An Annotation Scheme for Speech Reconstruction on a Dialog Corpus

Silvie Cinková[1], **Jan Hajič**[1], **Jan Ptáček**[1]

Abstract. This[1] paper presents the ongoing manual speech reconstruction annotation of the NAP corpus, which is a corpus of recorded conversations between pairs of people above family photographs, relating it to a more complex annotation scheme of the Prague Dependency Treebank family. The result of this effort will be a resource that will contain, on top of the audio recording of the dialog and its usual transcription, an edited and fully grammatical "reconstructed" dialog. The format and alignment with the original audio and transcription on one side and a similar alignment (linking) to a deep analysis of the natural language sentences uttered in the dialog on the other side will be such that the resource can serve as a training and testing material for machine learning experiments in both intelligent editing as well as in dialog language understanding. The resource will be used in the Companions project, but it will be publicly available outside of the project as well.

1 INTRODUCTION

The goal of the work described in this paper is to manually build gold-standard data for machine-learning tasks that involve automatic recognition of spontaneous speech and its "understanding" in a dialog system setting. So far, the overall performance of NLP systems that rely on ASR has been negatively affected by the fact that even the best possible ASR output is still hardly tractable for language-analysis tools, such as POS taggers, lemmatizers, parsers and semantic analyzers. These tools have been designed for written texts, and they cannot cope with the morphological and syntactic irregularities typical of spontaneous speech.

While the creation of rule-based language analysis tools specifically adapted to spontaneous speech seems difficult if not impossible (due to the unpredictability of the various speech artifacts in the much too faithful transcription as output by today's ASR systems), the employment of statistical machine-learning methods for automatic smoothing of the ASR output into a standard written text appears to be a challenging but a promising way to go.

The manual annotation of data for *speech reconstruction* will enable future machine learning experiments in various settings in order to either obtain grammatical sentences for further "classical" NL processing, or to "understand" them (i.e., to obtain their formal representation) directly.

This paper is a work-in-progress report on the acquisition, preparation, and manual annotation of the data. No statistical experiments have been performed yet, as the data is still sparse. Hence no quantitative evaluation can be given.

2 SPEECH RECONSTRUCTION IN THE COMPANIONS PROJECT

The NLP-part of the Companions project (www.companions-project.org), within which the speech-reconstruction effort is taking place, is supposed to create an NLP component of the whole dialog system that can analyze as well as generate text. The dialog should have a conversational, rather than task-oriented, character. Two domains have been selected for a demo system:

1. Health & Fitness Companion: the system assists the user in planning, pursuing, and reflecting a healthier way of living by evaluating the user's description of how the user spent the previous day (mainly with respect to diet and motion) and making suggestions for the current (or following) day.
2. Photo Companion: the system helps the user with browsing, tagging, and sorting of digital photographs. It also encourages the user to comment on the respective pictures. The research plan of the Companions project calls for heavy use of machine-learning. Therefore, (annotated) data is important, and an English and a Czech corpus of spoken dialogs for the two domains are being created. In this paper we will only describe the work on the corpus for the photo domain (called NAP after the Napier University in Edinburgh, where this corpus has been built), and we will refer to the annotation performed at the Charles University (CU), though the CU effort is just one of the alternatives[2]. However, the speech-reconstruction data seems to be universally useful, independently of the exact platform and NL tools used in/for the demo systems.

3 THE NAP CORPUS

The NAP corpus [1] currently consists of approx. 70 recording hours. Sixteen recording hours have been manually transcribed, out of which more than 140,000 tokens (in approx. 11,300 utterances) have been manually annotated with speech reconstruction. The dialogs were originally designed as conversations between a human and a (simulated) robot appearing on a computer screen in a lab, but later on the design

[1] Institute of Formal and Applied Linguistics, Computer Science School, Faculty of Mathematics and Physics, Charles University in Prague. Malostranske nam. 25, CZ-11800 Prague 1, Czech Republic. Email: {cinkova,hajic,ptacek}@ufal.mff.cuni.cz

[2] The other Companions annotation schemes are the procedures gathered within GATE [2] (automatic POS-tagging and lemmatization, shallow parsing, semantic labeling for information extraction, basic coreference, and named-entity annotation), dialog act annotation, and knowledge representation ([3] and others).

has been switched to – apparently more spontaneous – conversations between two humans in a relaxed environment (parks, restaurants, the respondents' homes, etc.). The material is rich in short turns but contains even longer narrative monologues as well as a pure social interaction.

4 ANNOTATION – STRATIFICATION AND FORMAT

The speech reconstruction is primarily understood as part of data preparation for deep parsing in the style of the *Prague Dependency Treebank* (PDT 2.0, [4]). The data is stored in a modified XML format called PML (*Prague Markup Language*, [5]), which is designed to capture and interlink all layers of the PDT-like annotation scheme. The annotation scheme stratifies the data into the following annotation layers (listed in ascending order; for illustration see Figure 1):

1. z-layer ("zero layer"): the output from automatic ASR in spoken data (not used with the NAP data). No markup is added; essentially, it is a single-cased stream of tokens, possibly consisting of the recognized non-speech events, such as pauses, possibly also coughs, laugh, sobbing or other emotional aspect (if provided by the recognizer).
2. w-layer ("word layer"): the manual literal transcription of the acoustic signal in spoken data (or the linear text in written data); each word obtains its unique ID during tokenization and it is regarded as a (w-)node. The raw manual transcripts of speech data also contain acoustic segmenting and synchronization with the original audio file. It is expected that this layer contains also a wide range of non-speech events, as introduced into the transcription by the human transcribers using the usual conventions.
3. m-layer ("morphological layer"): linear text in which each token acquires its unique ID and is regarded as an m-node. The m-nodes are linked to their corresponding w-nodes. The m-layer further provides sentence chunking, POS-tagging, and lemmatization. It is *on this layer that the manual speech reconstruction annotation takes place* (if the w-layer input is a speech transcript, not an - originally - written text). Manual sentence chunking is part of the annotation process. The POS tagging and lemmatization are performed, at the moment, by automatic tools, but will be later corrected by human annotators as well.
4. a-layer ("analytical layer") stands for the dependency-based shallow syntactic parsing. Each a-node has its unique ID and usually contains a single reference to the m-layer.
5. t-layer ("tectogrammatical layer") is the topmost and most abstract annotation layer within the scheme. Shaped as a transition between syntax and semantics, it reflects the underlying syntax ("deep grammar") of each sentence. Each sentence is represented as a dependency-based projective tree with nodes and edges. Only autosemantic words are represented as t-nodes, with references leading to all function words from the a-layer that affect the meaning of the given t-node in the given context. Rich semantic labeling and coreference annotation, as well as ellipsis restoration, are part of the tectogrammatical representation (annotation).

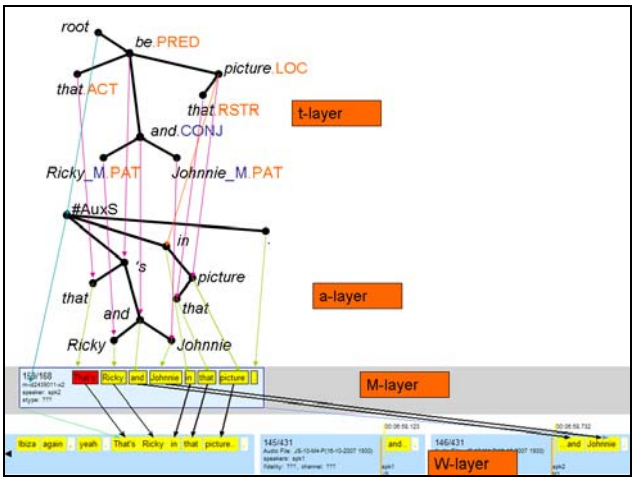

Figure 1. The PDT-style annotation layers

The raw manual transcriptions (i.e., the source for the w-layer annotation) are created by the Napier University team using the Transcriber annotation tool ([6], as adapted by Ircing [7]). Some basic non-speech events like laugh, cough, hesitation, etc., are preserved in the running text by means of special tags. Speaker identification is provided, along with acoustic segmenting and synchronization with the input audio file. These transcripts are converted into PML to become the w-layer data. The w-layer and the underlying audio are then loaded into MEd, the annotation tool for manual speech reconstruction. The conversion scripts that prepare the input for MEd automatically create and pre-annotate the m-layer by creating m-nodes from all w-nodes that are not tagged as non-speech events and m-segments from all w-segments, including reference arrows from the m-nodes to the corresponding w-nodes. They even take care of the tokenization by chopping all contracted forms into separate m-nodes (e.g., *"don't" -> "do" "n't"*). The annotator is supposed to manually check and correct the m-layer annotation in the MEd editor ([8]; see Figure 2).

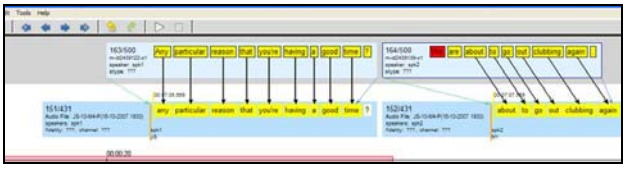

Figure 2. A MEd window with a sychronized audio track (bottom), the original transcription (middle), and the m-layer to be edited (the uppermost horizontal section).

5 BASIC ANNOTATION PRINCIPLES

The annotation of the m-layer resembles editing the transcription of an interview recording in order to be printable in a respectable journal or daily. The output must not only be intelligible but also grammatically correct and easy-to-read, which means cleansing the text from all speech-specific phenomena, such as disfluencies, incorrect word order, discourse-irrelevant non-speech events, and slips of the tongue. Besides, the annotator should make sure that:

- Only standard and orthographically correct variants of words are used.
- The punctuation is appropriate and consistent throughout the entire text.
- The capitalization conventions are applied in an appropriate way.

When performing the speech reconstruction editing, the annotator is supposed to follow two basic annotation principles:

1. **The Content-Preservation Principle:** the modifications of the original speech segments may not affect the content, or only minimally in case of uncertainties.
2. **The Minimal Modification Principle:** modifications are performed only when necessary to achieve written-text standard in the resulting text.

An m-segment is supposed to correspond to exactly one sentence that meets written-text standards. The annotator is allowed to merge or split the pre-generated m-layer segments to produce a good sentence, to move the indicators of m-segment start/end and to edit all nodes on the m-layer line. The annotator thus "smoothes" each sentence to meet written-text standards by means of the following modifications:

1. deletion
2. insertion (incl. punctuation)
3. substitution
4. word order change.

The tool does not allow the annotator to edit the original transcript on the w-layer, but he/she can make any suggestion of correction of the w-layer (the original transcript) by adding a comment field to the corresponding m-layer node.

Deletion

Some phenomena typical of spontaneous speech are systematically removed during the speech reconstruction. It is mainly:

1. discourse-irrelevant non-speech events
2. filler words and filler phrases
3. superfluous function words
4. reparandums and interregnums in restarts
5. repetitions (except rephrasing)
6. abandoned fragments.

Non-speech events like laughter, sobs, lip smacks as well as hesitation, agreement and disagreement *uh-huh* and *hmm* sounds are highlighted with special tags already in the manual transcription. During the conversion of the manual transcription into PML, they too become w-nodes. The conversion script, though, ignores tokens tagged as non-speech events when automatically generating the m-nodes corresponding to the respective w-nodes.

Example:

you can see the EE-HESITATION road in the distance there
⇨ You can see the road in the distance there.

Whenever the annotator recognizes a non-speech event as discourse-relevant, he/she must manually re-insert the m-node corresponding to the non-speech event and make the appropriate reference to the w-layer. This happens most often when the non-speech elements act as backchannels or answers to yes/no questions.

Examples (just m-layer):

Speaker 2:
It 's a picture of me in Ibiza.
Speaker 1:
UH

Speaker 1:
Okay?
Speaker 2:
UH

The same goes for typical fillers (e.g. *yeah, okay*) when they act as backchannels or answers to yes/no questions.

Relativizers such as *sort of* are normally not regarded as fillers even when modifying a verb (e.g. *we sort of went there*), but *like* is, except when introducing direct speech (e.g. *she was like, "shall we leave"*) and in exemplification. The decision is always up to the annotator.

The current convention says that sentences should better not start with coordinating connectives *(and, but, or, so)*. They should preferably be merged with the previous sentence (as long as the resulting sentence is syntactically, semantically, and stylistically acceptable[3]), or the connective should be deleted if it does not clearly put the adjacent sentences into the adversative or consecutive discourse relation *(but, so)*. The superfluous *and* is replaced with serial comma whenever possible *(me and John and Vicky ⇨ me, John, and Vicky)*. The decisions are again up to the annotators.

In restarts, only the new start coherent with the rest of the sentence is preserved. Reparandums (the original starts) and interregnums (editing expressions) are removed. Restarts are typically caused by a slip of the tongue *(wh - why were you in France ⇨ Why were you in France?)*, hesitation repetition *(So you like . . . you like drinking ⇨ You like drinking?)*, *instant correction (three, no wait four... ⇨ four)*. We are not yet quite consistent in the emphasis repetition *(a very very very nice...)*, but we tend not to regard it as a restart, and we decided to preserve it at the m-layer.

An abandoned fragment is a text span (one or several autosemantic words) that remained incomplete and it is not further referred to in the following text. Abandoned fragments are omitted at the m-layer. Fragments are to be held apart from incomplete sentences, which are preserved and end with ellipsis ('...'). We have found no universal criterion for distinguishing between the two of them. The rule of thumb is that:

1. in incomplete sentences, one can assume what sort of information has remained unsaid while in fragments one cannot
2. abandoned fragments are not further referred to in the discourse, unlike incomplete sentences.

[3] Too long clause coordinations are to be avoided. A complex sentence should not contain more than three coordinated clauses.

Insertion

The reconstructed text can contain lexical units that have not been pronounced (they do not occur at the w-layer) but are indispensable to constructing a grammatically as well as lexically correct sentence. Such lexical units are represented each with its own m-node inserted at the m-layer.

The inserted nodes typically stand for:

1. missing function words
2. unexpressed autosemantic words
3. punctuation.

Examples:

Same holiday ⇨ ***The*** *same holiday****.***

this picture you 've got Ahmed , Steve and Ian and myself ⇨ ***In*** *this picture you 've got Ahmed, Steve****,*** *Ian****,*** *and myself****.***

I 've just come out the shower and watching the sunset ⇨ *I 've just come out* ***of*** *the shower and* ***am*** *watching the sunset****.***

Questions are reconstructed in the following way: the verb-subject inversion is not obligatory. "Declarative questions" (like *You like drinking?)* are not to be transformed into regular questions with the verb-subject inversion and dummy-*do* (*Do you like drinking?*). On the other hand, missing subjects and auxiliary verbs are to be reconstructed (resulting in regular questions with the verb-subject inversion: *Want some more?* ⇨ *Do you want some more?*) Inserted auxiliary verbs must agree with the subject.

When a sentence is incomplete and the missing autosemantic word is obvious from the context, its insertion is allowed. The context can be either the verbal context or the knowledge the annotator has about the photographs discussed, or common knowledge (*on the right, let me see if I can remember her name.* ⇨ *On the right* ***is someone****, let me see if I can remember her name)*. However, not all ellipses must be necessarily restored since our annotation rules can handle most types of ellipsis on both syntactic annotation layers (the a- and t-layer).

In order to repair a sentence, e.g., when a part of the audio is unintelligible, the annotator can make small reformulations of the text, preferably by using vague autosemantic words such as *to be, to have, that person*, etc., or by repeating a relevant word used somewhere in the close context. The insertions of autosemantic words are not formalized in any way (cf. [11]). Not all unintelligible sequences are reconstructed. Sentences with 'unintelligible' text spans can occur. When substitution by a deduced text is impossible, the unintelligible text span should be represented by an m-node of type 'nontext' with the value 'unintelligible' in attribute 'type'. The corresponding nodes are again linked by a reference.

The reconstructed text must also have correct punctuation. The English punctuation shows a great deal of flexibility, compared e.g. to Czech, the more so that its use varies in different English-speaking regions. The use of the comma, hyphen, and semicolon seems to a large extent to be a matter of personal taste regarding cohesion and separation, even in printed material. What distinguishes the printing practice from e.g. private writing is the consistency in punctuation use kept throughout the entire document. Using a subset of the punctuation rules listed in [9] in combination with [10], the conventions imposed on speech reconstruction aim at gaining this consistency.

Substitution

The reconstructed text is supposed to contain only standard word forms. The lemma of the given lexical unit corresponds to the meaning it expresses. During the speech reconstruction annotation, the input word forms are checked and corrected whenever appropriate. The annotation manual lists contracted and genitive forms *('ll, 't, 'd, 's,* etc.*)* that are regarded as standard, as well as frequent contractions that are regarded as non-standard and are to be replaced with full forms by editing the m-node attribute 'form' (*'em, d'you, wanna*, *'cos,* etc.). Contractions that are not listed are implicitly regarded as non-standard and subject to the 'form' change. On the other hand, colloquial and low-standard lexical units are not stylistically 'upgraded'. For instance, the interjections *yeah, aye,* and *nope* are not replaced with *yes* and *no*.

Subject-verb concord is another frequent substitution issue. The grammatical concord in cases as *we was there* is restored.

When an obviously wrong word (e.g. a paronym) was used, the annotator is supposed to replace it by editing the 'form' as well as the 'lemma' attribute.

Example:

will you tell me ***who*** *were you last summer* ⇨
Will you tell me ***where*** *you were last summer?*

Word Order Change

All sentences at the m-layer must have a correct word order that makes the entire discourse fluent. E.g. the subject-verb inversion is cancelled in indirect questions:

I felt like asking where ***was the castle*** ⇨ *I felt like asking where* ***the castle was****.*

Topic-focus motivated fronting is though regarded as standard:

Jane I met at the university and lived with for a couple of years. ⇨ *Jane I met at the university and lived with for a couple of years.*

in rushed my husband Wilbur, yelling... ⇨ *In rushed my husband Wilbur, yelling...*

Other annotation issues

The annotation manual codifies spelling of numerical and non-alphabetical characters for frequent cases, e.g. amounts and currencies, Roman numerical indexes, etc. Made-up and argotic words are marked. The annotators can use a few types of annotators' comments to mark w-layer errors, doubts about orthography, etc. Figures 3, 4, and 5 show some editing examples.

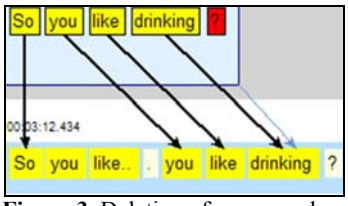

Figure 3. Deletion of a reparandum

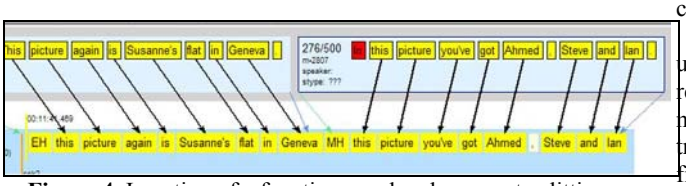

Figure 4. Insertion of a function word and segment splitting

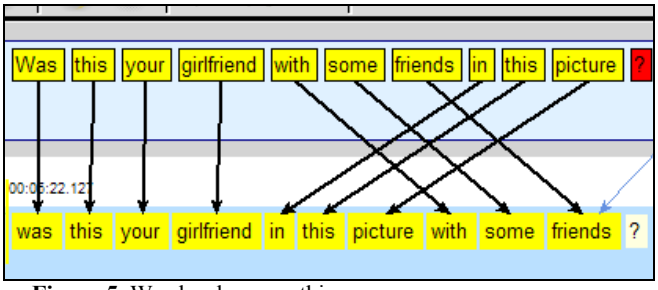

Figure 5. Word order smoothing

Double annotation has been launched. One file has been completed so far (5840 and 5594 tokens in the respective annotations). Other 6 files have just been assigned to the second annotator. Interannotator agreement monitoring is intended as soon as more data has been double-annotated, though a similar project ([11]) has shown that high interannotator agreement is not to be expected in this type of annotation.

6 RELATED WORK

The annotation of disfluencies in speech transcriptions itself is not a strikingly new idea. A lot of related research has been done in the last two decades on English. Heeman and Allen [12], [13] employed part-of-speech tagging to detect and correct speech disfluencies prior to parsing on the TRAINS dialog corpus [14] . Later on, Heeman [15] proposed a statistical language model that redefined the speech recognition problem. Heeman's model can detect and correct speech repairs, including their editing terms, and identify boundaries and discourse markers along with assigning and predicting POS tags, using the acoustic clues associated with speech repairs and phrase boundaries (silence, intonation). Corpora with dense speech disfluencies (recordings of stutterers) have been acquired and annotated [16], [17] to enable information extraction even from deficient speech.

The same goal, i.e., the extraction of meaningful utterances from spontaneous speech, is pursued by the Metadata Extraction (MDE) projects, originally supported by DARPA EARS [18], [19], [20], with the effort being extended beyond English to other languages (Czech, Mandarin Chinese, and Levantine Arabic). An MDE-annotated Czech corpus [21] is about to be released by the LDC (built upon a previous LDC release, [22]). All these annotation schemes make a difference between several types of disfluency and detect discourse markers as well as revisions and editing terms. What these annotation schemes have in common is that the annotators are not allowed to alter the text. In other words, the annotation is strictly designed to point out and classify the syntactic deviations from written-text standards, but the output is not required to meet written text standards. E.g., the annotators do not correct the word order. No punctuation conventions seem to be applied, either.

Since the speech reconstruction performed at our Institute is understood as a preparation step towards a tectogrammatical representation, whose basic unit is a sentence, the annotation must aim at creating a standardized text from the raw transcription. The annotators are therefore – apart from marking fillers etc. for deletion - allowed to substitute and insert words as well as to change the word order. On the other hand, the annotators do neither classify semantic nor syntactic relations between clauses (unlike MDE), and their decisions about sentence boundaries are based on stylistics rather than on prosody, although the annotators are obliged to listen to the acoustic segments as well. The m-segments do neither correspond to Sentence Units known from MDE nor to utterances, which can take the form of clauses or clause elements.

The annotation scheme of the PDT-style speech reconstruction[4] has been developed in parallel (and often in cooperation) with Fitzgerald and Jelinek [11], whose annotation provides a far more detailed classification of each alteration of the standardized text output; also, Fitzgerald and Jelinek incorporate argument structure labeling (along with a rough ellipsis restoration for argument structure reasons) straight into the speech reconstruction. This is all supplanted, in our case, by the linking of the higher levels of annotation, i.e., the a-layer (for surface syntax and the t-layer (for deep syntax and semantics), back to the speech reconstructed data at the m-layer of annotation.

7 CONCLUSIONS & DISCUSSION

This initial-stage project report presents developing and applying an annotation scheme for manual speech reconstruction above a corpus of spontaneous dialogs on family photographs, and it also gives an idea of how speech reconstruction is incorporated within the more complex PDT-style annotation scenario that spans from linear text to underlying syntax dependency trees. However, the manual speech reconstruction annotation is assumed to prove useful even outside this particular annotation scenario.

So far we have manually annotated approx. 140,000 tokens (somewhat more than 11,000 output sentences) since March 2008. The corpus obtained is expected to serve as a common input for further linguistic analysis within the Companions project, including the dialog act annotation.

[4] The first annotation experiments started on the Czech MALACH corpus, using a previous version of the MEd editor [23], [11]. Then the annotation scheme was adapted to the needs of English dialog annotation [24].

ACKNOWLEDGEMENTS

This work was funded in part by the Companions project (www.companions-project.org) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-034434, by the Czech Science Foundation under GACR grant number 405/06/0589, and by the projects of the Ministry of Education of the Czech Republic Nos. MSM0021620838, ME838, and LC536. The recordings and the transcripts have been acquired by the Napier University in Edinburgh, UK, in cooperation with the University of West Bohemia, Pilsen.

REFERENCES

[1] J. Bradley, O. Mival, and D. Benyon. A Novel Architecture for Designing by Wizard of Oz. In: *Proceeding of CREATE08, British computer Society, Covent Garden, London, 24-25 June 2008* (2008).

[2] K. Bontcheva, H. Cunningham, V. Tablan, D. Maynard, and H. Saggion. Developing Reusable and Robust Language Processing Components for Information Systems using GATE. In: P*roc. 3rd International Workshop on Natural Language and Information Systems (NLIS'2002), Aix-enProvence, France,*. IEEE Computer Society Pressnb(2002).

[3] R. Catizone, A. Dingli, H. Pinto, and Y. Wilks. Information Extraction tools and methods for Understanding Dialogue in a Companion. In: *Proc. Sixth International Conference on Language Resources and Evaluation (LREC08), Marrakech, Morocco* (2008).

[4] J. Hajič et al. *Prague Dependency Treebank 2.0.* CD-ROM. LDC, Philadelphia, USA (2006a).

[5] P. Pajas and J. Štěpánek. XML-Based Representation of Multi-Layered Annotation in the PDT 2.0. In: *Proceedings of the LREC Workshop on Merging and Layering Linguistic Information (LREC 2006)*, pp. 40-47 (2006).

[6] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman. Transcriber: a free tool for segmenting, labeling and transcribing speech. In: *First International Conference on Language Resources and Evaluation, Granada, Spain, May 1998*, pp. 1373-1376 (1998).

[7] P. Ircing: *COMPANIONS. Rules for annotation of English audio recordings using the Transcriber software*. Version 1, University of West Bohemia, Pilsen. Internal document (2007).

[8] P. Pajas and D. Mareček *MEd - an editor of interlinked multi-layered linearly-structured linguistic annotations*. **`http://ufal.mff.cuni.cz/~pajas/med`** (2007).

[9] R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik: *A Comprehensive Grammar of the English Language*. Longman, first published 1985 (2004).

[10] *Comma writing*. URL< http://owl.english.purdue.edu/>, quoted 2008-01-17.

[11] E. Fitzgerald and F. Jelinek. Linguistic Resources for Reconstructing Spontaneous Speech Text. In: *LREC 2008 Proceedings* (2008).

[12] P. Heeman and J. Allen, Tagging Speech Repairs. In: *ARPA Workshop on Human Language Technology*, Princeton, March 1994, pp. 187-192 (1994a).

[13] P. Heeman and J. Allen. Detecting and Correcting Speech Repairs. In: *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94)*, Las Cruces, June 1994, pp. 295-302 (1994b).

[14] P. A. Heeman and J. F. Allen. *The Trains Spoken Dialog Corpus*. CD-ROM, Linguistics Data Consortium, April 1995 (1995).

[15] P. Heeman. *Speech Repairs, Intonational Boundaries and Discourse Markers: Modeling Speakers' Utterances in Spoken Dialog.* Technical Report 673, U. Rochester, December 1997. Doctoral dissertation (1997).

[16] P. Heeman, A. McMillin, and J. S. Yaruss. An annotation scheme for complex disfluencies. In: *Proceedings of the 9th International Conference on Spoken Language Processing*, Pittsburgh (2006).

[17] P. Heeman, A. McMillin, and J. Scott Yaruss. Intercoder Reliability in Annotating Complex Disfluencies. In: *Proceedings of the 10th European Conference on Speech Communication and Technology*, Antwerp Belgium, August 2007 (2007).

[18] Y. Liu et al. Structural Metadata Research in the EARS Program. In: 2005 *IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, PA* (2005).

[19] S. Strassel. *Simple metadata annotation specification V6.2.* http://www.ldc.upenn.edu/Projects/MDE/Guidelines/SimpleM DE_V6.2.pdf. (2004) [quoted June 20, 2008].

[20] S. Strassel, J. Kolář, Z. Song, L. Barclay, and M. Glenn. Structural metadata annotation: Moving beyond English. In: *Procs. Interspeech Lisboa 2005*, ISCA, Bonn, pp. 1545-1548 (2005),

[21] J. Kolář, and J. Švec. Structural Metadata Annotation of Speech Corpora: Comparing Broadcast News and Broadcast Conversations. In: *Proc. LREC2008 , Marrakech, Morocco* (2008).

[22] 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).

[23] J. Hajič et. al. *Pražský závislostní korpus mluvené češtiny. Rekonstrukce standardizovaného textu z mluvené češtiny.* [Prague Dependency Treebank of Spoken Czech. Speech reconstruction]. Technical report UFAL MFF (TR-2006-33), Prague (2006b).

[24] S. Cinková and M. Mikulová. *Speech reconstruction for the syntactic and semantic analysis of the NAP/AAA corpus*. Annotation manual for annotators. **`http://ufal.mff.cuni.cz/~cinkova/speech/data/done/speechindex.htm`** (2008 – unpublished).