Spontal-N: A Corpus of Interactional Spoken Norwegian

Rein Ove Sikveland[(1)], Anton Öttl[(2)], Ingunn Amdal[(3)], Mirjam Ernestus[(2, 4)], Torbjørn Svendsen[(3)], and Jens Edlund[(5)]

[(1)] Department of Language and Linguistic Science
University of York, UK
[(2)] Radboud University Nijmegen, The Netherlands
[(3)] Department of Electronics and Telecommunications
Norwegian University of Science and Technology (NTNU), Trondheim, Norway
[(4)] Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands
[(5)] Department for Speech, Music and Hearing
KTH, Stockholm, Sweden

ros500@york.ac.uk, {anton.oettl, mirjam.ernestus}@mpi.nl, {ingunn.amdal, torbjorn}@iet.ntnu.no, edlund@speech.kth.se

Abstract

This paper presents Spontal-N, a corpus of spontaneous, interactional Norwegian. To our knowledge, it is the first corpus of Norwegian in which the majority of speakers have spent significant parts of their lives in Sweden, and in which the recorded speech displays varying degrees of interference from Swedish. The corpus consists of studio quality audio- and video-recordings of four 30-minute free conversations between acquaintances, and a manual orthographic transcription of the entire material. On basis of the orthographic transcriptions, we automatically annotated approximately 50 percent of the material on the phoneme level, by means of a forced alignment between the acoustic signal and pronunciations listed in a dictionary. Approximately seven percent of the automatic transcription was manually corrected. Taking the manual correction as a gold standard, we evaluated several sources of pronunciation variants for the automatic transcription. Spontal-N is intended as a general purpose speech resource that is also suitable for investigating phonetic detail.

1. Introduction

In this paper, we introduce Spontal-N, a new corpus of Norwegian that is being developed as a general purpose language resource for research on natural, interactional speech. It is suitable for research on spontaneous Norwegian speech, audiovisual cues in turn-taking, and last but not least, interference between mutually intelligible languages. To our knowledge, it is the first corpus of Norwegian in which the majority of speakers have spent significant parts of their lives in Sweden, and in which the recorded speech displays varying degrees of interference from Swedish.

The corpus consists of audiovisual studio recordings of four 30-minute dialogues, and a manual orthographic transcription of the entire material. Spontal-N also includes an automatically generated phonemic annotation for part of the corpus, and a phonemic reference transcription, which was created by a manual correction of randomly selected parts from the automatic transcription. The conventions used during transcription and annotation are documented in a protocol accompanying the corpus.

There is a current shortage of speech data for research on the Norwegian language. A national language resource bank is now in the start up phase, but the available resources are still limited, see e.g. the latest report from "Norsk Språkråd" (The Norwegian Language Council), (Svendsen et al., 2008). The Spontal-N corpus is thus a welcome supplement, especially for annotated spontaneous speech. Relevant other material for Norwegian includes the NorDiaSyn corpus (Johannessen et al., 2009), which focuses on Norwegian dialects. Since publicly available tools and resources for annotation are scarce for Norwegian, the procedures described in this paper for the generation of an automatic phonemic annotation may also be interesting for other languages with limited resources available.

In the following sections we present central aspects of the corpus and its development in more detail. In Section 2 we explain how the audiovisual material was obtained, and present the relevant technical specifications. In Section 3 we describe the contents and the structure of the manual transcriptions, and outline how effects of spontaneous speech were taken into account by adding extra-linguistic labels and tags to characterize non-dictionary items. We go on in Section 4 to explain how we made a forced alignment between the acoustic signal and phonemic pronunciations listed in a dictionary, in order to obtain an automatically generated phonemic transcription. We also explain how we enhanced the quality of this initial phonemic annotation by increasing the number of pronunciation variants listed in the pronunciation dictionary, and present results from an initial evaluation of the automatic annotations (based on comparisons to the reference transcription). We end this paper by pointing to on-going and future work on the corpus in Section 5.

2. Data Collection

2.1. The Speakers

Spontal-N features six native speakers of Urban East Norwegian (UEN), a regional accent spoken in Oslo and its surrounding regions (Kristoffersen, 2000). With one exception, all speakers have spent a significant number of years in Sweden (roughly corresponding to 25 - 75 % of their lifespans), and use Swedish for daily interactions. One of the speakers ('B') is bilingual with Swedish as the dominant language. All speakers have similar educational backgrounds. They signed a consent form in which they state that the material can be used for scientific research. More information about the speakers is presented in Table 1.

Recording (duration)	Speaker (age)	Gender	Norwegian background	Years in Sweden
1 (42 min)	A (60+)	F	Oslo	37
	O (63)	M	Oslo	18
2 (34 min)	B (45)	M	Oslo	33
	L (35)	M	Larvik	8
3 (33 min)	S (30-35)	M	Jevnaker	0
	L (35)	M	Larvik	8
4 (34 min)	T (35-40)	M	Larvik	18
	L (35)	M	Larvik	8

Table 1: Information about the speakers.

Norwegian and Swedish are mutually intelligible, and few speakers who have been exposed to both over longer stretches of time are successful in avoiding interference between the two languages. Consequently, the speech recorded for Spontal-N also contains a significant amount of interference from Swedish. This appears primarily from the use of words or phrases that do not exist in Norwegian and to some extent also from the use of Swedish intonation. In addition to this, Spontal-N contains some examples of code-switching to Swedish. This was (unintentionally) recorded as the speakers interacted with a Swedish experimenter outside the official experimental session.

Five informal ratings of speech samples from Spontal-N confirm that the degree of the speakers' exposure to Swedish is clearly audible, even in samples that are lexically and syntactically correct for Norwegian. As might be expected, the bilingual speaker was rated as displaying a low degree of interference. On the other hand, this speaker does produce some grammatical errors (e.g. assigning the gender of a Swedish noun to a Norwegian cognate).

2.2. The Recording Sessions

The material was recorded in a sound attenuated studio at KTH, Stockholm, in 2008, using the Spontal setup as documented in (Beskow et al., 2009). While a studio setting is necessary to guarantee a high quality of audiovisual data, it may also increase the speakers' awareness of being recorded, making it more difficult for them to relax and to interact naturally. Therefore the two speakers that were appointed as dialogue partners were either friends or acquaintances. Also, these pairs were informed that they could talk freely about anything they wanted. Each experimental session lasted 30 minutes and its structure was made explicit to the speakers prior to the recording (see Table 2).

The speaker pairs were informed that a box would be placed between them on the floor, and that they would be asked to explore it and its contents during the last ten minutes of the session. The box contained objects that could potentially generate curiosity: a surrealist hutch figure and a pencil sharpener among other things. The exploration task was presented as optional, in the sense that the speakers were explicitly told that they could continue the ongoing conversation, if this was what they preferred. For the official part of the recording, they were left alone in the studio, facing each other across a small table.

Time (min)	Description of activity	
< 0	adjustment of the microphones, instructions to speakers etc.	
0 - 10	*official part of recording*	free conversation between the two speakers
10 – 20		
20 – 30		speakers explore box
> 30	demounting of equipment etc.	

Table 2: Structure of the recording sessions.

2.3. Technical Specifications

As shown in Table 3 below, we used two sets of microphones (goose-neck and head-set microphones) for the recordings. This was done to achieve optimal recording quality, while keeping the degree of leakage from the other speaker to a minimum. A Phonic mixer console was used as a microphone pre-amplifier, and also to supply phantom power to the goose-neck microphones. The output of the console was connected to an M-Audio interface and recorded with Audacity (Mazzoni, 2008) in 4 channels 48 KHz/24 Bit linear PCM wave files on a 4 x Intel 2.4 MHz processor PC. To capture non-verbal interaction, two video cameras were placed with a good view of each subject's upper body, approximately level with their heads.

Microphones
2 Bruel & Kjaer 4003 omni-directional: 1m from speakers
2 Beyerdynamic Opus 54 cardioid: Head-mounted
Audio recording
4 channels at 48 KHz/24 Bit
Video recording
2 JVC HD Everio GZ-HD7 high definition video cameras
Resolution 1920x1080i ; Bitrate 26.6 Mbps

Table 3: Summary of technical equipment for audiovisual recordings.

3. Manual Orthographic Transcription

For the manual annotation of the recordings we used Praat (Boersma and Weenink, 2009). Each recording was annotated in a separate TextGrid file, consisting of five tiers (Figure 1). While the speech of each dialogue partner is transcribed on a separate tier, talk by the two experimenters is annotated on the same, third, tier. Finally, two tiers are reserved for noises and comments respectively. The latter allows the user of the corpus to quickly gain access to instances of extreme reduction and potentially omitted words. Uncertainties regarding the orthographic transcription, and any additional information that may be of relevance to the user, are also annotated on this tier. For example, if it is unclear whether a word is better interpreted as Norwegian or Swedish, a comment is placed on this level.

Temporally, each speech tier is divided into chunks, according to a list of requirements specified to facilitate the posterior phonemic alignment (for details see 4.1). As a general rule, chunks have a maximal duration of three seconds. To avoid artificial interruptions in the speech, all chunk boundaries are placed at natural pauses, which implies that some chunks exceed the three-second limit.

3.1. Conventions

3.1.1. Norwegian Standard Orthography

For Norwegian, two written standards coexist (Bokmål and Nynorsk). Urban East Norwegian as spoken in the recordings is well represented by Bokmål, and the orthographic transcriptions are kept in line with the norms of Bokmålsordboka[1]. Compared to other languages, the norms for written Norwegian are fairly permissive, in the sense that one often finds multiple entries for the same entity see e.g. (Kristoffersen, 2000) for the history behind this situation. As shown in Table 4, multiple entries can reflect different types of variation.

Level of variation	Norwegian	English
Morphology	a) prata b) pratet	talked
Lexicon	a) å ljuge b) å lyve c) å lyge	to lie
Phonetics	a) liksom b) lissom	sort of

Table 4: Multiple dictionary entries reflect different types of variation.

For orthographic transcription, multiple dictionary entries pose a challenge, since it is difficult to be consistent (keeping track of permissible forms). For spontaneous speech, it may even be problematic to distinguish acoustically similar forms. One solution would be to custom-make a standard by selecting one of the available forms, and using this consistently for the transcription of all variants. We discarded this approach in favour of including all permissible variants. If needed, these distinctions may be collapsed in retrospect.

[1] "The Bokmål Dictionary" is available at http://www.dokpro.uio.no/ordboksoek.html

Figure 1: Structure of transcription file.

3.1.2. Word Level Annotations

Non-dictionary words are marked as such, and we provide them with a spelling that is based on their pronunciation (e.g., a Norwegian past tense verb derived from English *catch* is transcribed as "kætsja\n"). If a word form markedly deviates from the form listed in the Bokmål dictionary, we transcribe the dictionary form, but add the nonstandard form in the tag. Thus, a clipped word "saks" (from "saksofon") is transcribed as "saksofon\n[form=saks]"). If a nonstandard pronunciation is used, we add the pronunciation in the tag. Pronunciation slips, i.e., pronunciations that are best classified as production errors, are specifically marked (e.g., "konsentrere\v[pron=kontstentrere]"). Finally, foreign words are transcribed in the original language, and are tagged accordingly (e.g. "påpetare\f[lang=SWE]"). Table 5 presents the complete set of word level tags used. Multi-word items (i.e. titles and names) are joined by underscores.

Tag	Use
\n	nonstandard item
\n[form=TEXT]	nonstandard form
\n[pron=TEXT]	nonstandard pronunciation
\n[pron=LANG]	pronunciation contains non-Norwegian phone(s)
\c	(not listed) compound
\-	interrupted word
\+	clipped compound (1st constituent)
\~	clipped compound (2nd constituent)
\f[lang=LANG]	foreign word
\v[pron=TEXT]	pronunciation slip
\o	onomatopoetic item
\x	uncertain transcription
Label	**Use**
[xxx]	unintelligible speech
[name=TEXT_TEXT]	names of persons
[title=TEXT_TEXT]	titles and non-person names

Table 5: Word level annotations.

3.1.3. Non-lexical Annotations

Since we want Spontal-N to be valuable for a wide range of users, the manual annotation also covers some extra-linguistic information. Clicks and percussives are consistently marked, as is the following set of fillers, interjections and/or response tokens: *eh, mhm, mh, hm, åh,* and *aha*. Further extra- and non-linguistic annotations include coughs, vocalizations, laughter and imitations. Since laughter and imitations often co-occur with speech, we also use interval marking (e.g. "[begin_imitation] TEXT [end_imitation]").

3.2. Some Data on the Corpus

Table 6 presents some initial information about the speech featured in Spontal-N. The counts are based on the orthographic transcription of all speakers, including the experimenters.

	Types	Tokens
Norwegian words	2035	18449
Swedish Words	362	1220
Other foreign words	76	202
Names and titles	153	424

Table 6: Some data on the corpus.

4. Automatic Phonemic Annotation

Our goal for the annotation is to make Spontal-N searchable for general purpose phonetic research. Apart from being less time-consuming than manual annotations, automatically generated phonetic annotations have the potential of being more consistent. Both manual and automatic annotations are prone to mistakes, but the latter have the advantage that the errors are predictable. For an assessment of various methods and a general discussion of applicability, see (Van Bael et al., 2007). The procedure described in the following sections is based on the procedure used for broad-phonetic annotation in the RUNDKAST project (Amdal et al., 2008). In addition to an orthographic transcription and an adequate representation of the corresponding acoustic signal, this procedure requires a pronunciation dictionary and acoustic models of the phonemes to be modelled. While the pronunciation dictionary is used to translate the orthographic transcription into a set of possible phonemic transcriptions, the acoustic models are needed to align the alternative transcriptions with the acoustic signal. The output of this procedure is the most likely transcription, relative to the acoustic models and the pronunciations listed in the dictionary. Since we believe that the currently available resources for Norwegian are inadequate to consistently annotate spontaneous speech automatically, we chose a less detailed, phonemic annotation. We have used SAMPA format symbols (Wells, 1997) with a core set based on the Norwegian SAMPA phoneme inventory[2].

4.1. Selection of Material

Since the quality of the alignment depends on the quality of the acoustic signal, speech chunks that contain noise are discarded from the automatic phonemic transcription. We have also chosen to exclude chunks containing any of the following:

- unintelligible speech
- laughter, coughing or imitations
- foreign words, titles or names
- interrupted words or onomatopoetic items

A final requirement for inclusion is that the chunk does not merely contain vocalizations, percussives, clicks or breaths. The amount of material that was initially available in the orthographic transcriptions (OT), and the amount that was left for the automatic phonemic transcription (APT) after the sifting, can be taken from Table 7.

	OT	APT
Number of chunks	7369	3377
Total duration	156 min	74 min
Avg. chunk duration	1.27 sec	1.32 sec
Avg. words per chunk	4.8	5.7

Table 7: Amount of material before and after sifting.

4.2. Outline of the Procedure

For the speech recognizer we used speaker- and context independent acoustic models. These HMM models were trained with the Hidden Markov Model Toolkit (HTK) (Young et al., 2006) on about 20 hours of continuous manuscript read Norwegian speech from about 900 speakers. This is a subset of the NST corpus in the Norwegian national language resource bank, (Svendsen et al., 2008). In addition to the set of Norwegian phonemes, the acoustic models also include a model for silences. This silence model was also used to capture clicks, percussives, vocalizations and breaths annotated in the orthographic transcriptions. For all automatic annotations we let the speech recognizer choose whether there was silence between words or not. No pronunciation probabilities were used since we have only a limited number of variants and no source to find/train these probabilities.

The lexicon for Spontal-N is mainly based on the (only) general Norwegian pronunciation dictionary available, NorKompLeks (Nordgård, 2000). The pronunciations in NorKompLeks are "typical" Urban East Norwegian and some pronunciation variants are present. It is fairly new compared to similar resources for other languages, and still lacks the quality assurance extensive use can give. We have therefore chosen to use a proprietor version kindly made available for the task by the company LingIT[3]. In addition to corrections and consistency checks of NorKompLeks, this version also includes names. There are about 330 000 words in total in the LingIT version we used.

[2] http://www.phon.ucl.ac.uk/home/sampa/norweg.htm

[3] http://www.lingit.no/

For the alignment, all words found in the orthographic transcription must be listed in the lexicon. For standard words, this is trivial, and for non-dictionary words new entries can be created. However, as outlined in 3.1.2, the orthographic transcriptions also contain information about nonstandard forms and pronunciations. The nonstandard forms were used as input to the alignment and therefore also listed as independent entries in the pronunciation dictionary. The nonstandard pronunciations were not used in the initial annotation, but logged as corpus-specific variants for later usage (see 4.3).

For the 1897 different words in the APT material, 1697 were found in NorKompLeks. For these, we included all pronunciation variants in the pronunciation dictionary. For the remaining words we used an automatic grapheme-to-phoneme (G2P) converter. This G2P system is based on the front-end of a text-to-speech synthesis system (see Svendsen et al., 2005), and outputs typical pronunciations. On average this resulted in 1.08 pronunciations per word for the Spontal-N vocabulary. The initial annotation was performed using this lexicon only.

4.3. Adding Pronunciation Variants to Improve the Annotation

While a lexicon of canonical pronunciations may be suitable for automatic phonemic transcriptions of carefully pronounced words (e.g., /el@r/ for the word "eller"), such a lexicon does not suffice for the transcription of spontaneous speech. In Spontal-N we find fundamental differences between the acoustic signals of many word tokens and their corresponding canonical forms. For example, "eller" may be reduced to /@/ in certain contexts, as shown in Figure 2.

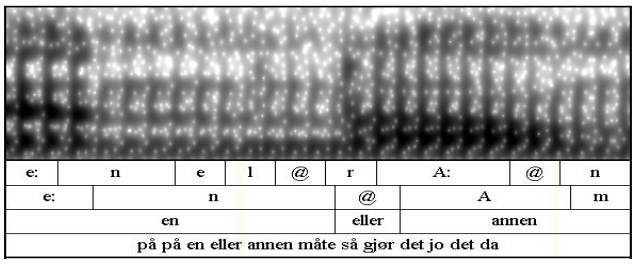

Figure 2: Initial automatic annotation (1st tier) vs. manual correction (2nd tier) of reduced speech.

To assess the hypothesis that adding pronunciation variants will give qualitatively better alignment, we experimented with two sources of pronunciation variation:

1. Corpus-based pronunciation variants
2. Co-articulation rules for word boundary effects

For both sources we added the new pronunciation variants to the ones used for the initial alignment. We then made a new alignment including both sources at a time as well as a combination.

4.3.1. Corpus-based pronunciation variants

During the manual transcription, a number of pronunciations deviating from the canonical pronunciations were collected. Some of these were annotated directly in the transcription files in a word-level tag; others (highly frequent pronunciations) were listed separately. For the vocabulary used in the initial annotation this resulted in 142 extra variants, raising the number of pronunciations per word to 1.16. This is not much, but there may be a significant effect on the automatic annotation, since highly frequent words are likely to have more variants, (Greenberg, 1999).

4.3.2. Co-articulation rules for word boundary effects

The text-to-speech system front-end used in 4.2 will output typical sentence-level pronunciations, i.e. taking cross-word co-articulation and some reductions into account. We extracted these general Norwegian co-articulation rules and added them to the speech recognition system. For example, for "er så", which was already listed in the dictionary as /{: r/ and /s O:/, we obtained the additional variant /{: S O:/. We also compensated for some NorKompLeks inconsistencies in word endings, by adding additional variants.

4.4. Evaluation of the Annotations

In order to evaluate the automatic annotations, we need a reference transcription (RT) for comparison. We therefore randomly selected 200 chunks from the initial automatic annotation, and corrected these manually (see Figure 2). Some information about the resulting RT in relation to the initial phonemic annotation (IA) is presented in Table 8.

	IA	RT
Number of chunks	3377	200
Total duration	74 min	5 min
Avg. chunk duration	1.32 sec	1.55 sec
Avg. words per chunk	5.7	6.8

Table 8: Initial annotation vs. reference transcription.

We evaluated the accuracy of the initial annotation and the three annotations presented in the previous section by comparing each of them to the manual RT. The evaluation is based on the minimal number of substitutions, deletions and insertions of symbols that are necessary to derive one transcription from the other without considering phonetic information. Since we calculate which changes must be done to the automatic annotation to arrive at the manual reference transcription, an insertion in Table 9 corresponds to the deletion of a phoneme. In the same way a deletion in the table corresponds to the insertion of a phoneme. Silence segments are counted as they may be involved in phoneme errors. Less than 1% of the errors were silence segment errors.

	Subs	Dels	Ins	Tot dis
Initial lex. based (IA)	6.5%	1.2%	12.2%	20.0%
Incl. corp. pron.var.	8.4%	2.6%	8.7%	19.7%
Incl. coart. pron.var.	6.1%	1.3%	11.4%	18.8%
Incl. all pron.var.	8.1%	2.7%	8.4%	19.2%

Table 9: Evaluation of automatic annotations.

The total number of disagreements between the initial annotation and the RT is 20.0%. This is similar to the results reported for Dutch in (Van Bael et al., 2007). As expected we can see that the majority (60%) of the transcription differences are due to reduced pronunciations in the manual (true) transcription. A closer inspection reveals that /@/ and /r/ are the most frequently reduced phonemes as has also been observed for Dutch. Adding the corpus specific variants did not give the expected boost in performance. More reduced forms are correct (fewer insertions in Table 9), but on the expense of more substitutions and deletions. Closer inspection of the differences is needed to improve this. The co-articulation rules gave a small improvement used alone, but not any added improvement in the combination with corpus specific variants. The number of pronunciations per word is still so low that we should not need pronunciation probabilities for a good result. The quality of the acoustic models may be insufficient and the results call for improvements, e.g., using corpus- or speaker adaptation.

5. On-going and Future Work

We are still working on improvements in the automatic phonemic annotation. For later versions we plan to adapt the acoustic models to the speakers, and to include data-driven pronunciation variation in the pronunciation dictionary. As we have found only scarce information on knowledge based pronunciation rules for Norwegian, we would like to investigate the effects of such resources for other languages. Dutch is one candidate since Norwegian and Dutch both are Germanic languages and there are far more resources available for spoken language technology.

Spontal-N (excluding the phonemic annotation) is currently being used in a 'Sound to Sense'[4] PhD study on the organisation of turns. This study will contribute to our understanding of how turn-transitions are projected and coordinated, with the use of different types of gestural and phonetic information.

Upon completion, Spontal-N will be made available to the research community.

6. Acknowledgements

The development of the Spontal-N corpus is funded by the EC-funded Marie Curie Research Training Network 'Sound to Sense' (S2S). The work described in this paper results from the collaboration between a number of institutions participating in S2S: the University of York, Royal Institute of Technology (KTH) Stockholm, Radboud University Nijmegen, and the Norwegian University of Science and Technology (NTNU) Trondheim. Thanks also to the 'Spontal' project (funded by the Swedish Research Council KFI, Grant for large databases, VR 2006-7482) for advice, the use of equipment, and technical assistance during the recordings.

[4] http://www.sound2sense.eu/

7. References

Amdal, I., Strand, O. M., Almberg, J. and Svendsen T. (2008). RUNDKAST: An Annotated Norwegian Broadcast News Speech Corpus. In *Proceedings of LREC 2008*, Marrakech, Morocco

Beskow, Jonas, Edlund, J., Elenius, K., Hellmer, K., House, D. and Strömbergsson, S. (2009). Project presentation: Spontal – multimodal database of spontaneous speech in dialog. In *Proceedings of FONETIK 2009*, Dept. of Linguistics, Stockholm University, Sweden.

Boersma, Paul & Weenink, David (2009). Praat: doing phonetics by computer (Version 5.1.14) [Computer program]. Retrieved August 5th 2009 from http://www.praat.org/

Greenberg, S. (1999). Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation. In *Speech Communication Volume 29,* pp. 159-176.

Johannessen, J. B., Priestley, J., Hagen, K., Åfarli, T. A., Vangsnes, Ø. A. (2009). The Nordic Dialect Corpus - an Advanced Research Tool. In *Proceedings of NODALIDA 2009. NEALT Proceedings Series Volume 4*. Odense, Denmark.

Kristoffersen, G. (2000). *The Phonology of Norwegian*. Oxford University Press.

Mazzoni, D. (2008). Audacity (Version 1.2.6) [Computer program]. Retrieved February 6th 2008 from: http://audacity.sourceforge.net/

Nordgård, T. (2000). NorKompLeks: A Norwegian computational lexicon. In *Proceedings COMLEX 2000*, Patras, Greece, pp. 89 – 92.

Svendsen, T., S. Spildo, J. O. Fretland, T. Breivik. (2008). Plan for etablering av norsk språkbank (in Norwegian). Report to Ministry of Culture. Available from http://www.sprakrad.no/Tema/IKT--sprak/Norsk-sprakbank/

Svendsen, T., I. Amdal, I. Bjørkan, P.O. Heggtveit, D. Meen, J.E. Natvig. (2005). Fonema - Tools for Realistic Speech Synthesis in Norwegian, In *Proceedings Norsig 2005*, Stavanger, Norway.

Van Bael, C., Boves, L., Heuvel, H. v. d. and Strik, H. (2007). Automatic phonetic transcription of large speech corpora. *Computer Speech and Language*, 21, pp. 652 – 668.

Young, S. et al. (2006), HTK version 3.4. [Computer program]. Retrieved December 14th 2006 from http://htk.eng.cam.ac.uk/

Wells, J.C. (1997). SAMPA computer readable phonetic alphabet. In Gibbon, D., Moore, R., Winski, R. (Eds.), *Handbook of Standards and Resources for Spoken Language Systems*. Berlin and New York: Mouton de Gruyter. Part IV, section B