Large Broadcast News and Read Speech Corpora of Spoken Czech

Josef Psutka[1], *Vlasta Radová*[1], *Luděk Müller*[1], *Jindřich Matoušek*[1], *Pavel Ircing*[1], *and David Graff*[2]

[1]University of West Bohemia, Department of Cybernetics, Univerzitní 8,
306 14 Plzeň, Czech Republic
{psutka, radova, muller, jmatouse, ircing}@kky.zcu.cz

[2] University of Pennsylvania, Linguistic Data Consortium, Philadelphia, USA
graff@ldc.upenn.edu

Abstract

This paper presents the first annotated and phonetically transcribed large speech corpora developed for spoken Czech. All corpora were collected during the last two years at the Department of Cybernetics, University of West Bohemia (UWB) in Pilsen. The first two collections are broadcast news, the third corpus is a high-quality read-speech database. This paper describes the collection conditions, annotation and phonetic transcription process related to each corpus. The basic phonetic and lexical characteristics of all corpora will be given and compared mutually.

1. Introduction

In the presented paper we describe our efforts with the acquisition, annotation and phonetic transcription of three large corpora designed for experimental work with continuous Czech speech recognition. Perhaps it is already known that Czech belongs to the family of Slavic languages that demonstrate interesting phenomena such as a high degree of inflection, a high degree of derivations (prefixes and suffixes) and a relatively free word order. These characteristics cause many problems with speech recognition [1]. For example, a 50,000-word vocabulary covers 99.3% of tokens in running English but only 90.5% of tokens in Czech. Even a 350,000-word vocabulary that has practically perfect coverage of English covers only 97.6% of Czech text.

On the other hand, Czech phonetics may seem a little simpler when compared to English. The Czech phoneme inventory consists of 44 distinct elements whose orthographic forms are quite regular [2]. There is a basic five-vowel system [a, e, i, o, u] with distinct long and short forms; the vowel system also includes three diphthongs [ow, aw, ew], and of these, the latter two occur only rarely, in words of foreign origin. There are, of course, some consonants not found in English, particularly the voiceless and voiced retroflex fricatives (which are represented orthographically by the single letter "ř", and in our phonetic alphabet as [rzh] and [rsh] – these segments are perhaps unique to Czech); also, the velar fricative [x], and the alveolar affricates [dz] and [c]. An important phonetic difference relative to English involves the rules in Czech for voicing assimilation of consonants across word boundaries: within a fluently spoken, continuous phrase (i.e. in the absence of a pause between words), the final consonant of a given word will be voiced or voiceless, in accordance with the voicing of initial segment of the following word. For example, the word sequence "*dvacet dva*" (Engl. twenty two), if uttered with a pause at the word boundary, would be transcribed phonetically as [dvacet_dva], but if uttered without a pause, the transcription would be [dvaced_dva]. This in itself is fairly common feature among various languages, but when we regard the combination of phonetic, morphological and syntactic features of Czech, we find, overall, some novel challenges for the standard methods of acoustic and language modeling.

Development of a large vocabulary speaker independent speech recognizer requires the preparation of a sufficiently large database of suitable utterances. During the past two years we were collecting, annotating and transcribing 3 large speech corpora. This work was performed with the financial support of CLSP, Johns Hopkins University, Baltimore, and the Grant Agency of the Czech Republic and with the help of LDC, University of Pennsylvania. The first two corpora are the Voice of America Broadcast News (UWB_B01) and the Czech TV&Radio Broadcast News (UWB_B02), the third corpus is the Czech Read-Speech Corpus (UWB_S01).

This paper is organized as follows: Section 2 describes the way in which all three corpora were acquired. A brief explanation will also be devoted to the selection of phonetically balanced and phonetically rich read sentences in the UWB_S01 corpus. Then, Section 3 provides information on the corpora annotation and Section 4 describes the phonetic transcription process using a set of phonetic rules and a vocabulary of exceptions. Section 5 gives some interesting phonetic and lexical statistics on all three corpora.

2. Data acquisition

The mentioned corpora were collected at the Department of Cybernetics, University of West Bohemia (UWB) in Pilsen during last two years as follows:

2.1. Voice of America Broadcast News Czech (UWB_B01)

Between February 9 and May 28, 1999, the Linguistic Data Consortium collected more than 30 hours of audio recordings from the Voice of America news service in Czech. The data files present in this corpus represented the daily broadcasts of 30-minute news programs. These files were transferred on a daily basis via ftp to the Department of Cybernetics, UWB Pilsen. At UWB, all recordings were audited to eliminate files that were unsuitable for transcription. This was necessary because during a portion of the collection period, the LDC had experienced problems in their reception of the VOA satellite broadcast, causing some files to have random but pervasive

intervals of noise, distortion, or total loss of signal. Overall, 62 files were retained for use. These consist of single-channel, 16 kHz, 16-bit linear sample data, and contain over 20 hours of transcribed speech.

2.2. Czech TV&Radio Broadcast News (UWB_B02)

The UWB_B02 is the Czech TV&Radio Broadcast News Corpus spanning the period February 1, 2000 through April 22, 2000. During this time news broadcasts on 3 TV channels and 4 radio stations were recorded. The whole corpus contains over 50 hours of audio stored on 347 audio files, which yield about 26 hours of pure transcribed speech. For the broadcast sources the following TV channels and radio stations were recorded:

TV: ČTV1 (22 files), NOVA (30), PRIMA (23)
Radio: RADIOŽURNÁL (138 files), PRAHA (57 files), VLTAVA (14), FREKVENCE1 (30)

The broadcast news does not contain weather forecasts, sports news and traffic announcements. The signal is single-channel, sampled at 22.05 kHz with 16-bit resolution.

2.3. Czech Read-Speech Corpus (UWB_S01)

The UWB_S01 corpus is a high-quality read-speech corpus consisting of the speech of 100 speakers (64 male and 36 female). Each speaker read 150 sentences, comprised of two subsets: an "adaptation" set of 40 sentences, which were identical for all speakers (to support speaker-adaptation modeling), and a "training" set of 110 sentences, in which no sentences were repeated across speakers (to support speaker-independent acoustic training). The database of text prompts from which the adaptation and training sentences were selected was obtained in an electronic form from the web pages of 3 Czech newspapers: *Lidové noviny*, *Mladá fronta Dnes* and *Právo*. The texts were obtained in a HTML format. The texts obtained were processed to divide each HTML file into particular articles, to transform each article to a uniform format, and to correct typing errors. For this reason a special vocabulary containing four categories of words (regular, phonetically exceptional, foreign words and abbreviations) was built and then used for sentence selection or the correction of typing errors. To assure that the text could be read aloud fluently, we limited the selection to sentences that ranged between 3 and 15 words in length, and contained no foreign vocabulary (since the pronunciation of foreign words tends to be irregular and quite variable). The conditioning of the text data was done automatically using special software tools [3]. The entire database of 32,983 text sentences with 333,884 tokens and 36,499 distinct words was then transliterated, again by an automated process, to strings of phonemic segments, in accordance with a set of phonological rules (see Section 4).

Special consideration was given to the selection of the 40 adaptation sentences, since these would need to provide a representative distribution of the more frequent triphone sequences (reflecting their relative likelihood in natural speech); the selection procedure is described in detail in [3]. The selection of training sentences consisted of 2 steps. First, a minimum set of sentences was selected so that it contained all the triphones occurring in the phonetically transcribed sentences. We call this set "necessary sentences". Then, the remaining sentences were selected with the goal of "naturally" balancing the relative number of triphones in the selected set of sentences [3].

The corpus was recorded in an office where only the speaker was present. Recordings were performed using the notebook IBM TP 760 ED owing to the very silent operation of this computer (it hasn't any fan). However, some noise from neighboring offices was sometimes heard in the recording room. Each sentence was recorded by two microphones simultaneously. A close-talking microphone (Sennheisser HMD410-6) yielded utterances of a high-quality, a desk microphone (Sennheisser ME65) recorded utterances including common office noise. The prompting/recording sessions yielded a total of 25 hour of speech, all of which was digitized into pairs of single-channel files at 44.1 kHz with 16-bit resolution.

Table 1 shows some basic numbers related to the described corpora. We appended also data of the fourth database denoted LN corpus that we had at our disposal. This text corpus was selected from newspaper articles of *Lidové noviny* (spanning period 1991 through 1995) and can be considered as a part of the Czech National Text Corpus (CNTC) (the whole CNTC comprises more than 100 millions tokens).

	UWB_B01	**UWB_B02**	**UWB_S01**	**LN**
#SEN	6,573	16,483	11,040	2,253,095
#TOK	108,097	233,959	93,327	33,216,772
#A_T/S	16.45	14.19	8.45	14.74
#D_W	21,256	31,936	18,262	658,549

Table 1: Basic numbers related to the developed corpora (where *#SEN* denotes the number of sentences, *#TOK* the number of tokens, *#A_T/S* determines average number of tokens per sentence and *#D_W* indicates the number of distinct words).

3. Annotation

The speech data of each corpus was annotated using special annotation software Transcriber 1.4.1 running under Windows NT. Transcriber is a tool for assisting the creation of speech corpora. It allows the manual segmenting, labeling and transcribing speech signals for later use in automatic speech processing. The package is currently available from the Linguistic Data Consortium (LDC) web site: http://www.ldc.upenn.edu/. Using the Transcriber, each file was divided into segments. Each segment contains usually a sentence or some of the non-speech events listed in Table 2.

AH	LOUD_BREATH
COUGH	PAPER_RUSTLE
DOOR_SLAM	SIGH
GRUNT	TONGUE_CLICK
LIP_SMACK	UNINTELLIGIBLE
MM	MOUSE_CLICK
PHONE_RING	MIKE_OVERLOAD
THROAT_CLEAR	REMOTE_ENGINE
UH	NOISE
UM	KNOCK_ON_MIC
CHAIR_SQUEAK	MUSIC
CROSS_TALK	BACKGROUND_MUSIC
ER	SIGNAL_MISSING
LAUGHTER	SILENCE

Table 2: List of the descriptors of non-speech events.

If a segment of the audio file contained no spoken content, but only an extended period of silence and/or one or more of the non-speech events listed above, it was marked as a "Nontrans" segment. Also, if there was spoken content, but this was accompanied by REMOTE_ENGINE or NOISE, this portion was also marked as "Nontrans"; all other segments with spoken content were marked as "Report". If particular words had pronunciations that were exceptions to the phonological rules enumerated for Czech (see Section 4) – i.e. foreign names, etc. – these were marked by a special prefix character in the transcription, so that the irregular pronunciations could be entered into the phonetic vocabulary later on. After the phonetic vocabulary was complemented, these special prefix characters were removed from the final transcripts.

- Spoken numbers were transcribed using their text forms.
- If an acronym was spoken as a word (e.g. "NATO") it was written as a word ("NATO"). If an acronym was spoken as a sequence of letters (e.g. "NSF") it was written according to the actual Czech pronunciation, either "N S F" or "EN ES EF".
- Non-speech events were indicated by a descriptor enclosed in square brackets. The descriptor was placed at the point at which the non-speech event occurred, e.g. "[PAPER_RUSTLE]". If a non-speech event overlapped a spoken lexical item, the descriptor was placed close to the item that was overlapped and the character "<" or ">" was appended to the descriptor depending on whether it was placed to the left or right of the co-occurring lexical item, e.g. "[<PAPER_RUSTLE]".
- If the waveform was truncated the symbol "~" was used to mark the incompletely spoken sentence. If a word was spoken incompletely due to the truncation, the spoken fragment of the word was delimited by "*", e.g. "Voice of *Ame* ~ ".

4. Phonetic transcription

Two different approaches are usually used to perform the phonetic transcription automatically. For languages like Czech both with a high degree of inflection (there are many forms for a single word) and with a high degree of derivation (usage of prefixes and suffixes) phonetic transcription by rules is preferred, since there are too many forms of words to be fitted in a dictionary. Phonetic transcription rules in a form like (1) are used to generate a sequence of phonemes from a sequence of letters [4]. Letter sequence A with left context C and right context D is transcribed as phone sequence B

$$A \rightarrow B \,/\, C\_D\,. \qquad (1)$$

For less inflectional languages (e.g. English) a dictionary-based approach to phonetic transcription is almost always used. In this conception a "phonetic dictionary", which contains phonetic transcriptions of every word, is employed. The phonetic transcription of a given word is then performed by looking it up in a dictionary.

Owing to the many international words appearing relatively often in a Czech text which are not subject to the phonetic rules of Czech we applied a combination of both approaches. We used about 50 rules together with a dictionary of "phonetic exceptions". This dictionary mostly consists of non-Czech words and currently contains about 620 stem-like word forms. One part of the phonetic rule inventory is also a small set of generalized rules describing the cross-word assimilation process that is typical for spoken Czech.

5. Phonetic and lexical statistics of collected corpora

This Section introduces some interesting phoneme and lexical statistics measured on the collected corpora. Table 3 gives the relative occurrences of ten most frequent phonemes including short and long pauses. Analyzing the phoneme attributes of all three corpora we verified the full coverage of the Czech phoneme inventory by the UWB_S01 corpus however we found that the phoneme [mg] (e.g. in the Czech word *tramvaj*, Engl. tram) in the UWB_B02 corpus and the phonemes [mg, dz] (*podzim*, Engl. autumn), and [dzh] (*džez*, Engl. jazz) in the UWB_B01 corpus did not occur at all.

UWB_B01		UWB_B02		UWB_S01	
Pho-neme	Relative occur. [%]	Pho-neme	Relative occur. [%]	Pho-neme	Relative occur. [%]
sp	13.74	sp	14.10	sp	14.09
e	7.77	e	7.77	e	7.72
o	6.22	o	6.27	o	6.61
i	5.61	a	5.59	a	4.93
a	5.52	i	5.25	i	4.86
s	4.24	s	4.03	sil	4.59
n	3.90	t	3.89	t	3.91
t	3.86	n	3.85	n	3.71
l	3.52	l	3.53	ii	3.65
r	3.39	r	3.46	s	3.60

Table 3: Relative occurrences of the ten most frequent phonemes (including short sp and long sil pauses) for each corpus.

Table 4 shows the number of within-word triphones (WiWoTrip), and the combined total of within-word and cross-word triphones (CrWoTrip), for each corpus. The union of distinct triphones across all three corpora totals 10,667 for the within-word triphones, and 22,099 for the within-word plus cross-word triphones. The numbers of triphones were computed after the HMMs of triphones were estimated and the best transcription was assigned to the uttered sentences (including the cross-word assimilation).

	UWB_B01	UWB_B02	UWB_S01
WiWoTrip	7,902	8,916	8,382
CrWoTrip	16,726	18,930	14,358

Table 4: Number of within-word triphones (WiWoTrip) and within-word plus cross-word triphones (CrWoTrip) for each corpus.

Table 5 shows the ten more frequent triphones for each corpus (drawn from the full inventory of within-word and cross-word triphones).

Table 6 shows the most frequent words in each corpus. They are mostly conjunctions (*a*, *i*) and prepositions (*na*, *v*, *že*, *o*). The rather different ranking of word frequencies in the UWB_S01 corpus is due, of course, to the fact that its content was carefully selected for phonological balance and was constrained by a limitation on sentence length.

UWB_B01		UWB_B02		UWB_S01	
Tri-phone	Rel. occur. [%]	Tri-phone	Rel. occur. [%]	Tri-phone	Rel. occur. [%]
p-r+o	0.36	p-r+o	0.38	p-r+o	0.44
v-j+e	0.28	v-j+e	0.27	e-nj+ii	0.35
e-nj+ii	0.27	s-k+ee	0.26	m-nj+e	0.31
i-c+k	0.26	m-nj+e	0.26	v-j+e	0.27
o-v+a	0.25	e-nj+ii	0.25	p-o+d	0.27
t-e+r	0.24	i-s+t	0.23	o-s+t	0.25
k-t+e	0.23	p-rsh+e	0.23	e-s+t	0.24
s-k+ee	0.22	ee-h+o	0.23	o-v+a	0.23
m-nj+e	0.22	o-s+t	0.22	s-p+o	0.22
i-s+t	0.22	p-o+d	0.21	p-rsh+e	0.22

Table 5: Relative occurrences of ten most frequent triphones.

UWB_B01		UWB_B02		UWB_S01		LN	
Word	Rel. occur. [%]	Word	Rel. occur. [%]	Word	Rel. occur. [%]	Word	Rel. occur. [%]
a	0.027	*v*	0.022	*je*	0.021	*a*	0.025
v	0.025	*a*	0.021	*se*	0.021	*v*	0.022
se	0.018	*se*	0.018	*v*	0.017	*se*	0.017
na	0.017	*na*	0.017	*to*	0.017	*na*	0.015
že	0.016	*že*	0.010	*na*	0.016	*je*	0.009
je	0.009	*o*	0.010	*a*	0.016	*že*	0.008
o	0.008	*je*	0.008	*o*	0.009	*o*	0.007
z	0.007	*z*	0.007	*i*	0.009	*z*	0.007
s	0.007	*to*	0.007	*že*	0.008	*s*	0.007
k	0.006	*s*	0.007	*by*	0.007	*set*	0.006

Table 6: Ten most frequent words for each corpus.

Since we want to use language models derived from the LN corpus for speech recognition experiments conducted on the other three corpora, we have to explore the OOV (out-of-vocabulary) rate. We need to know how many words from UWB_B01, UWB_B02 and UWB_S01 corpora did not occur in the LN corpus and are therefore not in the language model and cannot be recognized, see Table 7.

	UWB_B01	UWB_B02	UWB_S01
OOV rate 660K	1.33%	1.36%	0.65%
OOV rate 100K	4.75%	4.98%	3.46%
OOV rate 60K	7.28%	7.36%	5.60%

Table 7: OOV rate of developed speech corpora related to the LN text corpus.

OOV rates are quite low for the whole vocabulary of 660K words. In speech recognition experiments we are limited to approximately 60K or 100K words, because the state-of-the-art speech decoders are incapable of handling larger vocabularies. The resulting OOV rates for such vocabularies imply considerable difficulties with the construction of the large vocabulary recognizer for spoken Czech. A slightly lower OOV rate for the UWB_S01 corpus could be explained by the selection of read sentences from similar text sources like those in the LN text corpus. It should be noted that in the case of 660K vocabulary the most frequent OOV words are proper names, especially the names of people who were not well-known in the period when the LN corpus was collected, but became famous later (*Putin, Elian Gonzalez, Ocalan*), or the names of the reporters of particular TV or radio stations. In the case of 60K vocabulary the OOV words were quite frequent words.

6. Conclusion

The Voice of America Broadcast News Czech (UWB_B01) is now distributed by the Linguistic Data Consortium, see web site http://www.ldc.upenn.edu/catalog/LDC2000T53.html.The corpus consists of 62 speech files, corresponding transcripts and two vocabularies. The first vocabulary comprises individual words present in transcripts with their phonetic transcriptions. The second vocabulary respects as well eventual changes caused by the cross-word assimilation process. This means that these words are phonetically transcribed and listed also in a “cross-word version”. The complete corpus fills 6 CDs. The issue of the UWB_B02 corpus is now under preparation also in the LDC.

All corpora presented in this paper are currently used for training the acoustic models, which are then incorporated into the speech recognition system. Experiments performed on that system will help us solve the problems encountered in speech recognition of Czech. Many of the experiments results are valid not only for Czech, but for other highly inflectional languages as well.

7. Acknowledgement

Support for this work was provided by the Ministry of Education of the Czech Republic (Grants No.VS97159, No.MSM235200004 and No.ME293); and by the NSF Language Engineering Workshop at the Johns Hopkins University, Baltimore (NSF Grant No.IIS-9820687).

8. References

[1] Jelinek, F.: Proposal for Speech Recognition of a Slavic Language: Czech. CLSP, Johns Hopkins University, Baltimore, November 1997.

[2] Nouza, J., Psutka, J., Uhlíř, J.: Phonetic Alphabet for Speech Recognition of Czech. *Radioengineering.* Vol.6, pp. 16-20, December 1997.

[3] Radová, V., Psutka, J., Šmídl, L., Vopálka, P., Jurčíček, F.: Czech Speech Corpus for Development of Speech Recognition Systems. In: Workshop on Developing Language Resources for Minority Languages, Athens, May 2000.

[4] Psutka, J.: Communication with Computer by Speech. (in Czech), Academia, Prague, 1995.