Structural Metadata Annotation of Speech Corpora: Comparing Broadcast News and Broadcast Conversations

Jáchym Kolář, Jan Švec

University of West Bohemia – Faculty of Applied Sciences
Univerzitni 8, 306 14 Pilsen, Czech Republic
{jachym,honzas}@kky.zcu.cz

Abstract

Structural metadata extraction (MDE) research aims to develop techniques for automatic conversion of raw speech recognition output to forms that are more useful to humans and to downstream automatic processes. It may be achieved by inserting boundaries of syntactic/semantic units to the flow of speech, labeling non-content words like filled pauses and discourse markers for optional removal, and identifying sections of disfluent speech. This paper compares two Czech MDE speech corpora – one in the domain of broadcast news and the other in the domain of broadcast conversations. A variety of statistics about fillers, edit disfluencies, and syntactic/semantic units are presented. Among many others, we report the statistics indicating that disfluent portions of speech show differences in the distribution of parts of speech (POS) of their word content in comparison with the overall POS distribution. The two Czech corpora are not only compared with each other, but also with available statistics relating to English MDE corpora of broadcast news and telephone conversations.

1. Introduction

Structural metadata extraction (MDE) research aims to develop techniques for automatic conversion of raw speech recognition output to forms that are more useful to humans and to downstream automatic processes, such as speech summarization or machine translation. It may be achieved by inserting boundaries of syntactic/semantic units to the flow of speech, labeling non-content words like filled pauses and discourse markers for optional removal, and identifying sections of disfluent speech. For training of automatic MDE systems, adequately annotated speech corpora are required.

Several different annotation schemes have been presented for similar tasks. Earliest efforts include the manual for disfluency tagging of the Switchboard corpus (Meeter, 1995), Heeman's annotation scheme for the Trains dialog corpus (Heeman, 1997), and a syntactic-prosodic labeling system for spontaneous speech called "M" presented in (Batliner et al., 1998). For our work on spoken Czech, we have decided to adopt the "Simple Metadata Annotation" approach introduced by Linguistic Data Consortium (LDC) as part of the DARPA EARS program (Strassel, 2004). Originally, this standard was defined for English. Later efforts have extended the guidelines for use with Mandarin and Arabic (Strassel et al., 2005), however, the first complete non-English MDE corpus was Czech. The development of MDE annotation guidelines for spontaneous Czech as well as recording and annotation of the broadcast conversation corpus Radioforum (RF) was described in (Kolář et al., 2005).

We have recently annotated additional Czech data from another domain – broadcast news (BN). This new MDE corpus was created by enriching the existing Czech BN corpus (Radová et al., 2004) by MDE annotation. In this paper, we compare the two Czech corpora in terms of MDE statistics. This comparison is not only important for evaluating the complexity of the MDE task in particular genres, but also is interesting in terms of spoken discourse analysis, since to our best knowledge, there exists no similar large scale study for Czech or other Slavic languages. The insights provided by this study may help improve automatic MDE systems.

2. Speech Data

2.1. Broadcast News

The broadcast news (BN) data we used were from the Czech Broadcast News Corpus which is publicly available from the LDC (Radová et al., 2004). 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 broadcast companies include both public and commercial subjects. Therefore, the corpus contains news broadcasts presented in different styles – ranging from a very formal style to rather colloquial style typical for commercial broadcast companies that do not primarily focus on news. The whole corpus contains over 60 hours of audio recorded over the air, which yields about 26 hours of pure transcribed speech by 284 talkers (188 males and 96 females). The total word count is 234k. More details about corpus transcription are given in (Psutka et al., 2001).

2.2. Broadcast Conversations

The spontaneous speech database consists of 52 single channel recordings of a radio discussion program called *Radioforum* (RF), which is broadcast by Czech Radio 1 every weekday evening. Radioforum is a live talk show where invited guests (most often politicians) 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 total number of speakers in the whole corpus is 94 (77 males, 17 females). Because of the scope of the talk show, speakers younger than 30 years of age are rather rare. The total duration of pure transcribed speech is 24 hours, the transcripts contain 201k word tokens (lexemes). The recordings were acquired during the period from February 12, 2003 through June 26, 2003, the signal is single channel and recorded over the air. More information about the data is given in (Kolář et al., 2005).

3. Metadata Annotation

MDE 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. 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.

Both our MDE corpora have been annotated just by two annotators. Since Czech syntax is quite complex, "naïve" annotators could not be employed; at least some linguistic education is necessary and such annotators were difficult to find. The small number of labelers slowed down the annotation process, but supported annotation consistency. Submitted annotations were carefully checked for correctness by the authors of this paper who were allowed to make final decisions in questionable parts of annotation.

Because of budget constraints we were not able to get dual annotations for all transcripts, but evaluated the inter-annotator agreement on three dually-annotated recordings from the more difficult RF corpus. The agreement was measured in terms of the kappa statistic (Carletta, 1996). We got $K = 0.88$ for SU breaks and $K = 0.85$ for filler and disfluency labels. Given the complexity of the annotation task, these numbers seem to be very well acceptable. More details on the MDE annotation are given in (Kolář et al., 2005; Kolář, 2008) as well as on the project website `http://www.mde.zcu.cz`.

4. Metadata Statistics

We do not only compare the two Czech corpora with each other, but also with available numbers relating to English MDE corpora (Liu et al., 2006). The Czech BN corpus is compared with the English BN MDE corpus, and the RF corpus with the English conversational telephone speech (CTS) corpus. Quite a good match is expected for BN data, but when comparing Czech broadcast conversations with English telephone speech, we must also take into account particular speaking styles. Although both corpora contain spontaneous speech, broadcast conversations are more formal and less interactive than telephone speech.

4.1. Fillers

Fillers are words, short phrases, or non-verbal 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 transcription without losing "important" information about its content. Four types of fillers are distinguished within the MDE system: filled pauses (FP), discourse markers (DM), explicit editing terms (EET), and asides/parentheticals (A/P).

	RF	BN
% of words followed by FPs	3.8%	0.5%
Proportion of *EEs*	93.1%	N/A
Proportion of *MMs*	6.9%	N/A
% of words in DMs and DRs	1.6%	0.1%
Proportion of DMs	73.3%	46.7%
Proportion of DRs	26.7%	53.3%

Table 1: Filled pauses and discourse markers in Czech MDE corpora

4.1.1. Filled Pauses

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 only distinguished two types of Czech FPs: *EE* (most typical example 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 hesitation sounds that are phonetically more similar to consonants or mumble-like sounds, typically pronounced with a closed mouth). Note that this way of FP transcription was only employed in the RF corpus since the original transcriptions of the Czech BN corpus used the English notation for FPs although it is not convenient for Czech.

The top part of Table 1 reports numbers relating to occurrences of FPs. As expected, FPs are significantly more frequent in conversational than in broadcast news data. *EE* FPs are much more frequent than *MMs* – they represent more than 93 % of FPs. For comparison, 2.2 % of words is followed by an FP in the English CTS corpus, and 1.4 % in the English BN corpus. A relatively smaller number of FPs in English CTS data might be explained by three different factors. First, transcribers of the English database could have missed a number of FPs, since some of them are less audible and telephone data are more noisy. Second, Czech syntax is more complex than English, so that speaking in Czech represents a more complex mental process which may cause a higher number of hesitations. Third, talkers may hesitate by voice more when speaking in public, because in private conversations, people often do not care about being grammatically correct which makes speech planning easier. On the other hand, the larger percentage of FPs in English BN data is caused by the fact, that these data contain a larger proportion of speech having a relatively higher level of spontaneity. The reason is that Czech public radio and TV channels still use a significantly less interactive style than typical American broadcast news.

4.1.2. Discourse Markers

DMs are words or phrases, such as *you know,* that function primarily as structuring elements in 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. There exists a number of diverse definitions of DMs in the linguistics literature. Within MDE, we are only interested in DMs that can be interpreted as fillers and their potential cleanup does not lead to loss of "important" information for the reader.

MDE annotation also recognizes a special case of DMs – Discourse Response (DR). These are DMs employed to express an active response to what another speaker said, in addition to mark the structure of the discourse. For instance, a speaker may also initiate his/her attempt to take the floor. DRs typically occur turn-initially. However, DRs should not be confused with direct answers to questions. An example of a DR follows. It is presented in Czech as well as in its English translation. DRs are typed in boldface.

A: Já bych to tak udělal /.
B: **Hele** já si tím nejsem tak jistej /.
—
A: I'd do it that way /.
B: **Look** I'm not that sure about it /.

The bottom part of Table 1 shows numbers of words labeled as DMs or DRs. As with FPs, DMs are more common in conversational speech – just 0.1 % of words is labeled as DMs or DRs in our BN data. English MDE data contain more DMs – 4.4 % in CTS and 0.5 % in BN speech.

Another interesting statistic is the proportion of DMs and DRs in the Czech corpora. While in the RF corpus "non-DR" DMs prevail, DR subtype takes up over 53 % of all DMs in the Czech BN corpus. DMs are not frequently used by anchors in the studio, but rather by local reporters referring on actual events directly from their venues. These reporters typically react on questions coming from the studio and their interactive replies contain a number of DRs.

Overall, the most frequent DMs were *tak (*lit. *so)* – 17.0% of all DMs, *no (well)* – 13.2%, and *prostě (simply)* – 12.9%. The DM *tak* was the most frequent in both corpora (but more dominant in the BN corpus – 32.8% vs. 15.8%), while *no* came second in the RF corpus, and *prostě* came second in the BN corpus. Significant differences are also noticeable in DMs as *prostě (simply)*, *vlastně (actually)*, and *jaksi (somehow)* which are more frequent in the RF corpus. The reason is that DMs of the DR subtype prevail in the BN corpus, while the three mentioned DMs typically occur turn-internally.

Another interesting observation is that DMs containing a verb are much less frequent in Czech than in English. Although there exist some Czech equivalents of the common English DM *you know* (such as *víte* or *víte co*), only a minority of speakers use them. Note that all frequent Czech DMs consist of just one word.

4.1.3. Asides/Parentheticals

A/Ps occur when a talker utters a short side comment and then returns to the major 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. The two following transcriptions show examples of A/Ps displayed within curly braces.

The first example shows an aside:

A potom k němu přišel {moment musím si vypnout telefon} s tím velkým psem /.
—
And then he came to him {moment I must switch off my cell phone} with the big dog /.

The second example illustrates a parenthetical:

Občas se stane /, že člověk {nemělo by se to tedy stávat často} udělá chybu /.
—
Sometimes it happens /, that a man {well it shouldn't happen often} makes a mistake /.

The data indicate that A/Ps are on average approximately one word longer in Czech conversational speech than in BN speech (6.2 vs. 5.4 words). Furthermore, A/Ps are relatively frequent in the RF corpus (1.5 % of all words), but quite rare in BN speech (0.2 %). For comparison, the English CTS data contain only 0.3% of A/Ps, which supports the hypothesis that A/Ps are more frequent in conversational Czech than in conversational English. We believe that this is not only caused by differences in the situation of the speakers (political debates generally use a more complex language than telephone conversations), but also by the distinct nature of either language. Czechs like to verbalize thoughts intricately, while the typical English style is more straight.

4.1.4. Explicit Editing Terms

EETs only occur accompanying edit disfluencies. They are used by the speakers to signal that they are aware of the existence of a disfluency on their part. Basically, EETs can appear anywhere within the disfluency. The most common place of occurrence is right after the corrected part, but they can also occur, among others, after the correction. The following example shows a disfluent utterance containing an EET. The EET is displayed in boldface, other parts of the annotated disfluency will be explained in Section 4.2. below.

Tohle je naše [koherentní]* *EE* **nebo** <u>konzistentní</u> stanovisko /.
—
This is our [coherent]* *uh* **or** <u>consistent</u> statement /.

As expected, EETs were really rare. In total, they include just 0.08 % of words in the RF data, and 0.01 % in the BN data. This is similar to English, where EETs represent 0.05 % and 0.02 % of words, respectively. For both Czech corpora, we observed that the most frequent EET was *nebo (or)* (more than 66 % of all EETs). The second most frequent EET was *respektive ('or more precisely')* (3.7 %). Since average lengths of EETs are 1.2 and 1.1 words, respectively, it is possible to ratiocinate that one word EETs are strongly dominant in Czech.

	RF	BN
% of words followed by Edit IPs	2.0%	0.2%
% of words within DelRegs	2.8%	0.3%
% of DelRegs having Correction	83.8%	94.6%
% of DelRegs having EET	4.0%	3.5%
Avg. length of DelRegs (in words)	1.6	1.4
Avg. length of Corrections (in words)	1.6	1.5

Table 2: Statistics on edit disfluencies

4.2. Edit Disfluencies

Edit disfluencies are portions of speech in which a speaker's utterance is not complete and fluent. An 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), optional EET, and *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 spoken Czech. Their labeling is not extremely time consuming and the obtained data may be very useful.

For the illustration of an edit disfluency, we can use the example shown in Section 4.1.4. The following notation is used: DelRegs are displayed within square brackets, IPs are marked by *, EETs are typed in boldface, and corrections are underlined. The example presents a disfluency with a single IP, however, 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 explicitly tagging individual IPs. Since the MDE standard does not allow using nested disfluencies, all such cases are annotated using serial, non-nested DelRegs with multiple IPs.

4.2.1. Basic Statistics about Edit Disfluencies

Some statistics relating to edit disfluencies are presented in Table 2. As with fillers, we observed that disfluencies were much more frequent in the spontaneous corpus, where 2.8 % of words were labeled as within a DelReg. The percentage in Czech BN speech was just 0.2 %. Likewise, edit IPs were ten times more frequent in the RF corpus than in the BN corpus. In comparable English corpora, edit disfluencies were more frequent. DelRegs covered 5.4 % of words in the CTS, and 1.5 % in the BN data. Again, the explanation for this is in different speaking styles.

Another interesting numbers refer to occurrences of corrections and EETs within edit disfluencies. The proportion of DelRegs having a correction is dependent on the frequency of restart disfluencies, since this type of disfluency does not contain a correction. Since restarts are typical for spontaneous speech, the relative number of DelRegs having corrections is smaller for the RF corpus.

As expected, EETs were very rare in both corpora. Only approximately 4 % of all disfluencies contained an EET. It was also observed that short disfluencies predominated; the average length was around 1.5 words in both corpora. Another interesting observation was that corrections had almost the same average length as DelRegs.

Additional notable statistics are those referring to the total portion of data marked for the potential automatic cleanup. This number correlates with the complexity of the MDE task for a particular corpus. Hence, we summed words in DelRegs and fillers and compared the two Czech MDE corpora. As expected, the results were largely unequal – 9.8 % for the RF corpus and 1.1 % for the BN corpus. For comparison, in English MDE it was 17.7 % for the CTS and 3.8 % for the BN corpus.

4.2.2. POS Statistics of Edit Disfluencies

We have also analyzed DelRegs and corrections in terms of which parts of speech (POS) they contain. To our best knowledge, this is the first work studying the POS content of speech disfluencies in any language. Both Czech corpora were tagged using a state-of-the-art automatic morphological tagger based on the averaged perceptron (Spoustová et al., 2007). For Czech, we use a positional tagset where every tag is represented as a string of 15 symbols representing individual morphological categories. We only utilized the first position corresponding to the POS information. For either Czech corpus, we computed three POS distributions corresponding to the whole corpus, DelRegs, and corrections, respectively.

The relative frequencies of particular POSs are shown in Figure 1. The top chart represents the RF corpus, the bottom chart the BN corpus. The POS labels on the *x*-axis are sorted according to their relative frequencies in the RF corpus. The blue bars show that the corpora differ in overall POS distributions. The BN corpus contains significantly more nouns and adjectives, while the conversational corpus shows distinctively higher relative numbers of pronouns, adverbs, and conjunctions, and a slightly higher proportion of verbs. These observations may be clarified by differences in speaking styles. Broadcast news data consist of sentences that were prepared to be as informative as possible, and thus they contain a lot of nouns and adjectives. On the other hand, conversational language is characterized by a more complex way of locution. A great deal of complex and compound sentences logically implies a higher number of conjunctions, while numerous discourse markers having the form of adverbs cause the higher proportion of that POS type.

Despite the differences in overall POS distribution, both corpora show similar changes in this distribution when only the words in DelRegs are taken into account. The proportion of nouns, pronouns, and prepositions is increased, while verbs, adverbs, and adjectives are less frequent. The increased number of nouns in DelRegs can be explained by the fact, that disfluencies more frequently occur in more informative regions of utterances and nouns usually carry information more densely than, for instance, verbs. A higher frequency of prepositions is consequent, since prepositions are dependent on nouns.

Interesting variances may also be observed between DelRegs and their corrections. The most prominent difference is in the higher rate of adjectives within corrections. This fact indicate that speakers often put in the adjectives omitted in DelRegs. We can also observe slightly higher numbers of verbs and adverbs. On the other hand, nouns, con-

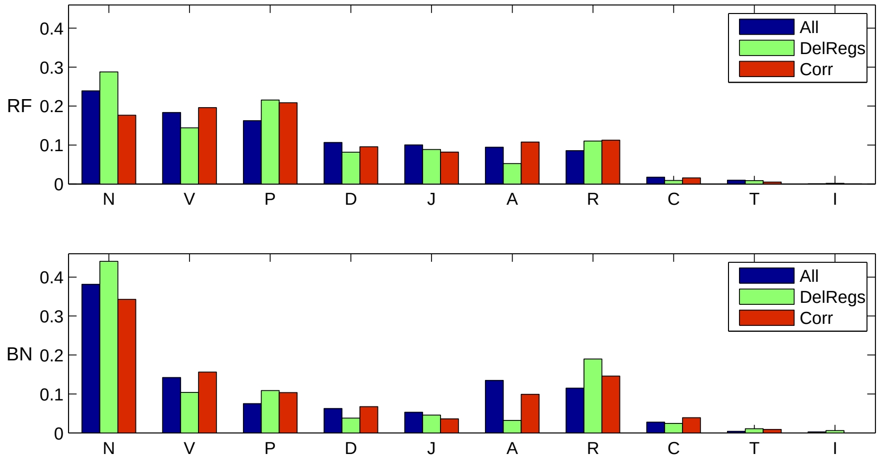

Figure 1: Relative frequencies of POS in all data, DelRegs, and corrections in both Czech corpora. The RF corpus is displayed in the top chart and the BN corpus in the bottom. (POS legend: N – Nouns, V – Verbs, P – Pronouns, D – Adverbs, J – Conjunctions, A – Adjectives, R – Prepositions, C – Numerals, T – Particles, I – Interjections)

junctions, and prepositions are less common than in DelRegs.

4.3. SUs

Dividing the continuous 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 usability by presenting transcribed text in small coherent chunks rather than long unstructured turns or stories. The MDE standard segments the flow of speech into utterance units called SUs (Syntactic/Semantic Units) that are classified according to their function within the discourse (Strassel, 2004). Because talkers often tend to use long continuous compound "sentences" in their speech, it is nearly impossible to identify the end-of-sentence boundary with consistency using only prosodic information. Thus, SUs divide the flow of speech into "minimal meaningful units" functioning to express one complete idea on the speaker's part.

SU symbols may be divided into two categories: sentence-internal (clausal and coordination breaks) and sentence-external (others). Sentence-external breaks are fundamental and indicate the presence of a main (independent) clause. Sentence-internal breaks are secondary; they signal 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. The following list shows all used SU symbols (breaks) along with brief descriptions of their function:

- /. – 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 there /,
  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 /.)

Besides other modifications, Czech MDE extended the original set of SU symbols by "//." and "//?". The double slashes indicate a strong prosodic marking on the SU boundary, as explained in detail in (Kolář et al., 2005).

Relative frequencies of all SU symbols are presented in Table 3. The data indicate that both corpora significantly differ in SU distributions. The symbol "/," is most frequent in the RF corpus, whereas "//." is strongly dominant in the BN corpus. This observation shows that complex and compound sentences are more common in spontaneous conversations, while prearranged broadcast news typically consists of statements with simpler syntax.

Further findings relate to differences between relative frequencies of single- and double-slash SU symbols. Overall, the double slash symbols are more frequent. The contrast is more distinctive in the BN data, where "//." is almost ten times more frequent than "/." The reason is that sentence boundaries are usually attentively prosodically marked by professional newscasters. The statistics regarding incomplete SUs indicate that incompletes are much more common in conversational speech. Furthermore, incomplete SUs interrupted by another speaker (/∼) are more frequent than the arbitrarily abandoned statements (/–). The latter type almost never appears in broadcast news.

The last group of SU statistics reflects average lengths of SUs. Overall, SUs in the RF corpus are longer than in the

Symbol	RF	BN	Symbol	RF	BN
/.	15.1%	6.7%	/-	0.4%	0.0%
//.	28.9%	**60.2%**	/∼	2.7%	0.6%
/?	0.7%	0.3%	/&	6.7%	2.9%
//?	3.4%	1.3%	/,	**42.2%**	28.1%

Table 3: Relative frequencies of particular SU symbols (both SU-internal and SU-external)

SU type	RF	BN	SU type	RF	BN
/.	12.6	7.8	//?	12.6	9.9
//.	16.1	13.7	/–	15.2	11.3
/?	11.5	8.7	/∼	11.4	10.2

Table 4: Average lengths (in words) of particular SU subtypes in both Czech corpora

BN corpus (14.5 vs. 13.0 words). In English, the CTS corpus has mean SU length 7.0 words and the BN corpus 12.5. The shorter length of segments in the CTS data can be explained by a large number of short answers and backchannels present in telephone conversations.

Table 4 reports average lengths of particular SU types in Czech data. The numbers indicate that statement SUs are longer than interrogative SUs. Further, double slash SUs having a strong prosodic marking are significantly longer than their one slash counterparts. This difference in length is more prominent in the BN data.

5. Summary

We have presented a comparison of Czech broadcast news and broadcast conversations in terms of MDE statistics relating to fillers, edit disfluencies, and SUs. The comparison can be used to evaluate the complexity of the MDE task in particular genres. Moreover, it provides interesting data for linguistic analyses of speech in the two domains. We have not only compared the two Czech corpora with each other, but also with the available numbers relating to English MDE corpora.

Among others, we have shown that the total proportion of filler words (i.e., the sum of all FPs, DMs, A/Ps, and EETs) is significantly higher in the RF corpus (6.97 % of words) than in the BN corpus (0.79 %). Likewise, edit disfluencies are much more frequent in the RF corpus (2.8 % of words within DelRegs in the RF and 0.2 % in the BN). We have also found that DelRegs and their corrections show differences in POS distributions in comparison with the general POS distribution. Regarding SU symbols, we have observed that clausal breaks are more frequent in the RF corpus which indicates that complex sentences are more common in talk shows than broadcast news. Furthermore, we have observed that SUs in conversational data are on average longer by 1.5 words.

Finally, we should mention that both described MDE corpora are planned to be made publicly available in the near future. Moreover, the RF corpus is currently being extended by 20 additional recordings.

6. Acknowledgments

This work was supported by the Ministry of Education of the Czech Republic under projects 2C06020 and ME909. We also thank Stephanie Strassel and Christopher Walker from the LDC for helping us understand all the subtle details of the MDE annotation, and Dagmar Kozlíková for her advice on detailed aspects of Czech syntax.

7. References

Anton Batliner, Ralf Kompe, Andreas Kiessling, Marion Mast, Heinrich Niemann, and Elmar Nöth. 1998. M = Syntax + Prosody: A syntactic–prosodic labelling scheme for large spontaneous speech databases. *Speech Communication*, 25:193–222.

Jean Carletta. 1996. Assessing agreement on annotation tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Peter Heeman. 1997. *Speech Repairs, Intonational Boundaries and Discourse Markers: Modeling Speakers' Utterances in Spoken Dialogs*. Ph.D. thesis, University of Rochester, New York.

Jáchym Kolář, Jan Švec, Stephanie Strassel, Christopher Walker, Dagmar Kozlíková, and Josef Psutka. 2005. Czech spontaneous speech corpus with structural metadata. In *Proc. Interspeech'05*, Lisbon, Portugal.

Jáchym Kolář. 2008. *Automatic Segmentation of Speech into Sentence-like Units*. Ph.D. thesis, University of West Bohemia, Pilsen, Czech Republic.

Yang Liu, Elizabeth Shriberg, Andreas Stolcke, Dustin Hillard, Mari Ostendorf, and Mary Harper. 2006. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on audio, speech, and language processing*, 14(5):1526–1540.

Marie Meeter. 1995. Dysfluency annotation stylebook for the Switchboard corpus. `ftp://ftp.cis.upenn.edu/pub/treebank-/swbd/doc/DFL-book.ps`.

Josef Psutka, Vlasta Radová, Luděk Müller, Jindřich Matoušek, Pavel Ircing, and David Graff. 2001. Large broadcast news and read speech corpora of spoken Czech. In *Proceedings of EUROSPEECH*, pages 2067–2070, Aalborg, Denmark. ISCA.

Vlasta Radová, Josef Psutka, Luděk Müller, William Byrne, Josef V. Psutka, Pavel Ircing, and Jindřich Matoušek. 2004. Czech Broadcast News Speech and Transcripts. Linguistic Data Consortium, CD-ROM LDC2004S01 and LDC2004T01, Philadelphia, PA, USA.

Drahomíra Spoustová, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. 2007. The best of two worlds: Cooperation of statistical and rule-based taggers for Czech. In *Proc. of the ACL Workshop on Balto-Slavonic Natural Language Processing*, Prague, Czech Republic.

Stephanie Strassel, Jáchym Kolář, Zhiyi Song, Leila Barclay, and Meghan Glenn. 2005. Structural metadata annotation: Moving beyond English. In *Proc. Interspeech'05*, Lisbon, Portugal.

Stephanie Strassel. 2004. Simple metadata annotation specification V6.2. `http://www.ldc.upenn.edu/Projects/MDE/Guidelines/SimpleMDE_V6.2.pdf`.