# Case Study: The AusTalk Corpus

Steve Cassidy, Dominique Estival and Felicity Cox

## Abstract

This chapter presents detail of the Annotation Task of the Big Australian Speech Corpus (Big ASC) project, in which AusTalk, a large audio-visual corpus of Australian English, was collected. We describe the scope of the task and its implementation and give an overview of the results so far. When complete, AusTalk will consist of 3 h of audio-visual recording from each of 1000 speakers of Australian English, across a wide range of tasks including scripted (read) speech, spontaneous speech and dialogue. The read speech of 100 participants has now been manually annotated but a challenge of the project was to produce transcriptions for the unscripted (spontaneous) speech data. We report on several avenues that have been explored for the automation of this task. We describe the annotation challenges, the processes that were adopted and the limitations of automated transcription.

## Keywords

Speech corpus · Australian English · Large corpora · Spontaneous speech

S. Cassidy (✉)
Department of Computing, Macquarie University, Sydney, NSW, Australia
e-mail: steve.cassidy@mq.edu.au

D. Estival
MARCS, University of Western Sydney, Sydney, NSW, Australia

F. Cox
Department of Linguistics, Macquarie University, Sydney, NSW, Australia

N. Ide and J. Pustejovsky (eds.), *Handbook of Linguistic Annotation*,
DOI 10.1007/978-94-024-0881-2_49

# 1 The Big ASC Project

The Big Australian Speech Corpus (Big ASC) is a collaborative project between 11 institutions, funded by the Australian Research Council (total budget of A$1.5M), with twin goals to (1) provide a standardized infrastructure for audio-visual (AV) recordings and (2) produce a large AV corpus of Australian English (AusE). The project planned to record up to 1000 geographically and socially diverse speakers in locations across Australia using 12 sets of standardized hardware and software (the Black Box) with a uniform and automated protocol (the Standard Speech Collection Protocol – SSCP) to produce the AusTalk corpus [5,20]. As of this publication, 90% of the data have been collected.

The overarching purpose of the project was to provide an extensible database to facilitate research charting the extent, degree, and detail of social, regional, ethno-cultural and stylistic variation in AusE [7,8,11–13] and to describe changes to the language since the collection of the outmoded ANDOSL corpus [15]. However, it was also designed to cater for a range of other research projects and applications in linguistics, speech science and language technologies. In Australia we have considerable research strengths in the speech sciences but a sufficiently large and current corpus to support this work was not available. The rationale for the corpus design was the imperative to cater for a range of different constituencies: phonetics, forensic studies, language technologies, linguistic analysis, audio-visual analysis. Thus, the project required an innovative solution to the demands of both high quality audio/video recording and field data collection, and it had to include both standardised read speech and elicited natural spontaneous speech. To this end 6 channels of audio (from 1 desk mic, 2 room mics, and 2 headset mics) and 2 channels of video (from 2 stereo cameras) were captured resulting in a rich dataset that can be used for a wide range of different purposes (see [5,20] for details).

AV corpora such as AusTalk are important for Natural Language Processing and Language Technologies in several respects. Not only does AusTalk provide audio and video data allowing research in audio-visual speech processing, such as the use of facial cues for speech recognition, but it contains data from a range of both read and spontaneous speech tasks. Specific tasks, such as the Interview, Map Task and Conversation tasks, provide data for the analysis of speech acts in dialogues, while the Read Story (the well-known "Arthur the Rat" passage modified to suit AusE) and Re-told Story tasks were designed for the study of differences between reading and spontaneous language. Details of these tasks are presented below.

Two important requirements for the ongoing utility of the Big ASC project were to make AusTalk widely available and to allow future contributions, including augmentation with further data and further annotations. Audio and video data are stored on a web-based repository that supports meta-data search with the ability to browse and download of the audio data. The data is now also available via the Alveo Virtual Laboratory [10] which supports the upload of new annotations which can then be published alongside the original data and annotations.

## 2 The AusTalk Corpus

When complete, the AusTalk corpus will comprise 3000 h of speech data from a total of 1000 AusE speakers, all having completed their primary and secondary schooling in Australia (but not necessarily having been born in Australia), a criterion ensuring inclusion of a range of speakers from various cultural backgrounds. 90% of the targeted data have now been collected at 14 different sites in major cities around Australia (Adelaide, Sydney, Perth, Brisbane, Melbourne, Hobart, Darwin) and in several regional centres. Data from more than 2300 sessions have been uploaded, comprising a total of 7.6 M files and around 20 TB of data.

### 2.1 Collection Protocol

As part of the anonymisation of the data, each participant was given a unique identifier linked to their name only in the off-line spreadsheet maintained by the Recording Assistants (RAs) at each site. The identifiers consist of a colour name followed by the name of an Australian animal. Each colour and animal also has a numerical value used to generate a short-form name for the participant. For example, participant *Gold - Fuscous Honeyeater* is also identified as *1_371*. We expect that most researchers will use the short-form numerical names, but we maintain the link to the longer animal names so that participants can identify their own contribution to the corpus and gain access to their own recordings.

Prior to the recording session each speaker completed an extensive online questionnaire to collect a comprehensive set of demographic, family, historical and language background data. Each speaker was recorded over three 1-hour sessions, separated by at least one week to capture natural variation in voice quality. Each session comprised a series of both read and spontaneous speech tasks to capture style shifting from highly formal word-list to more informal spontaneous conversation. In the third and final session, speakers were paired for two Map Tasks along the lines of [1] but re-designed for Australian English. The components of the corpus and the time taken for each task across the three recording sessions (S1, S2, S3) are shown in Table 1.

Spontaneous speech makes up approximately half of the collected data with a minimum of 40 min per speaker (Yes/No responses, Interview with RA, Re-told Story) and 40 min for 2 Map Task interactions with another participant as partner, followed by 5 min of conversation with that partner (see [5,20] for details).

All recordings were made on the Black Box, a dedicated computer with audio and video interfaces configured in a portable equipment rack that could be moved between sites if needed. Software on the Black Box was designed to run the collection protocol and display prompts simultaneously on dual screens - one for the RA running the session, and one for the participant being recorded. The software was responsible for management of the components listed in Table 1 by sequentially prompting for each word or sentence and directly recording the audio and video channels to disk. After each item was recorded, files were saved on disk and a metadata record (which

**Table 1** The AusTalk corpus components

| | Component (# Items, × Session) | Time per session (min) | Total time per speaker (min) |
|---|---|---|---|
| Read speech | Words (322 items, × 3: S1, S2, S3) | 10 | 30 |
| | Digit strings (12 items, × 2: S1, S2) | 5 | 10 |
| | Sentences (59 items, × 1: S2) | 8 | 8 |
| | Read story (1 item, × 1: S1) | 5 | 5 |
| Spontaneous speech | Yes/No answers (12 items, × 3: S1, S2, S3) | 3 | 10 |
| | Re-told story (1 item, × 1: S1) | 10 | 10 |
| | Interview (1 item, × 1: S2) | 15 | 15 |
| | Map Task (2 items, × 1: S3) | 20 | 40 |
| | Conversation (1 item, × 1: S3) | 5 | 5 |

included the time of recording and the text of the prompt) was written. The file names used to save the data were structured to include information about the item and some meta-data. For example, the file `1_207_1_11_002-ch6-speaker.wav` was recorded from speaker 1_207, in session 1, component 11, item 2 and contains audio from channel 6 (the speaker headset microphone). Files were grouped into a separate directory per component and these in turn were grouped by session and by speaker.

As a separate process, after each recording was made an MD5 checksum was calculated for each file and stored with the item metadata. The checksum enabled us to validate the data as it was uploaded to the central server or moved around to other storage locations.

During the recording process, video data from one or two of the stereo cameras was written to disk in raw format. For storage purposes, this data was compressed using the MPEG-4 codec through the open source ffmpeg software (http://www.ffmpeg.org) to generate a more manageable file size. Even so, some of the longer recordings resulted in a 2G+ video file. Once a session was complete, data from the entire session were uploaded to a central server via an automated script that interacts with a custom web application. As part of the upload process a manifest was first uploaded for each session followed by the data for each item. When complete, the server validated that all files were present and all checksums were correct. If errors were reported, the upload process could be re-run to capture any files that were missed the first time around; this occasionally occurred for large video files particularly when the upload was interrupted by network issues.

The central web server generated a series of upload reports. This allowed the RA at each site to verify the safety of their data and also facilitated tracking of progress by the project management team. The upload reports included the result of a validation process for each session so that any issues with missing or corrupted data could be identified and corrected quickly.

## 2.2 Quality Control

To ensure data quality as well as consistency across all the sites, several processes were implemented. First, prior to commencing data collection, all the RAs attended a 2 day training workshop where they practiced setting up the equipment and running through the recording sessions. Inevitable delays and changes in staffing lessened the positive impact of this centralised training to some extent and additional training was required when new RAs were recruited. Training was an essential factor in the process to ensure consistency of the data collection protocol. Second, each recording site made sample recordings that were checked by the management team for audio and video quality before the start of data collection at that particular site.

Third, there was continuous monitoring of data quality so that feedback and advice could be given to the RAs throughout the corpus collection. A Quality Control RA (QC-RA) employed at the central receiving site where the data was uploaded used a set of strict guidelines to check the quality of both audio and video data. To assist the site RAs and the QC-RA, we developed a utility to check the number of files and the presence of certain parameters, such as silence or loudness for audio, and frame skipping or brightness for video. The utility was run over the uploaded data for each a recording session and would alert the RAs to potential problems that could then be manually investigated.

The QC checks remain part of the metadata provided with the corpus and all the published data has ratings for video and audio quality (A, B, C or D with A as the highest quality and D as the lowest quality) associated with every component, with meaning as follows:

- A (A-OK)
- B (OK, but imperfect)
- C (bad, not acceptable)
- D (deficient or missing)

Any significant issues with data quality are noted in comments that are also included in the metadata.

# 3 The AusTalk Annotation Task

The Annotation Task could not be commenced until sufficient data had been collected and organized. In this section, we first delimit the scope of the Annotation Task, then describe the processes we have put in place and the annotations that have been produced before briefly discussing the main challenges we faced.

## 3.1 Scope

The original goal of the Big ASC project was to provide at least a base level of annotation (orthographic and phonemic – speech segment – transcription) for all the data collected. Given the volume of data and the limited budget, it was necessary to explore automated processes, while still providing high quality phonemic time-aligned manual annotation for a subset of the data. Hence, the approach we took for the annotation task was to consider using forced alignment for automatic phonemic segmentation and annotation of read speech and to explore the possibility of automatic orthographic transcription of spontaneous speech.

It was also important to create a storage environment that would facilitate the importation of newly created annotations (e.g. additional phonemic transcriptions, detailed phonetic transcriptions, intonation transcriptions, part-of-speech tagging) which could be contributed later by project partners or other researchers.

**Table 2** Number of speakers annotated at the phonemic and orthographic levels for read speech

| | 322 Words S1 | 322 Words S2 | 322 Words S3 | 59 Sentences S2 | Story (645 words) S1 |
|---|---|---|---|---|---|
| Manual orthographic | | | | 96 | |
| Manual phonemic transcriptions | | | | 96 | |
| Manual time-aligned phonemic praat TextGrid | 5 | 5 | 5 | 35 | 5 |
| Corrected time-aligned phonemic praat TextGrid | 13 | 13 | 13 | 9 | |
| Automatically generated MAUS TextGrids | 33 | 34 | 17 | | |

**Table 3** Number of speakers with orthographic transcription of spontaneous speech

| | Re-told story | Interview | Map task |
|---|---|---|---|
| Manual orthographic | 92 | 95 | 62 |

The result of the annotation task that has been possible within our budget is summarized in Tables 2 and 3. While this represents only a subset of the overall corpus, it provides us with a core of annotated data to support research and establishes the standard for further annotation work when funds become available.,As a result of the extensive preliminary work carried out to establish protocols for annotation we are now able to produce automatic forced-alignment phonemic transcriptions for all of the read speech in the corpus in addition to the manual annotations listed. Automated annotations will be added in due course. Together this amounts to around **8.7 h** of speech with aligned phonetic transcription and around **44 h** of spontaneous speech with orthographic transcription.

### 3.2 Training MAUS

Our goal was to make use of a forced-alignment tool to generate time-aligned phonemic annotations for this large data set. There are a number of such tools available now, notably the Penn Phonetics Lab Forced Aligner [21] and the Munich Automatic Segmentation (MAUS) system [16]. We chose to work with MAUS as we had links to the authors of this tool and they were keen to work with us to improve the quality of their aligner and extend its functionality to Australian English (AusE).

A forced aligner is a speech recognition engine used to match a known transcription to an acoustic signal. The orthographic transcription limits the possible interpretations of the acoustic signal allowing the tool to align words and/or phonetic segments with the input waveform. To perform well, the acoustic models in the speech recogniser must be first trained on data similar to that which will ultimately be processed. The language models must be tuned to accommodate the phonetic processes present in the language. MAUS was already trained on English but since there are distinct differences between the diverse varieties of English it was necessary to supply training data which would allow the models to be adapted for AusE. MAUS makes use of SAMPA *(Speech Assessment Methods Phonetic Alphabet)* as its phonemic transcription input in the training phase but SAMPA is dialect specific. We therefore had to create an AusE version of SAMPA (SAMPA-AUS) which contained the phoneme set specific to our AusE corpus. SAMPA-AUS is based on the phonemic transcription system for Australian English recommended by Harrington, Cox, and Evans [14]. Our first task in the annotation phase was to supply the MAUS development team with sufficient training data so that they could generate an AusE version of MAUS that could then be used to automate some of the data annotation.

A set of 100 diverse speakers from whom we had collected a complete data set was selected for the purpose of providing training data for MAUS, henceforth called the 'MAUS speakers'. These speakers would become the core set for the manual annotation work that would be conducted on the corpus.

The first task was to generate canonical phonemic transcriptions for each of the 59 read sentences. Based on each sentence elicitation prompt, we generated idealized citation-form phonemic transcriptions in SAMPA-AUS. This allowed for the creation of a small lexicon that had coverage of all words included in the sentences. Secondly, we generated an additional set of *connected speech phonemic transcription templates* to more closely reflect the connected speech used in the actual reading task for the 59 sentences. These connected speech transcription templates were created in a format suitable for use in the Transcriber annotation tool [2]. For each of the 100 MAUS speaker's sentences, a Transcriber compatible file was populated with the connected speech phonemic transcription template which was then hand-corrected by our annotation team with reference to the speaker's actual production. Transcriber was used to facilitate playback of the audio but no alignment took place in this case. In total, phonemic transcriptions for the 59 sentences for 96 speakers were checked and corrected. This was necessary to ensure that each speaker's individual sentence phonemic transcription accurately reflected the phonemes used in the actual speech data.. In some cases participants had not properly read the prompt so it was necessary to introduce new words into the individual transcriptions and revisit the citation form transcriptions to supplement the lexicon.

The Transcriber format files were then converted to Praat TextGrid format as required by the MAUS team for training their system. The ultimate new AusE model was then made available via the MAUS web interface (http://www.bas.uni-muenchen.de/Bas/BasMAUS.html) and via the downloadable MAUS software distribution.

### 3.3 Generating a Lexicon

One requirement for running forced alignment based on textual transcriptions is a pronunciation lexicon. From the earlier work where the sentences were painstakingly phonemically transcribed, we had generated a small lexicon (of phonemically transcribed words) which included vocabulary contained within the sentence set. We extended this lexicon to include items from the isolated word and digits elicitation tasks as well as the key landmarks/lexical items from the Map Task. In order to use MAUS on the wider set of data it was necessary to have an even broader coverage AusE lexicon. We therefore investigated a commercial provider who could offer a lexicon for research purposes, but it was not clear whether the conditions of use would allow us to make 'derived' forms from the pronunciation lexicon (such as a trained set of letter-to-sound rules) publicly available. Instead, we were fortunate to discover the typesetting files for an out-of-print Australian English dictionary, the Australian Learners Dictionary [9] and obtain permission from the copyright owner, Macquarie University, to publish the data for research use. The annotation team hand-corrected the dictionary phonemic transcriptions to reflect each word's pronunciation and ensure that the lexicon conformed to the transcription standards that had been adopted in the project. We were therefore able to extract a useful broad coverage pronunciation lexicon from the dictionary.

### 3.4 Correcting MAUS Annotations

The next stage was to make use of MAUS to generate automatic phonemic transcriptions of our read speech data. MAUS provides a web-based interface where audio files uploaded along with associated orthographic transcription are processed to generate Praat TextGrids containing time-aligned phonemic transcriptions. This interface is convenient for single files but since we have many thousands of files we required an automated process. The first approach was to write a script to send the audio files to the web service and store the results. The corpus meta-data was used to determine the prompt for each recording allowing us to send all of the read speech for a speaker to be processed. Unfortunately while this worked well it was very slow, taking a few days to process a batch of data. Fortunately, MAUS is also available as a downloadable package so we were able to run this locally and get a much better throughput – around 10 min per speaker for about 800 files.

Once the results of forced alignment were available, the annotation team began the laborious task of checking and correcting the annotations. Since the output of MAUS uses the Praat [4] TextGrid format, and since our annotation team was familiar with this tool, Praat was used. The task involved opening each of the MAUS TextGrid files and the associated audio file, checking and then correcting both the phonetic transcription and the positions of the segment boundaries. This is a very labour intensive task and has been the most time consuming part of the whole annotation process. However, it is significantly faster than annotating each file from scratch enabling us to generate high quality annotations for a much larger subset of the data. For researchers who intend to conduct detailed phonetic analysis, manual correction is mandatory.

This initial test phase for the MAUS aligner was run with all of the data from a single recording site (University of Canberra). Later, when we were able to run MAUS locally, we processed all of the 100 MAUS speakers' recordings and the annotation team has worked through correcting a subset of these. While we have only been able to hand correct a subset of the data, we will ultimately run MAUS over the entire corpus of read-speech (Words, Sentences, Digits, Read Story) to generate automatic annotations of the data we hold. Figure 1 (top) contains an example of a Praat TextGrid returned following MAUS processing and Fig. 1 (bottom) shows the corrected TextGrid with the boundary of the vowel onset moved left to align with the onset of voicing. The TextGrid tiers contain the orthographic representation of the word (tier 1), the canonical phonemic transcription (tier 2) and the time aligned phonemic representation for the speech segments (tier 3).

In addition to the checking/correcting of automatically generated data, a subset of data has been hand annotated from scratch resulting in a set of data that could be

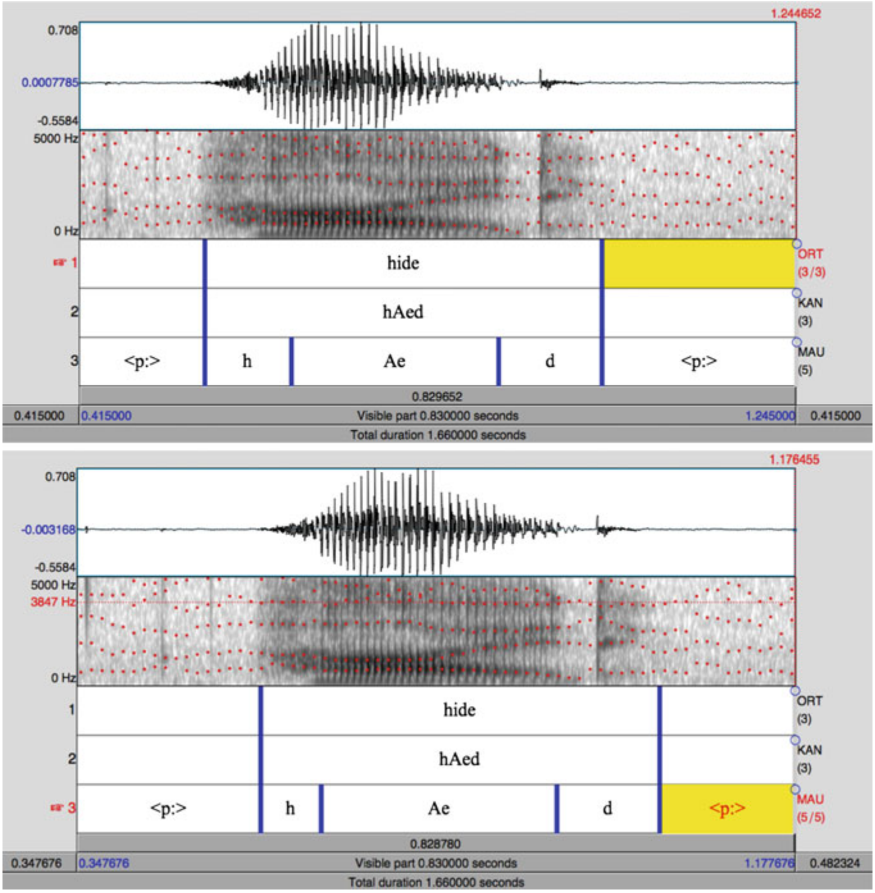


**Fig. 1** (*top*) A Praat TextGrid returned from MAUS for the word 'hide'. (*bottom*) The manually corrected TextGrid for the same item

directly compared with the automatically generated TextGrids. Because the MAUS automatic aligner was not ready to process data until quite late in the annotation process the team spent a large amount of time constructing time aligned annotations in Praat from scratch by hand. This has been one of the most time consuming components of the annotation process but has resulted in manually created phonemically aligned TextGrids for the full set of scripted Words, Sentences and Read Story reading tasks for five speakers. For these five speakers we have also manually created orthographic time-stamped transcriptions for the spontaneous speech (Re-told Story, Interview and Map Task). An additional 35 speakers have also had their full sentence set manually annotated in Praat with time aligned phoneme boundaries. In order to ensure that all annotators were consistent with themselves and with each other, a set of annotation guidelines was created and continually updated throughout the process.

These guidelines were used to ensure consistency for all TextGrid annotations and were used for both manual annotation and checking/correcting of automatically generated annotations. The guidelines will be made available for other researchers to ensure that standardisation of annotation principles continues into the future.

The corpus will ultimately contain three different types of time aligned phonemic annotations: manually generated, MAUS automatically created but manually corrected, and MAUS generated but uncorrected. These will be differentiated from each other so that researchers are aware of the origin of the annotations they are using and will be able to take the necessary steps to ensure the integrity of the data they are working with.

## 3.5 Transcription of Spontaneous Speech

The original goal of the project was to make use of automatic speech recognition technology to provide at least low-quality orthographic transcripts of the spontaneous speech tasks in the corpus: Interview, Re-told story, Map Task and Conversation. We originally started working with a European partner who indicated that they might be able to create orthographic transcriptions for us by adapting their speech recognition engine using some of our early transcribed data. We provided data from five speakers for this training task but the European team was unable to achieve usable performance from their engine. We then attempted to make use of a commercial desktop transcription system but again, the quality of the output was very poor and was judged not to be useful even as a low-quality transcription. A number of trials were undertaken to improve the quality of the output but none proved useful.

The eventual solution has been to make use of a low cost commercial manual transcription service in Australia. They have been able to provide us with high quality orthographic transcriptions of spontaneous speech that include time-stamps on every speaker turn and major pauses. One problem with commercial transcription services can be that they typically produce text for human consumption whereas we are particularly interested in automatic machine processing of our transcriptions. It was therefore important that speaker turns, timestamps and any non-lexical annotation added to the text were created in a consistent manner. We have been able to work with the transcription company to ensure consistency of transcription for our purposes. Spontaneous speech data from the Interview and the Re-told Story from 95 speakers has been processed this way. Here is an example of an exchange between an interviewer and interviewee.

> ***[00:06:15] Interviewer: Hmm. So when you interviewed people, was that all in Indonesian?***
> *[00:06:22] Interviewee: Yeah. It was all in Indonesian. Uhm, it was – uh, I've been doing Indonesian for eight years. So I was, uhm, I was able to converse fairly easily, but the problem with Indonesian is that while I had studied the formal language, there's a million dialects.*

An important component of the spontaneous speech data is the Map Task recordings. These involved two speakers playing an interactive game and will be of particular interest to dialogue researchers and those interested in spontaneous interactions. The annotators were able to manually complete 62 Map Task orthographic transcriptions (time-aligned by speaker turn) using the Transcriber tool.

## 3.6 Organisation and Publication

The end result of this work is a large collection of files and a number of different databases containing descriptive metadata. Since we didn't have the luxury of a long lead-in to this project, many technical decisions on storing and collecting data were made as the project progressed. The end result is that while we were careful to organise the various types of data well using standard formats and systems, there was still work to be done to incorporate the various parts into an integrated whole that could be published.

The final publication of the data was to take two forms: firstly a standalone website and secondly the submission of the data to the Alveo Virtual Laboratory – a new web based repository for Human Communication data in Australia [10]. Since the Alveo project started after the bulk of the data collection was complete we were unable to target it directly in our data organization; however since we were closely involved with its development, we were able to ensure that what we did was compatible with the emerging platform. In both cases we required an integrated version of the corpus meta-data that could be queried and browsed online. To this end we defined a data model based on RDF (the Resource Description Framework from the Semantic Web), a model that is well suited to meta-data representation.

Once the design of the data model was in place, scripts were written to interrogate the various data sources used to store data and meta-data as part of the recording process and bring them together into a unified whole. The different sources of data were:

- Participant meta-data from the web based questionnaire
- Descriptions of the sessions, components and items from the collection protocol (e.g. the prompt for each item)
- Item descriptors stored as XML with uploaded data
- QA ratings from spreadsheets
- Some additional participant data from anonymised RA spreadsheets not included in the questionnaire
- Audio and video file names and locations from the file system
- Annotation file names and locations from the various annotation tasks

All of these data sources are combined into a single RDF description for each item that references the participant metadata and the descriptors for each data file associated with the item. Figure 2 shows an excerpt from such a description. The descriptions generated conform to the requirements of Linked Open Data [3] in that

```
<http://id.austalk.edu.au/item/1_1216_1_5_001> a ausnc:AusNCObject ;
    ausnc:audience ausnc:individual ;
    ausnc:communication_context ausnc:face_to_face ;
    ausnc:componentName "digits" ;
    ausnc:interactivity ausnc:read ;
    ausnc:mode ausnc:spoken ;
    ausnc:speech_style ausnc:scripted ;
    austalk:cameraSN0 "11072149" ;
    austalk:cameraSN1 "11072158" ;
    austalk:component <http://id.austalk.edu.au/protocol/component/5> ;
    austalk:prompt "zero one two three" ;
    austalk:prototype <http://id.austalk.edu.au/protocol/item/5_1> ;
    austalk:session "1" ;
    austalk:version "1.5.2" ;
    dc:created "Thu Mar 22 10:19:15 2012" ;
    dc:identifier "1_1216_1_5_001" ;
    dc:isPartOf <http://id.austalk.edu.au/component/1_1216_1_5>,
        austalk:corpus ;
    dc:title "1_1216_1_5_001" ;
    olac:speaker <http://id.austalk.edu.au/participant/1_1216> .
```

**Fig. 2** A sample RDF description of a single item in the AusTalk corpus

everything that is described has a URL and that URL references a description of the entity (items, speakers, etc.).

These descriptors are then uploaded to an RDF database and are used to present a unified web-based view of the corpus. This is available at http://bigasc.edu.au and currently makes all of the audio data available after user registration. This website provides facilities to browse and search the meta-data, listen to recordings online and download data in batches.

The Alveo Virtual Laboratory [6] is a web-based repository for Human Communication data that provides a rich API to support building tools to query and analyse data that it holds. It was designed with the publication of the AusTalk corpus in mind and now holds the entire audio collection along with the associated meta-data. The Alveo API allows a richer set of search operations than the AusTalk website and is better tuned to support download of small and large subsets of the data [10]. Alveo supports links to Python and R environments for data analysis including the Emu/R toolkit for speech data analysis and visualization. The development of these tools is ongoing.

## 4 Conclusion and Future Work

Annotation is an important aspect of the Big ASC project and other similar projects for, without it, many of the applications and much of the proposed research could not be conducted. While the ideal of providing full annotations of 100% of the data will not be realised in this phase of the project, we are able to provide a full set of manually created time-aligned phonemic and orthographic transcriptions for read speech data for a selected number of speakers. Based on the work we have done with MAUS, we will also be able to provide automatically time-aligned phonemic transcriptions for all the read speech data. Manually generated orthographic transcriptions are available for a subset of the spontaneous speech data and these will be processed in MAUS to generate automatic time-aligned phonemic transcriptions.

The AusTalk data collection will continue in 2015 in order to complete the corpus containing 3000 h of AV data. Follow-on projects have already begun to collect data from different population groups (e.g. Chinese speakers in Canberra) and the analysis of AusTalk data is under way at other partner sites, e.g. video analysis for facial gestures [17–19] and close phonetic analysis of the isolated word list data. The AusTalk annotation task itself will continue until the data for the selected 100 MAUS speakers has been annotated as described above.

Meanwhile, the AusTalk corpus is now included in Alveo, a recent Australian collaborative project [6] that provides a platform for easy access to language, speech and other cognate databases along with integrated use of a range of analysis tools. This will allow the production of automated Part-of-Speech tagging and syntactic analyses as additional annotations for the corpus.

## References

1. Anderson, A.H., Bader, M., Bard, E.G., Boyle, E., Doherty, G., Garrod, S., Weinert, R.: The HCRC map task corpus. Lang. Speech **34**(4), 351–366 (1991)
2. Barras, C., Geoffrois, E., Wu, Z., Liberman, M.: Transcriber: development and use of a tool for assisting speech corpora production. Speech Commun. **33**(1–2), 5–22 (2000)
3. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. Int. J. Semant. Web Inf. Syst. **5**(3), 1–22 (2009). doi:10.4018/jswis.2009081901
4. Boesma, P., Weenink, D.: Praat: doing phonetics by computer (Version 5.1.05) (2009). http://www.praat.org/
5. Burnham, D., Estival, D., Fazio, S., Cox, F., Dale, R., Viethen, J., Wagner, M.: Building an audio-visual corpus of Australian English: large corpus collection with an economical portable and replicable Black Box. Paper presented at the Interspeech 2011, Florence (2011)
6. Burnham, D., Estival, D., Bugeia, P., Sefton, P., Cassidy, S.: Above and beyond speech, language and music: a virtual lab for human communication science (HCS vLab). NeCTAR (National eResearch Collaboration Tools & Resources) Virtual Laboratory (2012)
7. Butcher, A.: Levels of representation in the acquisition of phonology: evidence from 'before and after' speech. In: Dodd, B., Campbell, R., Worall, L. (eds.) Evaluating Theories of Language: Evidence from Disordered Communication, pp. 55–73. Whurr Publishers, London (1996)

8. Butcher, A.: Linguistic aspects of Australian aboriginal English. Clin. Linguist. Phon. **22**(8), 625–642 (2008). doi:10.1080/02699200802223535
9. Candlin, C., Blair, D.: Australian Learners Dictionary. National Centre for English Language Teaching and Research, Australia (1997)
10. Cassidy, S., Estival, D., Jones, T., Burnham, D., Berghold, J.: The alveo virtual laboratory: a web based repository API. Paper presented at the 9th language resources and evaluation conference (LREC 2014), Iceland (2014)
11. Cox, F., Palethorpe, S.: Regional variation in the vowels of female adolescents from Sydney. Paper presented at the ICSLP 1998, Sydney (1998)
12. Cox, F., Palethorpe, S.: The changing face of Australian English vowels. Varieties of English around the World: English in Australia, pp. 17–44. John Benjamins, Netherlands (2001)
13. Cox, F., Palethorpe, S.: The border effect: vowel differences across the NSW/Victorian border. In: Moskovsky, C. (ed.), Proceedings of ALS 2003 (2004)
14. Harrington, J., Cox, F., Evans, Z.: An acoustic phonetic study of broad, general, and cultivated Australian English vowels. Aust. J. Linguist. **17**, 155–184 (1997)
15. Millar, J. B., Dermody, P., Harrington, M., Vonwiller, J.: A national database of spoken language: concept, design, and implementation. Paper presented at the international conference on spoken language processing (ICSLP-90), Japan (1990). http://andosl.anu.edu.au/andosl/ANDOSLhome.html
16. Schiel, F., Draxler, C., Harrington, J.: Phonemic segmentation and labelling using the MAUS technique. Paper presented at the Workshop 'new tools and methods for very-large-scale phonetics research', University of Pennsylvania, Philadelphia (2011)
17. Sui, C., Haque, S., Togneri, R., Bennamoun, M.: A 3D audio-visual corpus for speech recognition. Paper presented at the SST2012, Sydney (2012a)
18. Sui, C., Haque, S., Togneri, R., Bennamoun, M.: Discrimination comparison between audio and visual features. Paper presented at the Asilomar 2012, Pacific Grove (2012b)
19. Togneri, R., Bennamoun, M., Sui, C.: Multimodal speech recognition with the AusTalk 3D audio-visual corpus. Tutorial at Interspeech 2014, Singapore (2014)
20. Wagner, M., Tran, D., Togneri, R., Rose, P., Powers, D., Onslow, M., Ambikairajah, E.: The big Australian speech corpus (The Big ASC). Paper presented at the 13th Australasian international conference on speech science and technology, Melbourne (2010)
21. Yuan, J., Liberman, M.: Speaker identification on the SCOTUS corpus. Paper presented at the Acoustics 2008 (2008)

# Annotations in the Nordic Dialect Corpus

Janne Bondi Johannessen

**Abstract**

In this chapter I focus on annotation in the Nordic Dialect Corpus, a dialect corpus that consists of dialectal speech from five closely related languages. There are two main types of annotation that are central: the annotation of speech itself, i.e. transcription, and the annotation of grammatical categories, i.e. tagging. Both are described and discussed, with a special focus on the success, or lack thereof, of some key choices.

**Keywords**

Linguistic basis · Speech corpus · Nordic languages · Transcription · Tagging · Maps

I would like to thank language engineer Kristin Hagen at the Text Laboratory, UiO, for her central role in the annotation part of the project and for her constructive comments on this paper. Also, the work of our colleagues, computer scientists Anders Nøklestad and Joel Priestley, has been indispensable. Too many people to be mentioned here have taken part in the transcription work – their names can be found here: http://www.tekstlab.uio.no/nota/scandiasyn/transcription.html.

J. Bondi Johannessen (✉)
The Text Laboratory and MultiLing, Department of Linguistics and Nordic Studies,
University of Oslo, 1102 Blindern, 0317 Oslo, UiO, Norway
e-mail: jannebj@iln.uio.no

N. Ide and J. Pustejovsky (eds.), *Handbook of Linguistic Annotation*,
DOI 10.1007/978-94-024-0881-2_50

## 1 Introduction

In this chapter I will discuss problems and solutions related to two types of annotation in the Nordic Dialect Corpus [15–19], which was launched at the end of 2011.[1] The corpus is designed to facilitate studies of linguistic variation; this is costly, but also rewarding for the linguists who use the corpus. Since the dialect corpus is a speech corpus, many of the challenges are related to transcription, which is one type of annotation focussed on here. Since the corpus is to be used for linguistic research, general searches in the corpus via grammatical categories had to be possible, so grammatical tagging is the second type of annotation that will be discussed. Grammatical tagging of speech is generally hard, since most taggers are trained on well-behaved written language that follows well-known and explicit norms. Further, since the dialect corpus consists of five languages, there are some additional challenges that we will comment on.

## 2 About the Nordic Dialect Corpus

The Nordic Dialect Corpus (NDC) was initiated by linguists from universities in six countries – Denmark, Faroe Islands, Finland, Iceland, Norway, and Sweden – within the research network Scandinavian Dialect Syntax (ScanDiaSyn). The aim was to collect lots of speech data and have them available in a corpus for easy access across all the Nordic languages. There were two reasons that this was a good idea: First, the Nordic languages are very similar to each other, and can to some extent be regarded as dialects of the same language. Second, the study of dialect syntax had suffered over the years, and the hope was that with lots of new material, new studies would emerge.[2]

The work started in 2006 and the corpus was launched in 2011. It covers five languages (Danish, Faroese, Icelandic, Norwegian, and Swedish). Most of the recordings have been done after 2000, but some additional Norwegian ones are older; from 1950–1980. There are altogether 228 recording places and 821 informants. All the recordings consist of conversations; at least one dialect speaker talking to either a research assistant or another dialect speaker.

The overall number of transcribed words is 2.8 million. The corpus has been very costly to build because of the manpower needed. As an example, transcribing the Norwegian part alone took 18 people to do, and more than 35 people have been

[1] URLs for the online tools and resources are provided after the list of references at the end of this chapter.

[2] Now that the project has ended, it is clear that the project has indeed led to a lot of new knowledge. Instead of mentioning all the studies here, I will simply point to the new web site *Nordic Atlas of Language Structure Online (NALS) Journal*, which has several tens of scholarly articles on various phenomena in morphology and syntax, with accompanying maps showing isoglosses that have never before been known.

**Table 1** The basic corpus statistics of NDC

| No. of languages | No. of informants | No. of words | No. of places | Year launched | Time of most recordings | No. of transcription types in Norwegian part |
|---|---|---|---|---|---|---|
| 5 | 821 | 2.8 mill. | 228 | 2011 | after 2000 | 2 |

involved in the recording work in Norway only, which included a lot of travel and organising. The Swedish recordings were given to us by an earlier phonology project, Swedia 2000. But all the way, several national research councils, Nordic funds, and individual universities, have contributed. The Text Laboratory at the University of Oslo (UiO) has been responsible for the technical development. The main numerical facts about the NDC are summarized in Table 1.

We know of no other speech corpus that has the combination that the NDC has of double transcriptions, easy search-interface, direct links to audio and video, maps, results handling, and availability on the web. There are other interesting resources with some of the features we have mentioned for other languages. The Scottish Corpus of Text and Speech contains over 4.5 million words, of which 23% is spoken, transcribed and linked to audio. The British National Corpus contains 10 million words of spoken English, which have been categorised into 28 different dialects. The sound files are transcribed orthographically and grammatically tagged, and many recordings, including naturalistic ones, have been made available recently. 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, with read sentences and meta-linguistic discussions. The C-ORAL-BRASIL I [25] is an informal spoken Brazilian Portuguese reference corpus available on DVD, transcribed and with audio and transcription aligned.

The NDC has been integrated in the Glossa corpus search system [11,14], which has user-friendly, yet advanced, options for searching and results handling, and with easy links between transcriptions and audio and video. Nothing more will be said about Glossa here, but wherever there are figures depicting searches or results, these are from that interface.

## 3 Annotation I: Transcription

### 3.1 Two Types of Transcription

In order for a speech corpus to be used, it is necessary to transcribe the spoken language into a written representation, where the conversations have to be transcribed word by word. To be able to search in a corpus, it has to be transcribed to a standard orthography.[3] All the recordings are therefore transcribed orthographically. However, the Nordic Dialect Corpus has been developed in order to facilitate linguistic studies in individual and dialectal variation. There are many linguistic purposes, not only phonological, but also morphological or syntactic ones, where it is desirable to have a phonetic transcription. Thus for all the recordings of Norwegian (for which there was sufficient funding) and for the dialect of Övdalian in Sweden (which is almost like a different language), we have also included phonetic or phonetic-like transcriptions. The corpus search interface makes it possible to search for a particular word or other sequence of words or parts of words by the orthographic or the phonetic transcription, or a combination of both.

## 3.2 The Transcription Process

The process of the two annotations in the Norwegian part of the corpus is described in this section. Each recording was phonetically transcribed manually by one assistant, while the output was proof-read by a different assistant, who checked the transcription against the audio. Then the text was run through a semi-automatic transliterator whose input was the phonetic transcription and its output orthographic transcription. A third assistant manually checked the output. Finally, a fourth person would proof-read the resulting orthographic transcription, checking it against the audio.

There were 18 part-time transcribers for the Norwegian part of corpus, consisting of 2,187,046 words, and 6 assistants doing the semi-automatic transliteration. They were all linguistics students who had read our extensive guidelines [16]; had learnt from each other; and cooperated and consulted each other. They were all expected to work in the same work place in order to ensure homogeneity in the transcriptions.

The other transcriptions were partly done at a national level, and partly in Oslo. The phonetic transcriptions follow national conventions, not the International Phonetic Alphabet. The conventions are described in Papazian and Helleland [30], they use only Latin letters. For Övdalian, the national Swedish orthography was used as the standard variant, while the orthography standardised by the Övdalian language council Råðdjärum was used as a "phonetic" transcription. To our knowledge, no other speech corpus contains double transcriptions. However, we would like to mention that a new Finland-Swedish dialect corpus—Talko—has adopted our tools; corpus design and interface, and even use two transcriptions (see Svenska Litteratursällskapet i Finland, in the reference list).

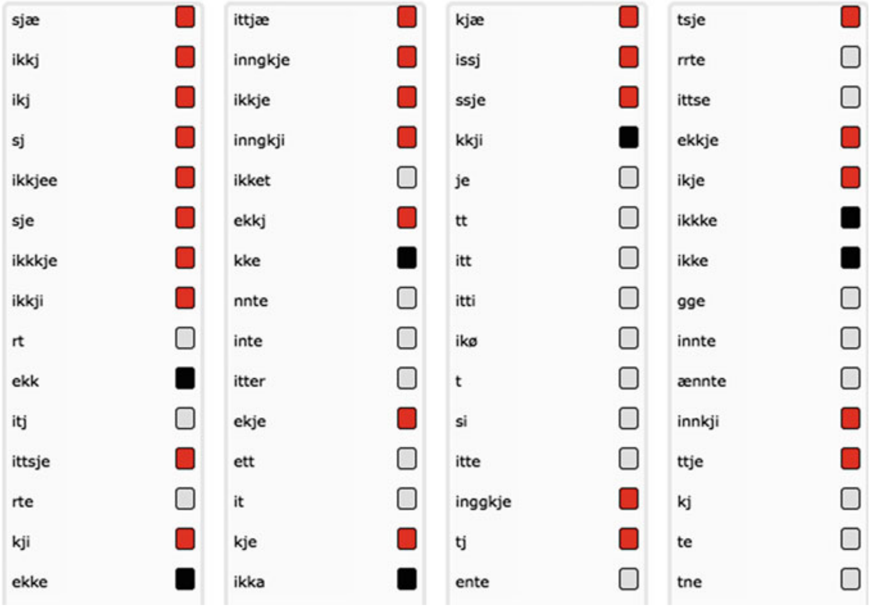
It is important that all words from the original phonetic transcription have an equivalent in the orthographic transcription. The two must be totally aligned word by word for the results to be used in the corpus search system.

[3] There are two Norwegian written norms, and for this corpus, we chose the *Bokmål* variant.

## 3.3 The Usefulness of Two Transcriptions for Corpus Users

The double transcriptions are extremely valuable. They make it possible to search for, for example, the Norwegian negator *ikke* 'not', and immediately get results for all pronunciations of this word: *ikke, innkje, inte, int, itte, itt* etc., as depicted in Fig. 1. The boxes accompanying the phonetic forms in Fig. 1 are blank to start with, but the corpus user can choose to colour each box separately, thereby getting a map that represents the different phonetic forms with different colour. By choosing for example red colour for all the fricative pronunciations /ç/ or /ʃ / and black colour for the velar stops /k/, a map can readily be produced, thus giving access to isoglosses (geographical borders for single language features) produced at an instant from spoken language data, as in Fig. 2. For a corpus aimed at dialect research, getting results in a map view is very useful. New knowledge on geographical variation can be depicted for almost any imaginable linguistic feature, as long as it is phonetically transcribed. The place of origin for each informant is located by GIS coordinates and the Google Maps API is used. Since every item in the corpus is connected to an informant, it means that for each word, string, piece of word or syntactic construction, there is a geographical location.

Without the phonetic transcription we would not have been able to find these dialect differences, and hence the new isoglosses. But the isoglosses also depend



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| sjæ | 🟥 | ittjæ | 🟥 | kjæ | 🟥 | tsje | 🟥 |
| ikkj | 🟥 | inngkje | 🟥 | issj | 🟥 | rrte | □ |
| ikj | 🟥 | ikkje | 🟥 | ssje | 🟥 | ittse | □ |
| sj | 🟥 | inngkji | 🟥 | kkji | ⬛ | ekkje | 🟥 |
| ikkjee | 🟥 | ikket | □ | je | □ | ikje | 🟥 |
| sje | 🟥 | ekkj | 🟥 | tt | □ | ikkke | ⬛ |
| ikkkje | 🟥 | kke | ⬛ | itt | □ | ikke | ⬛ |
| ikkji | 🟥 | nnte | □ | itti | □ | gge | □ |
| rt | □ | inte | □ | ikø | □ | innte | □ |
| ekk | ⬛ | itter | □ | t | □ | ænnte | □ |
| itj | □ | ekje | 🟥 | si | □ | innkji | 🟥 |
| ittsje | 🟥 | ett | □ | itte | □ | ttje | 🟥 |
| rte | □ | it | □ | inggkje | 🟥 | kj | □ |
| kji | 🟥 | kje | 🟥 | tj | 🟥 | te | □ |
| ekke | ⬛ | ikka | ⬛ | ente | □ | tne | □ |

**Fig. 1** Some of the phonetically transcribed variants of the negation *ikke* 'not'. Those that have been pronounced with a fricative have been coloured *red*, while those that have a velar stop have been coloured *black* (colour figure online)

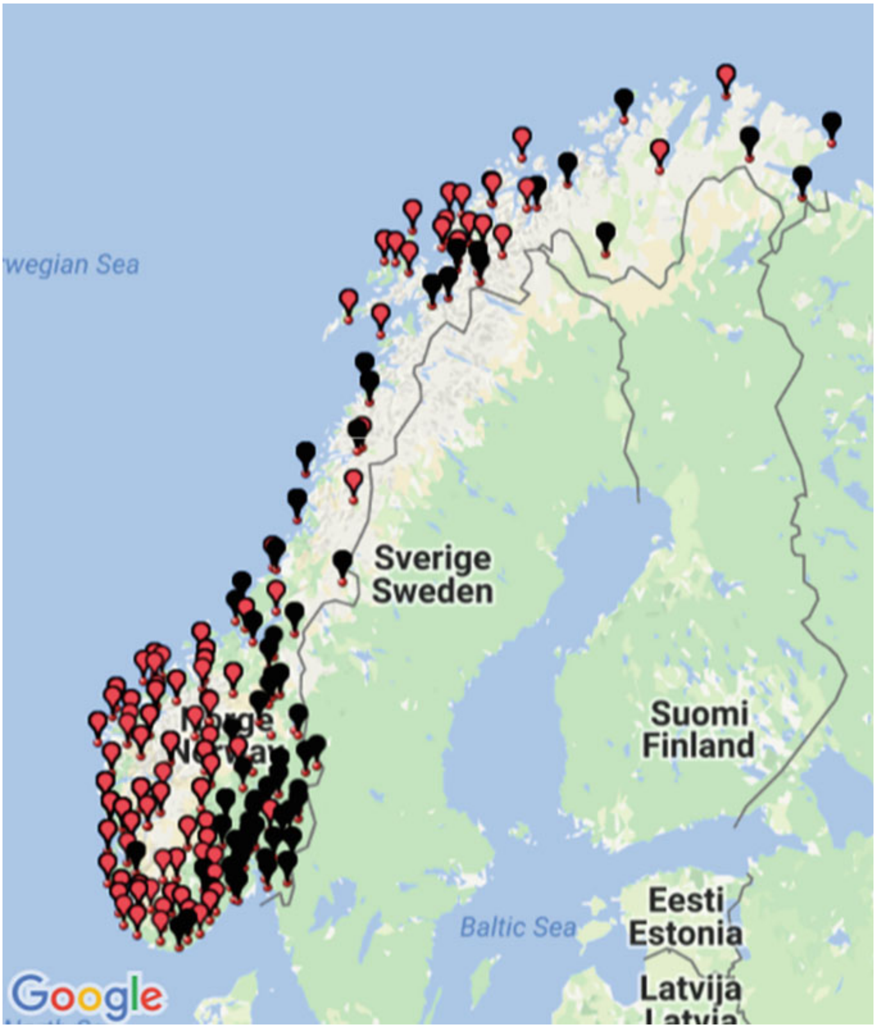
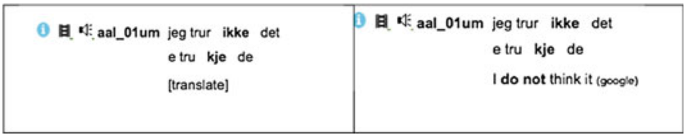


**Fig. 2** Map that shows results for fricative (*red*) and velar stop (*black*) pronunciations of *ikke*'not'. Clear isoglosses emerge from the map (screenshot from map generated by the Nordic Dialect Corpus, under a CC BY licence) (colour figure online)



aal_01um jeg trur **ikke** det
e tru **kje** de
[translate]

aal_01um jeg trur **ikke** det
e tru **kje** de
**I do not** think it (google)

**Fig. 3** One hit from a corpus search for the orthographic *ikke* 'not', depicting how both the phonetic and orthographic transcriptions are displayed. The result is shown before and after translation by Google Translate, which is integrated in the corpus search system

on the orthographic transcription, since it is exactly the pairing of the transcriptions that makes it possible to find the variation of one particular linguistic feature. The standard orthography also makes it possible to have the dialect results translated to English, by using a Google Translate API, see Fig. 3.

## 3.4 Transcription Software

For each language, transcription software was used that inserts time codes directly into the transcribed text at suitable intervals, enabling the transcription to be presented with its corresponding audio and video. Apart from most of the Swedish recordings, the other languages were transcribed by transcribers who were trained in Oslo, which ensured as uniform as possible a treatment of the different languages. Different software was used, but all transcriptions were adapted to the Transcriber XML format, which is also the interchange format used in the project. We mainly used Transcriber 1.5.1 for PC (see [2]). This has an intuitive user interface, and is fast and simple to use. It also has the advantage of offering the option of creating one's own macroes for various events such as laughter. The program further exports transcriptions to a nice HTML format. There are a couple of less attractive features, too. First, the PC version does not accept video. Second, overlapping speech can only be annotated for two people at a time.

A second type of software used in the project is the semi-automatic dialect transliterator, a program developed for the project at the Text Laboratory, UiO. It takes as input a phonetic text and an optional dialect setting. First, sets of text manually transliterated to orthography are used to provide a good basis for training the transliterator, enabling it to accurately guess the transliteration for further texts. The training process can be repeated, and the trained version can be used for similar dialects. Performing two types of transcriptions does not take twice the time of one, and is therefore much less costly than two fully manual transcriptions would have been. The transliterator can be used for any language, and has so far also been used for the Finland-Swedish corpus Talko.

A third type is the software developed in the project to fuse the two transcriptions and also to check that the phonetic and orthographic transcriptions are in fact totally aligned, typically after the tagging process.

## 3.5 Transcription Guidelines

Setting off time for the development of proper transcription guidelines is invaluable. The guidelines [16] for the Nordic Dialect Corpus were developed in close cooperation with the transcribers in frequent meetings in the initial months of the project. Even if we had experience from earlier transcription projects, such as that of the Corpus of Oslo Speech [13], the dialect project presented additional problems given that many dialects were further away from the orthographic norm, and that we had decided to have two transcriptions for Norwegian. Here we will discuss some of the problems we encountered and how we chose to solve them, but also other things that we think are important to be present in transcription guidelines.

The guideline starts with a gentle reminder of the practical challenges that are involved in transcription work, with some general advice about frequent breaks, and things like short-cuts using the keyboard etc. The document also contains detailed information about how to name files, where to put them, where the sound files are located, and instructions on how to use the transcription software. For transcriptions other than orthographic ones, the phonetic symbols and choices are explicitly described.

The transcription system we have followed is based on a system where alphabetic letters or letter combinations have the sound values from the Oslo dialect (assumed to be known by all readers). For example, the IPA symbol /u/ is represented by <o> in our transcription, while /o/ is <å>. There is a phonological distinction between long and short vowels in Norwegian, and short vowels are represented by double consonants, so that *oppvokst* 'grown up' is represented as: <åppvåkkst>. Consistency is central, so although many orthographic combinations would normally yield the same sounds, in the semi-phonetic transcription, one combination only has been chosen, so that the nasal velar is always represented by <ng> irrespective of the original orthographic version of the word: *tanke* 'thought' <tanngke>. There has to be guidelines for every speech sound, like syllable-carrying consonants: *gutten* 'the.boy' <gut'n>.

Names may represent unwanted identification of people and should be avoided. We have chosen to anonymise names, using $F_1 - F_n$ for female first names, $M_1 - M_n$ for males, and $E_1 - E_n$ for surnames. Non-linguistic sounds that may have some meaning in the conversations, such as laughter and yawning, should be marked. The same goes for sounds that seem to be linguistic, but whose meaning is not clear without further analysis. These are clicks of various kinds, which we have lumped together into two categories; front and back clicks. We have pre-identified some of these sounds and assigned them keyboard short-cuts, for quicker transcription. Further, citations and meta-linguistic comments are marked; they are simply put in inverted commas.

The semi-automatic dialect transliterator poses certain constraints on the transcriptions. The fact that the transcriptions will be translated to standard orthography and later automatically tagged, means that assimilations across words must be represented in separate words; this is also specified in the guidelines.

## 3.6 Transliteration Guidelines

There is a separate set of guidelines for the transliteration from phonetic script to standard orthography [22]. There are three main principles: (1) The standard orthography is always used. (2) Given the requirement for a complete word-alignment (for easy search-facilities when the transcriptions are put into the corpus), syntax is never standardised, but morphology is. (3) Two types of normalisation are marked especially, viz. words that are marked as foreign to the norm (tag = x), i.e., words not found in the standard dictionaries usually because they are loanwords or dialect words, and function words, i.e. grammatical words, that have been drastically translated to the norm (tag = o), in order for corpus users to be able to find all cases of, for example, a given subjunction independently of its phonological realisation in the various dialects.

The fact that standard orthography is used is not controversial. The difficult cases are dealt with by the choices accompanying the x and o tags (which we will discuss below). The fact that word order is never changed is also a straight-forward principle to follow, although the resulting text may look very strange with normalised orthography. However, the choice of normalising morphology also means introducing morphological distinctions that might not exist in a given dialect. (1)–(2) exemplify this. Many dialects do not have a case distinction in the third person plural pronouns, like the Botnhamn dialect (North Norway). As can be seen in the (a) examples, the subject in (1) and the object in (2) are represented by the same pronominal form, but the (b) examples reveal that they have different forms in the standard orthography.

(1)

| | | | | | |
|---|---|---|---|---|---|
| a. | **dæmm** | laggde | jo | sko | sjøll |
| b. | **de** | lagde | jo | sko | sjøl |
| | they | made | yes | shoes | themselves |

‘They made shoes themselves.’ (botnhamn_03)

(2)

| | | | | | |
|---|---|---|---|---|---|
| a. | førr | å | vie | **dæmm** | ut |
| b. | for | å | vide | **dem** | ut |
| | for | to | widen | them | out |

‘in order to widen them’ (botnhamn_03)

The x tag is used in order to be able to tell the corpus user that this word does not have a standard equivalent, i.e., is not found in the standard dictionary *Bokmålsordboka*. Such words are typically dialectal or loanwords. They are not translated to a normalised variant, since their meanings are often unclear to the transcribers and transliterators, but they are adapted with respect to morphology. Table 2 shows some examples of words that have been tagged with the x tag.

This tag has been used in several other corpora, too, and has been employed with success to, among other things, find English loan words in Norwegian [23], slang

**Table 2** Examples of words that have been tagged with the x tag. (<L> represents a retroflex flap, https://en.wikipedia.org/wiki/Retroflex_flap.) The question mark indicates that the meaning of this word is not known to the transcribers

| Phonetic | Normalised to | English |
|---|---|---|
| taimast | times | 'is timed' |
| løggLe | løgglig | ? (adjective) |
| nusstre | nustrig | ? (adjective) |
| bånfosst | barnefost | ? (noun) |
| smalamøki | smalamøkka | 'sheep droppings' |
| riffti | riftene | ? (noun) |
| kjårhæLær | kjårhæler | ? (noun) |

**Table 3** Examples of words that have been tagged with the o tag

| Phonetic | Transliterated to | English |
|---|---|---|
| så | enn | 'than' |
| vart | ble | 'became' |
| tå | av | 'of' |
| jå | hos | 'at' |
| kå | hva | 'what' |
| me | vi | 'we' |
| ekkå | noen | 'some' |

words amongst youths in Oslo [29], and detecting a written language bias in the vocabulary in the dictionary *Bokmålsordboka* [7].

The o tag is used to annotate function words that have a very different phonological form from the standard, but where the semantics is more or less the same. The reason for this choice is that the Nordic Dialect Corpus is also planned to be used for syntactic research. Function words are then important, and it is valuable to find all in one search, even if their form is different. These words are then translated to the equivalent or near-equivalent in the standard written language, and given the tag o for easy recognition. Some examples are given in Table 3.

The x and o tags make it possible to search for all the words that are tagged this way with one click.

## 3.7 Linguistic Knowledge as an Outcome of the Annotation

Just as the future use of the corpus by linguists has informed the annotation scheme (there is linguistic motivation for the two transcriptions, and the tags x and o), the converse is true. While developing the transcription guidelines it soon became clear

**Table 4** Some examples of interjections not found in the standard dictionaries

| Interjections not found in the standard dictionary | Meaning |
|---|---|
| eh | ‘I feel a distance to what is claimed’ |
| ehe | ‘I understand’ |
| heh | ‘I’m impressed’ |
| hm | ‘I wonder what you meant’ |
| m-m | ‘I deny the claim’ |
| mhm | ‘I understand’ |
| mm | ‘I agree/confirm’ |
| næ | ‘I’m surprised’ |
| u | ‘I’m impressed’ |

that the standard dictionaries are developed for the written language. Although many of the sounds that people utter while talking cannot be described as words, some sounds and sound combinations clearly have a stable meaning. These should typically be characterised as interjections, as they do not have a place inside the sentence. Having listened through a lot of speech in the recordings, we found a long series of new interjections, which have been included in the transcription guidelines [16] and are used in the corpus. We have given these interjections a standardised orthography. Some are shown in Table 4.

## 3.8 Ensuring High Quality

No formal evaluation of the transcription methods has taken place. However, we would like to emphasise that the whole process of developing the transcription standard was long and thorough, with frequent meetings between the transcribers (most were master students in linguistics) and the project leaders. The decisions therefore were made after long discussions on particular challenges, as well as a lot of trial and error in testing actual methods. One feature that was abandoned after this process was, for instance, the marking of stress. Although this feature is central in Norwegian phonology and varies systematically across the country, it turned out to be impossible for the transcribers to agree on what they heard. We had to conclude that this feature would never be annotated in such a way that it could be useful for researchers.

The transcription process included (as mentioned) proof reading by the transcribers of each others’ work with feed back, to ensure a consistent annotation practice. We add that the corpus user interface has a button for reporting bugs, including transcription errors, and these are regularly inspected, and the transcriptions corrected when necessary.

Finally, feed back from researchers show that the choice of having two types of transcriptions was a very good one, giving so many new options that were never possible in the past.

## 4 Annotation II: Grammatical Tagging

### 4.1 Grammatical Tagging of Five Languages

The transcriptions for the five languages have all been morphologically tagged with part of speech tags. Tagging speech data is different from tagging written data. Speech contains disfluencies, interruptions and repetitions, and there are rarely clear clause boundaries [1,10,12,38]. 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). In spite of this, we had to mostly use available written language taggers. These are not optimal for spoken language, but were the only ones available. Some of the taggers are statistics-based and some rule-based, and some are even a combination.

The Text Laboratory, UiO, has the responsibility for the tagging. Since the transcriptions have been tagged individually with taggers developed in other projects for the respective languages, each language has an individual tag set chosen by those who developed the taggers originally. The Danish transcriptions are lemmatised and POS tagged by a Danish Constraint Grammar Tagger [20] developed for written Danish, see Bick [3]. The Faroese transcriptions were first tagged with a Constraint Grammar Tagger for written Faroese, see Trosterud [34]. Since spoken Faroese has a lot of words that are not approved in written standard Faroese, about half of the material was manually corrected after the Constraint Grammar tagging. Finally a TreeTagger [33] was trained on the corrected material, and the rest of the transcriptions were tagged again.[4] The Icelandic transcriptions were first tagged with a tagger for written Icelandic, see Loftsson [24], and some manually corrected afterwards.[5] The orthographic Norwegian transcriptions were lemmatised and POS tagged by a TreeTagger originally developed for Oslo speech [27,28]. This speech tagger was trained on manually corrected output from the written language Oslo-Bergen tagger [8].[6] The TreeTagger gained an accuracy of 96.9% on the Oslo speech corpus, see Nøklestad & Søfteland [27,28]. The Swedish subcorpus was tagged by a modified version of the TnT tagger developed by Kokkinakis [21]. After having been manually corrected and retrained, a spoken language Swedish statistical HunPos tagger (Halácsy 2007) was developed at the Text Laboratory. The tagger was trained on the Swedish PAROLE corpus and the manually tagged orthographic Övdalian transcriptions.[7]

[4] The manual corrections of the Faroese tagger were done by Remco Knooihuizen at the Text Laboratory, UiO.

[5] The manual corrections of the Icelandic tagger were done by Gísli Rúnar Harðarson at the Text Laboratory, UiO.

[6] The manual corrections of the Norwegian speech tagger were done by Åshild Søfteland at the Text Laboratory, UiO.

[7] The manual corrections of the Swedish tagger were done by Piotr Garbacz at the Text Laboratory, UiO.

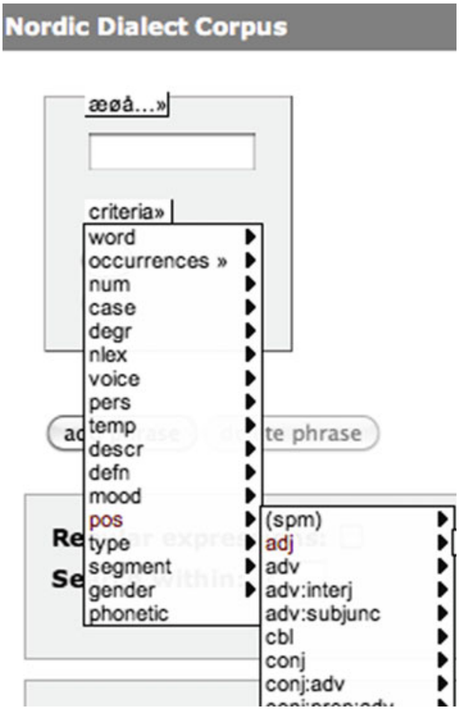


**Fig. 4** Querying for adjectives across all the languages in the corpus

The language sub-corpora thus have been tagged with different tag sets, but the tags have been standardised in the search system, making it possible to search for the same category across all the corpora, illustrated by a search for adjectives, in Fig. 4.

The search for adjectives results in hits like those in (3), among many others (and supplied with the Google Translate output):

(3)

a. nej # der kommer # **ældre** mennesker ind her (Danish)

no ... coming ... **older** people in here (google), (aarhus4)

b. ja teir eru teir eru eru **effektivir** í álopinum (Faroese)

Well they're they're are **effektivir** in álopinum (google),(fuglafjoerdur_f12)

c. ekki til að fara í svona rosalega **flotta** ferð býst ég við (Icelandic)

not to go so **awesome** trip I guess (google), (iceland_a1)

d. å de synns e e ufattele **gått** (Norwegian)

and I think this is incredibly **good** (google), (aal_01um)

e. før ig ir so kluvin ig wet it **siouv** (Övdalian Swedish)

or I am so ambivalent, I do not know **myself** (google), (aasen_35)

**Table 5** Some of the Danish and Swedish tags mapped to the standard

| Danish to standard | Swedish to standard |
|---|---|
| "GEN" => "poss" | "GEN" => "poss" |
| "IDF" => "indef" | "HP" => "subjunc" |
| "IMP" => "imp" | "I" => "interj" |
| "IMPF" => "pret" | "IE" => "inf-marker" |
| "IN" => "interj" | "IMP" => "imp" |
| "INDP" => "pron" | "IN" => "interj" |
| "INF" => "inf" | "IND" => "indef" |
| "INFM" => "inf-marker" | "IND/DEF" => "indef_def" |
| "KC" => "conj" | "INF" => "inf" |
| "KP" => "prep" | "In" => "interj" |
| "KS" => "subjunc" | "JJ" => "adj" |
| "LOC" => "" | "KN" => "conj" |
| "N" => "noun" | "KOM" => "comp" |
| "ND" => "" | "KON" => "subjunctive" |
| "NEU" => "neut" | "MAS" => "masc" |
| "NOM" => "nom" | "NEU" => "neut" |
| "NUM" => "det_quant" | "NN" => "noun" |
| "P" => "pl" | "NOM" => "nom" |

The mapping of tags from the individual tag sets to the common standard has mostly been straight-forward, as seen below in Table 5, where some of the categories from Danish and Swedish have been mapped to the standard ones.

As can be seen from Table 5, some of the categories, like Danish LOC, have not been transferred to anything, since we wanted a common, not too detailed tag set. Some tag sets have been more complicated to map. The Icelandic one is a case in point. There, the tags consisted of one-letter categories, which meant different things depending on which part of speech they belonged to. For example, if the POS was a verb, then "þ"=>"past participle", but if the POS was a noun, then "þ"=>"dative". We had to make a mapping script for this, as illustrated in (4).

(4)

```
    "noun"=> [{"k"=>"masculine","v"=>"feminine","h"=>"neuter","x"=>"unspecified"},
        {"e"=>"singular","f"=>"plural"},
        {"n"=>"nominative","o"=>"accusative","þ"=>"dative","e"=>"genitive"},
        {"g"=>"with suffixed definite article","-"=>""},
        {"m"=>"person name","ö"=>"place name","s"=>"other proper name"}],
    "verb"=> [{"þ"=>"past participle","n"=>"infinitive","b"=>"imperative",
    "f"=>"indicative","v"=>"subjunctive","s"=>"supine","l"=>"present participle"},
```

## 4.2 Some Problems Relating to the Tagging of Dialect Data

Given the written language bias of the taggers it is true to say that there is room for improvement with regard to all of them. Even the Norwegian TreeTagger, which was trained on speech (the Oslo dialect), is not performing perfectly. In this section some problems will be discussed, which are partly due to inherent difficulties relating to the fact that linguistic data, being dialects, are very varied, and partly to the fact that decisions were made that turn out in retrospect to have been somewhat unfortunate.

One problem is that the tagger is trained on the Oslo dialect, which is close to the written standard, while the dialects present more diverse word orders, which the tagger is not trained to recognise. This is illustrated in (5), where (5a) represents the standard word-order of constituent questions, with the verb as the second constituent of the main clause (known as V2 word order), while many dialects have non-V2 in constituent questions, as in (5b).

(5)

a. hva **liker** du å gjøre i fritida da ?

ko **lika** ru å jera i fritie ra ?

what **like** you to do in spare.time.the then

‘What do you like doing in your spare time then?’ (google), (aal_01um)

b. hva han **fikk** i den ?

ko hann **fe** i denn ?

what he **got** in it

‘what he got in it?’ (google), (aaseral_01um)

Another reason is that some words and discourse particles just do not exist in the standard language, like the discourse particle *sjø* ‘you see’ used in the areas in and around the city of Trondheim, see (6):

(6) jeg angrer faktisk litt på det sjøl **sjø**
e anngre fakktisk litt på de sjøL **sjø**
‘I regret actually a bit self **sea**’ (google), (alvdal_02uk)

In (6), the word *sjø* has been tagged as a noun, because of the homonymous word *sjø* ‘sea’ in the standard language (notice also the Google translation!), while a more correct tag would have been an adverb. In retrospect it would have been wise to have given this word a translation to an adverb like *vel* ‘well’, accompanied by an o tag (see the section on transliteration above). This would have given better tagging results. Alternatively, it might have been even better to train different taggers for different dialects – at least for different regions – so that the taggers would have been adapted to regional vocabulary and grammar.

Finally, we will mention a problem that we did not foresee for the Norwegian tagger, which is related to the fact that we have two transcription types. As mentioned, the two transcriptions have to be totally aligned at word-level, which is done by translating each dialect word to a standard orthographic word. The orthographic transcription is then tagged. However, the tagger collapses words that are regarded as set phrases, like *for_eksempel* (‘for example’). Since this process destroys the word-alignment, everything has to be checked and corrected afterwards. This additional step in the process would have been unnecessary had the tagger been differently trained. In addition, such collapsed phrases are bad for searching the corpus, since they do not show up as single words, in contrast to what the users probably assume.

The transcriptions in the NDC represent five different languages and have been tagged with five different taggers that were first trained on written languages, and then adapted, to a varying extent, to spoken language, and to dialects. Nivre & Grönqvist [26] achieved a respectable result of 95–97% (depending on tagset) for Swedish spoken language, and Nøklestad and Søfteland [27,28] achieved, as mentioned, 96.9% accuracy on their tagger for Norwegian Oslo speech. However, the NDC consists of dialects, and although the transcriptions have been standardised before the tagging, and most deviant words have been translated to standard words, there are still remaining features of the dialects that make them different from both written language and the language of the capital cities, regarding word order as well as discourse words. So although no evaluation has been performed on the general result of the taggers for the transcriptions in the NDC, their accuracy must be expected to be lower than the numbers reported for speech taggers used on less varied linguistic input. However, in spite of these problems, the NDC is definitely a morphologically tagged corpus, and very useful as such.

## 5 Reusability and Licensing of Software and Corpus

The Transcriber software is free of any licencing (see the web site). The semi-automatic dialect transliterator and the word-alignment checker are also freely available, by contacting the Text Laboratory, UiO. The corpus is accessible for searches from its the web site, but users must register for a password. The sound and video files are not freely downloadable due to legal restrictions in the Personal Data Act, but the transcriptions themselves are free. These are anonymous and any names have been removed.

## 6 Conclusion

Since this chapter has dealt with a speech corpus, the Nordic Dialect Corpus, it has discussed the special challenges and solutions that the spoken language represents. Transcription and grammatical tagging are the two most central annotation types for this kind of text.

I have shown that spoken language corpora that have corresponding phonetic and orthographic transcriptions give excellent options for the linguist to get out the variation that exists in the corpus. With geographical GIS-marking of all the informants, new isoglosses can be discovered almost *ad infinitum*. With such tools as the semi-automatic dialect transliterator described here, the overall cost is not as high as twice that of a single transcription. Using some extra tags, such as the o and x tags (marking a full translation of function words to a standard form, and a non-standard form for lexical forms, respectively), gives further options. The tagging of spoken language is challenging since taggers are usually trained on or developed for written language. Even with adaptions to spoken language, dialectal features represent a challenge, so that the taggers are not optimal for their task. Finally, since the dialects in the Nordic Dialect Corpus belong to five different languages, additional challenges turned up in the harmonisation of tag sets.

While one of the goals of this chapter has been to describe solutions to problems, another has been to describe choices that were less fortunate, or indeed not taken at all, leading to mistakes in the grammatical tagging of the dialects and causing occasional challanges for the use of the final corpus. Fortunately, although there have been some issues, the corpus is up and running, and is being used by several researchers and in several publications already (see for example the chapters and maps in Nordic Atlas of Language Structures Online).

## References

1. Allwood, J., Nivre, J., Ahlsén, E.: Speech management-on the non-written life of speech. Nord. J. Linguist. **13**, 3–48 (1990)
2. Barras, C., Geoffrois, E., Wu, Z., Liberman, M.: Transcriber: a free tool for segmenting, labeling and transcribing speech. In: First International Conference on Language Resources and Evaluation (LREC), pp. 1373–1376 (1998)
3. Bick, E.: PaNoLa - The Danish connection. In: Holmboe, H. (ed.) Nordic Language Technology, Årbog for Nordisk Sprogteknologisk Forskningsprogram 2000–2004 (Yearbook 2002), pp. 75–88. Museum Tusculanum, Copenhagen (2003)

4. Bokmålsordboka. 2005. Wangensteen, Boye (ed.). Oslo: Kunnskapsforlaget. http://www.nob-ordbok.uio.no
5. Christ, O.: A modular and flexible architecture for an integrated corpus query system. *COMPLEX'94*, Budapest (1994)
6. Evert, S.: The CQP query language tutorial. Institute for Natural Language Processing, University of Stuttgart, www.ims.unistutgart.de/projekte/CorpusWorkbench/CQPTutorial (2005)
7. Fjeld, R.V.: Talespråksforskningens betydning for leksikografien. In: Johannessen & Hagen, pp. 15–28 (2008)
8. Hagen, K., Bondi Johannessen, J., Nøklestad, A.: 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 vol. 19, pp. 31-48, University of Southern Denmark, Odense (2000)
9. Halácsy, P., Kornai, A., Oravecz, C.: Hunpos - an open source trigram tagger. In: Proceedings of the 45th Annual Meeting of the Asso- ciation for Computational Linguistics, volume Companion Volume, Proceedings of the Demo and Poster Sessions, pp. 209–212, Prague, Czech Republic. Association for Computational Linguistics (2007)
10. Jørgensen, F.: Automatisk gjenkjenning av ytringsgrenser i talespråk. In: Johannessen and Hagen (eds.), pp. 204–213 (2008)
11. Johannessen, J.B.: The Corpus Search and Results Handling System Glossa. Chung-hua Buddh. J. **25**, 87–104 (2012)
12. Johannessen, J.B., Jørgensen, F.: Annotating and parsing spoken language. In: Peter, J.H., Peter, R.S. (eds.) Treebanking for Discourse and Speech, pp. 83–103. København, Samfundslitteratur (2006)
13. Johannessen, J.B., Hagen, K. (eds.): Språk i Oslo. Ny forskning omkring talespråk. Novus forlag, Oslo (2008)
14. Johannessen, J.B., Nygaard, L., Priestley, J., Nøklestad, A.: Glossa: a multilingual, multimodal, configurable user inter-face. In: Proceedings of the Sixth International Language Resources and Evaluation (LREC'08. Paris: European Language Resources Association (ELRA) (2008)
15. Johannessen, J.B., Priestley, J., Hagen, K., Åfarli, T.A., Vangsnes, Ø.A.: The Nordic Dialect Corpus - an advanced research tool. In: Jokinen, K., Bick, E. (eds.) Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009. NEALT Proceedings Series, vol. 4 (2009a)
16. Johannessen, J.B., Hagen, K., Håberg, L., Laake, S., Søfteland og, Å., Vangsnes, Ø.: Transkripsjonsrettleiing for ScanDiaSyn (2009b)
17. Johannessen, J.B., Hagen, K., Nøklestad, A., Priestley, J.: Enhancing language resources with maps. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D., (eds.) Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), pp. 1081–1088. Paris: European Language Resources Association (ELRA) ISBN 2-9517408-6-7 (2010)
18. Johannessen, J.B., Priestley, J., Hagen, K., Nøklestad, A., Lynum, A., The Nordic Dialect Corpus. In: Calzolari, N., Choukri, K., Declerck, T., Ugur Dogan, M., Maegaard, B., Mariani, J., Odijk, J., (eds.) Proceedings of the Eighth International Conference on Language Resources and Evaluation. European Language Resources Association, pp. 3388–3391 (2012)
19. Johannessen, J.B., Vangsnes, Ø.A., Priestley, J., Hagen, K.: A multilingual speech corpus of North-Germanic languages. Raso and Mello (eds.) **2014**, 69–83 (2014)
20. Karlsson, F., Voutilainen, A., Heikkilä, J., Anttila, A. (eds.): Constraint Grammar. A Language-Independent System for Parsing Unrestricted Text. Mouton de Gruyter, Berlin (1995)
21. Kokkinakis, S.J.: En studie över påverkande faktorer i ordklasstaggning. Baserad på taggning av svensk text med EPOS. Ph.D. dissertation. Göteborg University (2003)