The Kachna L1/L2 Picture Replication Corpus

Helena Spilková*, Daniel Brenner†, Anton Öttl†, Pavel Vondřička†, Wim van Dommelen*, Mirjam Ernestus†,‡

*Norwegian University of Science and Technology
Trondheim, Norway
{helena.spilkova, wim.van.dommelen}@hf.ntnu.no
†Radboud University
Nijmegen, Netherlands
{dan.brenner, anton.oettl, pavel.vondricka, mirjam.ernestus}@mpi.nl
‡Max Planck Institute for Psycholinguistics
Nijmegen, Netherlands

Abstract

This paper presents the Kachna Corpus of Spontaneous Speech, in which ten Czech and ten Norwegian speakers were recorded both in their native language and in English. The dialogues are elicited using a picture replication task that requires active cooperation and interaction of speakers by asking them to produce a drawing as close to the original as possible. The corpus is appropriate for the study of interactional features and speech reduction phenomena across native and second languages. The combination of productions in non-native English and in speakers' native language is advantageous for investigation of L2 issues while providing a L1 behaviour reference from all the speakers. The corpus consists of 20 dialogues comprising 12 hours 53 minutes of recording, and was collected in 2008. Preparation of the transcriptions, including a manual orthographic transcription and an automatically generated phonetic transcription, is currently in progress. The phonetic transcriptions are automatically generated by aligning acoustic models with the speech signal on the basis of the orthographic transcriptions and a dictionary of pronunciation variants compiled for the relevant language. Upon completion the corpus will be made available via the European Language Resources Association (ELRA).

1. The Speakers

The speakers in the Kachna[1] Corpus were ten native Czech and ten native Norwegian speakers. Most speakers were university students, between 19 and 35 years of age. The speakers in each pair were either classmates or colleagues. Various dialects of Czech and Norwegian are represented in the recordings. Considerable English proficiency can be assumed for all speakers on the basis of their education and occupation (all are university students or employees in a company using English as the official work language). Speakers' confidence in being able to perform the replication task in English is also implicit in their decision to participate in the recordings. Details about individual speakers, their dialects, their onset of English exposure, and other information related to their use of English can be found in Table 1.

2. The Task and Instructions

The elicitation of spontaneous speech in an unnatural setting such as a recording studio is a difficult endeavour and has attracted the attention of a growing number of researchers. For a discussion of various tasks used for this purpose, see Ito and Speer (2006). The present corpus is based on a "picture replication" task. The advantage of using tasks in collecting spontaneous speech is that they enable a certain degree of control over discourse content and structure. Moreover, tasks distract the attention of speakers from the fact that they are being recorded, which might otherwise undermine the naturalness of their speech performance.

In the picture replication task, one speaker receives one of three detailed cartoon drawings[2] depicting humorous scenes. This speaker is instructed to describe it to the drawer, whose task is to replicate the picture as accurately as possible on a sheet of paper using a pencil. In order to encourage the active participation of the drawing speaker in the dialogue, an accompanying task was added: the sheet for the drawing contained five detail sections cut out from the original picture and the drawing speaker was instructed to identify their content and determine their location within the picture. Neither of the speakers could see the other's picture. The speakers were asked to use approximately 30 minutes for the task.

In Czech the standard language differs from the colloquial or dialectal varieties. In formal situations (e.g. in lectures, university examinations, and the media) the standard variety is considered appropriate, but for the Kachna recordings, speakers were explicitly encouraged to use their dialectal or colloquial varieties of Czech, i.e. to speak together the way they normally would in their everyday interactions. In Norway, however, the use of dialects is broadly accepted and encouraged, and dialects are considered appropriate in all social situations. Due to the status of Norwegian dialects, no special instructions regarding dialect were necessary for the Norwegian speakers.

The nature of the task requires active cooperation and interaction of speakers by asking them to produce a drawing

[1] *Kachna* ['kaxna] 'duck', from Czech. A duck appears in a picture used in the task.

[2] Accessible at: http://www.multimediamanufaktur.org/produkt_wimmel.htm

L1	Pair	Speaker	Age	Gender	Region	English Exposure Onset	Other English-related Information
Norwegian	1	a	28	M	Trøndelag	10	
		b	26	M	West Norway	bilingual	raised in Norway; mother is American
	2	a	22	M	Trøndelag, East Norway	5	holidays in English-speaking countries
		b	23	F	Trøndelag	7	holidays in English-speaking countries
	3	a	19	F	East Norway	10	2 yrs. in English-language high school
		b	19	M	East Norway	3	lived in USA from age 3 – 7.5; 2 yrs. in English-language high school
	4	a	22	M	West Norway	8	
		b	23	M	East Norway	8	
	5	a	25	F	East Norway, North Norway	10	
		b	25	F	East Norway	10	
Czech	6	a	20	M	S.W. Bohemia	10	
		b	20	M	S.W. Bohemia	6	
	7	a	21	F	Central Bohemia	11	1 yr. in USA
		b	20	M	Central Bohemia	10	
	8	a	21	F	Central Bohemia	11	two 14-day courses in UK, Ireland
		b	20	F	Central Bohemia	11	frequent short ($\sim$ 1 wk.) stays in UK
	9	a	21	M	Central Bohemia	3	work in English-speaking company (1 yr.)
		b	35	M	Central Bohemia	15	work in English-speaking companies (5 yrs.)
	10	a	21	F	East Moravia, Central Bohemia	9	
		b	20	F	Central Bohemia	13	

Table 1: Speaker Data, including native language (L1), pair and speaker identifiers, age (yrs.), gender ((M)ale, (F)emale), region(s) of speaker dialect, the onset of speaker's exposure to English (yrs. of age), and other information relevant to the speaker's English experience.

as close to the original as possible. Because all speaker pairs discuss details of the same pictures in similar discourse contexts, many of the same lexical items appear in several recordings. Moreover, the speakers often produce multiple repetitions of the same words or other elements.
In most cases, the describing speakers were more active, while the drawing speakers mainly asked questions for clarification. The speakers seemed to enjoy the task. This can be indirectly observed in the durations of the dialogues, as many of them exceed the recommended task length. A summary of recording durations is given in Table 2.

3. Audio Recording and Processing

The dialogues were recorded in a sound treated studio at the Department of Language and Communication studies, NTNU, Trondheim, for the Norwegian speakers, and at the Institute of Phonetics, Charles University, Prague, for the Czech speakers. The dialogues were recorded in stereo (one channel per speaker), with a sampling rate of 44.1 kHz and 16-bit quantisation. Table 3 lists the technical parameters for recordings made at the two studios.
The speakers were recorded using two boom-mounted microphones. Due to the activities involved in the replication task, the speakers were not expected to move significantly relative to the microphone, and the resulting recordings are quite good. Recording practices were consistent within each of the two studio locations, but the dimensions of the two recording studios were dissimilar. The size of the studio in Trondheim allowed for the speakers to sit several meters apart, back to back, each at a different table. The studio in Prague was much smaller, so it was necessary for speakers to sit next to each other at the same table, partially visually separated by a styrofoam board (obscuring one another's pictures from view during the task). As a result of these differing layouts, the recordings of the Norwegian speakers have channels very well separated while those of the Czech speakers have a strong cross-channel overlap.
As a consequence of the recording studio layout and the activities involved in the replication task, visual face-to-face interaction between the speakers was practically eliminated. The absence of visual contact forces speakers to rely fully on the auditory channel, much like in everyday telephone calls, which makes the recordings well suited for investigation into the use of acoustic information in oral communication.

L1	Pair	Language	Duration	Language	Duration
Norwegian	1	English	53	Norwegian	34
	2	English	48	Norwegian	38
	3	English	49	Norwegian	21
	4	English	57	Norwegian	34
	5	English	73	Norwegian	38
		Total:	280	**Total:**	164
Czech	6	English	41	Czech	29
	7	English	38	Czech	27
	8	English	31	Czech	30
	9	English	35	Czech	25
	10	English	32	Czech	41
		Total:	177	**Total:**	153

Table 2: Recording Durations (minutes)

	Trondheim	Prague
Audio format	wav	wav
Microphone	MILAB LSR-1000	AKG C 4500 B-BC
Preprocessing	high-pass filter (50 Hz)	
Sound card	Creative SB Live	Sound Blaster Audigy 4
Sampling Rate	44.1 kHz	44.1 kHz
Quantisation	16-bit	16-bit
Audio processing software	Adobe Audition V.2	Sound Audio Studio 8.0

Table 3: Recording and Audio Processing Specifications

4. Recording Sessions

The structure of the recording session was the same for all the participating speaker pairs. After the speakers were instructed in the task, the session started by recording the speakers performing the task in English (their second language) and proceeded until the speakers were satisfied with their achievement. The describing speaker is identified as speaker *a* in Table 1 above. After the first recording, a short refreshment break followed where the speakers could amuse themselves by comparing the model picture and the resulting drawing. Subsequently they carried out the same task in their native language. For this second recording, a new picture was provided and the speakers exchanged roles, so that the describer role is taken by the speaker identified as *b* in Table 1.

5. Orthographic Transcriptions

The recordings are accompanied by Praat (Boersma and Weenink, 2008) formatted Unicode text files providing transcriptions of the recorded speech and other sound events. The orthographic transcriptions were created by native-speaking researchers of the three languages in the corpus (author A.Ö. for Norwegian, D.B. for English, P.V. for Czech), and contain annotation of the words spoken by each speaker (including overlapping speech), breathing, laughter, and other noises present in the recording.

While we aim in our transcriptions to adhere to standard orthographic norms, this is not always straightforward. First, while some contracted forms such as English [ˈkɑːnt] 'can't' appear as pronunciation variants in most common dictionaries, many spoken colloquial forms such as [ˈamənə] meaning 'I'm going to' do not. Second, the writing systems for the three languages pose different problems for the transcriptions, and there are a few language-specific challenges. For example, Norwegian has two co-existing orthographic standards, Bokmål and Nynorsk. We adhere in our transcriptions to Bokmål norms, which allow for a good amount of variation. Words in the corpus varying greatly from this standard are tagged and their pronunciation noted. In Czech, colloquial varieties we find in our corpus often depart drastically from the written standard. A variety of forms may be used with varying degrees of relation to the standard written form. A delicate treatment of these various forms was required (see the accompanying documentation for details).

The orthographic transcriptions of each recording are composed of one tier for each speaker present in the recording. In a speaker's tier, we transcribe the speech of that speaker and also noises attributed to that speaker (such as coughing or swallowing). Intervals in the tier were chosen as minimal continuous stretches of speech for the speaker. Also contained in a speaker's tier are tags indicating properties of words or word sequences. For example, the Norwegian word *møkkagrep* 'pitchfork', in an English recording is transcribed `møkkagrep\f[Norwegian]`, where "\f" denotes a foreign word. Additionally, one tier is provided in which speaker-independent noises are indicated. These may be environmental sounds, or other sounds not clearly

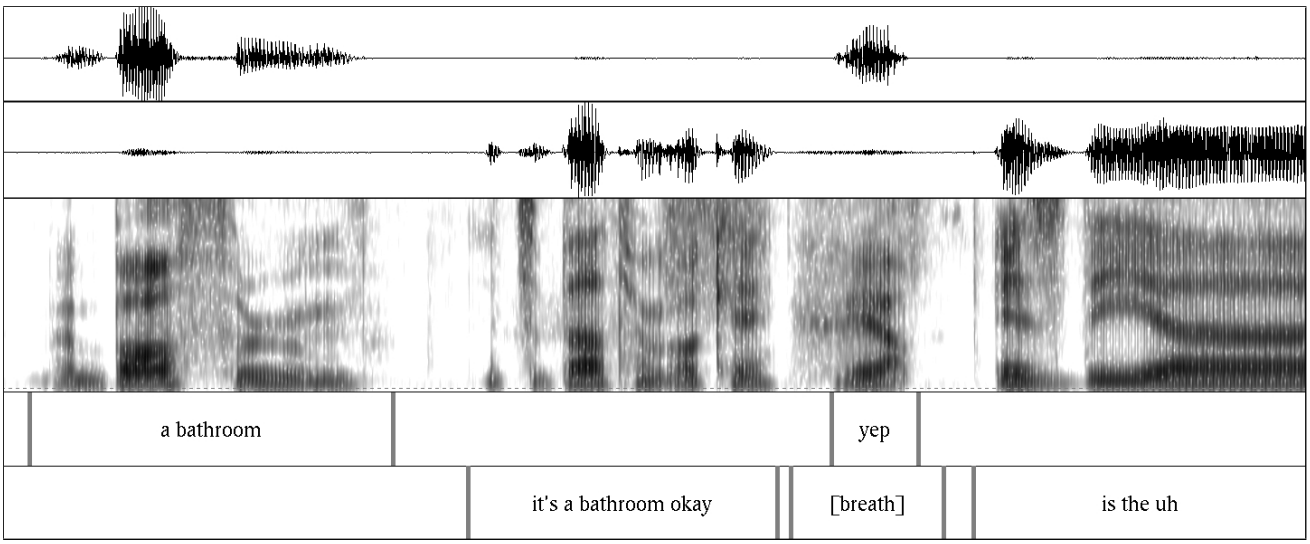

Figure 1: Sample Praat Display. Vertically from top: left channel waveform, right channel waveform, spectrogram, first speaker transcription, second speaker transcription (subsequent tiers are empty for this section of the recording).

attributable to a particular speaker. A "Notes" tier provides other miscellaneous information.

6. Phonetic Transcriptions

The automated phonetic alignment process aligns phonetic transcriptions of the words with the acoustic speech signal of the recording on the basis of the orthographic transcription, a pronunciation dictionary (lexicon), and acoustic phone models (Binnenpoorte, 2006). The inputs and output are shown in Figure 2.

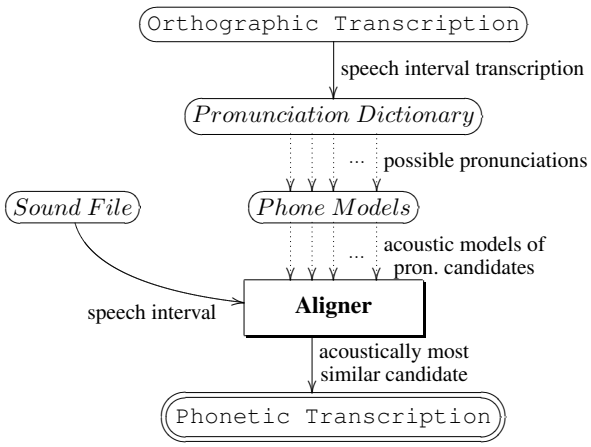

Figure 2: Input/Output Structure of the Phonetic Alignment Process

The pronunciation dictionary will be created in two steps. First, all word types in the subcorpora will be extracted and their canonical pronunciations will be referenced in established phonetically annotated databases. For example the English language section of the CELEX database (Baayen et al., 1995) will be used as the base for the English lexicon. Since spontaneous speech contains a large proportion of reduced forms, existing databases are insufficient for capturing all the acoustic realisations we find in the Kachna Corpus. It is necessary to expand the lexica to include entries for these reduced pronunciations. To accomplish this, we apply known reduction processes from the speech reduction and phonetic variation literature such as Shockey (2003) and Johnson (2004), *et alia*, to the unreduced pronunciation variants of words or word combinations. Additionally, we incorporate relevant observations of the transcribers during transcription, enabling the dictionary to be tailored to the recordings. The result is a list of possible pronunciation variants of each word or phrase, to be compared with the speech signal of the recordings.

For each speech interval in the corpus, the orthographic transcription is translated into possible phonetic transcriptions on the basis of the pronunciation dictionary. These phonetic transcriptions are compared with the corresponding acoustic signal from the original recording using phone models trained for the relevant language. An acoustic alignment process is then employed to assign each candidate transcription a score of acoustic similarity and the candidate with the highest score is selected. The final result will be a new tier in the text files in which phonetic symbols are aligned with the acoustic signal of the recording. This process is similar to that used in other phonetic alignment systems such as MAUS (Schiel et al., 1998).

7. Transcription Progress

The transcriptions of the Czech material are complete. The automatic alignment was carried out using the Prague Labeller (Pollák et al., 2007), a phone segmentation tool constructed on the basis of the Hidden Markov Model Toolkit (HTK) (Young et al., 2009) and implemented in a Praat environment (Boersma and Weenink, 2008). The labeller's basic integrated pronunciation lexicon has been extended based on the observations of the transcriber, as described above.

The word level alignment achieved with this method on the Czech material is reliable for searching in the corpus and for basic durational measurements. An accuracy evaluation

on 30 randomly chosen short intervals containing a total of 643 words (portions of overlapping speech and strong noise were excluded) has shown that 87% of word boundaries were placed within $\pm$30 ms of a reference segmentation. The alignment on the segmental level, however, can only be used as basic pre-segmentation of the material and the segment boundaries have to be manually adjusted. This is due to the large variability in casual speech, frequent occurrence of overlapping talk, and inadequate modelling of certain phenomena frequent in spontaneus speech (e.g. hesitation sounds, creaky voice).

The transcription of the Norwegian part of the corpus is scheduled for completion in April 2010. The automatic alignment will be carried out using a lexicon based on the NorKompLeks lexicon[3] (Nordgård, 2000). Phone models were trained on roughly 20 hours of continuous manuscript-read Norwegian speech from $\sim$900 speakers in the NST corpus from the National Language Resource Bank (Svendsen et al., 2008). Alignment will be performed via HTK.

English transcriptions are also in progress utilising a lexicon built on the CELEX Database (Baayen et al., 1995), phone models trained on the TIMIT Corpus (Zue and Seneff, 1988), and HTK for the alignment. Completion is anticipated May 2010.

8. Conclusion

The Kachna Corpus promises to be an important resource for study along several dimensions. The corpus is appropriate for the study of interactional features and speech reduction phenomena across native and second languages, as well as intra- and inter-speaker variability in these language contexts. The combination of productions in non-native English and in speakers' native language is advantageous for investigation of L2 issues while providing an L1 behaviour reference for all speakers. The balanced demographic profile of the speakers is also suitable for analysing a rich set of variables. The corpus is on schedule to be completed for presentation at LREC 2010, and will be offered to the European Language Resources Association (ELRA) for distribution.

9. Acknowledgements

Creation of the Kachna Corpus is funded by the European Union Marie Curie Research Training Network 'Sound to Sense' (S2S). It is the collaborative effort of the following member institutions: Norwegian University of Science and Technology (NTNU), Trondheim; Radboud University, Nijmegen; and Charles University, Prague. We would also like to express our gratitude to three anonymous reviewers for their thoughtful comments.

10. References

R. H. Baayen, R. Piepenbrock, and L. Gulikers. 1995. The CELEX lexical database [CD-ROM]. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

D. Binnenpoorte. 2006. *Phonetic transcriptions of large speech corpora*. Ph.D dissertation, Radboud University, Nijmegen, the Netherlands.

P. Boersma and D. Weenink. 2008. Praat: doing phonetics by computer (version 5.1). [Computer program], retrieved Oct. 21, 2009 from `http://www.praat.org/`.

K. Ito and S. R. Speer. 2006. Using interactive tasks to elicit natural dialog. In S. Sudhoff, editor, *Methods in Empirical Prosody Research*, pages 229–257. Walter de Gruyter, Berlin.

K. Johnson. 2004. Massive reduction in conversational American English. In K. Yoneyama and K. Maekawa, editors, *Spontaneous Speech: Data and Analysis. Proceedings of the 1st Session of the 10th International Symposium*, Tokyo, Japan. The National Institute for Japanese Language.

T. Nordgård. 2000. NorKompLeks: A Norwegian computational lexicon. In *Proceedings of COMLEX 2000, Workshop on Computational Lexicography and Multimedia Dictionaries*, pages 89–92, Patras, Greece.

P. Pollák, J. Volín, and R. Skarnitzl. 2007. HMM-based phonetic segmentation in Praat environment. In *Proceedings of the VIIth International Conference "Speech and Computer – SPECOM 2007"*, volume 1, pages 537–541, Moscow, Russia. MSLU.

F. Schiel, A. Kipp, and H. G. Tillmann. 1998. Statistical modelling of pronunciation: It's not the model, it's the data. In *Proceedings of the ESCA Workshop 'Modeling Pronunciation Variation for Automatic Speech Recognition'*, pages 131–136.

L. Shockey. 2003. *Sound Patterns of Spoken English*. Blackwell, Malden, MA.

T. Svendsen, S. Spildo, J. O. Fretland, and T. Breivik. 2008. Plan for etablering av norsk språkbank ('Plan for establishing a Norwegian speechbank', in Norwegian). Report to Ministry of Culture. Available from `http://www.sprakrad.no/Tema/IKT--sprak/Norsk-sprakbank/`.

S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. 2009. *The HTK Book*. Cambridge University Engineering Department.

V. Zue and S. Seneff. 1988. Transcription and alignment of the TIMIT database. In *Proceedings of the 2nd Meeting on Advanced Man – Machine Interface through Spoken Language*, pages 11.1–11.10.

[3] LingIT version, `http://www.lingit.no`.