Czech Spontaneous Speech Corpus with Structural Metadata

Jáchym Kolář [1], *Jan Švec* [1], *Stephanie Strassel* [2],
Christopher Walker [2], *Dagmar Kozlíková* [1], *Josef Psutka* [1]

[1]Department of Cybernetics, University of West Bohemia in Pilsen, Czech Republic
[2]Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, USA
{jachym,honzas,psutka}@kky.zcu.cz {strassel,chwalker}@ldc.upenn.edu

Abstract

This paper describes a Czech spontaneous speech corpus consisting of radio talk show recordings. As the first complete non-English MDE corpus, it has been annotated with structural metadata information beyond the words that is critical to both increasing transcript readability and allowing application of downstream NLP methods. Metadata annotation involves partitioning verbatim transcripts into syntactic/semantic units (SUs) that function to express a complete idea; and identifying fillers and edit disfluencies. Annotation guidelines for English metadata developed by Linguistic Data Consortium 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 propose some language-independent modifications including limited prosodic labeling at SU boundaries. Statistics about the structural metadata annotation present in the corpus and inter-annotator agreement numbers are also presented.

1. Introduction

Nowadays, the problem of automatically processing spontaneous speech is without a doubt one of the most important tasks for the HLT community, because spontaneous speech, as opposed to read speech, is the most natural form of human communication. When creating spontaneous speech corpora, standard annotation techniques designed for read speech data are inadequate, because the resulting transcripts would lack important structural information.

The reasons for annotating structural events in speech are straightforward. Raw streams of words do not convey complete information, because the structural information beyond the words (metadata) is equally important as the words themselves. 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.

Since spontaneous utterances are not as well-structured as read speech and written text, there exist a number of reasons why annotating structure by simply making reference to standard punctuation is inadequate. First, there are no agreed-upon rules for punctuating faulty syntactic structures, which are quite frequent in spontaneous speech. Second, punctuation marks are ambiguous; commas may indicate several different structural/syntactic events (e.g., clausal break, apposition, parenthesis, etc.). Third, even for written text, the rules for applying punctuation are quite variable; for instance commas are optional in many cases. Fourth, standard punctuation does not convey all structural information contained in spontaneous speech. Due to the online nature of spontaneous speech, a speaker's utterance is often not complete and fluent. Because dealing with these phenomena is crucial to spontaneous speech understanding, more precise annotation of disfluencies and other structural phenomena is required.

To this end, the Linguistic Data Consortium (LDC) has defined an annotation standard [1] as part of the DARPA EARS Metadata Extraction (MDE) program [2, 3]. Originally, this standard was defined for English. In this paper, we present a collaborative effort of the Department of Cybernetics, University of West Bohemia (UWB) Pilsen and LDC to create MDE data for a Slavic language – Czech. Czech is a good test bed for Slavic MDE, because it is probably the most explored Slavic language for ASR research.

2. Speech data and transcription

The UWB research group has gained its Czech spontaneous speech processing experience within the MALACH project [4, 5]. However, the testimonies of Holocaust survivors that comprise this corpus cannot be freely distributed, so a new corpus was recorded to support broader research on the problem of spontaneous Czech.

The current spontaneous speech database consists of 52 recordings of radio discussion program called Radioforum, which is broadcast by Czech Radio 1 every weekday evening. Radioforum is a live talk show where invited guests (most often politicians but also journalists, economists, doctors, teachers, soldiers, crime victims, and so on) spontaneously answer topical questions asked by 1 or 2 interviewers. The number of interviewees in a single program ranges from 1 to 3. Most frequently, 1 interviewer and 2 interviewees appear in the program. The material includes passages of interactive dialog, but longer stretches of monolog-like speech slightly prevail.

The recordings were acquired during the period from February 12, 2003 through June 6, 2003. The signal is single channel sampled at 44 kHz with 16-bit resolution. Typical duration for a single discussion is 33-35 minutes (shortened to 26-29 minutes after removing compact segments of telephonic questions asked by listeners, which were not transcribed). Verbatim transcripts with careful annotation of non-speech events were created in a standard way using Transcriber 1.4.2.

Czech belongs to the family of Slavic languages, which are highly inflectional and derivational and are characterized by a relatively free word order. Although the corpus was recorded from 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.

Colloquial Czech deviates from standard Czech (as defined by orthographic, morphological, lexical and syntactic rules by the Czech normative bodies). With respect to pronunciation variation, Czech is different from English and many other languages in that spelling rules for Czech are phonetically based. Therefore, colloquial Czech words have well-defined but different spellings than their standard variants. In other words, colloquial Czech has an orthographic written form. The difference between colloquial and standard Czech is most prominently displayed in the morphology - endings and prefixes are often changed.

Because common text corpora used for building language models usually do not contain colloquial word forms, it is advantageous to create a lexicon, in which colloquial word forms are mapped onto corresponding standard forms. Likewise, to enhance model robustness, word forms representing more orthographic variations of one word form are grouped together (e.g., *"socialismus"* and *"socializmus"*). Interested readers may consult [5] for more information. As is typical, we also marked personal names, place names, numbers and foreign words in the lexicon. Some basic statistics about the corpus are given in Table 1.

Table 1: *Czech Spontaneous Speech Corpus statistics*

#all tokens	225.3k	#speakers	94
#lexeme tokens	201.2k	- males	77
#fragment tokens	2.1k	- females	17
#unique words	25.3k	transcribed speech	24.0h
size of adj. lexicon	23.2k	non-overl. speech	23.1h

3. Structural metadata annotation

Metadata annotation can be viewed as a post-processing step applied to the standard transcription. It involves identification of a range of spontaneous speech phenomena (fillers and disfluencies) and insertion of syntactic/semantic breakpoints (SUs) to the flow of speech. Annotators not only work with the verbatim transcripts, but also listen to the audio and use prosody to resolve potential syntactic ambiguities.

When developing metadata annotation standard for Czech we followed the LDC guidelines for English [1]; however, it is not possible to simply translate and copy all conventions for English onto Czech. The rules must be adjusted to accommodate specific phenomena of the target language. Past experience indicates that acceptable inter-annotator agreement (IAA) can only be achieved in the context of rules grounded in "surface features" (based mainly on syntax and prosody); semantic features have not been reliable. In addition to the modifications motivated by specifics of Czech, we also propose some language-independent modifications.

To ease the annotation process we have developed new annotation software that is customized to reflect the particulars of the Czech annotation task. As with LDC's MDE Annotation Toolkit [6], the Czech tool allows annotators to highlight relevant spans of text, play corresponding audio segments and then record annotation decisions [7]. Since Czech syntax is quite complex, naïve annotators could not be employed; at least some linguistic education is necessary.

3.1. SUs

Dividing the continuous stream of words into sentence-like units is a crucial component of MDE annotation. Because speakers often tend to use long continuous complex sentences in spontaneous speech, it is nearly impossible to identify the end-of-sentence boundary with consistency using only prosodic information. A 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 (Sentential/Syntactic/Semantic Units) within the MDE task. Every word within the discourse is assigned to an SU (each word contained between two SU boundaries is considered part of the same SU), and all SUs must be classified according to their function within the discourse. Employed SU symbols (breaks) are the following:

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

Because our corpus is only single-channel, we do not use an SU symbol for backchannels. The SU symbols may be divided into two categories: sentence-internal (/& and /,) and sentence-external (others). Sentence-external breaks are fundamental and directly support the SU research task. They 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-level breaks may also appear after a short phrase that nonetheless functions as a "complete" sentence. Sentence-internal breaks are secondary and have mainly been introduced to support IAA.

3.1.1. Language-independent modifications

Compared to the original set of SU symbols, we made two significant changes that are not language-dependent. First, while the original set contains only one symbol for incomplete SUs, we decided to distinguish two types of incomplete SUs: /- indicates that the speaker abandoned the SU arbitrary, while /∼ indicates that the speaker was interrupted by another speaker. This distinction of incomplete turns is very useful, since their patterns differ significantly in prosody, semantics and syntax.

Second, in order to identify some "core boundaries" that could be both easier to detect automatically using prosodic cues, and also relevant for spontaneous discourse analysis, we added 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.

The additional annotation refinements do not seem to cause a corresponding growth in annotation complexity. A rule of thumb instructs annotators to use the double-slash SU symbols when in doubt. Measuring IAA on /. vs. //. on 3 dually-annotated recordings of total duration of 86 minutes, the following consistency results were achieved. If the two annotators

agreed on applying a statement break after a particular word, they applied the identical symbol (/. or //.) in 86.9% (617 out of 710) of the cases. Only taking into account the words followed by a statement break (i.e. ignoring all words with no statement label), the value of normalized (kappa) statistic [8] was $K = 0.69$. For comparison, for /. vs. //. vs. "other", we obtained $K = 0.85$. Higher IAA can be expected for broadcast news data.

Another proposed possibility, which has not yet been tested, is to use a 3-symbol system: /. for SU boundaries with no prosodic marking of finality; //. for SU boundaries with well-marked pitch fall and/or final lengthening, but with no noticeable pause; and ///. for SU boundaries with a noticeable pause. Note that, in contrast to ToBI-like systems, our system only involves labeling prosodic boundaries on SU boundaries, rather than on all word boundaries, which is much less time-consuming.

The next modification pertains to the pause threshold. In the English SimpleMDE V6.2 standard, in order to support IAA, the pause longer than 0.5 sec automatically induces the end of a speaker turn and thereby requires a corresponding SU-external break. But the 0.5 sec threshold is problematic, because some speakers produce long pauses in places where other speakers might produce filled pauses. Hence, we decided to drop the threshold rule and to rely solely on syntax. Likewise, we do not require the presence of a noticeable pause after incomplete (abandoned) SU breaks (/-) when the syntax provides an overt evidence of incompleteness.

3.1.2. Language-dependent modifications

Other modifications in SU annotation for Czech are motivated by differences between Czech and English. Comparing Czech sentence structure to English, the most distinctive difference (beyond the relatively free word order) is the possibility of subject omission. In English, subject dropping is only allowed in the second clause of a compound sentence when both clauses share the same subject; whereas in Czech, the subject (pronoun) can be dropped every time it is "understood" from context and/or from the form of a conjugated verb (predicate).

Thus, 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 /."* This phenomenon of subject dropping is typical for highly inflective languages.

Another important fact is that Czech syntax discriminates between compound sentences sharing a single common subject, and simple sentences with compound predicates (i.e. compound predication in a simple sentence). Compound predicates are defined as a "tight unit" of two or more predicate verbs predicating on the same subject. On the other hand, if the predicate verbs do not form such a "tight unit", a compound sentence is recognized. Unfortunately, there is not absolute agreement in the literature on the exact borderline between compound predicates and compound sentences. For our purposes, we have only considered the features of compound predicates that are clear. The compound predicate is recognized if: 1) The predicate verbs share a common constituent (e.g., object) – *"Nacpal /& a zapálil si dýmku /."*, lit. *"He filled /& and lit up his pipe /."*; or 2) The predicate verbs joined by a copulative conjunction have the same or very similar meaning *"Naši hosté často slaví /& a radují se /."*, lit. *"Our guests often rejoice /& and celebrate /."* While compound predicates did not motivate any SU breaks according to the initial version of our annotation guidelines, the current version instructs annotators to separate parts of compound predicates by a coordination SU break, because it leads to higher annotation consistency.

For the above stated reasons, subject dropping in the coordinated clause does not imply the use of the coordinating break (/&) alone, 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 clause (*"Robert do práce šel pěšky /. ale domů jel vlakem /."*, lit. *"Robert walked to work /. but (he) took the train home /."*). If both predicates share an auxiliary verb (i.e. it is dropped in the second clause), the clauses cannot stand alone and a coordination break is used (*"Zítra budu odpočívat /& a číst tvé básně /."*, lit. *"Tomorrow I will rest /& and read your poems /."*).

The /& symbol is also used when main clauses are joined by the syntactically primarily coordinating yet semantically often rather subordinating conjunction *"nebot'"*, lit. *"for"* – *"Šli jsme se koupat /& nebot' bylo krásné počasí /."*, lit. *"We went swimming /& for the weather was great /."* If the second clause of a compound sentence has an elliptical (dropped) predicate (*"Katka miluje kosatce /& ale Eva tulipány /."*, lit. *"Katka loves irises /& but Eva tulips /."*), the clauses are separated by /& as well. Other English rules for judging between /& and /. from [1] were retained (/& is used if there exists a subordinate clause depending on both main clauses or a compound sentence is structured with a non-continuous expression such as *"Not only ..., but also ..."*).

Moreover, we adjusted some rules dealing with SU-internal breaks to reflect Czech syntax. For instance, unlike in English, relative clauses are separated by clausal breaks (*"Jan /, který se narodil v Praze /, miluje Karlův most /."*, lit. *"Jan /, who was born in Prague /, loves the Charles bridge /."*).

3.2. Fillers

Four types of fillers are considered: filled pauses (FP), markers (DM), explicit editing terms (EET) and asides/parentheticals (A/P). Annotating fillers consists of identifying the filler word(s) and assigning them an appropriate label.

FPs are hesitation sounds used by speakers to indicate uncertainty or to control of a conversation while thinking what to say next. FPs usually vary a bit across languages. For Czech, in order to keep maximal annotation consistency (annotators are not consistent in distinguishing particular hesitation sounds), we distinguished just 2 types of FPs: *EE* (similar to English uh, er, eh) and *MM* (sequence of consonant-like sounds, most often *"mm"* or *"ww"*). *EEs* are much more frequent than *MMs*.

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. In our corpus, the most frequent DMs are *"tak"*, lit. *"so"* and *"no"*, lit. *"well"*. Compared to English, DMs containing a verb are less frequent. The most frequent from this group of DMs *"víte"*, lit. *"you know"*, is 22 times less frequent than the most frequent DM *"tak"*.

EETs are fillers occurring within the context of an edit disfluency. EETs are very rare. In our corpus, by far the most frequent one is *"nebo"*, lit. *"or"*.

A/Ps occur when the 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, 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.

Table 2: *Comparison of metadata annotation statistics for spontaneous Czech (CZ) and spontaneous English (EN) corpora*

	CZ	EN
Average length of a complete SU	12.3	10.9
- statements	12.4	11.1
- questions	11.3	8.1
Average length of an incomplete SU	9.4	3.9
Average distance between SU symbols	6.6	6.0
% of tokens within DelRegs	2.6%	7.0%
% of DelRegs being corrected	83.5%	N/A
% of tokens within A/Ps	1.5%	0.3%
% of tokens annotated as a DM	1.4%	4.5%
% of tokens annotated as an EET	0.1%	0.1%

Some very common words or short phrases, that can be denoted as "lexicalized parentheticals" (e.g. *"řekněme"* lit. *"say"*,*"myslím"* lit. *"I think"*) are not annotated as A/Ps. They usually lack the prosodic features that typically accompany A/Ps. In order to ensure a high IAA, a preliminary illustrative list of those "lexicalized parentheticals" was prepared. Their maximal length was restricted to 2 words.

As opposed to the English MDE corpora, A/Ps are relatively frequent in our corpus. Even with exclusion of "lexicalized parentheticals", 1.5% of all tokens were annotated as within an A/P.

3.3. 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. Edit disfluency consists 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 explicit editing terms (an overt statement from the speaker recognizing the existence of disfluency) and the correction (portion of speech in which speaker corrects or alters the DelReg). Whereas corrections are not explicitly tagged within the MDE project for English, we decided to label them in order to obtain relevant data for the further research of spontaneous Czech. Czech disfluencies have the same pattern as English. 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 Millera /.*
lit. *Our children love [the cat]* uh* ***or*** *the dog of Mr. Miller /.*

3.4. Metadata annotation statistics

Statistics relating to metadata annotation of our corpus and English Conversational Telephone Speech Corpus (part of Switchboard) are given in Table 2. All numbers listed in this table denote numbers of tokens or percentages of the total number of tokens in the corpus. The significant differences in the numbers are mainly caused by different nature of either corpus; the Czech one is a bit more formal and less interactive.

Relative frequencies of particular types of SU symbols in the Czech corpus are the following: /, (42.1% of all SU symbols), //. (28.8%), /. (15.0%), /& (6.7%), //? (3.3%), /∼ (3.0%), /? (0.7%), /- (0.4%). Measuring overall IAA on the same test data as described in Section 3.1.1. (13,026 tokens), we got $K = 0.88$ for SUs (all types) and $K = 0.85$ for other labels (i.e. for fillers, DelRegs and corrections). With respect to the complexity of our annotation task, the IAA numbers seem to be acceptable; $K > 0.7$ is claimed to be satisfactory value for tasks such as ToBI labeling or content analysis. For the key task annotation – "SU-boundary" vs. "no SU-boundary" – we got $K = 0.92$.

4. Conclusion and future work

In this paper, we have presented a Czech spontaneous speech corpus annotated with structural metadata. The metadata annotation is based on the LDC's "Simple Metadata Annotation Specification" for English. The original guidelines have been adjusted to accommodate specific phenomena of Czech syntax. In addition to the necessary language-dependent modifications, we propose some language-independent modifications including limited prosodic labeling at SU boundaries. Besides its importance to MDE research, this corpus is also useful for linguistic analysis of spontaneous Czech.

In the near future, we plan to metadata annotate a broadcast news corpus [9] and a corpus of ice-hockey transmission commentaries. We are also developing an automatic metadata extraction system for Czech. We believe that conclusions about MDE for Czech will be largely applicable to other Slavic languages and, more generally, to all highly inflective languages.

5. Acknowledgements

Support for this work was provided by the Ministry of Education of Czech Republic, projects No. LC536 and MSM235200004. Work by Linguistic Data Consortium to define the English Simple Metadata Annotation Specification was supported by the DARPA EARS (Efficient, Affordable, Reusable Speech-to-Text) Program.

6. References

[1] Strassel, S., "Simple Metadata Annotation Specification Version 6.2," http://www.ldc.upenn.edu/Projects/MDE/, 2004
[2] http://www.darpa.mil/ipto/programs/ears/
[3] Strassel, S., Miller, D., Walker, K., Cieri, Ch., "Shared Resources for Robust Speech-to-Text Technology," *EUROSPEECH 2003, Geneva, Switzerland*, 2003
[4] Psutka, J., Hajic, J., Byrne, W. "The Development of ASR for Slavic Languages in the MALACH Project," *IEEE ICASSP 2004, Montreal, Canada*, 2004
[5] Psutka, J., Ircing, P., Hajic, J., Radova, V., Psutka, J. V., Byrne, W., Gustman, S., "Issues in Annotation of the Czech Spontaneous Speech Corpus in the MALACH Project," *LREC 2004, Lisbon*, 2004
[6] Maeda, K., Strassel, S. "Annotation Tools for Large-Scale Corpus Development: Using AGTK at the Linguistic Data Consortium," *LREC 2004, Lisbon*, 2004
[7] http://www.mde.zcu.cz/
[8] Carletta, J., "Assessing agreement on annotation tasks: the kappa statistic," *Computational Linguistics, 22(2): 249-254*, 1996
[9] Kolar, J., Romportl, J., Psutka, J., "The Czech Speech and Prosody Database both for ASR and TTS Purposes," *EUROSPEECH 2003, Geneva, Switzerland*, 2003