Czech Broadcast Conversation Corpus

–Data Format Description–

The speech transcripts from the Czech Broadcast Conversation Corpus are stored in three different formats – TRS (Transcriber), QAn (Quick Annotator), and RTTM. The first one only represents standard speech transcripts for training and evaluation of automatic Speech-To-Text (STT) systems, while the other two formats also contain information about structural metadata (MDE).

We use the following file naming convention. All file names have the form "rfYYMMDD.format" where "rf" stands for Radioforum (name of the broadcast program), the following six digits indicate the date of broadcast, and the extension "format" corresponds to the data format of the particular file – "trs", "qan", or "rttm". Character encoding in all files is ISO-8859-2.

TRS (Transcriber) Format

The first format, TRS, is an XML-based format used by the well-known speech annotation tool Transcriber (http://trans.sourceforge.net/). The transcripts from the Czech Broadcast Conversation Corpus were created in Transcriber V1.4.1 (using DTD file "trans-13.dtd") but can be opened in newer versions of Transcriber as well.

The transcripts contain standard punctuation, but acceptable punctuation is limited to periods and question marks at the end of a sentence, and commas within a sentence. Capitalization is used for proper names but not at the beginnings of sentences. In addition to full words and punctuation, the TRS transcripts contain the following special events:

Table 1 Special events used in our ".trs" format

Event	Description
Speaker Noises:	
BREATH	Audible breaths
COUGH	Audible cough
LAUGH	Laughter
LIP-SMACK	Lip smack or tongue click
Other Noises:	
NOISE	Unspecified noise
BACKGROUND_SPEECH	Background or remote speech from other speakers
MUSIC	Music and jingles
Lexemes:	
EE-HESITATION	Filled pauses resembling Czech *é* or other vowel-like sounds
MM-HESITATION	Filled pauses resembling Czech *mm* or other consonant-like sounds
HM	Interjection expressing agreement

MH	Interjection expressing disagreement
UNINTELLIGIBLE	Unintelligible word

Note that the list of special events also includes some special lexemes. This setting is used only to make them more visible in the transcript.

Word fragments are tagged with a leading or trailing hyphen (e.g., fragments of *word* might be *–ord* or *wor–*).

Mispronounced words are marked by leading and trailing asterisks (e.g., when **lomocotive** is pronounced instead of *locomotive).*

QAn (Quick Annotator) Format

This format is the native format of the QAn (Quick Annotator) tool that was used to metadata annotate the corpus. It is based on the above described TRS format, which is extended by some special tags representing structural metadata.

The format uses two types of metadata tags: *SUs*, that are associated with interword boundaries, and *Labels*, that can span over one or more words. Thus, *Label* tags have the form of begin/end pairs, while SUs are only single tags.

The SU tags always start with "<mde:SU". Then, the tags have a mandatory attribute "type". The following types are used:

(a) SU-external symbols:

- "/." – Statement break *without* strong prosodic marking at boundary
- "//." – Statement break *with* strong prosodic marking at boundary
- "/?" – Question break *without* strong prosodic marking at boundary
- "//?" – Question break *with* strong prosodic marking at boundary
- "/-" – End of an incomplete (arbitrarily abandoned) SU
- "/~" – End of an incomplete SU interrupted by another speaker

(b) SU-internal symbols:

- "/," – Clausal break
- "/&" – Coordination break

(c) Interruption point symbol[1]:

- " * " – Interruption point within an edit disfluency (asterisk)

[1] Note that Interruption Points are included with SU tags only for the sake of format simplicity because same as SU boundary symbols, they are associated with interword boundaries. Strictly speaking, they should be in a separate group, but we did not want to introduce another group that would only include a single tag category.

In addition to the mandatory attribute, the SU tags may also contain the optional attribute "previous". This attribute indicates that the SU tag replaced a standard punctuation symbol (such as period or comma). This information is especially important for the annotation tool. If the annotator decides to delete an SU tag, the tool can display the original punctuation symbol again. For example, a tag with the "previous" attribute may be <mde:SU type="/." prev=","/>.

Furthermore, Interruption point tags may also receive the attribute "auto". If the tag looks like <mde:SU type="*" auto="1"/>, it indicates that the Interruption point tag was inserted automatically by the annotation tool at the right edge of the preceding Delreg.[2]

Labels

Label-type tags start with "<mde:Label". All Label tags have two mandatory attributes – "type" and "extent". The attribute extent may have two values – "begin" and "end", to indicate tag pairs. The attributes "type" may have the following values:

- "A/P" – Aside/Parenthetical
- "Backchannel" – Backchannel uttered by other speaker than
- "Correction" – Correction of previous Delreg
- "DM" – Discourse marker
- "DR" – Discourse marker of subtype "Discourse response"
- "Delreg" – Deletable region
- "EET" – Explicit editing term
- "FP" – Filled pause

An example of a Label tag is "<mde:Label type="DM" extent="begin" />". The following text serves as an example of a metadata annotated speech transcript in the ".qan" format:

```
<Event desc="BREATH" type="noise" extent="instantaneous"/> to
<mde:Label type="Delreg" extent="begin" /> bylo <mde:Label type="Delreg" extent="end" />
<mde:SU type="*" auto="1"/> <mde:Label type="Correction" extent="begin" /> bylo
<mde:Label type="Correction" extent="end" /> velkou zkouškou vládnoucí strany <mde:SU
type="/," prev=","/> protože
<Event desc="BREATH" type="noise" extent="instantaneous"/>
prakticky od roku devadesát šest v České republice neexistuje většinová vláda
<mde:SU type="//." prev=","/>
```

Note that the segments with overlapping speech from more than one speaker are not annotated for MDE.

RTTM Format

The RTTM format also provides information about structural metadata that enrich standard speech transcripts. The format described herein is based on the RTTM-format-v13 used for MDE in the EARS project. The original RTTM format could not be used in the exact form employed in the EARS project because annotation modifications[3] introduced in the Czech MDE annotation project

[2] Note that unlike English, Czech MDE does not use automatic interruption points before fillers.
[3] The modifications are described in the document "Structural Metadata Annotation for Czech: An Overview" that is also included in the corpus documentation.

had to be reflected. Note that the published RTTM files only contain description of those regions of data that have MDE annotation (i.e., sections with overlapping speech are not present in RTTMs).

The format uses object-oriented representation of the rich text data. There are four general object categories to be represented. They are STT objects, MDE objects, source (speaker) objects, and structural objects. Each of these general categories may be represented by one or more types and subtypes, as shown in Table 1. Note that the object subtypes that are generally allowed but do not appear in this corpus are marked with asterisks.

Table 2 Rich Text object types and subtypes

Type	Subtypes
Structural types:	
SEGMENT	<NA>
STT types:	
LEXEME	**lex**, **fp**, **frag**, **un-lex**[4], **interjection**, **mispronounced,** and **other***
NON-LEX	**laugh**, **breath**, **lip-smack**, **cough**, and **other***
NON-SPEECH	**noise**, **music**, **background_speech,** and **other***
MDE types:	
FILLER	**discourse_marker, discourse_response**[5], **explicit_editing_term, backchannel,** and **other***
EDIT	<NA>
CORRECTION	<NA>
IP	**edit**, **filler***, **edit&filler***, and **other***
SU	***/.*** ***//.*** ***/?*** ***//?*** ***/~*** ***/–*** [6]
CB	**coordinating**, **clausal**, and **other***
A/P	(none)
SPEAKER	(none)
Source information:	
SPKR-INFO	**adult_male**, **adult_female**, **child***, and **unknown***

Except for the static speaker information object [**SPKR-INFO**], each object exhibits a temporal extent with a beginning time and duration. (The duration of interruption points [**IP**] and clausal boundaries [**CB**] is zero by definition.)

[4] Un-lex is used to tag unintelligible words.
[5] By definition, discourse_response is a subtype of discourse_marker. They are listed on the same level herein only because the RTTM format does not allow to define "subsubtypes".
[6] Since there are more SU subtypes in Czech MDE than in the original standard, we rather use a symbolic instead of word representation of SU subtypes for the sake of simplicity.

These objects are represented individually, one object per record, using a flat record format with object attributes stored in white-space separated fields. The format is shown in Table 2.

Table 3 Object record format for RTTM objects

Field 1	2	3	4	5	6	7	8	9
type	file	chnl	Tbeg	tdur	ortho	stype	name	conf

where

`file` is the waveform file base name (i.e., without path names or extensions).

`chnl` is the waveform channel (e.g., "**1**" or "**2**").

`tbeg` is the beginning time of the object, in seconds, measured from the start time of the file.[7] If there is no beginning time, use `tbeg` = "**<NA>**".

`tdur` is the duration of the object, in seconds.[4] If there is no duration, use `tdur` = "**<NA>**".

`stype` is the subtype of the object. If there is no subtype, use `stype` = "**<NA>**".

`ortho` is the orthographic rendering (spelling) of the object for STT object types. If there is no orthographic representation, use `ortho` = "**<NA>**".

`name` is the name of the speaker. `name` must uniquely specify the speaker within the scope of the `file`. If `name` is not applicable or if no claim is being made as to the identity of the speaker, use `name` = "**<NA>**".

`conf` is the confidence (probability) that the object information is correct. If `conf` is not available, use `conf` = "**<NA>**".

This format, when specialized for the various object types, results in the different field patterns shown in table 3.

Table 4 Format specialization for specific object types

Field 1	2	3	4	5	6	7	8	9
Type	*file*	*chnl*	*tbeg*	*tdur*	*ortho*	*stype*	*name*	*conf*
SEGMENT	`file`	`chnl`	`tbeg`	`tdur`	**<NA>**	**eval** or **<NA>**	`name` or **<NA>**	`conf` or **<NA>**
LEXEME **NON-LEX**	`file`	`chnl`	`tbeg`	`tdur`	`ortho` or **<NA>**	`stype`	`name`	`conf` or **<NA>**
NON-SPEECH	`file`	`chnl`	`tbeg`	`tdur`	**<NA>**	`stype`	**<NA>**	`conf` or **<NA>**
FILLER **EDIT** **CORRECTION** **SU**	`file`	`chnl`	`tbeg`	`tdur`	**<NA>**	`stype`	`name`	`conf` or **<NA>**
IP **CB**	`file`	`chnl`	`tbeg`	**<NA>**	**<NA>**	`stype`	`name`	`conf` or **<NA>**

[7] If `tbeg` and `tdur` are "fake" times that serve only to synchronize events in time and that do not represent actual times, then these times are tagged with a trailing asterisk (e.g., `tbeg` = **12.34*** rather than **12.34**).

A/P SPEAKER	file	chnl	tbeg	tdur	**<NA>**	**<NA>**	name	conf or **<NA>**
SPKR-INFO	file	chnl	**<NA>**	**<NA>**	**<NA>**	stype	name	conf or **<NA>**

The following table shows mapping between the QAn and the RTTM format for the events that are defined in both formats.

Table 5 Mapping between QAn and RTTM annotation

QAn Type	RTTM Type	RTTM Subtype
/.	**SU**	**/.**
//.	**SU**	**//.**
/?	**SU**	**/?**
//?	**SU**	**//?**
/~	**SU**	**/~**
/–	**SU**	**/–**
/,	**CB**	**clausal**
/&	**CB**	**coordinating**
*****	**IP**	**edit**
A/P	**A/P**	**<NA>**
Backchannel	**FILLER**	**backchannel**
Correction	**CORRECTION**	**<NA>**
DM	**FILLER**	**discourse_marker**
DR	**FILLER**	**discourse_response**
Delreg	**EDIT**	**<NA>**
EET	**FILLER**	**explicit_editing_term**
FP	**LEXEME**	**fp**