The Czech Broadcast Conversation Corpus

Jáchym Kolář, Jan Švec

Department of Cybernetics, Faculty of Applied Sciences,
University of West Bohemia, Univerzitní 8, CZ-306 14 Plzeň, Czech Republic
{jachym,honzas}@kky.zcu.cz

Abstract. This paper presents the final version of the Czech Broadcast Conversation Corpus released at the Linguistic Data Consortium (LDC). The corpus contains 72 recordings of a radio discussion program, which yield about 33 hours of transcribed conversational speech from 128 speakers. The release not only includes verbatim transcripts and speaker information, but also structural metadata (MDE) annotation that involves labeling of sentence-like unit boundaries, marking of non-content words like filled pauses and discourse markers, and annotation of speech disfluencies. The annotation is based on the LDC's MDE annotation standard for English, with changes applied to accommodate phenomena that are specific for Czech. In addition to its importance to speech recognition, speaker diarization, and structural metadata extraction research, the corpus is also useful for linguistic analysis of conversational Czech.

1 Introduction

Spoken language corpora are important for training and testing automatic speech recognition and understanding systems. For widespread languages such as English and Mandarin, many spoken language resources from various domains are publicly available, but for smaller languages, such as Czech, the speech resource availability to researchers is limited. Although the two major language resource publishers, the Linguistic Data Consortium (LDC) at the University of Pennsylvania and the European Language Resources Association (ELRA) offer Czech broadcast news [1,2] and prompted speech corpora [3,4] in their catalogs, no Czech conversational speech resources have been publicly available.

This has been a significant handicap for Czech researchers since the problem of automatically processing conversational speech is without a doubt one of the most important tasks in the field. Hence, in order to support broader research on the problem of conversational Czech, we have decided to create and publish a new speech corpus of broadcast conversations. The broadcast conversation genre was selected because of the easy data acquisition as well as for its increasing popularity in the speech processing community – for example, some current large research projects (such as GALE) include automatic translation of broadcast conversations [5].

The Czech Broadcast Conversation Corpus not only contains audio recordings and standard transcripts; the important additional value of this corpus lies in its "structural metadata" annotation. This annotation involves partitioning verbatim transcripts into sentence-like units (SUs) that function to express a complete idea; and identifying fillers and edit disfluencies. The structural information is critical to both increasing

human readability of the transcripts and allowing application of downstream NLP methods (e.g., machine translation, summarization, parsing), which are typically trained on fluent and formatted text.

The corpus will be released by the LDC as catalog numbers LDC2009S02 (audio) and LDC2009T20 (transcripts and structural metadata annotations) in summer 2009. The remainder of this paper describing the final version of the corpus is organized as follows. Section 2 describes the audio data, Section 3 presents details about speech transcription, Section 4 is devoted to structural metadata annotation, and Section 5 provides a summary and conclusions.

2 Audio Data

The broadcast conversation speech database contains recordings of a radio discussion program called *Radioforum*, which is broadcast by Czech Radio 1 (CRo1) every weekday evening. *Radioforum* is a live talk show where invited guests (most often politicians but also journalists, economists, teachers, soldiers, crime victims, and so on) spontaneously answer topical questions asked by one or two interviewers. The number of interviewees in a single program ranges from one to three. Most frequently, one interviewer and two interviewees appear in a single show. The material includes passages of interactive dialog, but longer stretches of monolog-like speech slightly prevail. Because of the scope of the talk show, all speakers are adults. Although the corpus was recorded from a public radio where standard (literary) Czech would be expected, many speakers, especially those not used to talking on the radio, use colloquial language as well. Literary and colloquial word forms are often mixed in a single sentence. The usage of colloquial language, however, is not as frequent as in unconstrained informal conversations.

The number of transcribed shows has been increased over the last year. The final release of the corpus contains 72 recordings acquired over the air during the period from February 12, 2003 through June 26, 2003. The signal is single channel, sampled at 22.05 kHz with 16-bit resolution. Typical duration of a single discussion is 33–35 minutes (shortened to 26–29 minutes after removing compact segments of telephonic questions asked by radio listeners, which were not transcribed). In total, the duration of the audio data is 40 hours, which yield approximately 33 hours of pure transcribed speech. The total number of speakers in the whole corpus is 128; male speakers are more frequent than females (108 males, 20 females).

3 Transcripts

The goal of the transcription phase was to produce precise time-aligned verbatim transcripts of the broadcast recordings. The data were manually transcribed in the Transcriber tool [6] using the careful transcription approach. The transcripts were created by a large number of annotators. To keep them maximally correct and consistent, all submitted transcripts were manually checked by a senior annotator.

The transcripts contain speaker turn labels – time stamps and speaker IDs were recorded at each speaker change. Overlapping speech regions were also labeled; within

Table 1. Corpus size (* – 'lexemes' include words, special interjections, and filled pauses; 'tokens' include lexemes plus speaker and background noises.)

Number of shows	72	Number of unique words	30.5k
Total duration	40.0h	Total number of speakers	128
Duration of transcribed speech	33.0h	— males	108
Total number of tokens	306.6k	— females	20
Total number of lexemes*	292.6k	Number of speaker turns	8.0k

these regions, each speaker's speech was transcribed separately (if intelligible). Regions of unintelligible speech were marked with a special symbol. To break up long turns, breakpoints roughly corresponding to "sentence" boundaries within a speaker's turn were added. The transcripts contain standard punctuation, but acceptable marks were limited to periods and question marks at the end of a sentence, and commas within a sentence. Capitalization was used for proper names, but not for the beginnings of sentences (unless they start with a proper name). Word fragments and mispronounced words were also tagged. In addition to words and punctuation, the transcripts contain special tags marking speaker noises (BREATH, COUGH, LAUGH, and LIP-SMACK), other noises (MUSIC, BACKGROUND-SPEECH, and unspecified NOISE), and "inarticulate" interjections expressing agreement (HM) and disagreement (MH).

Special attention was paid to transcription of filled pauses (FPs). FPs are hesitation sounds used by speakers to indicate uncertainty or to keep control of a conversation while thinking what to say next. In order to support maximal annotation consistency, we distinguished only two types of Czech FPs: *EE* (most typical example of EE is an FP similar to long Czech vowel *é*, but this group also includes all hesitation sounds that are phonetically closer to vowels), and *MM* (all FPs that are phonetically more similar to consonants or mumble-like sounds, typically pronounced with a closed mouth).

The overall size of the corpus in terms of a number of different measures is presented in Table 1. Among others, note that the number of distinct words in the vocabulary created from the corpus transcripts is quite large given the size of the corpus. Czech, same as other Slavic languages, is highly inflectional, and thus uses an extremely large number of distinct word forms.

4 Structural Metadata (MDE) Annotation

4.1 Annotation Approach

The structural metadata (MDE) annotation can be viewed as a post-processing step applied to the standard transcription. Structural information is critical to both increasing human readability of the transcripts and allowing application of downstream NLP methods, which typically require a fluent and formatted input. Because spontaneous utterances are not as well-structured as read speech and written text, annotating structure by simply making reference to standard punctuation is inadequate. Hence, several different schemes have been proposed for annotation of typical spontaneous speech phenomena. Earliest efforts include the manual for disfluency tagging of the Switchboard corpus [7], Heeman's annotation scheme for the Trains dialog corpus [8], and a

syntactic-prosodic labeling system for spontaneous speech called "M" presented in [9]. Recently, Fitzgerald and Jelinek [10] presented a new annotation scheme for spontaneous speech reconstruction.

For our work on conversational Czech, we have decided to adopt the "Simple Metadata Annotation" approach introduced by the LDC as part of the DARPA EARS program [11]. Originally, this standard was defined only for English. Later, the authors proposed to extend the guidelines for use with Mandarin and Arabic [12], but because of the premature termination of the EARS project, these efforts ended up as early as during the pilot annotation tests and no reasonably-sized corpora have been created to this day. When developing particular annotation rules for Czech structural metadata annotation, the LDC's annotation guidelines for English were taken as the starting point, with changes applied to accommodate specific phenomena of Czech. In addition to the necessary language-dependent modifications, we further proposed and applied some language-independent modifications. The annotation involves insertion of sentence-like unit breakpoints (SUs) to the flow of speech and identification of a range of spontaneous speech phenomena (fillers and disfluencies). The following subsections briefly describe individual annotation subtasks. A more detailed description is given in [13].

SUs Dividing the stream of words into sentence-like units is a crucial component of the MDE annotation. The goal of this part of annotation is to improve transcript readability and processability by presenting it in small coherent chunks rather than long unstructured turns. Because speakers often tend to use long continuous compound sentences in spontaneous speech, it is nearly impossible to identify the end-of-sentence boundaries with consistency using only a vague notion of a "conversational equivalent" of the "written sentence" definition; strict segmentation rules are necessary. One possible solution is to divide the flow of speech into some "minimal meaningful units" functioning to express one complete idea on the speaker's part. These utterance units are called SUs (Sentence-like/Syntactic/Semantic Units) within the MDE task. The SU symbols are the following:

/. – Statement break – end of a complete SU functioning as a declarative statement
(*Theresa loves irises /.*)
/? – Question break – end of an interrogative
(*Do you like irises /?*)
/, – Clausal break – identifies non-sentence clauses joined by subordination
(*If it happens again /, I'll go home /.*)
/& – Coordination break – identifies coordination of either two dependent clauses or two main clauses that cannot stand alone
(*Not only she is beautiful /& but also she is kind /.*)
/- – Incomplete (arbitrarily abandoned) SU
(*Because my father was born there /, I know a lot about the /- They must try it /.*)
/∼ – Incomplete SU interrupted by another speaker
(*A: Tell me about /∼ B: Just a moment /.*)

The SU symbols may be divided into two categories: sentence-internal (/& and /,) and sentence-external (others). Sentence-external breaks are used to indicate the presence of a main (independent) clause. These independent main clauses can stand alone

as a sentence and do not depend directly on the surrounding clauses for their meaning. Sentence-internal breaks are secondary and have mainly been introduced to support inter-annotator agreement. They delimit units (clauses) that cannot stand alone as a complete sentence.

We did not use the identical set of SU symbols as originally defined in [11] but introduced two significant modifications. First, the original set contains only one symbol for incomplete SUs, but we decided to distinguish two types of incomplete SUs: /– indicating that the speaker abandoned the SU arbitrarily; and /∼ indicating that the speaker was interrupted by another speaker. This distinction of incompletes is useful since their patterns differ significantly in prosody, semantics, and syntax. Second, in order to identify some "core boundaries", we introduced two new symbols: //. and //? — the double slashes indicate a strong prosodic marking on the SU boundary, i.e. pause, final lengthening, and/or strong pitch fall/rise.

Other modifications in the SU annotation pertain to differences between Czech and English. One example of a difference affecting the SU annotation is the possibility of subject dropping in Czech. In English, subject dropping is only allowed in the second clause of a compound sentence when both clauses share the same subject. In Czech, the subject (pronoun) can be dropped every time it is "understood" from the context and/or from the form of a conjugated predicate (verb). Since the conjugation of the verb includes both person and number of the subject, it is possible to say for instance *Běžím /.*, lit. *(I am) running /.* As a result, subject dropping in the coordinated clause does not imply the use of the coordinating break (/&), as is the case for English. Instead, we separate the coordinated clauses with an SU-external break, even if the subject is present in the first clause and dropped in the second (*Robert do práce šel pěšky /. ale domů jel vlakem /.*, lit. *Robert walked to work /. but (he) took the train home /.*).

Fillers Four types of fillers are labeled: filled pauses (FPs, already described in Section 3), discourse markers (DMs), explicit editing terms (EETs) and asides/parentheticals (A/Ps). Annotating fillers consists of identifying the filler word(s) and assigning them an appropriate label.

DMs are words or phrases that function primarily as structuring units of spoken language. They do not carry separate meaning, but signal such activities as a change of speaker, taking or holding control of the floor, giving up the floor or the beginning of a new topic. Frequent examples of Czech DMs are *tak* (lit. *so*) or *no* (*well*). Unlike English, DMs containing a verb (such as *you know*) are less frequent. We also labeled a DM subtype – Discourse Response (DR). DRs are DMs that are employed to express an active response to what another speaker said, in addition to mark the discourse structure. For instance, the speaker may also initiate his/her attempt to take the floor. DRs typically occur turn-initially.

EETs are fillers only occurring within the context of an edit disfluency. These are explicit expressions by which speakers signal that they are aware of the existence of a disfluency on their part. EETs are quite rare. In our corpus, by far the most frequent one is *nebo* (lit. *or*). The further filler type, A/P, occurs when a speaker utters a short side comment and then returns to the original sentence pattern (e.g., *And then that last question* {*it was a funny question*} *came up /.*). Strictly speaking, A/Ps are not fillers,

Table 2. Structural metadata statistics

Total number of SUs	21.7k	Tokens in DelRegs	2.9%
Average length of complete SUs	13.7	DelRegs having correction	88.2%
- statements	13.9	Tokens in A/Ps	1.5%
- questions	11.6	Tokens annotated as DMs	1.5%
Average length of incomplete SUs	10.2	Tokens annotated as EETs	0.1%

but because as with other filler types, annotators must identify the full span of text functioning as an A/P, they are included with fillers in the MDE definition. Some very common words or short phrases, that can be denoted as "lexicalized parentheticals" (e.g. *řekněme*, lit. *say*) are not annotated as A/Ps.

Edit disfluencies Edit disfluencies are portions of speech in which a speaker's utterance is not complete and fluent. Instead, the speaker corrects or alters the utterance, or abandons it entirely and starts over. In MDE, edit disfluencies consist of the deletable region (DelReg, speaker's initial attempt to formulate an utterance that later gets corrected), interruption point (IP, the point at which the speaker breaks off the DelReg with an EET, repetition, revision or restart), optional EET (an overt statement from the speaker recognizing the existence of a disfluency), and correction (portion of speech in which the speaker corrects or alters the DelReg). Whereas corrections had not been explicitly tagged within the MDE project for English, we decided to label them in order to obtain relevant data for further research of disfluencies. An example of a disfluency follows (* denotes IP, DelReg is displayed within square brackets, EET is typed in boldface, and correction is underlined):
Naše děti milují [kočku] EE* ***nebo*** *psa pana Bergera /.*
lit. *Our children love [the cat]* uh* ***or*** *the dog of Mr. Berger /.*

Table 2 displays some interesting numbers relating to the metadata annotation of the corpus. Among others, the numbers indicate that statement SUs are on average longer than interrogative SUs, and that, as expected, complete SUs are slightly longer than incomplete SUs. Furthermore, note that the number of DelRegs that were corrected is quite high, indicating that false starts (the disfluencies that do not have a correction) are not that frequent. The numbers also show that EETs are very rare, while DMs and A/Ps are more frequent.

4.2 Annotation Tool and Formats

When creating MDE annotations, annotators not only work with the verbatim transcripts, but also listen to audio and use prosody to resolve potential syntactic ambiguities. To ease the annotation process, a special software tool called QAn (Quick Annotator) was developed. It allows annotators to highlight relevant spans of text, play corresponding audio segments, and then record annotation decisions. A screenshot of the tool is displayed in Figure 1. The tool can be freely downloaded from the project website `http://www.mde.zcu.cz/`.

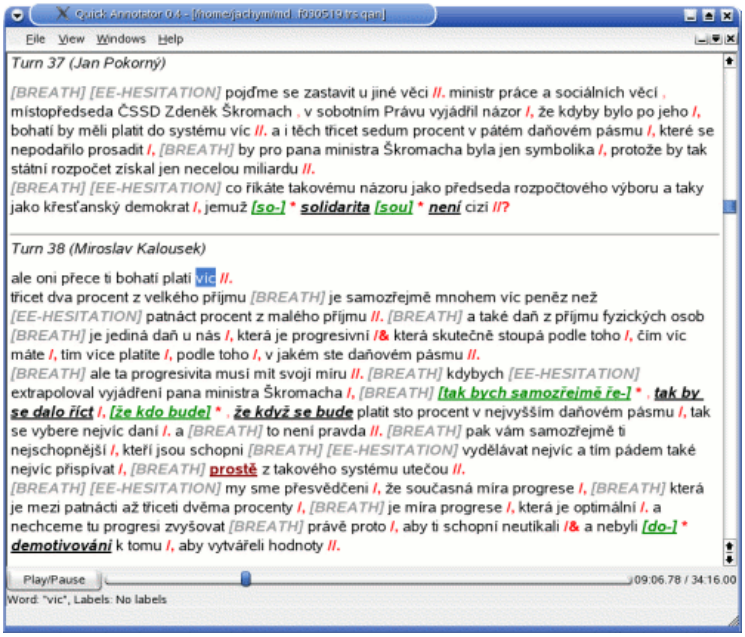

Fig. 1. QAn – the MDE annotation tool

In the corpus release, the structural annotations are provided in two different formats. The first is the format used by QAn. It is based on the XML-based Transcriber format, which is extended by special tags representing structural metadata information. The format uses two types of metadata tags: SUs associated with interword boundaries, and Labels spanning over one or more words (i.e., these are begin/end pairs).

The second format is RTTM. It is based on the RTTM-format-v13 specification that was used for storing MDE annotations in the EARS project. The format uses object-oriented representation of the rich text data. There are four general object categories to be represented: speech-to-text objects, MDE objects, source (speaker) objects, and structural objects. Except for the speaker information object, each object exhibits a temporal extent with a beginning time and duration. Note that the duration of interruption points and clausal boundaries is zero by definition. The objects are represented individually, one object per record, using a flat record format with object attributes stored in white-space separated fields.

5 Conclusion

In this paper, we have presented the final version of the Czech Broadcast Conversation Corpus that will be released by the LDC in summer 2009. The corpus was created in order to support broader research on the problem of conversational Czech. It contains 72 recordings of a broadcast discussion program, which yields about 33 hours of pure

transcribed speech from 128 adult speakers. The annotations not only include verbatim transcripts and speaker information, but the additional value of this corpus is that it also contains structural metadata annotations capturing important information about sentence-like units, fillers, and edit disfluencies. The metadata annotation is based on the LDC's standard for English, but the original guidelines had to be adjusted to accommodate specific phenomena of Czech syntax. In addition to the necessary language-dependent modifications, we further proposed and applied some language-independent modifications. We believe that besides its importance to speech recognition, speaker diarization, and MDE research, the corpus will also be useful for linguistic analysis of conversational Czech.

Acknowledgments

This work was supported by the Ministry of Education of the Czech Republic under projects 2C06020 and ME909. The authors also thank Stephanie Strassel, Christopher Walker, Dagmar Kozlíková, and Vlasta Radová for their help with the creation of the corpus.

References

1. Psutka, J., Radová, V., Müller, L., Matoušek, J., Ircing, P., Graff, D.: Voice of America (VOA) Czech broadcast news audio and transcripts. Linguistic Data Consortium Catalog No. LDC2000S89 and LDC2000T53, Philadelphia, PA, USA (2000)
2. Radová, V., Psutka, J., Müller, L., Byrne, W., Psutka, J.V., Ircing, P., Matoušek, J.: Czech Broadcast News Speech and Transcripts. Linguistic Data Consortium Catalog No. LDC2004S01 and LDC2004T01, Philadelphia, PA, USA (2004)
3. ELRA: Czech SpeechDat(E) database. Catalog Reference ELRA-S0094 (2001)
4. ELRA: GlobalPhone Czech. Catalog Reference ELRA-S0196 (2006)
5. Zheng, J., Wang, W., Ayan, N.F.: Development of SRI's translation systems for broadcast news and broadcast conversations. In: Proc. Interspeech 2008, Brisbane, Australia (2008)
6. Boudahmane, K., Manta, M., Antoine, F., Galliano, S., Barras, C.: Transcriber: A tool for segmenting, labeling and transcribing speech. `http://trans.sourceforge.net`
7. Meeter, M.: Dysfluency annotation stylebook for the Switchboard corpus. `ftp://ftp.cis.upenn.edu/pub/treebank/swbd/doc/DFL-book.ps` (1995)
8. Heeman, P.: Speech Repairs, Intonational Boundaries and Discourse Markers: Modeling Speakers' Utterances in Spoken Dialogs. PhD thesis, University of Rochester, NY, USA (1997)
9. Batliner, A., Kompe, R., Kiessling, A., Mast, M., Niemann, H., Nöth, E.: M = Syntax + Prosody: A syntactic–prosodic labelling scheme for large spontaneous speech databases. Speech Communication **25** (1998) 193–222
10. Fitzgerald, E., Jelinek, F.: Linguistic resources for reconstructing spontaneous speech text. In: Proc. LREC'08, Marrakech, Morocco (2008)
11. Strassel, S.: Simple metadata annotation specification V6.2. `http://www.ldc.upenn.edu/Projects/MDE/Guidelines/SimpleMDE_V6.2.pdf` (2004)
12. Strassel, S., Kolář, J., Song, Z., Barclay, L., Glenn, M.: Structural metadata annotation: Moving beyond English. In: Proc. Interspeech'05, Lisbon, Portugal (2005)
13. Kolář, J.: Automatic Segmentation of Speech into Sentence-like Units. PhD thesis, University of West Bohemia, Pilsen, Czech Republic (2008)