Speech reconstruction for the syntactic and semantic analysis of the NAP/AAA corpus

Silvie Cinková, Marie Mikulová, 2007-12-27
UFAL MFF UK

Version 1.1

last revision: 2008-02-25

Table of content

Motivation

Unprepared spontaneous speech breaks many rules by which written texts are constituted. The speakers, for instance, often get the syntax wrong; they mispronounce or confuse lexical units, and their use of ellipsis as well as of deictic words and connectives is abundant, compared to standard written texts. Despite most spontaneous oral communications not meeting the basic written-text standards, the mutual understanding among humans does usually not get harmed. Posing no problem for humans, spontaneous speech is yet very difficult to handle for machines, whose analytical tools have been designed for written-language processing.

The only way out seems to lead through machine learning: to process enough data for the machine to learn how to tell apart noise from relevant structures and how to restore the commonest types of ellipses to be able to analyze the spoken data with the tools already available. To build such data means smoothing the speech transcription to meet the usual written-text standards by re-chunking and re-building the original segments into grammatical sentences with acceptable word order, proper morphosyntactic relations between words and appropriate wording, while preserving links to the original transcription.

This document is primarily meant to serve as a manual for human annotators of spoken data of the English NAP/AAA corpora, which comprise conversations above photographs between a computer avatar simulated by a human and a human who believes himself to be talking to a real robot. This manual is largely based on a similar manual used for annotation of The Prague Dependency Treebank of Spoken Czech (PDTSC) (Hajič et al., 2006). The (few) differences between both annotation routines are explicitly stated in the text. Some orthographical rules listed here may appear self-understood to native speakers, who are supposed to perform this annotation. However, this text seeks to ensure a reasonable output quality even if it were necessary to employ Czech annotators.

1 Introduction

Speech reconstruction provides an interface between speech recognition and linguistic text analysis on any level. It is motivated by the fact that the analytical tools currently available have been designed for written language, and thus they hardly cope with the irregularities of spontaneous speech. The purpose of the speech reconstruction is to make the spontaneous speech meet written text standards before it is tagged and parsed. The speech reconstruction annotation is done manually and is based on audio transcripts. The FGD-based annotation, which has been performed in PDTSC and will be used in the Napier and AAA Corpora, is divided in several mutually interlinked layers (from bottom to top; cf. Figure 1).

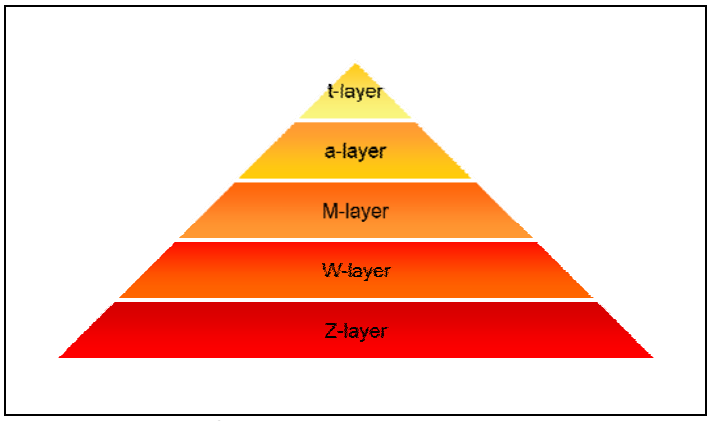

Figure 1 ***Hierarchy of the annotation layers in FGD***

For speech corpora the annotation layers are defined as follows:

- Z-layer: ASR transcript. Basic unit: token, represented as z-node with attribute values.
- W-layer: manually corrected ASR transcript or a manual transcript. Basic unit: token, represented as w-node with attribute values.
- M-layer: speech reconstruction (includes defining sentence boundaries), morphological tagging, lemmatization ("m-lemmas"). Basic unit: word (includes numbers and punctuation), represented as m-node.
- A-layer: surface-syntax annotation. Basic unit: sentence, represented as a rooted dependency tree with a-nodes and edges.
- T-layer: tectogrammatical annotation ("underlying syntax", "semantics"). Basic unit: sentence, represented as a rooted dependency tree with t-nodes and edges.

Figure 2 illustrates the multi-layered annotation of several speech segments (w-layer) merged into a sentence at m-layer and further parsed at a- and t-layer. The a- and t-layer representations in this picture are only roughly sketched in the Word editor.

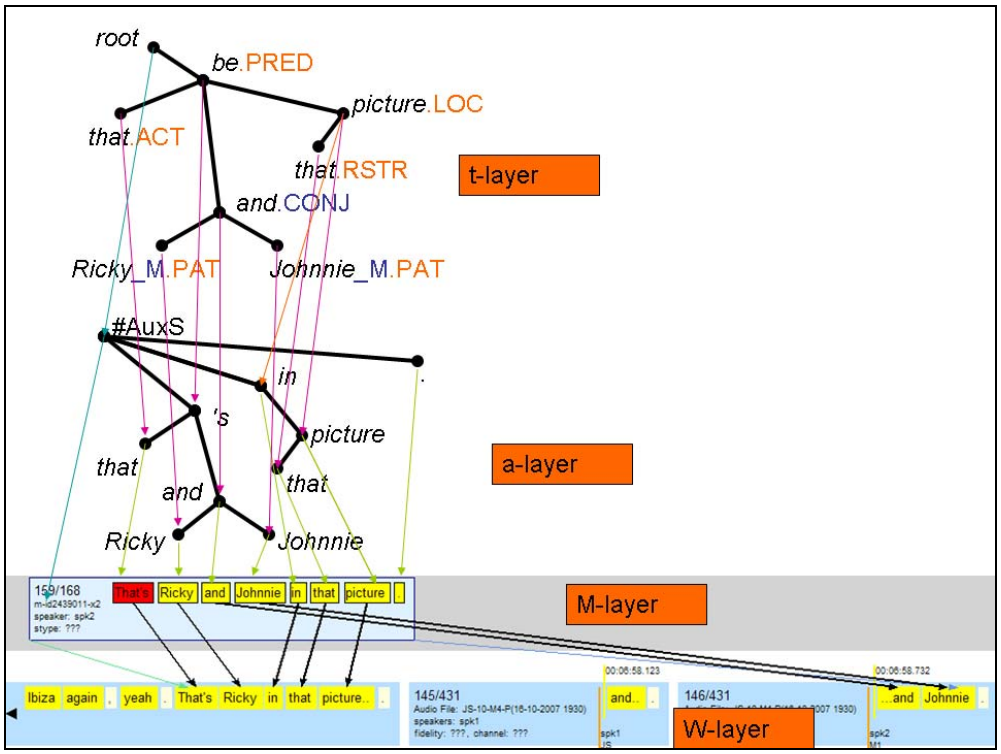

Figure 2 ***A sentence annotated at m-, a- and t-layer***

2 Basic annotation principles

The annotation of m-layer resembles editing the transcription of an interview record for print. The output must not only be intelligible but also grammatically correct and easy-to-read. For the purpose of the speech reconstruction, a text conforms to **written-text standards** when:

- The text does not contain any discourse-irrelevant non-verbal events: laughing, coughing, background noise etc.
- Speech-specific phenomena are removed.
- The text is fluent.
- The word order and syntax are correct.
- Only standard and orthographically correct variants of words are used.

When performing the speech reconstruction, 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.

2. **The Minimal Modification Principle:** modifications are performed only when necessary to achieve the written-text standard in the transcribed speech.

2.1 Annotation layers

The purpose of this document is to describe the speech reconstruction at m-layer. Other procedures performed at m-layer, such as lemmatization and morphological tagging, are not described here. Similarly, a separate document summarizes the conventions of the manual transcription/manual correction of automatic ASR (Ircing, 2007). The topmost layer (t-layer) and the guidelines for its manual annotation have been described in Cinková et al. (2006) for English (just some parts of annotation) and in Mikulová et al. (2006) for Czech (complete annotation). A-layer has been explicitly specified only for Czech (Hajič et al., 1999). So far there has been no a-layer specification for English since English a-layer is just automatically generated when parsing the automatic t-layer for manual annotation.

When performing the speech reconstruction, we usually work with w- and m-layer.

2.1.1 Z-layer

Z-layer is the uncorrected ASR output gained from the original audio file. This layer is not necessary when the manual transcription is made straight from the audio.

2.1.2 W-layer

W-layer is the manual audio transcription or a manual correction of an automatic ASR transcript. It captures the utterances of the speaker including all slips of the tongue, coughing, laughter, pauses etc., as well as background noises. The basic units of w-layer are the so-called events. The most important events are **content events**. They are divided into three types:

- recognized word forms (tokens, w-nodes of type **`w`**)
- recognized non-speech events (w-nodes of type **`nonspeech`**)
- recognized background noise (w-nodes of type **`background`**)

The events (the w-nodes) are segmented in **turns**, which are primarily defined by being produced by one speaker. However, the cases of overlapping speech allow for one turn to be associated with several speakers.

2.1.3 M-layer

M-layer is constituted by **reconstructed speech** (speech smoothed to meet written text standards) with subsequent morphological tagging and lemmatization. The finalized m-layer is ready for parsing.

The basic m-layer units are lexical units (word forms, numbers and punctuation) represented by m-nodes of type **`m`**. Discourse-relevant non-speech events (e.g. *UHUH* as an expression of agreement) are captured as m-nodes of type **`nontext`**. Non-speech events irrelevant for discourse (e.g. loud breathing, pauses or laughter in the middle of a

sentence) are not represented at m-layer. The m-nodes are manually grouped into sentences represented by the elements **s**.

2.2 Relations between the m-layer- and w-layer units

The differences between the input segments of the manual transcription (captured by w-layer) and the standardized sentences (captured by m-layer), i.e. the modifications performed at m-layer, are visualized by the relations between the two layers and their units.

Every m-node that corresponds to a w-node contains a reference to that w-node.

The most important references between m-layer and w-layer are the links between the m-nodes of type **m** (representing tokens at m-layer) and the w-nodes of type **w** (representing tokens at w-layer). The following statement is true for the relations between m-nodes of type **m** and w-nodes of type **w**:

An m-node of type m does not have to contain any reference to w-layer.[1]

An m-node of type **m** that contains no reference to w-layer is called **inserted m-node**. It represents the annotator's own insertions (see 5.4.1.10, Insertion).

A w-node of type w does not have to be referred to from any node at m-layer.[2]

A w-node of type **w** which is not referred to from any m-node at m-layer represents a lexical unit that was classified as discourse-irrelevant by the annotator. It is called **deleted node** (see 5.4.1.1, Deletion).

The order of the m-nodes of type **m** at m-layer does not have to correspond to the order of the w-nodes of type **w** at w-layer.

Different node order at the two layers indicates word order changes performed by the annotator at m-layer (see 5.4.3, Word order changes).

There are two other reference types:

- references from the m-nodes of type **nontext** (which represent discourse-relevant non-speech events) to the w-nodes of type **nonspeech** (which represent non-speech events).
- references from the m-nodes of type **nontext** (which represent discourse-relevant non-speech events) to the w-nodes of type **background** (which represent background noise)

[1] In other words: a token can be added at m-layer during the annotation.
[2] In other words: a token can be deleted at m-layer during the annotation.

The following statements are true for the relations between the m-nodes of type **nontext** and the w-nodes of types **nonspeech** or **background**, respectively:

An m-node of type nontext does not have to contain any reference to w-layer.

An m-node of type **nontext** that does not contain any reference to w-layer stands for a discourse-relevant non-speech event that has not been captured at w-layer; e.g. a particular stress on a word, whispering etc.

A w-node of type nonspeech does not have to be referred to from any node at m-layer.

A w-node of type **w** which is not referred to from any m-node at m-layer represents a non-speech event that was regarded by the annotator as either discourse-irrelevant or already reflected by the means of written text.

The **type** attribute values will be inserted automatically as soon as the manual annotation is finished.

The following **type** attribute values have been considered so far:

A. Types of references from the m-nodes of type **m** to the w-nodes of type **w:**

- **basic**: the form of the m-node is identical with the token of the w-node or there was only an orthographical modification (see 5.1, Orthographical modifications)
- **num**: orthographical modifications of numbers (see 5.3.3.1, Numbers and digits)
- **substitution**: the form or the lemma of an existing m-node was altered, and it is different from the one of the corresponding w-node (substitution was performed – see 5.4.2, Substitution)

B. Types of references from the m-nodes of type nontext to the w-nodes of types nonspeech and background:

- **nonspeech**

One m-node can contain references to more than one w-node.

Example:

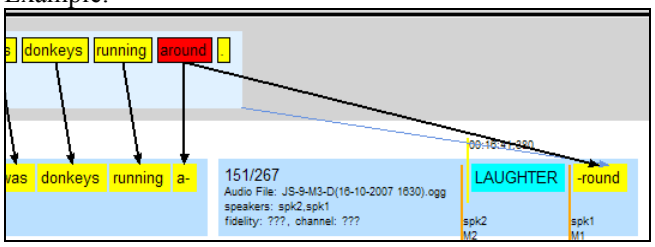

Figure 3

2.3 Sentence attributes

The m-nodes at m-layer are segmented into sentences, which are represented by the so-called **s-elements**.

Each s-element contains two (non-typed) references to w-layer: to the first and to the last content event that belong to the reconstructed sentence.

The references from the s-element (`w-begin.rf`, `w-end.rf`) indicate the span of w-layer that was used as the input for the given reconstructed sentence (see 4, Sentence segmentation)
Each s-element has the following attributes:

- **`w-speaker.rf`**: identification of the speaker. This attribute value is added automatically after the manual annotation is finished.
- **`is_modified`**: this attribute indicates whether the sentence represented by the given s-element has been altered with respect to the corresponding w-layer segment(s). This attribute is added automatically after the manual annotation is finished.

The Czech manual annotation also employs the attribute **`stype`**, which roughly classifies the sentences as speech acts. The attribute values are listed in Table 1:

`information`	information, a discourse-relevant sentence
`instruction`	instruction, order, request
`question`	inquiry
`confirmation`	agreement of the second speaker with what has just been said
`surprise`	surprise, astonishment of the second speaker over what has just been said
`disbelief`	a sentence that signals disbelief of the second speaker towards what has just been said
`repetition`	the same thought rephrased
`other`	yet another type

Table 1 Values of the attribute stype (PDTSC only)

This attribute is currently not considered for the annotation of the NAP/AAA data used in the Companions project. The reason is that this data will be subject to separate dialog-act (DA) annotation especially tailored for a dialog manager. The basic units in the DA annotation are **utterances**. Utterances do not always correspond to sentences at m-layer. An utterance can consist of a sequence of sentences as well as merely of a sentence part.

2.4 The annotation procedure

The following annotation procedure has been set for the speech reconstruction:

1. The annotator reads the entire manual speech transcription captured by w-layer.

2. When something in the text is unclear or ambiguous, the annotator listens to the corresponding part of the original audio recording.
3. The annotator performs the sentence segmentation (see 4, Sentence segmentation).
4. The annotator smoothes each sentence to meet written-text standards by means of the following modifications:
    a. deletion
    b. insertion
    c. substitution
    d. word order change
5. The annotator makes sure that each m-layer segment start (green arrow) and m-layer segment end (blue arrow) are placed properly (for more detail see 4.2, Indicating sentence borders in spontaneous speech). Node references are to be checked as well.
6. The annotator reads the entire reconstructed text (the resulting m-layer) and adds further modifications whenever necessary.

3 The annotation tool

The speech reconstruction of the NAP/AAA corpora makes use of MEd ("m-layer editor"), the annotation tool developed for PDTSC by Petr Pajas and David Mareček. MEd is an annotation tool in which linearly-structured annotations of text or audio data can be created and edited. The tool supports multiple stacked layers of annotations that can be interconnected by links. It is available under the GNU General Public License and can be downloaded from **http://ufal.mff.cuni.cz/~pajas/med/index.html**. Its development has been financed by the following projects:

- GA AV ČR 1ET101120503
- Center for Computational Linguistics (LC536, 2005-2009)
- GAČR (GA405/06/0589, 2006-2008)
- EU - Companions IST-FP6-034434

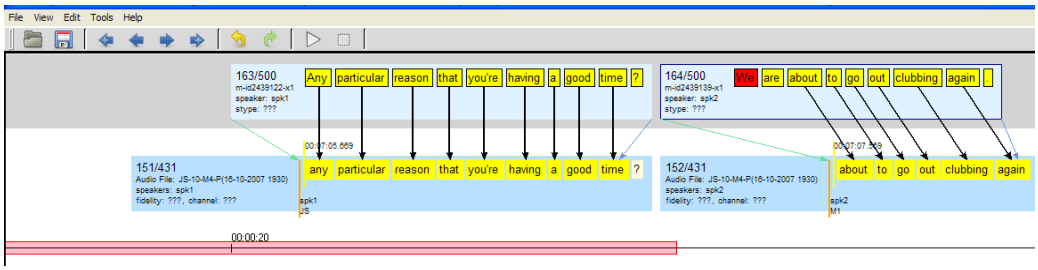

Figure 4 ***A MEd window with the w- and m-layers.***

A detailed documentation to MEd will be available on the MEd project page (see above).

A list of MEd commands is to be found in Appendix 1.

4 Sentence segmentation

Neither the segments automatically created at z-layer nor the segments manually created at w-layer correspond necessarily to sentences. The actual sentence segmentation occurs only at m-layer as a part of the speech reconstruction.

An m-layer sentence must be discourse-relevant. Discourse-irrelevant segments are not reflected by the m-layer. For more detail see 4.3.5, Discourse-irrelevant and content-less .

4.1 Clause types

The sentence segments created at m-layer during the speech reconstruction are supposed to meet written-text standards. The resulting sentence can even be incomplete (for instance an unfinished thought), but it must be shaped as one of the four clause types described in the tectogrammatical annotation manual (Cinková et al., 2006), Section *Verbal and verbless clauses*. It is the following types:

- verbal clause (possibly containing an ellipsis)
- subject-case clause
- vocative clause
- interjectional clause

or their combination.

The following types are **not to be smoothed** into finite-verb clauses at m-layer:

- Exclamatory clauses modified by a restrictive relative clause:

  The clothes she wears! (= Look at the terrible/beautiful clothes she wears!)

- Exclamatory prepositional phrases beginning with *of all* and expressing strong disapproval:

  Of all the impudence!

- Exclamatory constructions with *what a:*

  What a beautiful day!

- Exclamatory adjective phrases expressing approval or disapproval:

  Very interesting!

- Exclamatory noun and adjective phrases expressing approval or disapproval:

  Charming couple!
  Excellent meal!

- Exclamatory noun phrases conveying a warning or e.g. alarm or frustration after a period of forgetfulness:

  Fire!
  The police!
  The cake!
  My husband's birthday!

- Scornful exclamatory clauses consisting of a noun phrase, generally a personal pronoun in the subject or object case, followed by *and* and another noun phrase with a matching possessive pronoun:

  You and your statistics! (= It's funny how much you rely on and over-use statistics!)

- noun phrases acting as yes/no questions (including invitations and offers):

  Any luck, Ron?
  New hat?
  More coffee?

- Noun phrases functioning as assertions or conveying of information:

  False alarm.
  No luck.

- Noun phrases in the sense of imperative:

  Attention, please! (= pay attention)
  Surgeon, immediately! (= bring/call a surgeon)
  Another coffee, if you don't mind. (= make/bring another coffee)
  Scissors, somebody! (= bring scissors)

- Noun phrases with the force of wh-questions:

  Your name?

- Adjective phrases with the force of yes-no question:

  Boring?

- Adverbials:

  In Praha, at five o'clock.

Examples:

{A: You all look like having a good time.}

B: All having a good time again. ➞ *We are all having a good time again.*

aye. same holiday. ➞ *Aye. The same holiday.*

right. ➞ *Right.*

from what I can remember... ➞ *From what I can remember...*

The recurrent types of segments and the recommended way of their reconstruction are described in 5.4, Text modifications.

4.2 Indicating sentence borders in spontaneous speech

The sentence at m-layer is a sequence of m-nodes determined by the so-called s-element. This sequence corresponds to one or several spans of recognized content events at w-layer. This span is determined by two references to w-layer.

The arrows determining the start and the end of a segment should include the corresponding discourse-irrelevant non-speech events and deleted words. This is especially to observe in manual segment splitting and in automatic pre-segmentation in overlapping speech! The automatic pre-segmentation often lets the arrows point to the sync or even to the speaker's mark rather than to the first/last node (Figure 5). This is **not** to be corrected manually provided the node immediately following sync or the speaker's mark really is the one to be referred to.

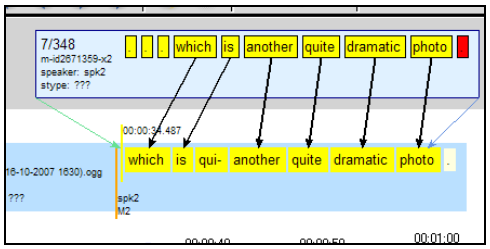

Figure 5 The arrow points automatically to the sync mark.

NB! The check of segment starts/ends is laborious but very important. The segment start and segment end are to be checked and adjusted as the last step when smoothing a sentence, along with the reference arrows pointing from the respective m-layer nodes to the corresponding w-layer nodes.

Examples:

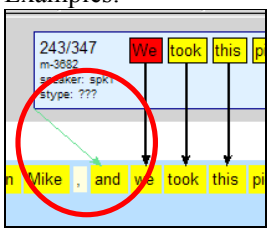

Figure 6

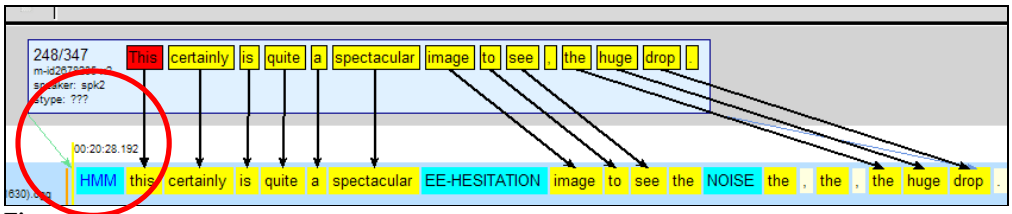

Figure 7

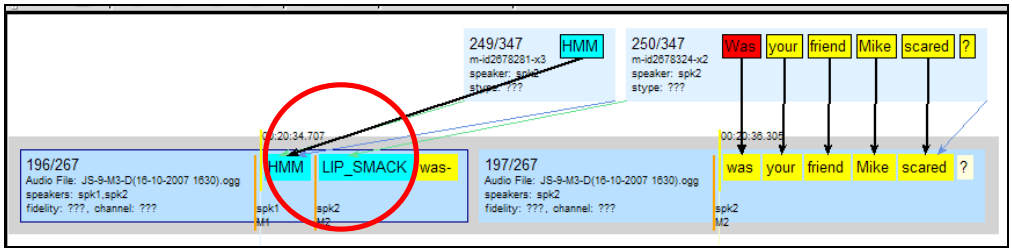

Figure 8

Each s-element contains two references to w-layer: the reference to the first and the reference to the second content event that was used to constitute the given sentence at m-layer.

References from two different s-elements can cross in the cases of overlapping speech (see section 4.3.4, Overlapping speech). There can be content events at w-layer that were not used to constitute any sentence at m-layer (see section 4.3.5, Discourse-irrelevant and content-less).

4.3 Determining sentence and clause borders

4.3.1 Clause borders

The following principle is applied when determining the borders of clauses:

- **The Principle of the longest possible clause:** the clause is supposed to include as many potential sentence elements as possible to remain syntactically as well as semantically acceptable.

Examples:

that particular place <EH> had two balconies ➞ *That particular place had two balconies.*

mostly the music <EM> again, drinking ➞ *Mostly the music and, again, drinking.*

4.3.2 Sentence borders (connecting clauses in complex sentences)

When connecting clauses into complex sentences, too long clause coordinations are to be avoided. A complex sentence should not contain more than three main clauses. When the speaker speaks continuously without pause or dropping intonation, repeatedly using the connective *and*, his turn has to be chunked into several shorter complex sentences.

The superfluous *and* (or *then* etc.) are to be deleted from the m-layer (see also 5.4.1.5, Superfluous connectives). When restructuring the m-layer segments (sentences, s-elements) with the intention to delete a superfluous connective while splitting a segment, the connective-to-be-deleted belongs at the first position of the new segment.

Examples:

Again, this is the Forth Rail Bridge but a bit further away, there was lots of boats out that day, there was... | and it was a really, really nice day, as I said, | and there was also, there's a... | and we took the boat from South Queensferry | and Port Edgar is near South Queensferry | and there's lots of boats moored up in Port Edgar so there was lots of people out. ➞ *Again, this is the Forth Rail Bridge but a bit further away, there were lots of boats out that day. It was a really nice day, as I said, and we took the boat from South Queensferry. Port Edgar is near South Queensferry. There're lots of boats moored up in Port Edgar so there were lots of people out.*

I'm at the top of the Abbey, I've climbed up to the top of the Abbey and you can just see the weather vein at the top and from this you can just see some of those war look-out posts on the other side of the Island. ➞ *I'm at the top of the Abbey. I've climbed up to the top of the Abbey. You can just see the weather vein at the top and from this you can just see some of those war look-out posts on the other side of the Island*.

who's that in the photo is that you and who else ➞ *Who's that in the photo? Is that you? Who else?*

4.3.3 Discourse-relevant fragments

When a discourse-relevant sentence has evidently not been finished, be it deliberately or due to an interruption by another speaker, the utterance is to be left incomplete and **the end of the segment is to be marked with horizontal ellipsis** (three dots, ...) added as three one-dot segments behind the last word.

Figure 9 shows the annotation of incomplete sentences. The interruption of a speaker by another speaker that causes a regular turn shift is in MEd visually distinguished from overlapping speech (cf.4.3.4, Overlapping speech).

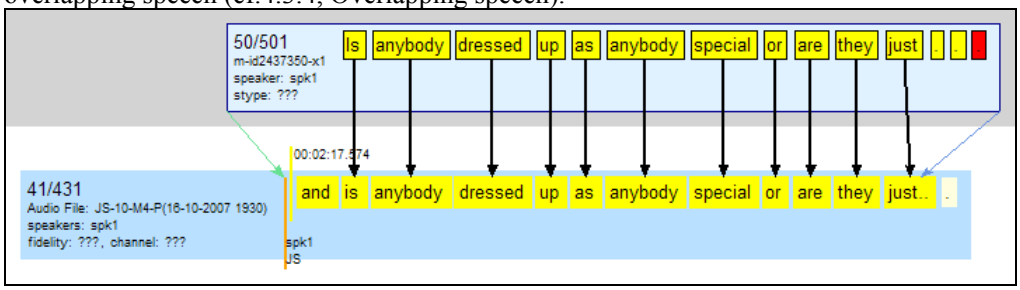

Figure 9

When the second speaker interrupts the first speaker by completing the first speaker's sentence in his own way and the first speaker accepts the completion by not completing the utterance in his own way, the completion is marked with ellipsis at the beginning. The first word in the completion is not capitalized (Fig. 5).

Example

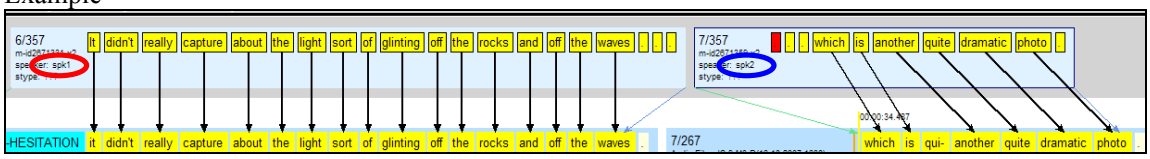

Figure 10

When the first speaker completes his utterance despite the second speaker's interruptive completion, the segments uttered by the first speaker are merged to make a sentence as complete as possible. This sentence may still remain incomplete, though, which is indicated by the ellipsis at the end of the sentence. The interruption that came in between does not start with an ellipsis, and its first letter is capitalized (Figure 11).

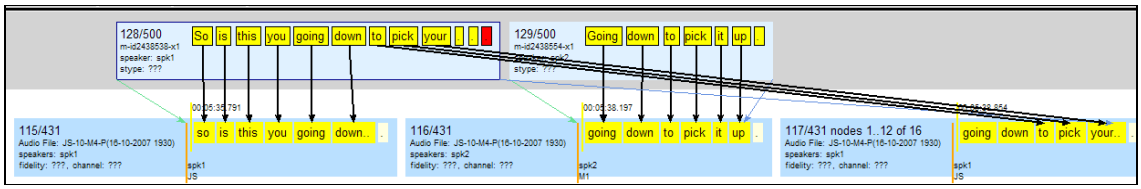

Figure 11

4.3.4 Overlapping speech

When two (or more) speakers make different utterances simultaneously the annotator is supposed to merge the discontinuous segments into semantically and syntactically complete utterances. In MEd overlapping speech differs graphically from turn switches by displaying the utterances of both speakers in the same segment (marked with blue at w-layer, Figure 12)

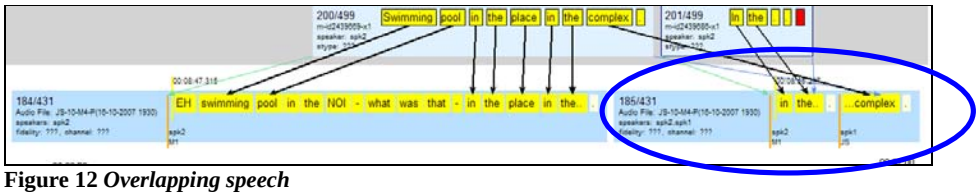

Figure 12 ***Overlapping speech***

The discontinuous segments uttered by one speaker are to be merged as far as possible. In marginal cases the speech overlapping can be discourse-relevant on its own. The annotation resolution for this case (to be used extremely sparingly!!!) is preserving the discontinuity of the respective segments and marking the overlapping speech at m-layer by inserting additional three dots at the beginning of each segment being discontinuous at its beginning, such as:

Example:

A: Swimming pool in the place in the... B: ... in the... A: ... complex.

The segments introduced by the three dots do not start with a capital letter.

4.3.5 Discourse-irrelevant and content-less events

W-nodes of discourse-irrelevant events like restarts, slips of the tongue, hesitations, etc. have no correspondent m-nodes at m-layer. However, they are normally included in the m-layer segments (see Figure 13) since they contribute to the final shape of the respective sentences.

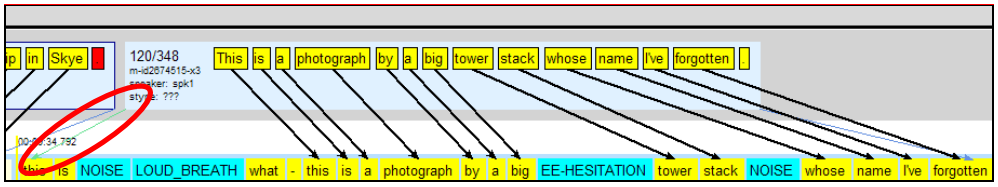

Figure 13 Noise and loud breath as discourse-irrelevant content event (included within a segment)

There are yet discourse-irrelevant events at w-layer that cannot be easily associated to any segment as they are separated from the speech events by long silence, much noise or more than two non-speech events. These events (even if they are tiny fragments of text) have no semantic content and should not be included into the segments at m-layer. There is in fact no sharp line between a content-less event and e.g. a reparandum enclosed by hesitation sounds. The decision is up to the annotator. The general recommendation, based on a 500-sentence sample of the NAP corpus, is that apparent content-less events are rather rare and virtually everything from w-layer should be included into m-layer segments.

Examples:

<UH-HUH> <LOUD_BREATH> *well* <UH-HUH> *so* <NOISE> ➜ Ø

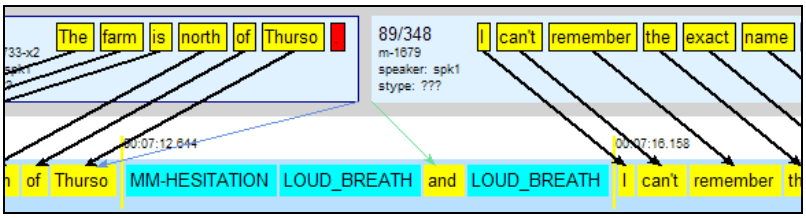

Figure 14 Hesitation and loud breath as content-less events (not part of a segment)

5 Text modifications

The most essential part of the annotation is various types of modifications of the input transcription at w-layer performed in order to make the text conform to written-text standards. There are two basic modification types:

- orthographical modifications (see 5.1, Orthographical modifications)
- substantial modifications (see 5.4, Substantial modifications)

5.1 Orthographical modifications

Under orthographical modifications we understand:

- deletion of discourse-irrelevant non-speech events (see 5.2, Deletion of discourse-irrelevant non-speech events)
- orthographical corrections (see 5.3, Orthographical issues)
- transcription by non-alphabetic characters

5.2 Deletion of discourse-irrelevant non-speech events

Non-speech events (e.g. breathing, coughing) are consequently recorded at w-layer (the list of marks of non-speech events recorded at w-layer see Table 2 *Marks of non-speech events*). Only discourse-relevant non-speech events are preserved at m-layer (see Annotation of discourse-relevant non-speech events). Most non-speech events are yet discourse-irrelevant and are deleted from the m-layer.

NB: Non-speech backchannels like UH-HUH and HMM are discourse-relevant unless they interrupt the first speaker in the middle of a sentence and the first speaker continues the sentence after the backchannel. They are to be restored at m-layer as nodes of the type **`m-nontext`**, not followed by a node representing the full stop. The form of the backchannel (UH-HUH, HMM) as displayed at w-layer should be filled in the attribute **`type`**.

Examples:

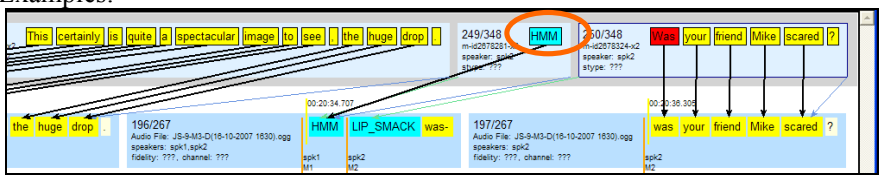

Figure 15

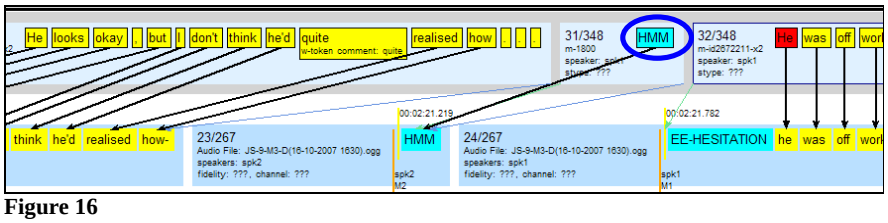

Figure 16

but:

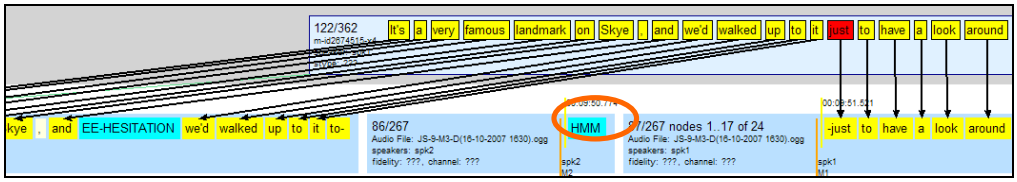

Figure 17

The m-layer contains no m-nodes for discourse-irrelevant non-speech events.
There are no references from the m-layer to discourse-irrelevant non-speech events.

Table 2 ***Marks of non-speech events***

MM-HESITATION	filled pause – closed mouth	**MH**
EE-HESITATION	filled pause - opened mouth	**EH**
LOUD_BREATH	exceptionally loud breath	**LB**
HMM	agreement - closed mouth	**HM**
MM	disagreement	**MM**
UH-HUH	agreement – open mouth	**UH**
NOISE	other noise – instantaneous	**NOI**
NOISE	other noise – beginning	**NOB**
NOISE	other noise – end	**NOE**
UNINTELLIGIBLE	unintelligible fragment	**UN**
LAUGHTER	laughter	**LA**
LIP_SMACK	loud lip smacking	**LS**

Examples:

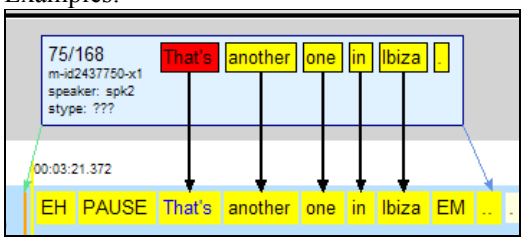

Figure 18

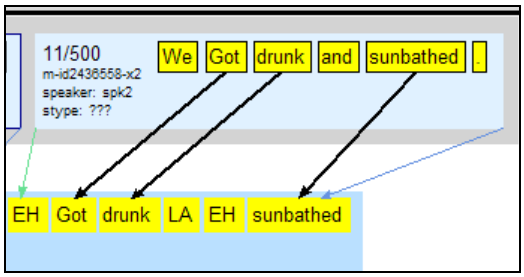

Figure 19

5.3 Orthographical issues

The output text at m-layer is supposed to meet written-text standards (unlike the text at w-layer).

Two things are especially important at this point:

- punctuation (see 5.3.1, Punctuation)
- capitalization (see 5.3.2, Capitalization)

5.3.1 Punctuation

The m-layer output text is supposed to contain correct punctuation.

An inserted m-node representing a punctuation mark contains no reference to w-layer.

Example:

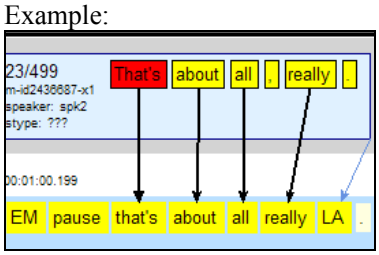

Figure 20

The English punctuation shows a great deal of flexibility, compared e.g. to Czech. The use of the comma, hyphen and semicolon is to a large extent 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 throughout the entire document. Being stricter than e.g. the punctuation rules listed in Quirk et al. (2004), the conventions imposed on speech reconstruction aim at gaining this consistency.

5.3.1.1 Hyphen and colon

Hyphens are not used in sequences adverb-participle when modifying a noun phrase (not **a well-established system* but *a well established system*). However, numerals like *twenty-five, twenty-fifth* are obligatorily spelled with hyphen, and so are noun modifiers consisting of an entire adjective-noun phrase (*a short-wave radio*), a prepositional phrase (*an on-the-go list*) or an adjectival phrase (*a Czech-English dictionary*). Each part of the hyphenized cluster is reflected as one m-node respectively, including the hyphen(s). Even words containing prefixes like *re-, non-* etc. (*non-profit*, *re-invention*), when spelled with hyphen, are to be recognized as three nodes (prefix + hyphen + word). These words are therefore preferably not spelled with hyphen at m-layer (where acceptable).
The same applies to colon occurring between two digits. Each digit as well as the colon itself are represented as one m-node, respectively.
When the hyphen/colon is present already at w-layer, the m-node with hyphen refers to it.

Examples

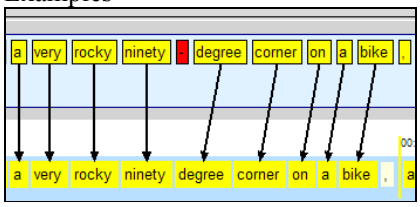

Figure 21

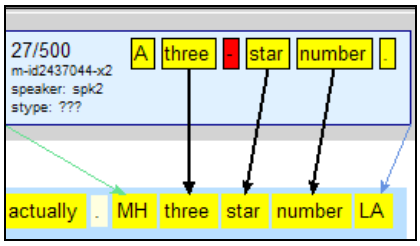

Figure 22

5.3.1.2 Diaeresis

Diaeresis is not to be used (not **naïve*, **coöperation* but *naive, cooperation*)

5.3.1.3 Comma

This section makes a difference between the use of the comma between clauses and between other syntactic elements[3].

Included parenthetic clauses are always separated by commas.

He is, I think, a teacher.

The comma is not normally used to separate independent[4] clauses unless they are linked by a coordinator. Use colon or semicolon instead. Colon is used when the second unit is in apposition to (a part of) the first unit and no coordinator (e.g. *and*) can be inserted in between:

I've just had some good news: I've been offered a job.

Semicolon is used when a coordinator could be inserted between the two clauses:

Schoolchildren have adopted the fund as one of their favourite charities; their small contributions have enabled the fund to reach its target.

The comma between two coordinated independent clauses linked by the coordinator *and, or, neither, nor, but* and *yet* is to be inserted when both clauses have their expressed subjects, respectively.

The work was pleasant, and the hours were short.

We are thinking of buying a radio, but we haven't made up our minds.

The comma between two independent clauses linked with a coordinator is to be omitted when one of the clauses does not have the subject expressed:

You can sit at my desk and write your letters.

I enjoy tennis but don't play it very often.

When only two independent clauses are coordinated and they are linked with the coordinator *and, or, nor, neither, so* or *yet*, the comma is appropriate between them (unless one of them or both have unexpressed subjects). When three and more clauses are coordinated and only the fore last and the last clause are linked with a coordinator, the comma should be inserted between each clause pair. This is called **serial comma**:

Prices fell, interest rates fell, and employment figures rose.

---

[3] Many example sentences originate from the web page "Using Commas" maintained by the Purdue University Online Writing Lab and from Quirk et al. (2004).

[4] An independent (main) clause is a clause that is neither a wh-clause, nor an infinitive or a participial clause, nor is it introduced by a subordinating conjunction, e.g. *if*.

Exception: when the clauses are short and semantically as well as structurally parallel, commas are inserted even without a coordinator present. Such coordination typically consists of three clauses:

I must, I can, I will.

When chunking the segments into clauses and clauses into sentences, please note: when no coordinator is used between two clauses that are linked into one sentence at w-layer and they are not closely semantically related, split them into two different sentences (s-elements, segments)!

**I went there, I can't remember anything more.* ➜ *I went there. I can't remember anything more.*

We generally do not insert any punctuation mark if we coordinate two units within a sentence other than independent clauses, using a coordinator:

The movie was long and boring.

I argued for implementation of the report and against any further discussion.

When three and more units within a sentence other than independent clauses are coordinated without using a coordinator between the fore last and the last unit, the comma must be inserted between each pair:

He walked with long, slow, steady, deliberate strides.

When three and more units within a sentence other than independent clauses are coordinated using a coordinator between the fore last and the last unit, the comma must be inserted between each pair, even the one with the coordinator:

**She slowly, carefully and deliberately moved the box*[5].
She slowly, carefully, and deliberately moved the box.

Comma is also to be used before *etc*.

The difference between coordination and the hypotactic relation of adjectives is to be preserved:

long, steady strides = long and steady strides
a good black coffee = a black coffee that is good

[5] Here, as in many other places, is to be particularly stressed that not inserting the comma before the coordinator is not generally false but is to be avoided merely for the sake of consistency in the annotated document.

Clause elements like subject, object and predicate complement may not be separated from the verb by any punctuation, no matter which form they take (a phrase, a clause etc.) unless they are direct speech in quotation marks.

Direct speech in quotation marks is always separated from the introducing sentence by a comma. Ordinary quotation (of any kind, not only direct speech) is enclosed by double quotation marks while a quotation inside a quotation is enclosed by single quotation marks:

"I heard 'Keep out' being shouted," he said.

She enjoyed the article "Cities are for walking."

Note that commas and full stops are **inside** the quotation marks.

Dependent clauses and long prepositional phrases (more than two words) **preceding** the main clause are separated with a comma:

Having finished the test, he left the room.

The sun radiating intense heat, we sought shelter in the cafe.

To get a seat, you'd better come early.

While I was eating, the cat scratched at the door.

Because her alarm clock was broken, she was late for class.

If you are ill, you ought to see a doctor.

When the snow stops falling, we'll shovel the driveway.
After the test but before lunch, I went jogging.

A dependent clause **following** the main clause is **not** to be separated by a comma:

She was late for class because her alarm clock was broken.

The cat scratched at the door while I was eating.

Exception: In rare cases the speaker might wish to indicate that an adverbial is very loosely attached to the sentence, expressing a circumstance parallel to the main predication rather than modifying it. Sometimes two typologically similar adverbials (temporal, spatial, etc.) seem to be in apposition, as in e.g.:

We'd realised he'd broken his collar bone, but he didn't realise he'd punctured his lung until two days later, when he was having trouble breathing.

Discourse-relevant introductory words like *yes, in fact, however, moreover* should be separated with a comma. (Discourse-irrelevant introductory words like *so* (not meaning consequence) and *well* are to be deleted.)

Yes, the package should arrive tomorrow morning.

However, you may not be satisfied with the results.

Adverbs at the start of a sentence are always followed by a comma:

Frankly, I don't like it.
Slowly, he left the room.

Cases like the one below should be regarded as an apposition of *it* and a postponed phrase/clause, and a comma should be inserted. This usage of *it* is not to be regarded as superfluous deixis.

*We'd gone down to visit friends and I was really struck by how beautiful it was**,** looking out to the sea from Dunbar.*

Almost all punctuation rules named above can be broken whenever their observance would cause ambiguities, look funny or make the reader stumble.

5.3.2 Capitalization

Capitalization rules are to be observed in the entire output m-layer text, especially:

- **capitalizing at the start of a sentence**
- **capitalizing in proper names.**

Example:

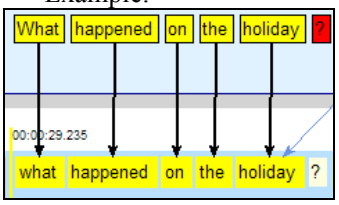

Figure 23

5.3.3 Transcription of non-alphabetic tokens

5.3.3.1 Numbers and digits

Numeric information is recorded as words at w-layer. M-layer is supposed to capture them as usual in standard writing. This section gives general guiding on how to write numeric information as well as advice for a few specific contexts.
Generally, cardinal numbers up to and including *100* **except in dates, years, measuring units, money sums and in mathematical contexts like equations and formulas** are spelled in words. So are ordinal numbers.

Two-digit cardinal as well as ordinal numbers are hyphenised when spelled in words: *twenty-one, twenty-first,* not **twentyone*, **twenty one,* etc. Each digit and the hyphen are represented by one m-node, respectively.

Example

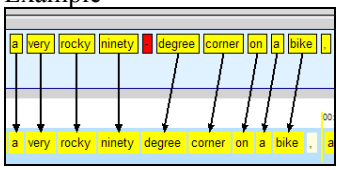

Figure 24

Even higher numbers are spelled in words when acting as premodifiers. The hyphen between the last number and the premodifying noun is to be omitted (generally in English!!!): (*a twenty-five day limit*, not **a 75-day-tour, a five-day limit*). The hyphenisation is ignored in many-digit numbers: *three hundred and sixty degree turn*. Then, each token corresponds to one m-node; e.g. *three hundred and fifty-five* corresponds to 6 m-nodes. Note that fractions in premodifications follow their own rules: *a three-quarter majority* but a *two thirds majority*.

Thousands are separated by space and never by comma (only one of the three possible spellings – space, comma, period - is to be used), even in financial contexts: *£ 100 000*, not **£ 100,000*. The entire number corresponds to one m-node and the space is to be preserved in the value of the attribute "form".

Amounts of money are always spelled as numbers, **the currency always preceding the number**: *£ 2*, not **2£,*2 £*. The currency and the amount correspond each to one m-node, respectively.

Decimal numbers are spelled with a period (*zero point seven five* ➞ *0.75*, not **0,75*). The entire decimal number corresponds to one m-node.

Dates are to be written with **ordinal numbers preceding the name of the month, no comma preceding the year number** (though there are many other correct variants): *the third of June* or *June the third nineteen oh seven* ➞ *3rd June 1907*. The ordinal number

consisting of one or two digits and the appropriate ending is spelled together and corresponds to one m-node.

Clock time encoding depends on the speaker's formulation: expressions containing *past/to* are spelled in words **including** the hour number: *quarter past two* ➞ *quarter past two, quarter to ten* ➞ *quarter to ten, half past eight* ➞ *half past eight, twenty-five to one* ➞ *twenty-five to one*. Also expressions containing *o'clock* are spelled entirely in words: *five o'clock* ➞ *five o'clock*. Expressions containing only numbers and the expression *zero hour* are written as digits, with hours separated from minutes and minutes from seconds by a colon: *two fifteen* ➞ *2:15, seventeen ten* ➞ *17:10*. Each number and the colon are represented as one m-node, respectively.

Fractions are to be written as fractions, not with decimal digits: *two and a half* ➞ *2½*, not *2.5*. The entire fraction corresponds to one m-node.

Names of **decades** when not meaning someone's age are spelled as digits: *in the early nineteen-eighties* ➞ *in the early 1980s* (with *1980s* in one node) but: *a woman in her early thirties* ➞ *a woman in her early thirties*.

Numbers in **appositive** use are written in numbers. The pronounced word "number" is omitted unless it is part of a proper name: *line number nine* ➞ *line 9* but *two drops of Chanel No. 5.* Roman numbers are used in ruler's appositional indexes: *Henry the eighth* ➞ *Henry VIII*. Note that common nouns like *section, bloc, figure, table, type*, etc. start acting like proper nouns (no article inserted and capitalization) when they get an appositive numeral: *Can I see Number 8/Room 104/Section 1.2/Chart 3a?*

Results of sports matches etc. correspond to three m-nodes and are written in digits separated by colon. Even the m-node for colon refers to the corresponding w-node.

Example:

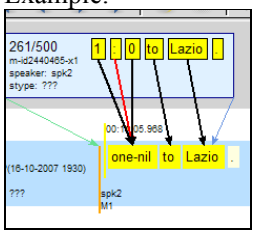

Figure 25

5.3.3.2 Other non-alphabetic tokens

Units of measurement are usually encoded with their abbreviations and symbols[6] when not in premodification: *a three meter gap* ➞ *a three-meter gap*, *in two thousand five hundred meters* ➞ *in 2500 m*, *fifty percent of women* ➞ *50 % of women*. When a unit of measurement is used with a natural number smaller than 100, which would normally be spelled in words, the numbers should be written in digits: *other three meters* ➞*other 3 m*.

The number and the unit of measurement correspond each to one m-node, respectively.

Mathematical operators are encoded with their symbols:

equals ➞ =

divided by ➞ ÷

minus ➞ -

plus ➞ +

times or *multiplied by* ➞ ×

the (square) root of ➞ √

5.3.4 Foreign expressions and proper names

When the speaker uses a foreign expression that is not common in English or its original spelling is still respected (e.g. *faux pas*) the spelling in the original language is preserved even if e.g. an English ending is added.

Example:

they were drinking retsina. ➞ *They were drinking retsina*.

the shop specialized in oolongs. ➞ *The shop specialized in oolongs*.

two geisha appeared. ➞ *Two geisha appeared*.

Meta-language: Sometimes the speaker's pronunciation of a certain foreign word is the subject of the utterance; e.g. someone makes a comment on someone else's bad pronunciation or the word's unusual form etc. Then the word should be spelled phonetically and enclosed by double quotes. This should also be noted in the annotator's comment of type `metalanguage`. Actually, foreign expressions are supposed to be spelled correctly already in w-layer input (the transcription).

[6] Here the speech-reconstruction manual for English differs from Czech, where the units of measurement are spelled in words. They appear to be generally spelled in words in Czech, and they are declined (which is the possible reason why they are commonly spelled in words).

5.3.5 Ad-hoc and unknown words

Words (except proper nouns) that are not commonly used in a real-world language community, occasional words and that like are to be enclosed by double quotes (the example has been taken from the novel *Clockwork Orange* by Anthony Burgess).

Example:

Alex and his droogs indulge in ultraviolence, dratsing and tolshocking and kicks in the yarblockos. ➞ *Alex and his "droogs" indulge in "ultraviolence", "dratsing" and "tolshocking" and kicks in the "yarblockos".*

5.3.6 Abbreviations

The entire abbreviation corresponds to a single node (unlike in spelled words, see 5.3.7, Spelled words) and the abbreviation should be written correctly, i.e. with correct capitalization.

Example:

IBM ➞ *IBM* *ai be em* ➞ *IBM*

5.3.7 Spelled words

The letters of spelled words are written as capital letters separated with spaces (not as phonetically transcribed spelling syllables, e.g. *b*, not *beeh* or anything like that). Each letter corresponds to one m-node.

Example:

my name is John. j o h n. ➞ *J O H N.*

5.3.8 Slips of the tongue

Slips of the tongue are corrected at m-layer.

Examples:

here comes the loconotive. ➞ *Here comes the locomotive.*

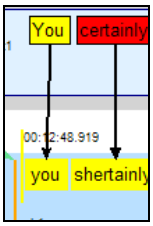

Figure 26

5.4 Substantial modifications

The most important modifications are the so-called **substantial modifications** of the transcribed input text, which effect significant differences between the transcription and the reconstructed text.

There are the following types of substantial modifications:

- **Modifications of lexical units** (see 5.4.1, Modifications of lexical units):
  - deletion (see 5.4.1.1, Deletion)
  - insertion (see 5.4.1.10, Insertion)
  - substitution (see 5.4.2, Substitution)
  - word order changes (see 5.4.3, Word order changes)

- **Annotation of discourse-relevant non-speech events** (see 5.5, Annotation of discourse-relevant non-speech events).

NB! The respective modification types are illustrated on isolated sentence segments. The actual annotation requires considering the context of the entire reconstructed text.

The modification terminology is based on the reconstruction process going in the direction from the input transcribed text at w-layer towards the output reconstructed text at m-layer.

5.4.1 Modifications of lexical units

This section describes the modifications that affect primarily lexical units, i.e. mainly the relations between tokens at w-layer (w-nodes of type `w`) and lexical units at m-layer (m-nodes of type `m`).

5.4.1.1 Deletion

The reconstructed text contains only discourse-relevant lexical units. They are lexical units that carry meaning and contribute to the content of an utterance. Discourse-irrelevant lexical units as well as entire text spans coming with the input transcription are identified and deleted from the reconstructed m-layer output.

A w-node (of type `w`) representing a discourse-irrelevant lexical unit has no corresponding node at m-layer.

M-layer contains no reference to a w-node (of type `w`) that represents a discourse-irrelevant lexical unit at w-layer.

Mainly the following lexical units are regarded as discourse-irrelevant:

- filler words (see 5.4.1.2, Filler words)
- filler phrases (see 5.4.1.3, Filler phrases)
- superfluous deictic words (see 5.4.1.4, Superfluous deictic words)
- superfluous connectives (see 5.4.1.5, Superfluous connectives)
- superfluous and wrong function words (see 5.4.1.6, Superfluous or wrong function words)
- reparandums and interregnums (see 5.4.1.7, Reparandums and interregnums)
- repetitions (see 5.4.1.8, Repetitions)
- fragments (see 5.4.1.9, Fragments)

5.4.1.2 Filler words

Filler words are semantically empty lexical units. The speaker uses them when hesitating what to say or searching for the right word. Also parasitic words are fillers. Tag questions are **not** regarded as fillers.

Typical fillers: *right, basically, well, so, yeah, okay, like, kind of/sort of*[7], etc.

Example:

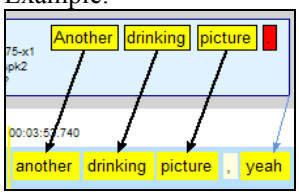

Figure 27

NB: *yeah, okay* etc. are **not** fillers when they primarily act as responses to yes/no questions.

[7] Sometimes *kind of/sort of* really means a comparison. Then it of course should be preserved. The rule of the thumb is that it should be preserved when an unusual combination of collocates would occur. Cf: *It was Macleods who built that castle. My friend is a Macleod and we ~~sort of~~ went there to have a look around.* (=*We went there to have a look around*. Omission is OK.) × *The castle sort of sits on its own on an island.* (Castles usually do not sit. Omission gives a funny sentence.) The decision is a subtle one, and it is clearly a matter of personal judgment.

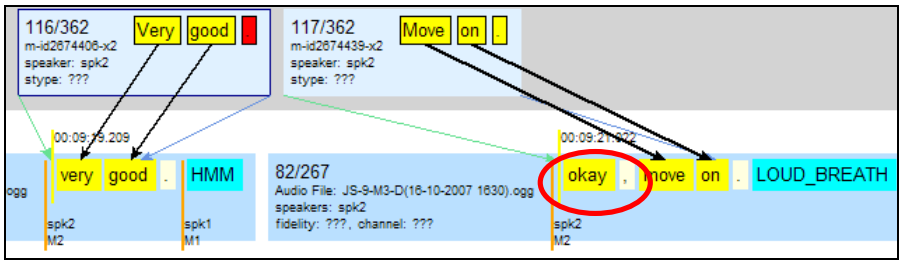

Figure 28

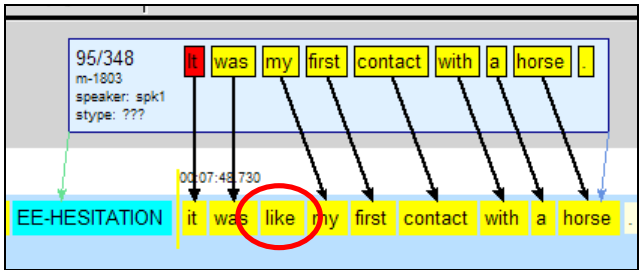

Figure 29

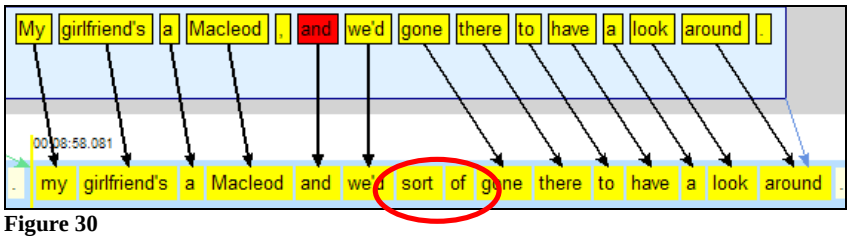

Figure 30

5.4.1.3 Filler phrases

Under filler phrases we understand parenthetical constructions like *I think*, *God beware*, *see*, *the hell*, etc. They can be preserved when they do not destroy the structure of the sentence (a decision up to the individual judgment of the annotator).

Examples:

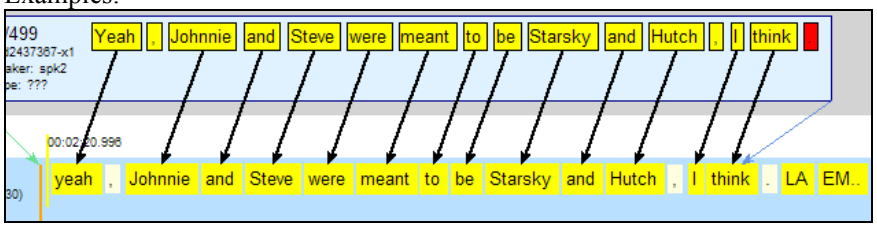

Figure 31

Cf:

yeah, Johnnie and Steve I think were meant to be Starsky and Hutch. ➞ *Yeah, Johnnie and Steve were meant to be Starsky and Hutch.*

5.4.1.4 Superfluous deictic words

Deictic words are superfluous and should be deleted when preceding a noun without denoting a given named entity (selecting it from a set of similar entities etc.). Then it is entirely used either because the speaker is hesitating before pronouncing the name or indicating that the given entity has been mentioned before.

Examples:

I told that Mary that I did not want to. ➞ *I told Mary that I did not want to*.

5.4.1.5 Superfluous connectives

And at the beginning of a sentence as false coordination with preceding text and *so* as a false consequence of preceding text are the most common superfluous connectives. *And* is often to be deleted. Connectives with a more specific meaning than just coordination, especially *but*, *or*, *therefore* etc. should generally be preserved if they are not obviously discourse-irrelevant. *So* is a slightly tedious case. W-layer annotators (transcribers), who are familiar with the original audio input, were asked to make a difference between *so* as a filler at the beginning of the sentence and *so* as a connective meaning consequence. The occurrences of *so* that the transcribers classified as fillers are followed by a comma. They are supposed to be deleted at m-layer. The m-layer annotators do not have to respect their decisions, but they should take them into consideration.

Generally, a sentence starting with a connective should be merged with the preceding sentence if the resulting sentence is not too long.

Example

I normally take photographs of people or stuff happening. But this is one of these things that you just, when you catch a feeling somewhere, you just take a few snapshots of it. ➞ *I normally take photographs of people or stuff happening, but this is one of these things that you just you just take a few snapshots of when you catch a feeling somewhere.*

And is to be deleted where superfluous in enumerations and usually at the beginning of questions. *So* is **not** to be deleted at the beginning of questions if it is not obviously a hesitation word.

Examples:

Me and John and Mary ➞ *Me, John and Mary*

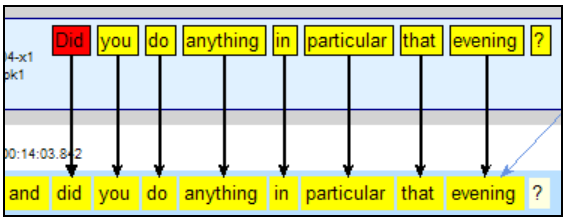

Figure 32

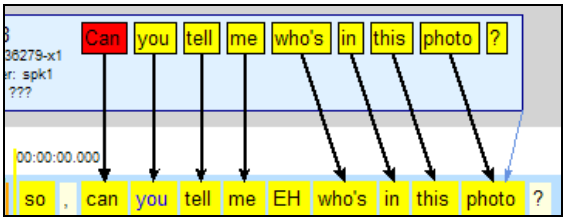

Figure 33

5.4.1.6 Superfluous or wrong function words

Sometimes the speaker can happen to use too many (or wrong) auxiliary verbs, determiners or connectives.

Examples:

this is me and Jess and Tim and Lucy. ➞ *This is me, Jess, Tim and Lucy.*

he was went there. ➞ *He went there.*

they were approaching to the house. ➞ *They were approaching the house.*

I will go there on the Sunday. ➞*I will go there on Sunday.*

him Robert I don't like. ➞ *Robert I don't like.*

5.4.1.7 Reparandums and interregnums

Sometimes the speaker changes his mind in the middle of an utterance and decides to reformulate it, abandoning the original start and restarting the speech event (or the entire utterance) in a different way. The original start is called **reparandum**. The restart is often (but not necessarily) preceded by the so-called **interregnum**. Interregnum is a speech event that indicates that the preceding utterance is going to be abandoned, and it introduces the restart. It can be verbal (*sorry, no*, etc.) or non-verbal (*erm, eh*, etc.). The restart is a correction of the reparandum. Both reparandums and interregnums are deleted at m-layer.

Example:

I want two no sorry three pieces of this.
Reparandum: *two*

Interregnum: *no, sorry*
Restart: *three*

➜ *I want three pieces of this.*

There are several types of events that effect a reparandum:

a. **stammer**.

Example:

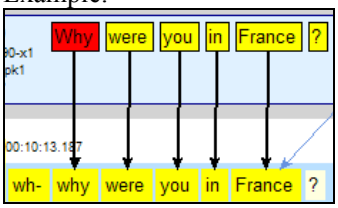

Figure 34

b. **repeating the same word - hesitation**.

Examples:

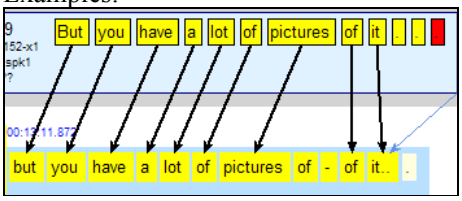

Figure 35

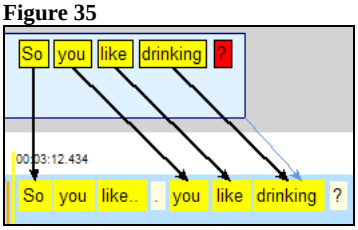

Figure 36

c. **repeating the same word - emphasis**.

Example:

That was really really fun. ➜ *That was really fun.*

d. **correction**.

Single words as well as larger text spans can be corrected. Sometimes a reparandum can also be regarded as a fragment of a quite different utterance.

Example:

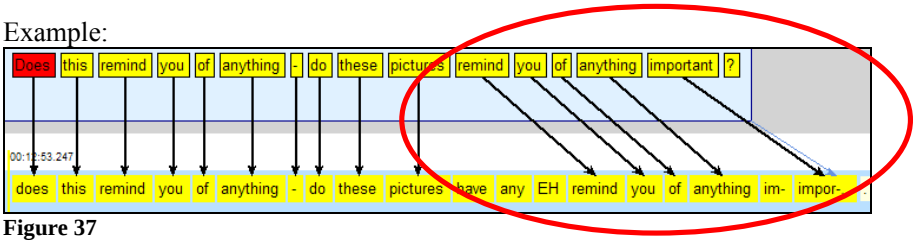

Figure 37

5.4.1.8 Repetitions

Repeated text spans in which the repetition does not have any special meaning for the discourse are deleted at m-layer.
NB: Rephrasing of a sentence in the following sentence is not regarded as a repetition.

Example:

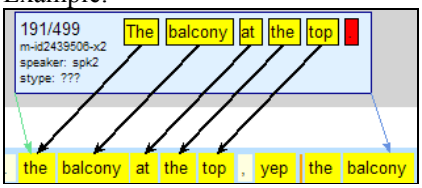

Figure 38

Cf. a case of rephrasing:

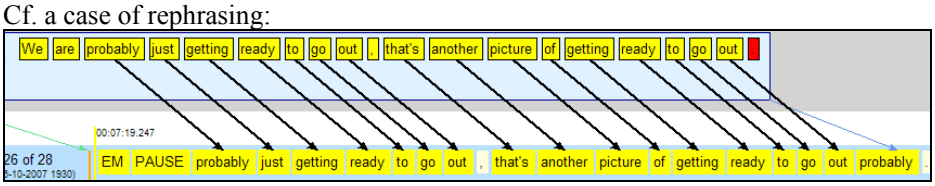

Figure 39

5.4.1.9 Fragments

A **fragment** is a text span (one or several autosemantic words) that remained incomplete and is not further referred to in the following text. Fragments are to be held apart from incomplete sentences (see 4.3.3, Discourse-relevant fragments). A good rule of thumb is that in incomplete sentences one can assume what sort of information has remained unsaid while in fragments one cannot.

Example:

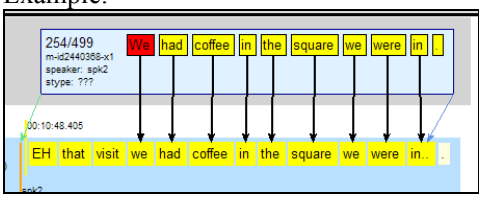

Figure 40

5.4.1.10 Insertion

The reconstructed text can even contain lexical units that have not been pronounced (they do not occur at w-layer) but they 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 m-layer.

M-layer can contain (inserted) m-nodes (of type `m`) that represent lexical units missing at w-layer.
The m-node (of type `m`) that represent a lexical unit missing at w-layer does not contain any reference to w-layer.

The inserted nodes typically stand for:

- missing function words (see 5.4.1.11, Missing function words)
- unexpressed autosemantic words (see 5.4.1.12, Missing autosemantic words).

NB: Even punctuation is inserted as m-nodes. See 5.3.1, Punctuation.

5.4.1.11 Missing function words

Missing function words are to be inserted. Function words are most often missing at the beginning of a sentence or a clause.

Some cases of such reduction are acceptable in spoken language but must be corrected to meet the written standard, e.g. the definite article missing in front of *same*: *same holiday* ➞ *the same holiday*

Examples:

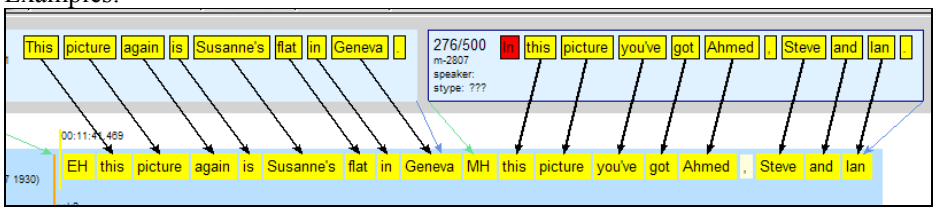

Figure 41

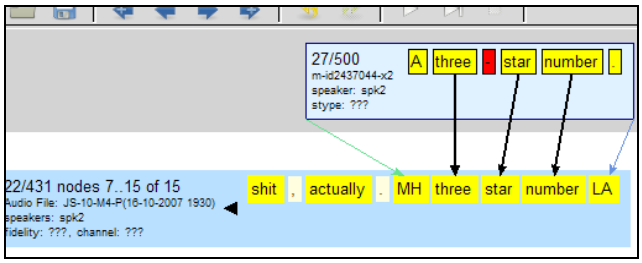

Figure 42

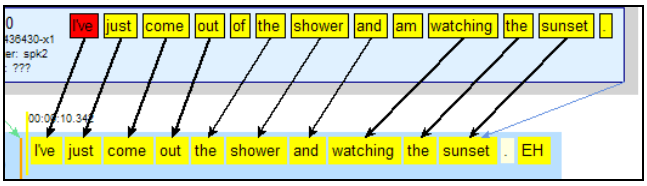

Figure 43

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 (*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 be grammatically concord with the subject. As a rule, when an auxiliary verb is omitted, its subject is missing, too. See 5.4.1.12, Missing autosemantic words to learn more about determining the morphosyntactic categories of missing personal pronouns.

5.4.1.12 Missing autosemantic words

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. The first example does not presuppose any non-verbal knowledge:

Example:

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.*

In the second example (below), two words are to be inserted to restore the sentence, one (the finite verb) being restorable without any knowledge of the non-verbal context, but one actually presupposing that the annotator has seen the picture to be able to decide between the plural and the singular of the inserted first person pronoun[8].

[8] The issue of pronoun restoration will be subject to changes as soon as the input text comes interlinked with the photographs.

Example:

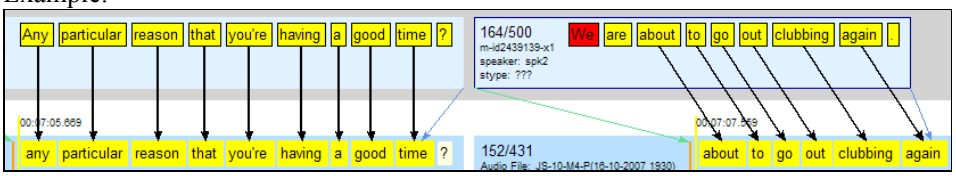

Figure 44

The case of the missing verb shows that apparent prepositions are in fact often particles belonging to omitted phrasal verbs. The sentence ought then to be restored by completing the finite part of that particular phrasal verb:

About to go out clubbing ➞ *We **are about** to go out clubbing.*

The restoration of missing subjects is somewhat tricky as it requires the knowledge of the photographs, which are not attached to the input text. At present, the annotator can insert the best-fitting pronoun according to his personal judgment. However, he should not be inconsistent within a span that clearly focuses one picture; e.g. the inserted pronouns should not imply that the speaker was and was not in the picture at the same time. In many cases, the switch of the photograph is obvious from the text.

Missing subjects are restored in finite verb forms:

A: What happened on that holiday?

B: Got drunk... eh... [laughter] *sunbathed.* ➞ *I got drunk and sunbathed.*

Sometimes the answer is syntactically governed by an *–ing* form. An *–ing* form at this position can be interpreted as:

- present participle with elided auxiliary verb (this can be identified by the progressive verb tense in the question):

  A: UH - then what are you doing?

  B: Playing cards./Being silly. ➞ *I am playing cards./ I am being silly.*

  Here, both the most likely subject and the auxiliary verb are inserted.

- nominalization:

A: And what's the next picture of?

B: EH the next picture is go-karting in France. ➞ *The next picture is go-karting in France.*

A: Then what was the reason for - was there a special event? ➞ *Then what was the reason for it - was there a special event?*

B: I think it was probably just another Saturday night, going out. ➞ *I think it was probably just another Saturday night, going out.*

A: What's this? ➞ *What's this?*

B: Opening of a new gallery. ➞ *Opening of a new gallery.*

Nothing is to be inserted.

- a participial clause[9].

A: This one is in Majorca... ➞ *This one is in Majorca.*

B: UH. ➞ *UH.*

A: Our holiday picture. ➞ *Our holiday picture.*

A: EM, drinking beer again [laughter] *strangely enough.* ➞ *Drinking beer again, strangely enough.*

A: Is there any particular reason that you like this photo? ➞ *Is there any particular reason that you like this photo?*

B: Just having a beer, having a good time. ➞ *Just having a beer, having a good time.*

A: And wh- what was happening in that picture? ➞ *What was happening in that picture?*

B: Watching TV. Making some kind of comment on the TV. ➞ *Watching TV. Making some kind of comment on the TV.*

So far nothing is inserted as the annotators cannot see the pictures and the context itself does usually not provide enough clues to disambiguate the unexpressed subjects.

Komentář [SC1]: Access to the pictures ought to be gained, otherwise no co-reference annotation will be possible in these positions.

5.4.2 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 speech reconstruction the input word forms are checked and corrected whenever appropriate. When an obviously wrong word was used the annotator is supposed to replace it.

[9]The difference between a nominalization and a subjectless participial clause is that the object of the event expressed by the *–ing* form of a transitive verb is introduced by the preposition *of* in nominalizations (see the previous example) while it is prepositionless in participial clauses. The default interpretation of *–ing* forms in intransitive verbs is that they are nominalizations. The difference between a nominalization and a subjectless participial clause is kept in case the annotators gain access to the pictures the speakers are discussing. Subjects of the recognized participial clauses would be inserted.

The form and the lemma of an m-node (of type `m`) does not necessarily correspond to the token of the corresponding w-node (of type `w`).

5.4.2.1 Form change in a lexical unit

Non-standard and incorrect word forms as well as typing errors of the transcribers are corrected at m-layer.

The following forms are regarded as non-standard and should be replaced with standard forms:

- *'em* ➞ *them*
- *a-* in front of a present participle should be removed; e.g: *a-goin'* ➞ *going*
- *-n'* at the end of a present participle should be replaced with *-ng*; e.g.: *goin'* ➞ *going*
- *'cos, 'cause* ➞ *because*
- *'cept* ➞ *except*
- *'deed* ➞ *indeed*
- *'tween* ➞ *between*
- *o'* ➞ *of* (not in *o'clock* and in names like *O'Hara*)
- *gotta, gonna, wanna* ➞ *got to, going to, want to*
- *c'mon* ➞ *come on*
- *d'you* ➞ *do you*

etc.

On the other hand, the common contracted forms of negator and auxiliaries are regarded as standard and should not be replaced with the full forms. The contracted negator (*n't*) should be represented by an extra node. The auxiliary form, when reduced, is not to be completed:

can't ➞ *ca* *n't*
won't ➞ *wo* *n't*
shan't ➞ *sha* *n't*
ain't ➞ *ai* *n't*
don't ➞ *do* *n't*
etc.

Similarly, the contracted forms *'d, 'll, 'm, 's, 're, 've* should not be replaced by the full forms but are to be represented by an extra node (including the apostrophe):

she'd ➞ *she* *'d*

I'm ➞ *I* *'m*

we've ➞ *we* *'ve*

you're ➞ *you* *'re*

let's ➞ *let* *'s*

The genitive 's is also to be represented by its own node, and so is the apostrophe alone when indicating the genitive of a word ending with *-s, -z, -x*:

Linda's paper ➞ *Linda* *'s* *paper*

Max' dog ➞ *Max* *'* *dog*

citizens' preferences ➞ *citizens* *'* *preferences*

Eventually, *'n'* is to be represented by an extra node:

fish 'n' chips, fish'n'chips ➞ *fish* *'n'* *chips*

rock'n'roll ➞ *rock* *'n'* *roll*

In the speech reconstruction of the NAP/AAA dialogs special attention must be paid to **subject-verb number concord**. Grammatical concord is an area where (formal) written text and spontaneous speech show significant differences. A grammatically inadequate form is to be substituted with the form that renders the appropriate morphosyntactic categories: *we was there* ➞ *we were there.*

The spoken language is more affected by the so-called proximity principle: the verb number agrees with the number of the immediately preceding noun rather than with the actual clause subject, especially when the subject is rendered by an indefinite pronoun or an indefinite numeral. E.g. sentences like the following are acceptable in informal English (Quirk et al., 2004):

Lots of stuff is going to be thrown away.
A lot of papers are going to be thrown away.

Some frequent cases found in the data, though, cannot even be explained by the proximity principle:

There was lots of people out.

To make things simple, all subject-verb number disagreements except the semantically conditioned ones (e.g. in collective nouns - *the police were busy* etc.) are to be corrected at m-layer.

Examples:

Lots of stuff is going to be thrown away. ➞ *Lots of stuff are going to be thrown away.*

A lot of papers are going to be thrown away. ➞ *A lot of papers is going to be thrown away.*

There was lots of people out. ➞ *There were lots of people out.*

5.4.2.2 Lemma change in a lexical unit

Lemmas are changed in the following cases:

- sometimes the speaker uses a paronym (a phonetically but not semantically similar word) of the word he intended to use, or a partial synonym which is not appropriate for the given context: *I am sorry, I didn't mean to consult you.* ➞ *insult you*. *The house had a good isolation* ➞ *insulation*.
- incorrect valency: *typical for* ➞ *typical of*
- the current lemma causes confusion in co-reference. Either a wrong pronoun was chosen or it refers too far back in the discourse and the antecedent becomes too difficult to trace back.

Both the attribute "form" and the attribute "lemma" are edited.

Note: expressive and vulgar words are never substituted with stylistically neutral words.

Example:

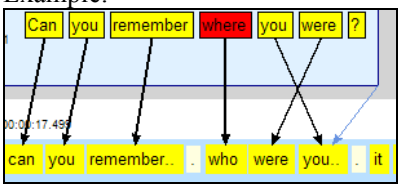

Figure 45

5.4.2.3 Substitution of an unintelligible text span with deduced text

Unintelligible text spans (represented by the w-nodes of type `nonspeech` at w-layer marked as `unintelligible`) are to be reconstructed whenever possible, at least by means of very general wording. All inserted m-nodes (of type `m`) that represent the deduced text contain references to the mark `unintelligible` at w-layer.

Example:

m-layer: *Have you faced such problems before?*
w-layer: *Have you faced* `<unintelligible>` *problems before?*

When the missing text is impossible to deduce, the rules listed in 5.5, Annotation of discourse-relevant non-speech events, are to be observed.

5.4.3 Word order changes

All sentences at m-layer must have a correct word order that makes the entire discourse fluent.

The order of the nodes at m-layer does not have to correspond to the order of the nodes at w-layer.

Example:

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

Topic-focus-motivated word order shifting like fronting is 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...*

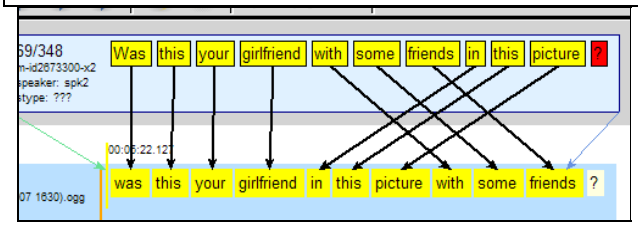

Figure 46

5.5 Annotation of discourse-relevant non-speech events

The reconstructed text is primarily not supposed to contain non-speech events. Discourse-irrelevant non-speech events are therefore deleted without compensation (see 5.2, Deletion of discourse-irrelevant non-speech events).

Discourse-relevant non-speech events are primarily captured at m-layer by the means typical of written text, e.g. punctuation.

The following means are used for marking discourse-relevant non-speech events:

- exclamation mark (emphasized sentences)
- dash (longer pause)
- single quotes enclosing the word (irony in a word). In a multi-word expression only the first word has left single quotes and the last word right single quotes.
- word order, topic-focus articulation (emphasis on a word)

Example

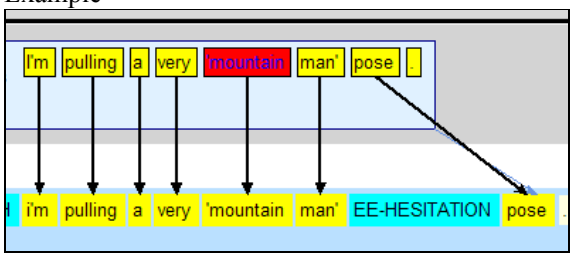

Figure 47

There are, however, other discourse-relevant non-speech events that cannot be captured by the means used in written text, e.g.: ironic laughter, whisper, sudden voice raising etc.

Discourse-relevant non-speech events can also be captured by a special type of m-node (type `nontext`).

The m-node (type **`nontext`**) has the attribute `type`, in which the annotator gives a description of the non-speech event he considers to be discourse-relevant in his/her own wording.

Examples of values of the attribute `type`:

jiggling
hesitating (silence)
probably nodding
started to whisper
gave a whistle

The m-node of type `nontext` is always part of a sentence (though it may be the only node in an s-element)

If the corresponding non-speech events are recorded at w-layer (w-nodes of type `nonspeech` and `background_begin`), they are referred to from the relevant m-nodes of type `nontext`.

Even **unintelligible text spans** can be discourse-relevant. An unintelligible text span recorded at w-layer by a w-node of type `nonspeech` with the value `unintelligible` in attribute `desc` is primarily to be replaced by deduced text (see 5.4.2.3, Substitution of an unintelligible text span with deduced text). When the substitution is impossible (the text cannot be deduced) it 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.

5.6 Meta-language

Words are used as meta-usage when it is their meaning, form or sound that is the actual subject of the discourse. The word (or other lexical unit) used as meta-language is typically introduced by nouns that indicate meta-language use: inscription, word, text, question, notion, term, expression, utterance, meaning, sense etc. Also verbs like *mean, denote, render, express, pronounce* and *spell* indicate meta-language.

Words in meta-language use are usually enclosed by single quotes and marked with the annotator's comment `metalanguage` (no text is needed as the attribute value). In phrases, clauses and sentences the annotator's comment is only added to the syntactically governing node of the entire cluster.

Examples:

The word "robot"`[annotator's comment = metalanguage]` *comes from Czech.*

5.7 Transcription errors

The reconstruction makes a difference between transcription errors and speakers' errors. While speakers' errors are corrected by the reconstruction at m-layer, transcription errors ought to be removed straight at w-layer. As the m-layer annotator cannot alter w-layer himself he is supposed to mark the error by the annotator's comment of type **`w-token`**. He treats the node as correct. Errors in nodes deleted at m-layer are to be noted in the comment of type **`w-token`** in the next m-node. See also 5.8, Annotator's comment.

Example:

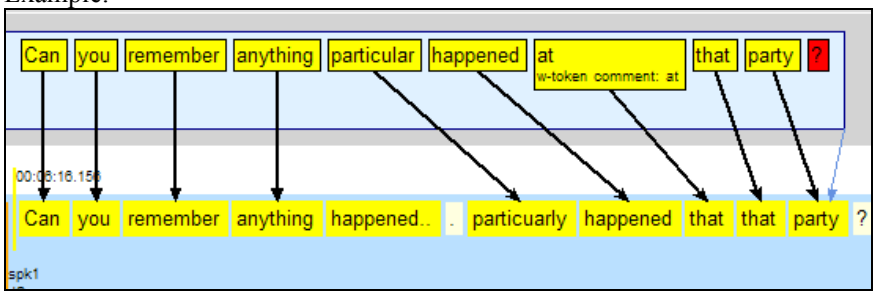

Figure 48

5.8 Annotator's comment

For technical reasons, each m-node contains the optional attribute `comment,` which is meant for recording of different comments that concern the annotation. There are four types of comments to choose from:

- **`metalanguage`**: (see 5.6, Meta-language)
- **`form:`** uncertainty in the lemma or form. Used especially with foreign expressions and ad-hoc words. Free text allowed.
- **`w-token`**: error at w-layer. See 5.7, Transcription errors. Free text required.
- **`other:`** other comment. Free text required.

References

- Silvie Cinková, Jan Hajič, Marie Mikulová, Lucie Mladová, Anja Nedolužko, Petr Pajas, Jarmila Panevová, Jiří Semecký, Jana Šindlerová, Josef Toman, Zdeňka Urešová, Zdeněk Žabokrtský: Annotation of English on the tectogrammatical level. Tech. Report 35 UFAL MFF UK. Prague, 2006.
- Comma writing: URL< http://owl.english.purdue.edu/>, quoted 2008-01-17
- Jan Hajič, Jarmila Panevová, Eva Buráňová, Zdeňka Urešová, Alla Bémová (in cooperation with Jiří Kárník, Jan Štěpánek, Petr Pajas): Annotations at analytical level. Instructions for annotators. UFAL MFF UK, Prague, 1999.
- Jan Hajič, Marie Mikulová, Martina Otradovcová, Petr Pajas, Petr Podveský, Zdeňka Urešová: *Prague Dependency Treebank of Spoken Czech. Speech Reconstruction.* ÚFAL/CKL Technical Report TR-2006-33, ISSN 1214-5521. Prague, 2006 (in Czech, original title: Pražský závislostní korpus mluvené češtiny. Rekonstrukce standardizovaného textu z mluvené řeči).
- Pavel Ircing: COMPANIONS. Rules for annotation of English audio recordings using the Transcriber software. Version 1, 5. 10. 2007. Západočeská univerzita v Plzni. Unpublished.
- Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský: Annotation on the tectogrammatical level in the Prague Dependency Treebank. Annotation manual. Tech. Report 30 ÚFAL MFF UK. Prague, 2006.
- Marie Mikulová: *Speech Reconstruction in the Prague Dependency Treebank of Spoken Czech. Annotation Manual.* Unpublished manuscript, 2007. (in Czech, original title: Rekonstrukce standardizovaného textu z mluvené řeči v Pražském závislostním korpusu mluvené češtiny. Manuál pro anotátory.)
- Randolph Quirk, Sydney Greenbaum, Geoffrey Leech and Jan Svartvik: A Comprehensive Grammar of the English Language. Longman, 2004, first published 1985.

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, along with GA-CR 405/06/0589, ME838 and MSM 0021620838.

.

Index

M

N

O

P

Q

R

S

T

T

U

V

W

Y

Z