Structural Metadata Annotation for Czech: An Overview

Jáchym Kolář

April 8, 2009

Contents

1 Introduction

This text presents structural metadata (MDE) annotation for spoken Czech. It mainly serves as a documentation for the release of Czech MDE corpora at the Linguistic Data Consortium (LDC). If you are not familiar with structural MDE annotation of speech as introduced within the EARS project, we strongly recommend to read the original English MDE guidelines [1] first. The Czech MDE annotation guidelines are based on the original English standard, but the annotation rules were modified in order to accommodate particularities of our task. The modifications are both language-dependent and language-independent.

Most of the text in this overview was taken from the thesis [2]. The text is organized as follows. Section 2 describes the used audio data and briefly outlines how the data were transcribed. Section 3 overviews approaches to annotating structure of spontaneous utterances and states the reasons why we chose to adopt the structural metadata annotation approach. Section 4 presents the structural metadata annotation guidelines for Czech. Section 5 provides a summary.

2 Speech Data

2.1 Czech Broadcast News Corpus

The broadcast news (BN) speech data we used for this work were taken from the Czech Broadcast News Corpus. This corpus was recorded at UWB and is publicly available from the Linguistic Data Consortium (LDC) [3]. The corpus is 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 60 hours of audio stored on 342 waveform files, which yield more than 26 hours of pure transcribed speech.[1] The recordings do not contain weather forecasts, sport news, and traffic announcements. The signal is single channel. It was originally sampled at 44.10 kHz with 16-bit resolution, but for the official release, the waveforms were downsampled to 22.05 kHz. Basic numbers about the BN corpus are listed in Table 1. Details about orthographic transcription of this corpus were given in [4].

2.2 Czech Broadcast Conversation Corpus (Radioforum)

2.3 Audio Data

The newly recorded broadcast conversation (BC) speech database consists of 72 recordings of a 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 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 one show. The material includes passages of interactive dialog, but longer stretches of monolog-like speech slightly prevail.

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.

[1] Because of copyright issues, only 284 of the 342 recorded shows yielding 22.8 hours of transcribed speech could have been published at the LDC.

Table 1: Basic numbers about Czech Broadcast News corpus (BN) and Czech Broadcast Conversation Corpus (BC)

	BN	**BC**
Number of shows	342	72
Number of word tokens	234.2k	292.6k
Number of unique words	31.9k	30.5k
Duration of transcr. speech	26.7h	33.0h
Total number of speakers	284	128
— male speakers	188	108
— female speakers	96	20

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 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). Some basic numbers about the corpus are presented in the third column of Table 1.

2.4 Speech Transcription

The recorded shows were manually transcribed based on detailed annotation guidelines. The goal of the transcription was to produce precise time-aligned verbatim transcripts of the audio recordings. The transcription guidelines for this Czech corpus were based on the guidelines published in [4, 5]. However, the original guidelines were adjusted to better accommodate specifics of the recorded spontaneous speech corpus. For example, in order to increase inter-labeler consistency, the number of tags for labeling speaker and background noises was significantly reduced. For the same reason, we also changed the rules for transcription of filled pauses because the original rules had been too vague. The filled pause issue is discussed in more detail below in Section 4.1.1.

Among others, the transcription guidelines instructed annotators how to deal with the following phenomena:

- *Speaker turns* – a corresponding time stamp and speaker ID are inserted every time there is a speaker change in the audio.
- *Turn-internal breakpoints* – to break up long turns, breakpoints roughly corresponding to "sentence" boundaries within a speaker turn are inserted.
- *Overlapping speech* – an overlapping speech region is recognized when more than one speaker talks simultaneously; within this region, each speaker's speech is transcribed separately (if intelligible).
- *Background noises* – *[NOISE]* tags are used to mark noticeable *background* noises.
- *Speaker noises* – speaker-produced noises are identified with one of the following tags: *[BREATH]*, *[COUGH]*, *[LAUGH]*, *[LIP-SMACK]*.
- *Filled pauses* – filled pauses produced by a speaker to indicate hesitation or to maintain control of a conversation are transcribed either as *[EE-HESITATION]* or as *[MM-HESITATION]*, based on their pronunciation.

- *Interjections* – certain interjections typically used as backchannels or to express speaker's agreement or disagreement are transcribed using the *[HM]* (agreement) and *[MH]* (disagreement) tags.
- *Unintelligible speech* – regions of unintelligible speech are marked with a special symbol.
- *Numbers* – all numerals are transcribed as complete words.
- *Mispronounced words* – mispronounced words (reading errors, slips of the tongue) are transcribed in the spelling corresponding to their pronunciation in the audio (i.e., the incorrect pronunciation is represented[2]) and marked with a special symbol.
- *Word fragments* – the pronounced part of the word is transcribed and a single dash is used to indicate point at which word was broken off.
- *Punctuation* – standard punctuation (limited to commas, periods, and question marks) is used to enhance transcript readability.

The verbatim transcripts of this corpus were created by a large number of annotators. To keep them maximally correct and consistent, all submitted annotations were manually revised.

3 Annotation of Spontaneous Speech Structure

In the previous section, the creation of time-aligned *verbatim* transcripts was described. This annotation is usually sufficient for training and testing standard ASR systems. However, raw streams of words do not convey complete information because the structural information beyond the words (metadata) is equally important as the words themselves. As mentioned in the Introduction, 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.

This thesis is focused on automatic generation of *rich* speech transcripts[3] which not only contain words but also boundaries of sentence-like units. Thus, the key problem is how to annotate sentence boundaries in speech. In principle, there are two basic options. The first option is to use standard punctuation, whereas the other is to employ a special annotation scheme tailored for spoken language. Although the former approach is quite convenient for read speech, its convenience for spontaneous speech is at least questionable.

Because 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 for many applications. 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. Spontaneous utterances are often incomplete or disfluent. Since dealing with the specific spontaneous speech phenomena is crucial to spontaneous speech understanding, more precise annotation of disfluencies and other structural phenomena is required. On the other hand, we must take into account that special annotation of spontaneous speech phenomena is extremely labor-intensive and thus expensive.

[2] Unlike English, this is possible in Czech since spelling rules are phonetically based.

[3] Besides sentence boundaries, such rich transcripts may also include speaker diarization (who speaks when), disfluency annotation, or other structural information.

3.1 Related Work on Spontaneous Speech Annotation

Several different annotation schemes has been presented for similar annotation tasks. Earliest efforts include the Meteer manual for disfluency tagging of the Switchboard corpus [6]. A detailed annotation scheme for the Trains dialog corpus was proposed by Heeman [7]. His annotation scheme included labeling of intonational boundaries within the ToBI framework, identification of discourse markers, and a very detailed annotation of speech repairs.

Within the VERBMOBIL project, Batliner et al. [8] presented a syntactic-prosodic labeling system for large spontaneous speech databases called "M". This annotation was only based on word transcripts; the annotators did not have access to audio recordings. Using a rough syntactic analysis, each word in a turn was assigned to one of 25 "M" classes. Depending on the target task, these 25 classes were grouped either into 3 main M classes (M3 - clause boundary, M0 - clause internal, MU - boundaries that cannot be determined without listening to audio or knowing particular pragmatic context), or into 5 syntactic classes (S0 - no boundary, S1 - boundary after a particle, S2 - phrase boundary, S3 - clause boundary, S4 - sentence boundary). An apparent drawback of this approach is that the annotation scheme is very complex and requires experienced linguist annotators.

At the time we started our work on this task, there was no similar annotation system for Czech. In addition, there was almost no published work focusing on syntax of conversational Czech. The exception was a treatise by Müllerová [9]. This monograph described syntactic phenomena specific for conversational Czech, surveyed various types of speech repairs, and briefly discussed Czech discourse markers. Although Müllerová's work is interesting, it does not offer any clear clues to explicit annotation of sentence-like units in spoken Czech. The definitions provided in this work are too vague to be applicable to NLP tasks; the author only studied spoken Czech in terms of a qualitative linguistic description.

3.2 Structural Metadata Annotation Approach

For our work, we have decided to adopt the "Simple Metadata Annotation" approach [1], which was introduced by LDC as part of the DARPA EARS (Efficient, Affordable, Reusable Speech-to-Text) program [10]. This annotation was defined for the EARS Metadata Extraction (MDE) subtask [11]. The goal of MDE is to create automatic transcripts that are maximally readable. This readability may be achieved in a number of ways: creating boundaries between natural breakpoints in the flow of speech; flagging non-content words like filled pauses and discourse markers for optional removal; and identifying sections of disfluent speech.

The word "simple" in the name of the approach only emphasizes a contrast with an early MDE definition known as "Full MDE". Since a pilot annotation study had revealed a number of problems with consistency of the "full" annotation, LDC developed the "simple" definition that eliminated some annotation tasks entirely and simplified others. As a result, the current MDE annotation can be performed by non-linguist annotators with reasonable consistency.

4 Structural Metadata Annotation for Czech

Originally, the structural MDE annotation standard was defined for English. When developing structural metadata annotation guidelines for Czech, we tried to follow the LDC guidelines for English as much as possible. However, it would not be correct to simply translate and copy all conventions from one language to another. Individual rules must be adjusted to accommodate specific phenomena of the target language. The language-dependent modifications are mainly

based on the description of syntax of Czech compound and complex sentences as given by [12]. We also used two other Czech syntax handbooks [13, 14].

In the following text, all illustrative examples are presented in Czech and then in their English translations. Note that, because of significant differences between Czech and English, it is often impossible to present a good verbatim translation. We tried to use English translations that best illustrate the linguistic phenomena of interest. Also note that SU symbols in all English translations are not displayed as annotated based on the English MDE standard, but rather illustrate their placement with respect to the Czech guidelines. Furthermore, the examples do not contain standard punctuation but only SU symbols. All examples are typed in a typewriter font. If an example represents a conversation, the speakers are distinguished using capital letter IDs (A:, B:).

In all examples in this section, we use a notation that is very similar to the notation used in the original English guidelines [1]. The notation is the following:[4]

Fillers:

word	–	"word" is a filler (discourse marker or explicit editing term)
{ word }	–	"word" is an Aside/Parenthetical

Edit Disfluencies:

[word] *	–	"word" is a Deletable Region; * denotes an interruption point (IP)
word	–	"word" is the corrected portion of a disfluency

SUs:

/.	–	statement SU break
/?	–	question SU break
/-	–	incomplete SU break – arbitrarily abandoned
/∼	–	incomplete SU break – interrupted
/&	–	coordination break
/,	–	clause break
⊘	–	no break at a place where one might be expected

The remainder of this section is organized as follows. Section 4.1 describes annotation of fillers, Section 4.2 presents annotation of edit disfluencies, and Section 4.3 is devoted to annotation of syntactic-semantic units (SUs).

4.1 Fillers

Fillers are words, short phrases, or non-verbal hesitation sounds that do not alter the propositional content of the utterance in which they are inserted. Their characteristic feature is that they do not depend on identities of surrounding words. In general, fillers are those parts of the utterance which could be removed from its transcript without losing any "important" piece of information. Four types of fillers are distinguished within the MDE system:

[4] In this introductory section, we just list the employed MDE symbols. Their meaning is explained below in corresponding sections.

- Filled Pauses (FP),
- Discourse Markers (DM),
- Explicit Editing Terms (EET),
- Asides/Parentheticals (A/P).

Annotating fillers consists of identifying the filler words and assigning them an appropriate label. Note that in contrast with English MDE, we do not insert automatic IPs before fillers in Czech MDE. However, their automatic addition is very easy to implement for users who would wish to have all fillers marked with preceding IPs.

4.1.1 Filled Pauses

FP is a non-verbal hesitation sound produced by speakers (either intentionally or not) to indicate uncertainty or to keep control of a conversation while thinking what to say next. In general, FPs can appear anywhere in the flow of speech. By their definition, they make no contribution to the semantic proposition of the utterance. Thus, FPs should not be confused with certain interjections that function to express agreement or disagreement, or as backchannels (such as English *uh-huh*).

An important (and also very interesting) fact about FPs is that they vary across languages. For instance, FPs in American English are known as *uh* and *um*, while Japanese speakers use *ahh*, *ano*, or *eto*, and French talkers most frequently vocalize a sound similar to *euh* [15].

No thorough linguistic study on FPs has been conducted for Czech. Consequently, there is no general agreement on how to transcribe them in text – their transcription differs corpus to corpus. For instance, Kaderka and Svobodová [16] propose to distinguish ten non-verbal sounds, six of which correspond to FPs[5] (*e*, *ee*, *eee*, *eh*, *ehm*, *em*). The first three FPs from this list differ only in length, the other three differ in phonetic makeup. My opinion is that to distinguish six different FPs is too many.

When developing transcription rules for FPs, one must be aware of the fact that there is always a trade off relation between transcription accuracy and consistency. If we choose too many categories, annotators will not be consistent in their recognition. On the other hand, too broad categories might cluster FPs that are completely different, both phonetically and functionally.

In order to be able to design annotation guidelines for Czech FPs, we spent a lot of time listening to Czech spontaneous speech recordings. Based on this experience, we decided to distinguish the following two FP categories:

- *EE* – this FP category is most typically represented by sounds similar to Czech *é*, but in my annotation guidelines, it also includes all hesitation sounds that are phonetically closer to vowels than consonants – for example, sounds similar to Czech long vowel *á* also may function as FPs. Also note that EEs are sometimes accompanied with a creaky voice quality.
- *MM* – this FP category contains all hesitation sounds that are phonetically more similar to consonants or mumble-like sounds. The most frequent hesitation sound from this group is similar to *mmm*. Another not infrequent example of an *MM* is an FP resembling a lengthened Czech consonant *v*. *MMs* typically pronounced with a closed (or almost closed) mouth – openness of mouth is also a good feature distinguishing *MMs* from *EEs*.

[5] They do not discriminate between interjections altering content and hesitations in their guidelines.

Overall, *EEs* are significantly more frequent than *MMs*. Experience with annotation of the two Czech corpora presented herein indicates that these two categories very well cover a vast majority of all FPs occurring in spontaneous Czech. Moreover, our annotators felt comfortable with using these two FP labels. Besides the positive experience, the number of recognized FP categories is also in line with the number of FP categories in American English.

The only problematic instances of FPs in terms of this transcription approach are those similar to *emm*. In such FPs, vowel-like and consonant-like components immediately follow each other. Since such FPs are really rare in spontaneous Czech, we decided not to introduce a special tag for them. However, we had to prepare instructions specifying their transcription. Annotators were instructed as follows. Only the *MM* symbol is used when the vowel-like component is much shorter than a dominant consonant-like component. By analogy, only the *EE* tag is used when the vowel-like component is strongly dominant. When both components are strong, the FP is transcribed using both symbols as *EE MM*. This notation is also used when instances of *EE* and *MM* appear separated by a pause. An example of an FP-annotated utterance follows.

To je *EE* jenom *EE MM* jeho sen /.

This is *EE* just *EE MM* his dream /.

Since we did not allow annotators to transcribe FPs using other words or symbols than *EE* and *MM*, the MDE annotation of FPs was in principle performed during the verbatim transcription stage. However, the annotators in the MDE annotation stage had the right to insert or change FP symbols. We find this two-pass annotation setup useful because FPs are quite often missed by human transcribers.

Note that the presented novel approach to annotating FPs was only used for the RF corpus. The verbatim transcripts of the BN corpus had been created earlier using an FP annotation based on different guidelines. These guidelines paid only little attention to description of FP types. The vast majority of FPs in this corpus were transcribed as *ERs*.[6] Annotators could also use another English FP transcription, *UM*, but this symbol occurs only several times in the transcripts – apparently because FPs of this type are very rare in Czech.

4.1.2 Discourse Markers

DMs are words or phrases, such as the well-known "*you know*", that function primarily as structuring elements 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 beginning of a new topic. There exists a number of diverse definitions of DMs in the linguistic literature. Within MDE, we are only interested in such DMs whose presence in the utterance is unnecessary and whose cleanup do not lead to loss of "important" structuring information. Thus, structuring units such as *"Za prvé, ..." ("First, ...")* do not receive DM labels. If multiple DMs occur in succession, each DM is tagged separately, rather than labeling one long DM spanning over all successive DM instances. An example of DM annotation follows.

Tak já *jako* nevím /.

So I *like* don't know /.

For any language, it is not possible to create a closed list of possible DMs. The of use DMs is dependent on a dialectal variation and rhetorical style of a particular talker. The list of popular DMs in Czech includes: *dobře (well), jako (like), jaksi (sort of), no (well), podívejte se (you*

[6] *Er* is a British variant of American *uh*.

see), prostě (simply), tak (so), takže (thus), tedy (then), víte (you know), víte co (you know what), vlastně (actually), v podstatě (basically), among others.

Some of the frequent DM words and phrases also have other literal meanings, which sometimes makes identification of DMs more difficult. For example, it is often difficult to decide whether the word *takže* serves as a DM or not. When annotating instances of this word, one must analyze whether the speaker intented to express relation to his/her preceding proposition, or to mark a discourse boundary. Another ambiguous word is *jako*, as illustrated in the following example. In the first sentence, it expresses a comparison, while in the second, it functions as a DM.

```
Je rychlý jako blesk /.
vs.
To jako není nic neobvyklého /.

He is fast like lightning /.
vs.
It is like nothing unusual /.
```

Besides general DMs, the MDE annotation system also recognizes its special case – 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, a speaker may also initiate his/her attempt to take the floor. DRs typically occur turn-initially. Importantly, DRs should not be confused with direct answers to questions. Distinction between DRs and direct responses to questions is discussed below in Section 4.3.16. An example of a DR follows.

```
A: Já bych to tak udělal /.
B: Hele já si tím nejsem tak jistej /.

A: I'd do it that way /.
B: See I'm not that sure about it /.
```

4.1.3 Asides/Parentheticals

Asides and parentheticals occur when the speaker utters a short side comment and then returns to the original sentence pattern. Asides are comments on a new topic, while parentheticals are on the same topic as the main utterance. For annotation purposes, asides and parentheticals are not distinguished but treated as a single filler type. A/Ps are often prosodically marked. Speakers usually pause or shift their intonation. 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 in the guidelines. An example of an A/P follows.

```
Potom k němu přišel { moment musím si vypnout telefon } s tím velkým
psem /.

Then he came to him { moment I must switch off my cell phone } with the
big dog /.
```

Some very common Czech words or short phrases that can be denoted as "lexicalized parentheticals" (e.g., *řekněme (say),myslím (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. In addition, the maximal allowed length of a lexicalized parenthetical was limited to two words.

An important restriction of A/Ps within our MDE guidelines is that they cannot occur as SU-initial or SU-final. Such grammatical parentheticals occurring not in the middle of an SU but on their onset or end should rather be separated by an SU symbol (clausal break or an SU-external break).

Je to sto hlasů i s tím ministrem /, jak jsme dneska četli /.

It's one hudred votes including the minister /, as we read today /.

4.1.4 Explicit Editing Terms

Another type of fillers, EET, may only occur accompanying an edit disfluency. EETs are explicit expressions by which speakers signal that they are aware of the existence of a disfluency on their part. Basically, they can appear anywhere within the disfluency, but most frequently occur right after the end of the reparandum. EETs are rather rare in actual conversational language. Typical Czech EETs are e.g., *nebo (or), či (or), spíše (rather), vlastně (actually)*, or *chtěl jsem říct (I wanted to say).*

Tohle je naše [koherentní]* *EE spíše* konzistentní stanovisko /.

This is our [coherent]* *EE rather* consistent statement /.

4.2 Edit Disfluencies

Edit disfluencies are portions of speech in which a speaker corrects or alters his/her utterance, or abandons it entirely. Herein, the phases of an edit disfluency are referred to as 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 a disfluency), and correction (portion of speech in which speaker corrects or alters the DelReg). Whereas corrections were not explicitly tagged within the English MDE project, we decided to label them in order to obtain relevant data for further research of spontaneous Czech. Their labeling is not very time consuming and the obtained data may be very useful – some typical correction patterns may be learned. An example of an edit disfluency follows:

Naše děti milují [kočku]* *EE vlastně* psa pana Krause /.

Our children love [the cat]* *EE actually* the dog of Mr Kraus /.

Moreover, it often happens that a speaker produces several disluencies in succession, either as serial or nested. In case of serial disfluencies, we simply mark the maximal extent of the disfluency as a single DelReg with multiple IPs that are explicitly tagged.

Ale [ta * myšl- * ten * ten zlej]* ten podivnej pocit to se nedá dobře popsat /.

But [the * ide- * the * the bad]* the strange feeling it can't be described well /.

Nested disfluencies (some component of the disfluency is disfluent itself) are more difficult to annotate. To keep annotation as simple as possible, the MDE standard does not allow using nested disfluency labels, so that all such disfluencies must be annotated using just simple, non-nested DelRegs. The following example shows a correction that contains an additional disfluency.

Přijel jsem [do Brna]* do [Plz-]* Plzně dnes ráno /.

I arrived [to Brno]* to [Pil-]* Pilsen today morning /.

Since Czech disfluencies have the same pattern as English, the rules about complex disfluencies from [1] can basically be directly applied to Czech. Thus, we do not survey all particular rules for annotating complex disfluencies herein because interested readers may consult the original English guidelines.

4.3 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 boundary with consistency using only a vague notion of a "conversational equivalent" of a written sentence – strict segmentation rules are necessary. Past experience with similar annotation problems indicates that acceptable inter-annotator agreement (IAA) can only be achieved in the context of rules grounded in "surface features", i.e. mainly syntax and prosody. Semantic features have not proved to be reliable.

One possible solution to the "conversational sentence" definition problem is to divide the flow of speech into "minimal meaningful units" functioning to express one complete idea on the speaker's part. It means that we divide the stream of words wherever it is grammatically possible and meaningful. The resulting units are either shorter or equally long as sentences in standard writing. These smaller units also seem to be convenient for downstream automatic applications. For example, speech translation applications usually prefer to process shorter segments [17].

The target utterance units are called SUs within the MDE task. In the MDE definition, the abbreviation SU may stand for one of the following possibilities: Sentential/Syntactic/Semantic/Slash Unit. 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 are classified according to their function within the discourse. The following list shows the employed SU symbols (breaks) along with brief descriptions of their meaning:

- /. – 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 of either two dependent clauses or 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 /.`)

In contrast to the English MDE, we do not use an SU symbol for backchannels because both Czech corpora are single-channel. Therefore, backchannels that do not overlap with words

uttered by the dominant speaker, and thus can be captured in a single-channel transcript[7], are treated as a special type of a filler.[8] In the illustrative examples below, we also use a special symbol "⊘" which denotes "no break" at places where one might be expected. This symbol is only used for illustration purposes herein, and does not occur in real MDE annotations.

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. In many cases, these breaks would be represented in standard writing with end-of-sentence punctuation. Sentence-internal breaks are secondary and have mainly been introduced to support IAA. However, it should be noted that it is important to have these symbols in the MDE annotations since some future task may require them to be automatically detected. Sentence-internal breaks delimit units that are smaller than a main clause and cannot stand alone as a complete sentence. In standard writing, these breaks often correspond to commas.

In SU annotation, the fundamental problem is to determine when to insert a new SU boundary and when to place two segments within the same SU. External breaks are inserted between SU boundaries, internal breaks (if exist) may further refine each SU. Besides a few exceptions[9], candidate locations for both sentence-internal and sentence-external SU labels are usually boundaries between two adjacent clauses. Thus, the key problem is to recognize the type of each clause boundary.

The above presented set of SU symbols corresponds to the original MDE standard. However, we did not use it as it was originally defined but introduced two significant modifications. Both modifications are language-independent. First, the original set contains only one symbol for incomplete SUs, but we propose to distinguish two types of incomplete SUs: /- — indicating that the speaker abandoned the SU arbitrary; and /∼ — indicating that the speaker was interrupted by another speaker. This distinction of incompletes 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 based on prosodic cues, and also relevant for spontaneous discourse analysis, 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. The additional annotation refinement does 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. 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 proposed guideline modifications did not only include changes in the SU symbol set. Another 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.

[7] Overlapping backchannels are treated as noises since they cannot be explicitly transcribed within the dominant speaker turn.

[8] Since these non-overlapping backchannels are extremely rare in the corpus, we did not present their annotation in a separate section.

[9] These are mentioned in the following sections.

Likewise, we do not require the presence of a noticeable pause after incomplete (abandoned) SU breaks (/-) when syntax provides an overt evidence of incompleteness.

The following subsections provide descriptions of particular rules for SU annotation. To keep the description reasonably long, we only present the most important examples – especially those that emphasize differences between Czech and English. Full annotation guidelines may be found at `http://www.mde.zcu.cz`.[10]

4.3.1 Short Stand-alone Phrases Not Containing Verbs

SUs do not necessarily have to contain a verb. Even though some phrases do not constitute grammatically complete sentences, they may function as a complete utterance. To identify them correctly, annotators must be sure that the phrases are not syntactically connected with the previous SU.

```
Vítejte u Radiofóra /.  Hosté Jan Novák poslanec a Pavel Kučera stínový
ministr obrany /.


Welcome to Radioforum /.  Guests Jan Novák a deputy and Pavel Kučera a
shadow minister of defense /.
```

These stand-alone phrases often occur following a question – talkers sometimes repeat the question's topic to establish common ground before answering.

```
A: Jsou pro vás Glasgow Rangers těžkým soupeřem /?
B: No Rangers /.  My jsme s losem spokojeni /.


A: Are Glasgow Rangers a tough opponent for you /?
B: Well Rangers /.  We are satisfied with the draw /.
```

Another examples are headline news in broadcast news data which should also be annotated as individual SUs even if they do not form a complete clause.

```
Mušaraf neoficiálním vítězem pákistánských prezidentských voleb /.
Německo ochromeno stávkou železničářů /.  Madelaine Albrightová
v exkluzivním interview pro Českou televizi /.


Musharraf the unofficial winner of Pakistan's presidential vote /.
Germany paralyzed by rail strike /.  Madelaine Albright in an exclusive
interview for the Czech TV /.
```

All these rules are identical to the corresponding English rules.

4.3.2 Juxtaposition of Clauses

In general, juxtaposition means an absence of linking elements in a group of words that are listed together. As juxtaposition of an "introductory clause", we understand a connection of two main clauses that cannot be classified using any of the standard semantic relations defined by normative Czech grammar (copulative, disjunctive, etc.). The second clause is syntactically and

[10] The complete annotation guidelines are available only in Czech.

semantically determined by the preceding clause, but there is no formal syntactic relationship. Thus, it is a kind of parenthetical clause in terms of grammar. In most of the cases, such clauses could be connected using the Czech conjunction *"že (lit. that)"* without any change of meaning. In English, this phenomenon does not have a separate rule since, unlike Czech, dropping of the conjunction *that* is standard. Czech guidelines instruct annotators to separate the clauses in juxtaposition using a clausal break.

```
Já vím /, vy to nemáte rád /.
I know /, you don't like it /.
```

4.3.3 Quotations

Since no quotation marks are used within the MDE annotation, direct or indirect quotations impose clausal breaks. The quote and its attribution typically form a single SU. This rule is identical to the corresponding English rule.

```
Půjdu tam /, řekl David /.
Martin řekl /, že tam nepůjde /.
I'll go there /, David said /.
Martin said /, that he wouldn't go there /.
```

If the quote is long and the quoted portion of the utterance contains several sentences, additional SUs are recognized, as shown in the following example:

```
Ale premiér řekl /, nikdy jsem ho neviděl /.  Já toho člověka vůbec
neznám /.  Tahle aféra je směšná /.
But the prime minister said /, I have never seen him /.  I don't know
that man at all /.  This affair is ridiculous /.
```

4.3.4 Idiomatic Expressions

Similarly to English, frozen idiomatic expressions are not separated by any SU symbols even if they contain multiple finite verbs.

```
To je ⊘ prašť ⊘ jako uhoď /.
It is ⊘ hit ⊘ or punch /.
```

4.3.5 Independent Subordinate-like Clauses

A complete SU may also be composed of stand-alone independent clauses starting with subordinate conjunctions. In these cases, the subordinate conjunctions basically functions as particles rather than conjunctions.

```
Protože toto je opravdu jednoduché /.
Because this is really easy /.
```

4.3.6 Parcelation

In spontaneous speaking, speakers often do not precisely plan the structure of their utterances in advance. As a result, we sometimes observe discontinuous appending of additional utterance constituents. The talker composes several successive elliptic utterance units, which typically have separate focal accents. This phenomenon is referred to as parcelation in Czech literature. If this parcelation is strong (which is typically recognized from short pauses between constituents), the utterance is segmented into multiple SUs.

```
Chceš s ním mluvit /?  Sama /?  Beze svědků /?
Do you want to speak with him /?  Alone /?  Without witnesses /?
```

4.3.7 Appositions

Apposition is a grammatical construction in which two elements are placed side by side, with one element serving to define or refine the other. In standard Czech writing, these constituents are typically separated by a comma, however, in the MDE annotation, we do not separate them by any SU breaks.

```
Daniel Mach ⊘ ředitel místní školy ⊘ je můj přítel /.
Daniel Mach ⊘ the local school principal ⊘ is my friend /.
```

In most cases, noun phrases form appositions. However, based on the broad definition of appositions (as defined by Vladimír Šmilauer), verbs, adverbs, or even clauses may form appositions, too. If a clause embedded in another clause appears in a apposition, we separate it by a clausal break.

```
Řeka se kroutí /, tedy tvoří meandry /, blízko u lesa /.
The river twirls /, thus it forms meanders /, close to the wood /.
```

We should also mention special introductory phrases such as *"to jest (*lit. *that is)"* or *"to znamená (*lit. *it means)"* (i.e., frozen phrases containing a finite verb), which frequently accompany clausal appositions. In terms of MDE, these phrases should not be understood as clauses but rather as introductory particles. As a result, they do not motivate any SU breaks. An illustrative example follows.

```
Řeka se kroutí /, to znamená ⊘ tvoří meandry /, blízko u lesa /.
The river twirls /, it means ⊘ forms meanders /, close to the wood /.
```

In contrast to the previous examples, if a clause that seems to be appositional is not embedded in another clause and may stand alone, it should be annotated as an independent SU. This is in agreement with the rule of thumb for problematic decisions – "segment wherever it is possible!".

```
Důkazy byly takové /, že soudy je osvobodily /.  To znamená zbavily je
toho obvinění /.
There were such evidence /, that the court set them free /.  It means
they found them not guilty /.
```

4.3.8 Freely Adjoined Sentence Parts

Freely adjoined sentence parts, which are typically separated by commas and introduced by such particles as "a sice" or "a to" (lit. "and that") in Czech, do not motivate any SU breaks.

Konečně dal gól ⊘ a to kupodivu hned Dominiku Haškovi /.

He finally scored a goal ⊘ and that for a wonder against Dominik Hašek /.

4.3.9 Anacolutha

An anacoluthon in spoken language can be defined as an abrupt change of syntax within an utterance. In other words, an utterance begins in a way that implies a certain logical resolution, but concludes differently from the form grammar leads us to expect. An example of an anacoluthon in Czech is a disagreement between subject and predicate within a clause. Although anacolutha can also be used as a purposeful stylistic virtue, they more frequently occur as a consequence of an unintentional grammatical fault in conversational language. In the Czech corpora, anacolutha often occur in the vicinity of parentheticals and asides, as shown in the following example.

Pokud se skupina států {a nejsou to jenom Spojené státy je to také Velká Británie a další} **rozhodnou** použít sílu /, tak ...

If the group of states {and it's not just the United States it's also Great Britain and others} **decide** to use power /, then ...

In the example above, to match the singular subject *"skupina (group)"*, Czech grammar strictly requires to use the singular *"rozhodne (decides)"* instead of the plural *"rozhodnou (decide)"*. However, the speaker got confused by using a plural form in the parenthetical, and continued to use the plural form in the completion of the original message.

Annotators were instructed not to use any special annotation for these "small" anacolutha, pretending that grammar in the disturbed utterances is correct. However, note that anacolutha should not be confused with edit disfluencies.

4.3.10 Tag Questions

Tag questions are short phrases added to the end of a statement in order to appeal to the listener to give feedback. In terms of MDE, the statement plus the added phrase form a single SU that should be labeled as interrogative. The added phrase is separated from the preceding statement using a clausal break. This rule is identical to the corresponding English rule.

To si děláte legraci /, že jo /?

You must be joking /, aren't you /?

However, if intonation gives a clear clue that the added phrase does not function as a question, the whole SU is labeled as a statement and the added phrase is labeled as a DM.

Přišli tam včera *jo* /.

They came there yesterday *yeah* /.

4.3.11 Rhetorical Questions

Rhetorical questions receive standard question SU labels. This rule is identical to English.

```
I my můžeme vyhrát /.  Není snad míč kulatý /?
Even we can win /.  Isn't the ball round /?
```

4.3.12 Embedded Questions

When a question is embedded in a larger carrier clause, SU type is assigned according to the function of the whole utterance, and not according to the embedded question. Embedded questions most frequently occur in quoted direct speech. This rule is also identical to the corresponding English rule.

```
Zeptala se /, přidáš se k nám (?)  /.
She asked /, will you join us (?)  /.
```

4.3.13 Incomplete SUs

When a speaker's utterance does not express a complete thought, an incomplete SU is recognized. Boundaries of the incompletes are labeled with either "/–" or "/∼". If the utterance is interrupted and cut short by another speaker, then the "/∼" symbol is used. On the other hand, if the speaker abandons his/her utterance arbitrarily, the SU is annotated as "/–". It implies that "/∼" may only occur at a turn boundary, whereas "/–" may also occur as turn-internal. In standard text, "/–" may correspond to ellipses (. . .). The first example illustrates the use of "/∼":

```
A: Pokud vložíte dostatek peněz do /∼
B: Ale to není jen otázka peněz /.
A: If you put enough money into /∼
B: But it is not just a matter of money /.
```

The second example illustrates the use of "/–":

```
Jeho boty vypadaly jako /- On je divnej kluk /.
His shoes looked like /- He is a weird guy /.
```

4.3.14 Distinguishing Incomplete SUs and Restart Disfluencies

Incomplete SUs are sometimes difficult to distinguish from restart disfluencies, which do not receive incomplete SU labels, but are annotated as DelRegs. The distinction between these two phenomena within the MDE standard is based on the following rules. In comparison with the original MDE standard for English, the rules for Czech are slightly more complex.[11] Our experience indicates it does not lead to a significant decrease of IAA, and the accuracy of annotation is increased.

[11] In the original MDE standard, incomplete SUs are only recognized if "a speaker is interrupted or when the speaker trails off, failing to complete the utterance within a turn". Thus, incomplete SUs can only occur at the end of a speaker's turn.

1. Restart disfluency may never appear as turn-final. In such cases, the incomplete utterance is always identified as an incomplete SU.

2. If the speaker immediately restructures the interrupted utterance and continues speaking on the same topic, restart disfluency is recognized. On the other hand, if he/she does not return to the incomplete message, an incomplete SU is recognized.

3. Incomplete SUs of the "/–" type must always contain either one or more SU-internal breaks (/& or /,), or "useful information" that is not repeated in the same turn. However, this does not mean that the occurrence of an SU-internal break within an incomplete utterance automatically implies the use of "/–". When the SU-internal break occurs in a very short introductory phrase such as *"víte /, že (you know /, that)"*, it is possible to annotate it as a DelReg (if other necessary conditions are met).

4.3.15 Turns with Missing Onsets

Turns whose onsets are missing in the verbatim transcripts, or whose onsets are transcribed within the immediately preceding overlapping speech section, are annotated in the same way as if their onsets were present. Note that the overlapping speech regions in the verbatim transcripts were not used for MDE annotation, and thus they do not contain any MDE symbols. An example of such a turn follows. The first line in the example corresponds to an overlapping speech region (both A and B speak), in the second line, the speaker B continues the utterance that was started in the overlapping region.

```
A: Překvapil vás.  B: Na druhou stanu překvapil ~ (OVERLAP)
B: ~ i mě /.  Já jsem k tomu už názor vyjádřil /.
```

```
A: He surprised you.  B: On the other hand he surprised ~ (OVERLAP)
B: ~ me as well /.  I have already stated my opinion on this /.
```

4.3.16 Direct Responses Expressing Agreement or Disagreement

Direct responses to questions expressing speaker's agreement or disagreement such as *ano (yes), ne (no), jo (yeah), m-hm (uh-huh)* typically form a complete SU.

```
A: Bude to hotové do příštího týdne /?
B: Ano /.  Pevně v to doufám /.
```

```
A: Will it be finished by next week /?
B: Yes /.  I strongly hope so /.
```

If a subordinate clause having an explanatory function is attached to words expressing agreement or disagreement, a clausal break is used.

```
A: Zkusíte to /?
B: Ne /, protože už je příliš pozdě /.
```

```
A: Will you try that /?
B: No /, because it's too late now /.
```

One must also be aware of the fact that words that often express agreement or disagreement may also function as discourse markers. The discriminative rule for these ambiguities says that

both agreement and disagreement words must *always* be preceded by a question. Otherwise, a DM is recognized. Although this simplification is not absolutely accurate in terms of discourse analysis, it was introduced in this simplified form in order to support IAA. The use of this rule is illustrated in the following example presenting a part of a fictitious dialog.

A: Myslím /, že se to stane /.
B: Ano /?
A: Ano /. Ten příkaz už je podepsaný /.
B: *Ano tak* to je problém /.
A: *Ano* ⊘ je to opravdu nepříjemné /.
B: *Takže* oni přijdou /? a *jo* ⊘ odnesou všechno /?
A: Ano /. Je mi to líto /.

A: I guess /, that this will happen /.
B: Yes /?
A: Yes /. The order has already been signed /.
B: *Yes so* it's a problem /.
A: *Yes* ⊘ it is really bothersome /.
B: *So* will they come /? and *yeah* ⊘ take everything /?
A: Yes /. I am sorry /.

Note that nonverbal sounds such as *uh-huh* may also function as direct responses to yes/no questions. In that case, they also form a complete SU.

A: Je to v pořádku /?
B: HM /. Pojďme dál /.

A: Is it ok /?
B: Uh-huh /. Let's move on /.

4.3.17 Subordinate Clauses within Complex Sentences

Dealing with complex and compound sentences represents one of the most important parts of the MDE annotation. This section describes annotation of complex sentences that contain some kind of subordination. Subordinate clauses cannot themselves constitute a complete SU because they depend on the rest of the sentence and thus may not stand on their own; they are semantically linked to their main clauses. Subordinate clauses are separated by clausal breaks within MDE.

Já ti tu adresu dám /, když mi zavoláš /.

I will give you the address /, if you call me /.

If there is not just a single subordinate clause but two subordinate clauses dependent on the same independent clause and joined by coordination, a coordination break is used to separate these two dependent clauses. An SU-external break cannot be applied since neither of these subordinate clause can stand on its own without changing meaning of the whole statement.

```
Já ti tu adresu dám /, když mi zavoláš /& nebo pošleš email /.
I will give you the address /, if you call me /& or send me an email /.
```

Unlike English, relative clauses are separated by clausal breaks in the Czech MDE. This adjustment reflects Czech syntax which requires to separate relative clauses by commas, regardless whether they are restrictive or not. If we did not use clausal breaks for relative clauses, the MDE transcripts would be less transparent for the annotators.

```
Daniel /, který se narodil v Praze /, miluje Karlův most /.
Daniel /, who was born in Prague /, loves the Charles bridge /.
```

4.3.18 Compound Sentences

Compound sentences consist of two ore more main (independent) clauses joined by coordination. As described above, the goal is to divide the compound sentences within a spoken discourse into "minimal meaningful units" functioning to express a complete idea. It means that we split independent clauses into two complete SUs every time they can stand alone (i.e. they do not depend on each other for completion of an idea). The potential break point is the interword boundary right before the coordinating conjunction as shown in the following example.

```
Adam hraje tenis /. a Robert cvičí jógu /.
Adam plays tennis /. and Robert practices yoga /.
```

However, not all cases are that clear as the one in the example above. In some cases, coordinated main clauses cannot be split into two independent SUs. In such cases, a coordination break is used instead of an SU-external symbol. In English MDE, this situation most frequently arise when the second coordinate clause has a dropped subject. In English, subject dropping is only allowed in the second clause of a compound sentence when both clauses share the same subject. It implies that such compound sentences cannot be divided into two SUs because coordinated main clauses with dropped subjects do not form syntactically encapsulated units, and thus cannot stand alone. See the difference in the following illustrative example.

```
I love volleyball /. but I hate playing with beginners /.
vs.
I love volleyball /& but hate playing with beginners /.
```

However, this rule cannot be applied to Czech. In contrast with English, Czech subjects (pronouns) can be dropped every time they are "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 just "*Běžím /.*" which means "*(I am) running /.*" This phenomenon of subject dropping is typical for highly inflective languages.

For the above stated reason, 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 /.
Robert walked to work /. but (he) took the train home /.
```

In the Czech MDE, a coordination break is used for separation of coordinated main clauses in the following cases:

1. The compound sentence is structured with a non-continuous expression such as *sice – ale (though – but), buď – nebo (either – or)*, or *nejen – ale i (not only – but also)*.

   ```
   Ona je nejenom krásná /& ale také je laskavá /.
   Not only she is beautiful /& but also she is kind /.
   ```

2. The second coordinate clause is elliptical and cannot stand alone.

   ```
   Katka miluje kosatce /& ale Eva tulipány /.
   Katka loves irises /& but Eva tulips /.
   ```

3. There exists a subordinate clause that is dependent on both main clauses.

   ```
   Když byl hotov /, zavřel okno /& a sedl si na postel /.
   When he was finished /, he closed the window /& and sat on the bed /.
   ```

4. Main clauses are joined by the syntactically primarily coordinating yet semantically often rather subordinating conjunction *neboť (for)*.

   ```
   Šli jsme se koupat /& neboť bylo krásné počasí /.
   We went swimming /& for the weather was great /.
   ```

The rule No. 1 was adopted from the English MDE. The rules No. 2 and 3 are not explicitly mentioned in the English guidelines, however, the correct annotation of these phenomena should be the same as for Czech. The last rule in the list is specific for Czech.

4.3.19 Coordinate Questions

When two questions are coordinated within a compound sentence, both of them receive the question label.

```
Půjde Robert do divadla /?  a Adam zůstane doma /?
Will Robert go to the theater /?  and will Adam stay at home /?
```

However, it is very common to drop an auxiliary verb in the second interrogative clause, which, in contrast, induces the use of a coordination break.

```
Bude Robert v divadle /& a Adam doma /?
Will Robert be in the theatre /& and Adam at home /?
```

4.3.20 Compound Predicates

Another important fact influencing the Czech MDE 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). The compound predicate is 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 the MDE purposes, we only recognize those compound predicates that can be identified based on unambiguous features. Within Czech MDE, the compound predicate is recognized if:

1. The predicate verbs share a common constituent (e.g., object).

   `Nacpal /& a zapálil si dýmku /.`
   `He filled /& and lit up his pipe /.`

2. The predicate verbs joined by a copulative conjunction have the same or very similar meaning.

   `Naši hosté často slaví /& a radují se /.`
   `Our guests often rejoice /& and celebrate /.`

While compound predicates did not motivate any SU breaks according to the initial version of the annotation guidelines, the current version instructs annotators to separate parts of compound predicates by a coordination SU break. A preliminary analysis showed that the redefined annotation rule supported inter-annotator agreement.

5 Summary

We have described structural metadata annotation for Czech. The structural metadata annotation is based on the LDC's "Simple Metadata Annotation Specification", originally defined for English. The original guidelines were adjusted to accommodate specific phenomena of Czech syntax. These adjustments mostly affected annotation of SUs. We also proposed and used a novel approach to transcribing and annotating filled pauses in Czech, distinguishing vowel-like (*EE*) and consonant-like (*MM*) sounds.

In addition to the necessary language-dependent modifications, we applied some language-independent modifications refining the original annotation scheme. The most important language-independent modifications are the following:

- We distinguish two types of incomplete SUs: The incompletes interrupted by another speaker (/–) and the incompletes arbitrarily abandoned by the same speaker (/∼).
- We introduced limited prosodic labeling at SU boundaries that distinguish complete SUs with strongly prosodically marked boundaries (//. and //?) and SUs without strong prosodic marking at the boundary (/. and /?).
- Because we only have single channel data in which non-overlapping and intelligible backchannels are very rare, we do not annotate backchannels as separate SUs. Intelligible backchannels by speakers not holding the floor are annotated as special fillers within the dominant speaker turn.
- We dropped the rule that a pause longer than 0.5 sec automatically induces the end of a speaker turn and thereby requires a corresponding SU-external break. In such cases, our SU segmentation rules rely solely on syntax.
- We do not automatically insert IPs before fillers. IPs are inserted automatically only at the right edges of DelRegs.

Acknowledgments

We thank Stephanie Strassel and Christopher Walker from the Linguistic Data Consortium for helping us understand all the subtle details of the MDE annotation, and Dagmar Kozlíková for valuable advice about some detailed features of Czech syntax. This work was supported by the Ministry of Education of the Czech Republic under projects No. 2C06020 and ME909. The views expressed are those of the authors, and not the funding agencies.

References

[1] S. Strassel, "Simple metadata annotation specification V6.2," http://www.ldc.upenn.edu/Projects/MDE/Guidelines/SimpleMDE_V6.2.pdf, 2004.

[2] J. Kolář, "Automatic segmentation of speech into sentence-like units," Ph.D. dissertation, University of West Bohemia, Pilsen, Czech Republic, 2008.

[3] V. Radová, J. Psutka, L. Müller, W. Byrne, J. V. Psutka, P. Ircing, and J. Matoušek, "Czech Broadcast News Speech and Transcripts," Linguistic Data Consortium, CD-ROM LDC2004S01 and LDC2004T01, Philadelphia, PA, USA, 2004.

[4] J. Psutka, V. Radová, L. Müller, J. Matoušek, P. Ircing, and D. Graff, "Large broadcast news and read speech corpora of spoken Czech," in *Proceedings of EUROSPEECH*. Aalborg, Denmark: ISCA, 2001, pp. 2067–2070.

[5] J. Psutka, L. Müller, J. Matoušek, and V. Radová, *Mluvíme s počítačem česky (Speaking with Computer in Czech)*. Prague: Academia, 2006.

[6] M. Meeter, "Dysfluency annotation stylebook for the Switchboard corpus," ftp://ftp.cis.upenn.edu/pub/treebank-/swbd/doc/DFL-book.ps, 1995.

[7] P. Heeman, "Speech repairs, intonational boundaries and discourse markers: Modeling speakers' utterances in spoken dialogs," Ph.D. dissertation, University of Rochester, New York, 1997.

[8] A. Batliner, R. Kompe, A. Kiessling, M. Mast, H. Niemann, and E. Nöth, "M = Syntax + Prosody: A syntactic–prosodic labelling scheme for large spontaneous speech databases," *Speech Communication*, vol. 25, pp. 193–222, 1998.

[9] O. Müllerová, *Mluvený text a jeho syntaktická výstavba (The Syntax of Spoken Text)*. Praha: Academia, 1994.

[10] S. Strassel, D. Miller, K. Walker, and C. Cieri, "Shared resources for robust speech-to-text technology," in *Proc. EUROSPEECH 2003*, Geneva, Switzerland, 2003.

[11] Y. Liu, E. Shriberg, A. Stolcke, B. Peskin, J. Ang, D. Hillard, M. Ostendorf, M. Tomalin, P. Woodland, and M. Harper, "Structural metadata research in the EARS program," in *Proc. IEEE ICASSP*, Philadelphia, USA, 2005.

[12] K. Svoboda, *Souvětí spisovné češtiny (Compound Sentences in Standard Czech)*. SPN Praha, 1970.

[13] M. Grepl and P. Karlík, *Skladba češtiny (Syntax of Czech)*. Votobia, 1998.

[14] M. Grepl, Z. Hladká, M. Jelínek, P. Karlík, M. Krčmová, M. Nekula, Z. Rusínová, and D. Šlosar, *Příruční mluvnice češtiny (Handbook of Czech grammar)*, P. Karlík, M. Nekula, and Z. Rusínová, Eds. Lidové noviny, 2003.

[15] M. Erard, "Just like, er, words, not, um, throwaways," *New York Times*, vol. (Jan 3, 2004), 2004, available from http://www.speech.sri.com/press/nyt-jan03-2004.html.

[16] P. Kaderka and Z. Svobodová, "Manuál pro přepisovatele televizních diskusních pořadu (Guidelines for annotators of broadcast discussions)," *Jazykovědné aktuality*, no. 3-4, pp. 18–51, 2006.

[17] D. Hillard, "Automatic sentence structure annotation for spoken language processing," Ph.D. dissertation, University of Washington, Seattle, WA, USA, 2008.