriously, and gone to great lengths to annotate turn taking, overlaps, interruptions etc. This choice has slowed down the transcription process considerably, but we think has also added to the general value of the corpus.

A picture of a dialogue sequence with informants is shown in figure 11.

3 The Search Interface

3.1 Limiting Search w.r.t. Informants

It is possible, and indeed easy, to limit the search to subgroups of informants. One main choice is between types of recording; free dialogue vs. semi-formal interview:

Figure 2. Limiting searches w.r.t. recording type.

Furthermore, it is possible to limit the subgroup of informants according to all the informant variables, such as gender, age, place of residence, place of birth, work, educational background:

Figure 3. Choosing subgroups of informants.

Ticking off some of the boxes will lead to the popping up of new and more detailed ones. The idea behind the gradually more specific choices is to keep each interface no more complex than the user needs, while at the same time allowing even

[4] Ordinary cough resulting from illness is not annotated.

advanced, complex searches to have a user-friendly interface.

In the figure above, the various boxes expands into new menus (e.g. those that refer to place of birth or residence) or to new boxes for numbers (e.g. for age). Figure 4 shows how having ticked off the box for *yrke* ('work') – also found in the figure above – has expanded the choice with several more subcategories, for types such as *håndverk/yrkesfag* ('trade'), *service*, *kontor* ('office'), frie yrker ('free trades'). (The categories have been adopted from the state agency Statistics Norway.)

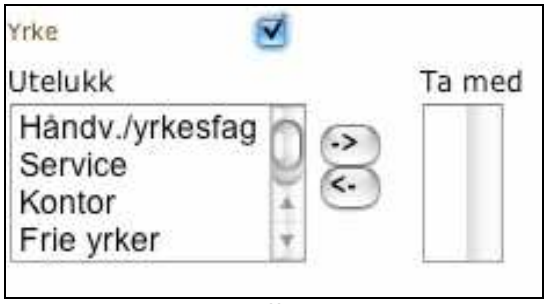

Figure 4. Ticking off a choice such as yrke *('work'), expands the choice into subtypes.*

3.2 Limiting Search w.r.t. Linguistic Criteria

A maximum level of user-friendliness has been attempted at all levels, given that the users will generally be non-technical linguists who are opposed to going through a long period of learning how to use such tools. We support the ideas advanced by Johannessen, Hagen and Nøklestad (2000), in which regular expressions for any kind of simple or complex search are to be avoided for non-technical users. User-interfaces for machine-human dialogue should be based on boxes and menus, not complicated query languages. Below is a search interface of the simplest kind – for just one or two words:

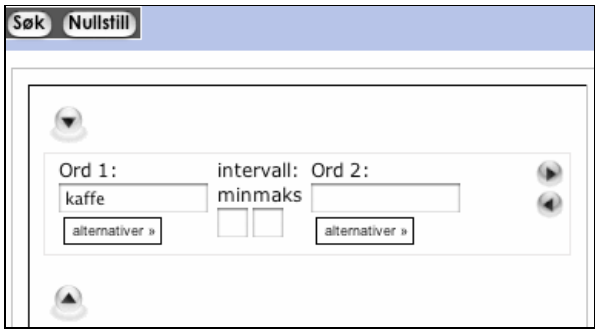

Figure 5. Search interface for linguistic strings.

In order to increase the number of search words, the user clicks on the arrow on the right hand side, and more boxes will appear. In order to search for alternative words, the user clicks on the arrows below, to get more boxes along that dimension.

By pulling down a menu at a word, more options will appear. Since the corpus is part-of-speech (POS) tagged, one option is to choose part of speech (*ordklasse*). It should be noted that POS can also be chosen without an accompanying specification of a word or part of word, giving the user a frequently wanted search option. In this respect it is superior to many other corpora, whether written or spoken language ones. Thus, a user can choose, for example, to get all the nouns in the corpus.

The advantage of this search option is unquestionable. For a linguist who studies the behaviour of a particular part of speech in its context, being able to get a concordance of all instances of that category, gives great opportunities for empirical exploration.

It is of course possible to specify only parts of words, such as the beginning, the middle, or the end. Given that the corpus is lemmatised, it also possible to specify a search for all words belonging to the same paradigm, by choosing 'lemma'. Below is an example of how to choose POS with no specified word or string of letters.

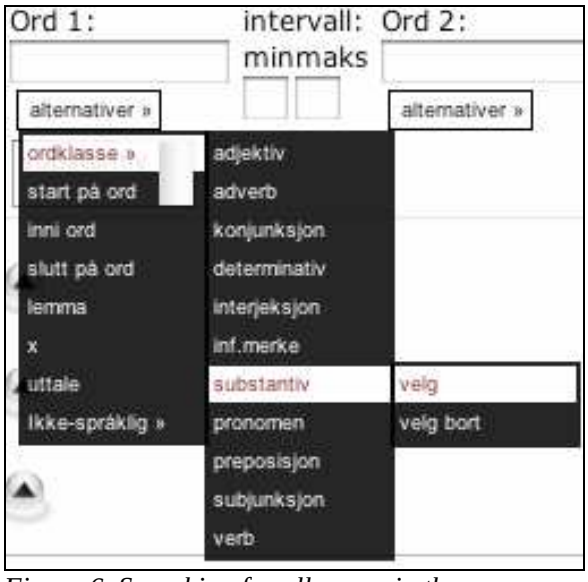

Figure 6. Searching for all nouns in the corpus.

Notice in Figure 6 that we have chosen the *velg* option ('choose'). We could also have chosen the negative *velg bort* ('exclude'), a useful feature to exclude a part of speech from a particular search context. Below are shown a very small subset of the resulting 85233 hits for this search.

nei nei på	**stadion**	
r) * oppå	**tribunen**	?
a jeg sto i	**mål**	da
de en bra	**jobb**	det var
let var litt	**forskjell**	fra # den c
len der m	**rævkampen**	mot Hviter
npen mot	**Hviterussland**	
or mange	**stadion**	tar jeg
ar det på	**stadion**	i hvert fall
e ord) en	**kompis**	av meg so

Figure 7. Results from a search on all nouns.

Below is an example where we have chosen to search for all instances of the irregular verb *være* ('be'), regardless of inflection, followed by a preposition. This time, we have written the infinitive form (the dictionary look-up form) of the verb in the first box, and made sure we have chosen the alternative *lemma* from the menu. The second box is empty, but the alternative POS *preposition* has been chosen from the menu.

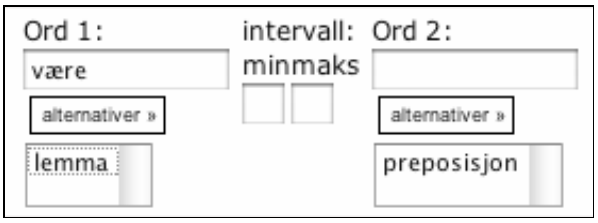

Figure 8. Search for all inflectional forms of the verb være *('be') followed by a preposition.*

The corpus yields 3135 results, some of which are shown below:

har du noen gang	**vært i**	Tromsø (u
jeg har	**vært i**	Tromsø ja
ksom alltid hadde	**vært der oppe**	jeg
r en kamerat som	**er fra**	Tønsberg
uten at Lyn	**var i**	finalen en
ja de # de	**er i**	semien de

Figure 9. Some results from searching for the lemma være *('be') followed by a preposition.*

4 Comparison with Other Speech Corpora

It is instructive to compare the NoTa corpus with other speech corpora. Such corpora are generally expensive to develop, and more so if they are to have a variety of different features. For this reason existing speech corpora do not necessarily have as many advanced features as their developers and users would have liked.

In this section, we will compare the NoTa corpus with three other Scandinavian speech corpora: the Swedish Gøteborg Spoken Language Corpus (GSLC), the Danish BySoc Corpus, and a small dialect corpus of Norwegian (Talesøk). We will also compare it with the British National Corpus (BNC), possibly the most widely known speech corpus available, and the Scottish Corpus of Text and Speech (SCOTS), a new speech corpus with many nice features. See the Reference section for all URLs.

The corpora vary somewhat in size (up to 2 million words, except for the BNC, which is 10 million words), but they have in common that they have been updated after 2000, and that they all aim at a wider audience of non-technical experts.

The table does not reflect reality in every detail: We have ticked off "yes" for multimedia representations in the SCOTS corpus, although the texts in that corpus vary w.r.t. this variable. Also, we have written "no" for tagged transcriptions in Talesøk, since the tagging that exists for that corpus are not available from the main search interface.

	NoTa	Talesøk	GSLC	BySoc	BNC	SCOTS
Transcription linked to audio	Yes	Yes	No	No	No	Yes
Transcription linked to video	Yes	No	No	No	No	Yes
User-friendly search without regular expressions	Yes	Yes	No	No	No	No
Possible to limit informant selection	Yes	Yes	Yes	Yes	Yes	Yes
Overlaps/ turntaking annotated	Yes	Yes	Yes	Yes	No	Yes
Transcription as standard orthography (or slightly modified)	Yes	Yes	Yes	Yes	Yes	Yes
POS tagged	Yes	No	No	No	Yes	No
POS tags can be used as the only search expressions	Yes	–	–	–	No	–

Figure 10. A comparison between the NoTa corpus and five other speech corpora.

The table shows that the NoTa corpus compares favourably with the other corpora w.r.t. the variables we have chosen. This is of course related to the fact that the NoTa corpus is the newest one, and we have been able to learn from the other corpora. Also, the general technical advances have made it possible to offer features that would have been unthinkable only a few years ago. We have chosen variables that have been important to us as developers. However, we think that these features are important to many other researchers, too.

5 Access

Corpus search via the corpus web site (see Reference section for URL) is available for all researchers. Information about how to get a password is also given there.

The corpus can also be downloaded to see the full transcriptions and view and listen to the full recordings. Furthermore, full-scale versions can be downloaded for other purposes, such as language technology research and development. Contact information is given on the web site.

6 Conclusion and Future Work

We believe that we have developed a speech corpus that will be valuable to linguists as well as technologists, both due to its technical features and its contents. Its main use will, we think, be the corpus with its user-friendly web interface, but the transcriptions, audio files, and tools and resources developed as part of the project will all be useful for other researchers.

There are mainly two paths that we plan to follow in the future. One is to syntactically parse the corpus. So far, some preliminary work has been done w.r.t. pre-processing (see Johannessen and Jørgensen 2006, Jørgensen 2007).

The other path we hope to follow is expanding the corpus. We are expanding it at the moment by adding more speech material of young urbans via cooperation with the project UPUS. We are also planning to add material from other big cities (Bergen, Trondheim, Tromsø), and dialect material from rural areas. The latter task has started in connection with cooperation within the Nordic Centre of Excellence in Microcomparative Syntax, NORMS, and the ScanDiaSyn network.

We also hope to evaluate the corpus.

Figure 11. An example of dialogue in a multimedia window.

References

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

BySoc Corpus: Danish Vernacular – Dansk Talesprog. http://www.id.cbs.dk/~pjuel/cgi-bin/BySoc_ID/index.cgi?EeNnGg

Bødal, Anne Marit, Hilde Cathrine Haug, Ingunn Indrebø Ims and Signe Laake**.** To appear. Dilemma ved ortografisk transkripsjon. In Johannessen and Hagen (eds.).

Gøteborg Spoken Language Corpus: http://www.ling.gu.se/projekt/tal/index.cgi?PAGE=3

Hagen, Kristin. 2005. Transkripsjonsveiledning for NoTa-Oslo. Ms. The Text Laboratory, Univ. of Oslo. http://www.tekstlab.uio.no/nota/oslo/index.html

IMS Corpus Work Bench: http://www.ims.uni- stuttgart.de/projekte/CorpusWorkbench/OldDocus/FAQ.html

.Johannessen Janne Bondi, and Kristin Hagen (eds.). To appear: *Språk i Oslo*, Novus forlag, Oslo.

Johannessen, Janne Bondi, Lars Nygaard, Kristin Hagen, Hanne Gram Simonsen. 2005. Transkripsjon i et talespråkskorpus Paper presented at MONS 11, Bergen.

Johannessen, Janne Bondi, Kristin Hagen and Anders Nøklestad. 2000. A Web-Based Advanced and User Friendly System: The Oslo Corpus of Tagged Norwegian Texts. In Gavrilidou, M., G. et al. (eds.) *Proceedings, Second International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, 1725-1729.

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

Jørgensen, Fredrik. To appear. Ytringer, setninger, fragmenter og feiltyper. In Johannessen and Hagen (eds.).

Nordic Centre of Excellence in Microcomparative Syntax, NORMS. http://norms.uit.no/

NoTa Norwegian Speech Corpus. The Text Laboratory, ILN, University of Oslo. http://www.tekstlab.uio.no/nota/oslo/index.html

Nygaard, Lars. 2007. *Glossa – The Corpus Explorer, Version.0.9*.http://www.hf.uio.no/tekstlab/glossa.html

Nøklestad, Anders and Åshild Søfteland, UiO. To appear. Manuell morfologisk tagging av NoTa-materialet med støtte fra en statistisk tagger. In Johannessen and Hagen (eds.).

Oslo-Bergen Tagger. http://omilia.uio.no/obt/

Scandinavian Dialect Syntax, ScanDiaSyn. http://uit.no/scandiasyn

Scottish Corpus of Text and Speech. http://www.scottishcorpus.ac.uk/

Spoken Dutch Corpus (CGN).

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

Talesøk.
http://helmer.aksis.uib.no/talekorpus/Hovedside.htm

Trascriber:
http://trans.sourceforge.net/en/presentation.php

UPUS.
http://www.hf.ntnu.no/hf/adm/forskning/prosjekter/UPUS

Wangensteen, Boye. 2005. Edited: *Bokmålsordboka: definisjons- og rettskrivningsordbok*. Kunnskapsforlaget, Oslo.