Syntactic annotation of transcriptions in the Czech Academic Corpus: Then and now

Barbora Hladká and Zdeňka Urešová

Charles University in Prague

Institute of Formal and Applied Linguistics

{hladka, uresova}@ufal.mff.cuni.cz

Abstract

Corpus annotation plays an important role in linguistic analysis and computational processing of both written and spoken language. Syntactic annotation of spoken texts becomes clearly a topic of considerable interest nowadays, driven by the desire to improve automatic speech recognition systems by incorporating syntax in the language models, or to build language understanding applications. Syntactic annotation of both written and spoken texts in the Czech Academic Corpus was created thirty years ago when no other (even annotated) corpus of spoken texts has existed. We will discuss how much relevant and inspiring this annotation is to the current frameworks of spoken text annotation.

1 Motivation

The purpose of annotating corpora is to create an objective evidence of the real usage of the language. In general, it is easier to annotate written text – speech must be recorded and transcribed to process it whilst texts are available "immediately"; moreover, written texts usually obey standard grammar rules of the language in questions, while a true transcript of spoken utterances often does not.

The theoretical linguistic research considers the language to be a system of layers (e.g. the Government and Binding theory (Chomsky, 1993), the Functional-Generative Description of the language (Sgall, Hajičová, Panevová, 1986)). In order to be a valuable source of linguistic knowledge, the corpus annotation should respect this point of view.

The morphological and syntactic layers of annotation represent a standard in today's text corpora, e.g. the Penn Treebank[1], the family of the Prague Dependency Treebanks[2], the Tiger corpus[3] for German, etc. Some corpora contain a semantic annotation, such as the Penn Treebank enriched by PropBank[4] and Nombank[5], the Prague Dependency Treebank in its highest layer, the Penn Chinese[6] or the Korean[7] Treebanks. The Penn Discourse Treebank[8] contains discourse annotation.

It is desirable that syntactic (and higher) annotation of spoken texts respects the written-text style as much as possible, for obvious reasons: data "compatibility", reuse of tools etc.

A number of questions arise immediately: How much experience and knowledge acquired during the written text annotation can we apply to the spoken texts? Are the annotation instructions applicable to transcriptions in a straightforward way or some modifications of them must be done? Can transcriptions be annotated "as they are" or some transformation of their inner structure into a written text structure must precede the annotation? The Czech Academic Corpus will help us to find out the answers.

2 Introduction

The first attempts to syntactically annotate spoken texts date back to the 1970s and 1980s when the Czech Academic Corpus – CAC (Králík, Uhlířová, 2007)[9] and the Swedish Talbanken (Nilsson, Hall, Nivre, 2005) appeared. Talbanken[10] was annotated with partial phrase structures and grammatical functions, CAC with dependency-based structures and analytical functions. Thus both corpora can be regarded as belonging to the pioneers in corpus linguistics, together with the paper-only “Quirk corpus” (Svartvik, Quirk, 1980); computerized later as the London-Lund Corpus).[11]

During the last twenty years the work on creating new treebanks has increased considerably and so CAC and Talbanken have been put in a different light, namely with regard to their internal formats and annotation schemes. Given that, transformation of them became necessary: while the Talbanken’s transformation concerned only the internal format, transformation of CAC concerned both internal format and annotation scheme.

Later, more annotated corpora of spoken texts have appeared, like the British Component of the International Corpus of English (ICE-GB, (Greenbaum, 1996))[12], the Fisher Corpus for English (Cieri *et al*., 2004), the Childes database[13], the Switchboard part of the Penn Treebank (Godfrey *et al*., 1992), Corpus Gesproken Nederlands (Hoekstra *et al*., 2001)[14] and the Verbmobil corpora.[15] The syntactic annotation in these corpora is mostly automatic using tools trained on written corpora or on a small, manually annotated part of spoken corpora.

The aim of our contribution is to answer the question whether it is possible to annotate speech transcriptions syntactically according to the guidelines originally designed for text corpora. We will show the problems that arise in extending an explicit scheme of syntactic annotation of written Czech into the domain of spontaneous speech (as found in the CAC).

Our paper is organized as follows. In Section 3, we give a brief description of the past and present of the Czech Academic Corpus. The compatibility of the original CAC syntactic annotation with a present-day approach adopted by the Prague Dependency Treebank project is evaluated in Section 4. Section 5 is the core of our paper. We discuss phenomena typical for spoken texts making impossible to annotate them according to the guidelines for written texts. We explore a trade-off between leaving the original annotation aside and annotating from scratch, and an upgrade of the original annotation. In addition, we briefly compare the approach adopted for Czech and those adopted for other languages.

3 The Czech Academic Corpus: past and present (1971-2008)

The idea of the Czech Academic Corpus (CAC) came to life between 1971 and 1985 thanks to the Department of Mathematical Linguistics within the Institute of Czech Language. The original purpose of the corpus was to build a frequency dictionary of the Czech language. The corpus has been morphologically and syntactically annotated manually.

The discussion on the concept of academic grammar of Czech, i.e. on the concept of CAC annotation, finally led to the traditional, systematic, and well elaborated concept of morphology and dependency syntax (Šmilauer, 1972). By the mid 1980s, a total of 540,000 words of CAC were morphologically and syntactically manually annotated.

The documents originally selected for the CAC are articles taken from a range of media. The sources included newspapers and magazines, and transcripts of spoken language from radio and TV programs, covering administrative, journalistic and scientific fields.

The original CAC was on par with it peers at the time (such as the Brown corpus) in size, coverage, and annotation; it surpassed them in that it contained (some) syntactic annotation. CAC was used in the first experiments of statistical morphological tagging of Czech (Hajič, Hladká, 1997). After the Prague Dependency Treebank (PDT) has been built (Hajič *et al.*, 2006), a conversion from the CAC to the PDT format has started.

The PDT uses three layers of annotation: morphological, syntactic and "tectogrammatical" (or semantic) layers (henceforth m-layer, a-layer and t-layer, respectively). The main goal was to make the CAC and the PDT compatible at the m-layer and the a-layer, and thus to enable integration of the CAC into the PDT. The second version of the CAC presents such a complete conversion of the internal format and the annotation schemes. The overall statistics on the CAC 2.0 are presented in Table 1.

Style	**Form**[16]	**#docs**	**#sntncs (K)**	**#tokens (K)**
Journalism	w	52	10	189
Journalism	s	8	1	29
Scientific	w	68	12	245
Scientific	s	32	4	116
administrative	w	16	3	59
administrative	s	4	2	14
Total	w	135	25	493
Total	s	44	7	159
Total	**w&s**	**180**	**32**	**652**

Table 1 Size of the CAC 2.0 parts

Annotation transformation is visualized in Figure 1. In the areas corresponding to the corpora, the morphological annotation is symbolized by the horizontal lines and syntactic annotation by the vertical lines.

Conversion of the originally simple textual comma-separated values format into the Prague Markup Language (Pajas, Štěpánek, 2005) was more or less straightforward.

Morphological analysis of Czech in the CAC and in the PDT is almost the same, except that the morphological tagset of CAC is slightly more detailed. Semi-automatic conversion of the original morphological annotation into the Czech positional morphological tagset was executed in compliance with the morphological annotation of PDT (Hana *et al.,* 2005). Figure 1 shows that morphological annotation conversion of both written and spoken texts was done. The only major problem in this conversion was that digit-only tokens and punctuation were omitted from the original CAC since they were deemed linguistically “uninteresting”, which is certainly true from the point of view of the original CAC’s purpose to give quantitative lexical support to a new Czech dictionary. Since the sources of the CAC documents were no longer available, missing tokens had to inserted and revised manually.

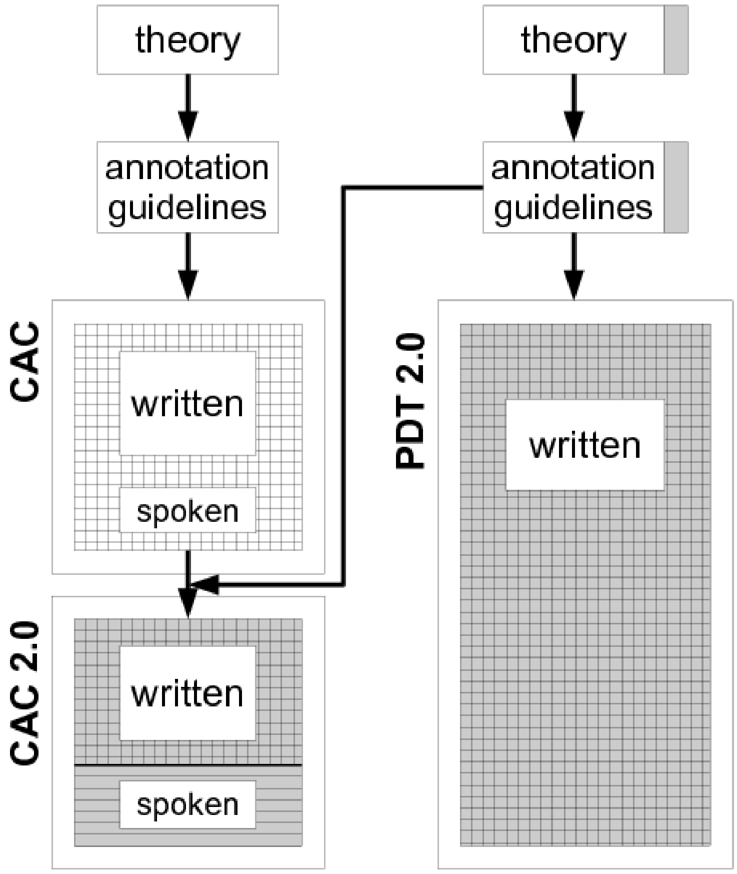

Figure 1 Overall scheme of the CAC conversion

Syntactic conversion of CAC was more demanding than the morphological one. In a pilot study, (Ribarov *et al*., 2006) attempt to answer a question whether an automatic transformation of the CAC annotation into the PDT format (and subsequent manual corrections) is more effective than to leave the CAC annotation aside and process the CAC's texts by a statistical parser instead (again, with subsequent manual correction). In the end, the latter variant was selected (with regrets). No distinction in strategy of written and spoken texts annotation transformation was made. However, syntactic annotation of spoken texts was eventually excluded from the CAC 2.0 (see the missing vertical lines in the spoken part of the CAC 2.0 area in Figure 1). Reasons for this are explained in detail in the following two sections.

4 Syntax in the CAC and the PDT

4.1 Syntactic annotation in the CAC

The syntactic analysis of Czech in the CAC and in the PDT is very much alike, but there are phenomena in which the CAC syntactic annotation scenario differs from the PDT, even though both corpora are based on the same linguistic theory (Šmilauer, 1969), i.e. on the dependency grammar notion common to the "Prague school" of linguists since the 1930s.

However, the syntactic annotation differs between the two corpora. The CAC works with a single syntactic layer, whereas the PDT works with two independent (although interlinked) syntactic layers: an analytical (syntactic) one and a tectogrammatical one (a-layer and t-layer, respectively). In this paper, we are referring to the a-layer of the PDT in our comparisons unless specifically noted for those elements of the tectogrammatical annotation that do have some counterpart in the CAC.

The CAC annotation scheme makes a substantial distinction between two things: surface syntactic relations within a single clause as well as syntactic relations between clauses in a complex sentence. These two types of syntactic information are captured by two types of syntactic tags (see Figure 2).

(a) Word-level (intra-clausal) syntactic tag is a max. 6-position tag consisting of digits and spaces assigned to every non-auxilliary ("autosemantic") word within a single clause, representing the intra-clausal dependency relations (e.g., the "5231 1" at the word "*je*").

(b) Clause-level (intra-sentential) syntactic tag is a max. 8-position tag assigned to the first token of each clause in a complex sentence, representing the status (and possible dependency) of the given clause within the given (complex) sentence ("911" at "*Tak*" in Figure 2).

The CAC thus annotates not only dependency relations within a single clause but also dependency relations within a complex sentence.

A description of the 6-position and the 8-position tags is given in Tables 4 and 5, respectively. Ribarov *et al*. (2006) gives a detailed description.

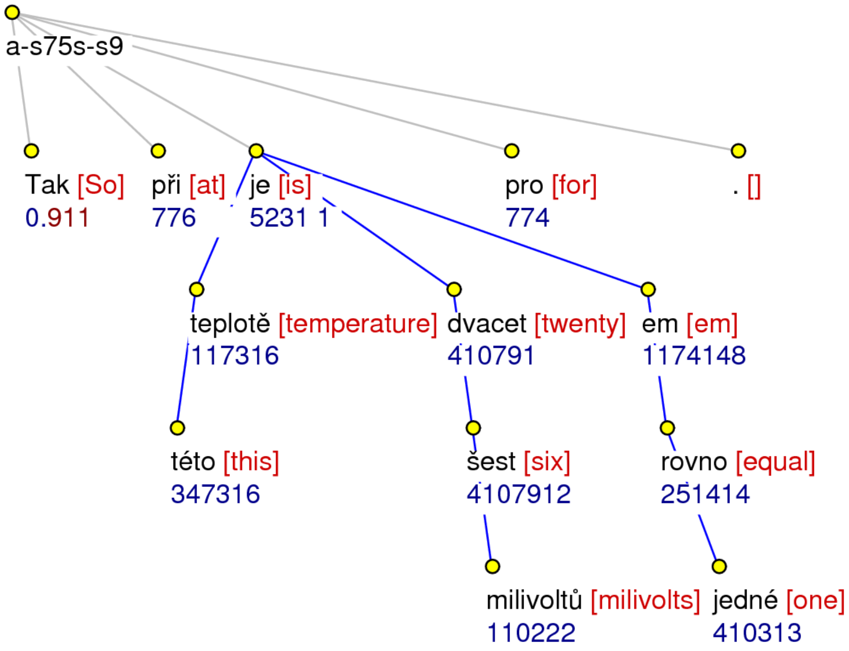

Figure 2 Syntactic structure and morphosyntactic tags in the original CAC

4.2 Syntactic annotation in the PDT

The PDT a-layer annotates two main things: a *dependency structure* of the sentence and *types* of these dependencies.

Representation of a structure of the sentence is rendered in a form of a dependency tree, the nodes of which correspond to the tokens (words and punctuation) that form the sentence. The type of dependency (subject, object, adverbial, complement, attribute, etc.) is represented by a node attribute called an "analytical function" (*afun* for short; the most frequent values of this attribute are listed in Table 6; see also Figure 3 for a full example of annotated sentence).

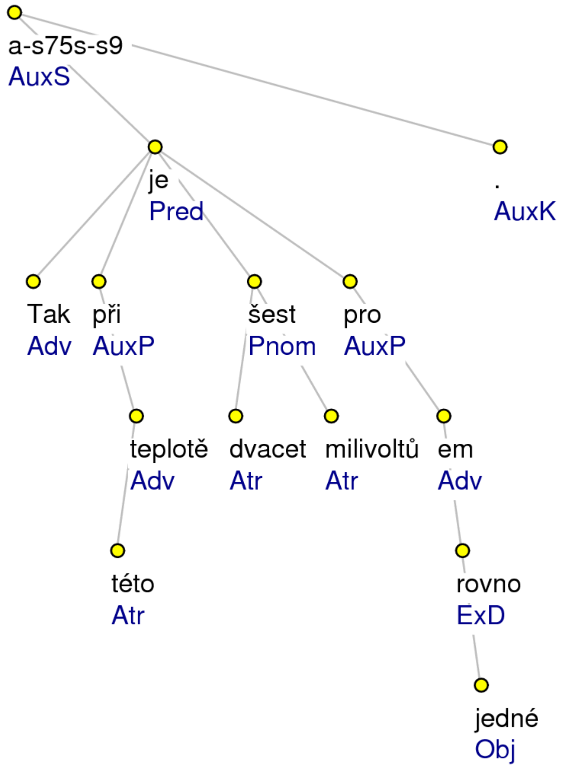

Figure 3 PDT-style annotation of the sentence from Figure 2

4.3 CAC vs. PDT

Comparing the CAC and the PDT syntactic annotation scenarios, we can see that the annotation of the major syntactic relations within a sentence is very similar, from similar adaptations of the theoretical background down to the high-level corpus markup conventions. For example, in both corpora the predicate is the clausal head and the subject is its dependent, unlike the descriptions we can find in the traditional Czech syntactic theory (Šmilauer, 1969). Another (technical) similarity can be found in the way the dependency types are encoded. In both corpora, the dependency type label is stored at the dependent. No confusion arises since the link from a dependent to its governor is unique.

However, the list of differences is actually quite long. Some are minor and technical: for example, in the PDT an "overarching" root of the sentence tree (marked AuxS) is always added, so that all other nodes appear as if they depend on it. Some differences are more profound and are described below.

We are not going to list all the differences in individual syntactic labels - they can be found easily by confronting Tables 4 and 5, but we would like to draw the readers' attention to the main dissimilarities between the CAC's and the PDT's syntactic annotation scenarios.

4.3.1 Punctuation

The first difference can be observed at first glance: in CAC no punctuation marks can be found (as mentioned in Section 3). While some might question whether punctuation should ever be part of syntax, in computational approaches punctuation is certainly seen as a very important part of written-language syntax and is thus taken into account in annotation (for important considerations about punctuation in spoken corpora, see Section 5).

4.3.2 Digits

CAC leaves out digital tokens, even though they are often a valid part of the syntactic structure and can plausibly get various syntactic labels as we can see in the PDT annotation, where nothing is left out of the syntactic tree structure.

4.3.3 Prepositions and function words

The next most significant difference is in the treatment of prepositions (or function words in general, see also the next paragraphs on conjunctions and other auxiliaries). Whereas CAC neither labels them nor even includes them in the dependency tree, PDT at the a-layer, reflecting the surface shape of the sentence, makes them the head of the autosemantic nodes they "govern" (and labels them with the AuxP analytical function tag). The CAC way of annotation (rather, non-annotation) of prepositions is, in a sense, closer to the annotation scenario of the underlying syntactic layer (the t-layer) of the PDT, It is also reflected in the adverbial types of labels (column 2 in Table 4) – these would all be labeled only as Adv at the (surface-syntactic) a-layer of the PDT, but at the (deep) t-layer, they get a label from a mix of approx. 70 functional, syntactic and semantic labels. Unfortunately, only seven such labels are used in the CAC, resulting in loss of information in some cases (adverbials of aim, accompaniment, attitude, beneficiary, etc.); the same is true for certain subtypes of time and location adverbials, since they are not distinguished in terms of direction, location designation (*on*/*under*/*above*/*next to* and many other), duration, start time vs. end time, etc.

4.3.4 Conjunctions

Further, subordinating as well as coordinating conjunctions get only a sentential syntactic tag in the CAC (if any), i.e. they are labeled by the 9-position tag but not by the word-level, intra-clausal syntactic tag. In PDT, subordinating and coordinating conjunctions get assigned the analytical function value AuxC and Coord, respectively, and they are always included in the syntactic tree. For subordinating conjunctions, the CAC approach is again in some ways similar to the annotation scenario of the tectogrammatical layer of PDT – dependencies between clauses are annotated but the set of labels is much smaller than that of t-layer of the PDT, again resulting in a loss of information. For coordination and apposition, the difference is structural; while CAC marks a coordination element with a specific label ('1' in column 6 of a word-level tag and in column 8 of the clause-level tag, see Tables 4 and 5), PDT makes a node corresponding to the coordination (apposition) a virtual head of the members of the coordination or apposition (whether phrasal or clausal). CAC thus cannot annotate hierarchy in coordination and apposition without loss of information, while PDT can.

4.3.5 Reflexive particles

In CAC, reflexive particles *se*/*si* are often left unannotated, while PDT uses detailed labels for all occurrences. Lexicalized reflexives (AuxT in the PDT), particles (AuxO) and reflexive passivization (AuxR) and also certain (yet rare) adverbial usages (Adv) are not annotated in the CAC at all. The only case where CAC annotates them is in situations where they can be considered objects (accusative or dative case of the personless reflexive pronoun *sebe*).

4.3.6 Analytic verb forms

In CAC, no syntactic relation is indicated for auxiliary verbs, loosing the reference to the verb they belong to; in the PDT, they are put as dependents onto their verb, and labeled AuxV to describe their function.

4.3.7 Special adverbs and particles

In PDT, there are also syntactic labels for certain type of "special" adverbials and particles, such as *raději* [*better*]*, zřejmě* [*probably*]*, také* [*also*]*, přece* [*surely*]*, jedině* [*only*]. In CAC, dependencies and syntactic tags for these tokens are missing.

Other differences in both syntactic scenarios will be described in the next section since they are related to spoken language annotation.

5 CAC syntactic annotation of spoken utterances

Current Czech syntactic theory is based almost entirely on written Czech but spoken language often differs strikingly from the written one (Millerová, 1994).

In the CAC guidelines, only the following word-level markup specifically aimed at the spoken utterance structure is described:

- non-identical reduplication of a word (value '7' in column 6),
- identical reduplication of a word (value '8' in column 6),
- ellipsis (value '9' or '0' in column 6).

Words *jsou* [*are*] and *ta* [*the*] (Table 2) represent a non-identical reduplication of a word; that is why they have been assigned the value '7' (as described above), while *je* [*is*], *jednom* [*one*], *to* [*the*] and *nic* [*nothing*] represent an identical reduplication of a word, i.e. they get the value '8' ("identical reduplication of a word"). The description does not quite correspond to what a closer look at the data reveals: '7' is used to mark a reparandum (part of the sentence that was corrected by the speaker later), while '8' is used to mark the part that replaces the reparandum (cf. also the “EDITED” nonterminal and the symbols “[“, “+” and “\” in the Penn Treebank Switchboard annotation (Godfrey *et al*., 1992). Ellipsis (the value ‘9’) was assigned to the words *trošku* [*a bit*] and *té* [*to the*].

> *A to jsou trošku, jedna je, jedna má světlou budovu a druhá má tmavou budovu, ony jsou umístěny v jednom, v jednom areále, ale ta, to centrum, patřilo té, bylo to v bloku Univerzity vlámské, a já jsem se ptala na univerzitě, na, v Univerzitě svobodné, že, no a to přeci oni nevědí, to nanejvýš, to prostě jedině, když je to Univerzita vlámská, tak o tom oni přece nemohou nic vědět, a nic.*
>
> (Lit.: *And they are a bit, one is, one has a light building and the second has a dark building, they are placed in one, in one campus, but the, the center, it belonged to the, it was in a bloc of the Flemish University, and I asked at the University, in, at the Free University, that, well, and that surely they don’t know, it at most, it simply only, if it is the Flemish University, so they surely cannot know anything, and nothing.*)

Table 2 Transcript of a Czech spoken utterance (from CAC)

However, our sample sentence contains more phenomena typical for spoken language than CAC attempts to annotate, for example:

- unfinished sentences (fragments), with apparent ellipsis: *A to je trošku…* [*And they are a bit…*],
- false beginnings (restarts): *jedna je, jedna má* [*one is, one has*],
- repetition of words in the middle of sentence: *jsou umístěny v jednom, jednom areále* [*they are placed in one, in one campus*],
- redundant and ungrammatically used words: *ony jsou umístěny v jednom…, univerzitě, na,v Univerzitě svobodné,…* [*, they are placed in one… at the University, in, at the Free University,*],

- redundant deictic words: *…ale ta, to centrum…* [*…but the, the center…*],
- intonation fillers: *no* [*well*],
- question tags: *na Univerzitě svobodné, že* [*at the Free University, that*],
- redundant conectors: *když je to Univerzita vlámská, tak to o tom* [*if it is the Flemish University, so they surely cannot know anything*],
- broken coherence of utterance, „teared“ syntactic scheme of proposition: *ale ta, to centrum, bylo to v bloku* [*but the, the center, it belonged to the, it was in a bloc*],
- syntactic errors, anacoluthon: *přeci nemohu nic vědět, a nic*. [*surely (I) cannot know anything, and nothing*].

The CAC syntactic scenario does not cover these phenomena in the guidelines (and tag tables), and even if some of them would easily fall in the reparandum/repair category (such as the phrase *jedna je*, *jedna má* [*one is, one has*]), which is seemingly included, it does not annotate them as such. Moreover, these are just some of the spoken language phenomena, taken from just one random utterance; a thorough look at the spoken part of the CAC reveals that most of the well-known spoken language phenomena, e.g. grammatically incoherent utterances, grammatical additions (as an afterthought), redundant co-references or phrase-connecting errors (Shriver, 1994, Fitzgerald, 2009), are present in the texts but left unnoticed.

In comparison, however, the PDT covers none of these typical spoken structures in the text annotation guidelines (the main reason being that it does not contain spoken material in the first place). Thus, at the surface-syntactic layer (the a-layer) of the PDT, there are only limited means for capturing such spoken phenomena.

For example, words playing the role of fillers could get the analytical function AuxO designed mostly for a redundant (deictic or emotive) constituent.

Many phenomena typical for spoken language would get, according to the PDT guidelines, the analytical function ExD (Ex-Dependent), which just “warns” of such type of incomplete utterance structure where a governing word is missing, i.e. it is such ellipsis where the dependent is present but its governing element is not.

In Figure 4, we present an attempt to annotate the above spoken utterance using the standard PDT guidelines. The "problematic" nodes, for which we had to adopt some arbitrary annotation decisions due to the lack of proper means in the PDT annotation guidelines, are shown as dark squares. For comparison, we have used dashed line for those dependency edges that were annotated in the CAC by one of the spoken-language specific tags (values '7', '8', '9' in the column 6 of the original annotation, see above at the beginning of Sect. 5),

Most of the square-marked nodes do correspond well to the PDT labels for special cases which are used for some of the peripheral language phenomena (ExD, Apos and its members, several AuxX for extra commas, AuxY for particles etc.).

It can also be observed that the dashed lines (CAC spoken annotation labels) correspond to some of the nodes with problematic markup in the PDT, but they are used only in clear cases and therefore they are found much more sparingly in the corpus.

6 Reconstruction of spoken utterances

Given the main principles of the a-layer of PDT annotation (no addition/deletion of tokens, no word-order changes, no word corrections), one would have to introduce arbitrary, linguistically irrelevant rules for spoken material annotation with a doubtful use even if applied consistently to the corpus. Avoiding that, transcriptions currently present in the CAC could not be syntactically annotated using the annotation guidelines of the PDT, thus the latest published version of CAC - CAC 2.0 – consists of written texts both morphologically and syntactically annotated and spoken texts morphologically annotated only.

We plan to complete the annotation of the spoken language transcriptions, using the scheme of the so-called "speech reconstruction" project (Mikulová *et al*., 2008), running now within the framework of the PDT (for both Czech and English).[17] This project will enable to use the text-based guidelines for syntactic annotation of spoken material by introducing a separate layer of annotation, which allows for "editing" of the original transcript and transforming it thus into a grammatical, comprehensible text. The "edited" layer is in addition to the original transcript and contains explicit links between them at the word granularity, allowing in turn for observations of the relation between the original transcript and its syntactic annotation (made "through" the edited text) without any loss. The scheme picks up the threads of the speech reconstruction approach developed for English by (Fitzgerald, Jelinek, 2008).

Our sample sentence (listed in Table 2) transformed into a reconstructed sentence, into three separated sentences in fact, is presented in Table 3 (The bold marking means changes, and parentheses indicate elements left out in the reconstructed sentence.). After the reconstruction, the sentences can be annotated according to the PDT guidelines (Figure 5).

A ***(to)*** *jsou trošku* ***rozdílné,(jedna je,)*** *jedna má světlou budovu a druhá má tmavou budovu.* ***(, ony)*** *Jsou umístěny* ***(v jednom,)*** *v jednom areále, ale* ***(ta,)*** *to centrum* ***(, patřilo té,)*** *bylo* ***(to)*** _v bloku Univerzity vlámské_***(,)*** *a já jsem se ptala na* ***(univerzitě, na, v)*** _Univerzitě svobodné_***.(, že, no a to přeci oni nevědí, to nanejvýš, to prostě jedině,) K***_dyž je to Univerzita vlámská, tak o tom oni přece nemohou nic vědět_ ***(, a nic).***

(Lit.: *And they are a bit* ***different****, one has a light building and the second has a dark building. They are placed in one campus, but the center* ***(, it belonged to the, it)*** *was in a bloc of the Flemish University, and I asked at the* ***(University, in, at the)*** _Free University_***.(, that, well, and that surely they don't know, it at most, it simply only,) If*** _it is the Flemish University, so they surely cannot know anything_***(, and nothing)***.)

Table 3 Spoken utterance after reconstruction (parentheses deleted, changes in boldface)

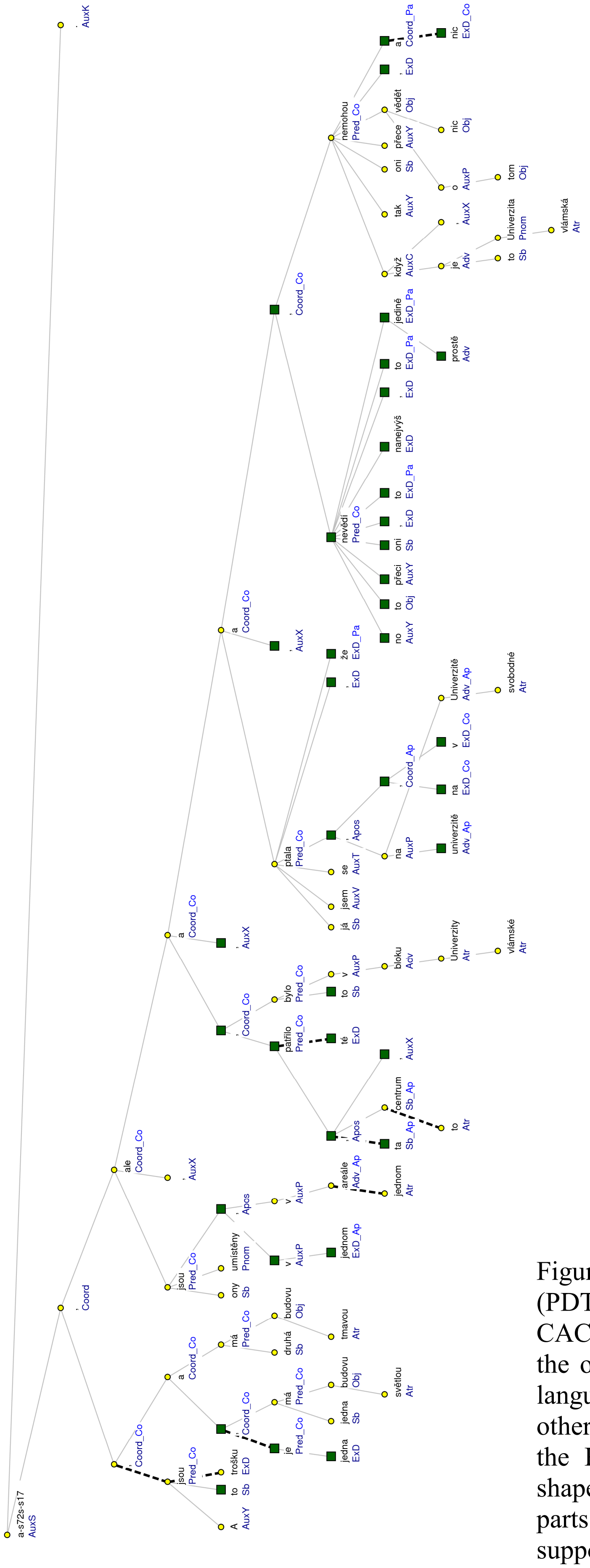

Figure 4: A syntactic annotation attempt (PDT-guidelines based) at the sample CAC sentence. The dashed edges are the only ones containing some spoken-language specific CAC annotation, the others correspond as close as possible to the PDT annotation scenario. Square-shaped nodes highlight the problematic parts (phenomena with no explicit support in the PDT annotation guidelines)

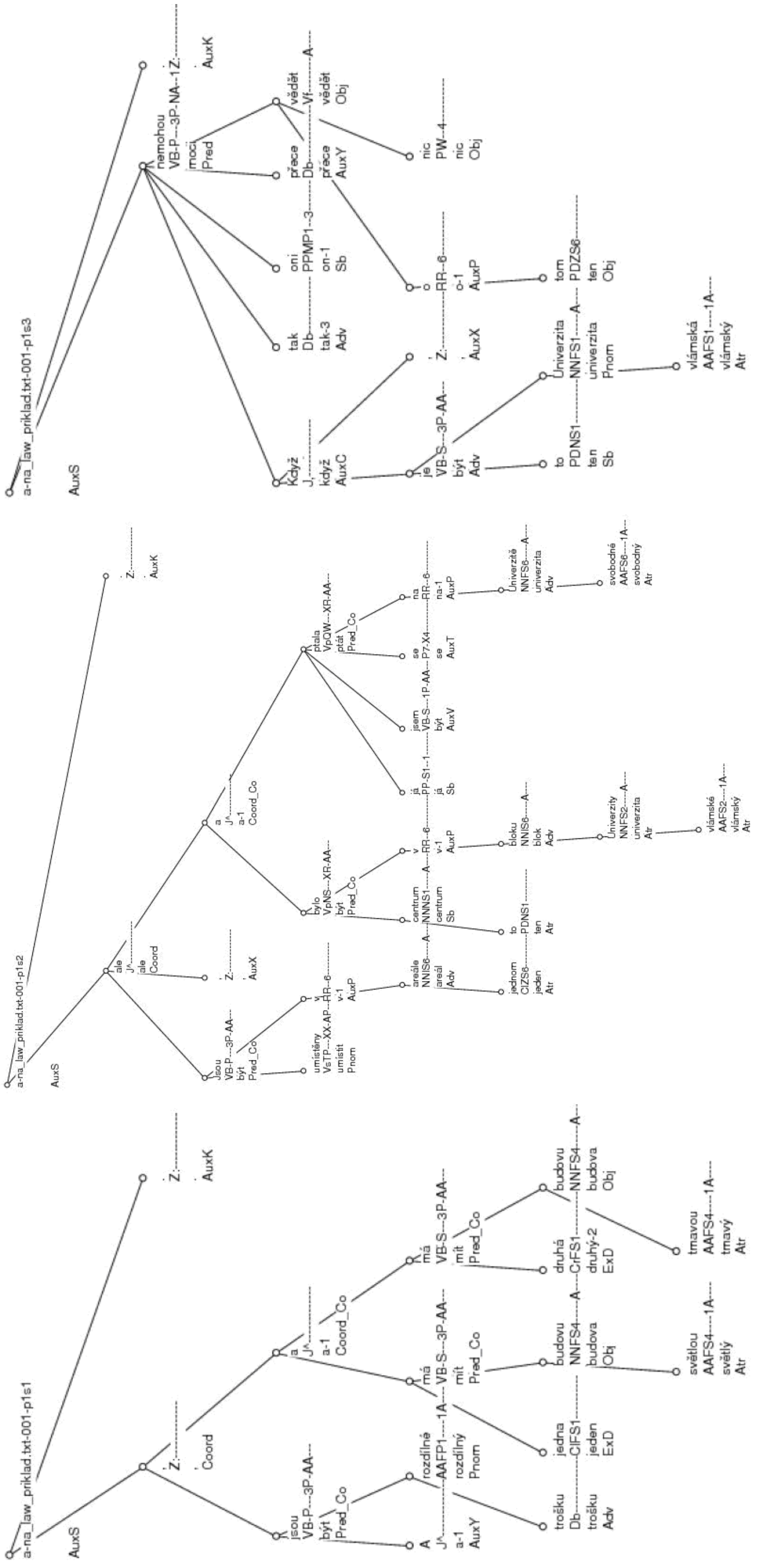

Figure 5: Syntactic trees after the original utterance reconstruction

Dependency relation		Dependency subtypes	Governor				Other	
			Direction		Offset			
1		2	3		4	5	6	
Tag	Desc.		Tag	Desc.			Tag	Desc.
1	Subject	Values specific to the dependency relation (see column 1)	+	Right	Distance between words (two digit string: for ex. 01 denotes neighboring word)		1-6	Coordination types
2	Predicate		-	Left			7,8	Repetitions (for the spoken part)
3	Attribute						9, 0	Ellipses
4	Object							
5	Adverbial							
6	Clause core							
7	Trans. type							
8	Independent clause member							
9	Parenthesis							

Table 4 Main word-level syntactic tags in the Czech Academic Corpus

Clause ID		Clause Type		Subordination (dep.) type		Governing clause/word			Clausal relation	
						Gov. noun	Gov. clause			
1	2	3		4		5	6	7	8	
		Tag	Desc.	Tag	Desc.				Tag	Desc.
Two-digit id (unique within a sentence: for ex. 91 denotes the first sentence		1	Simple			One-digit relative position of a noun modified by the clause Attributive clauses only	Two-digit id of the governing clause		1	Coordination
		2	Main						2	Parenthesis
		3	Sub-ordinated	1	Subject				3	Direct Speech
				2	Predicate				5	Parenthesis in direct speech
				3	Attribute				6	Introductory clause
				4	Object				8	Parenthesis, introductory clause
				5	Local				!	Structural error
				...	….				...	etc.

Table 5 Clause-level syntactic tags in the Czech Academic Corpus

Analytic function	Description
Pred	Predicate
Sb	Subject
Obj	Object
Adv	Adverbial
Atr	Attribute
Pnom	Nominal predicate, or nom. part of predicate with copula *to be*
AuxV	Auxiliary verb *to be*
Coord	Coordination node
Apos	Apposition (main node)
AuxT	Reflexive tantum
AuxR	Reflexive,neither Obj nor AuxT (passive reflexive)
AuxP	Primary preposition, parts of a secondary preposition
AuxC	Conjunction (subordinate)
AuxO	Redundant or emotional item, ‘coreferential’ pronoun
ExD	A technical value for a node depending on a deleted item (ellipsis with dependents)
Aux.., Atv(V),..	Other auxiliary tags, verbal complements, other special syntactic tags

Table 6 Dependency relation tags in the Prague Dependency Treebank

7 Conclusion

Courage of the original CAC project’s team deserves to be reminded. Having the experience with the present spoken data processing, we do appreciate the initial attempts with the syntactic annotation of spoken texts.

However, the design of the original corpus proved to be inconsistent with today’s demands on annotated corpora, even in the very basic requirements, such as the presence of all tokens from the original text or transcript. The syntactic scheme, while providing valuable insight in the then-current state-of-the-art in dependency theory, was only partial from today’s point of view. Therefore, we cannot conclude that two syntactic annotation schemes, even if based on the same theory, are convertible to each other with only minor changes; on the contrary, we have shown that in fact it was more effective to annotate syntax (as opposed to morphology) from scratch, and that the problems of spoken material have yet to be fully resolved (perhaps with the help of the new direction in speech reconstruction annotation).

Notes

[1] http://www.cis.upenn.edu/~treebank/

[2] http://ufal.mff.cuni.cz/pdt2.0

[3] http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/

[4] http://verbs.colorado.edu/~mpalmer/projects/ace.html

[5] http://nlp.cs.nyu.edu/meyers/NomBank.html

[6] http://www.cis.upenn.edu/~chinese/ctb.html

[7] http://www.cis.upenn.edu/~xtag/koreantag/#Treebank

[8] http://www.seas.upenn.edu/~pdtb/

[9] http://ufal.mff.cuni.cz/rest/cac.html

[10] http://w3.msi.vxu.se/~nivre/research/talbanken.html

[11] When these annotation projects began in the 1960s, there were only two computerized manually annotated corpora available: the Brown Corpus of American English and the LOB Corpus of British English. Both contain written texts annotated for part of speech. Their size is 1 mil. tokens.

[12] http://ice-corpora.net/ice/index.htm

[13] http://childes.psy.cmu.edu/grasp/

[14] http://lands.let.kun.nl/cgn/ehome.htm

[15] http://verbmobil.dfki.de/

[16] Either written (w) or spoken (s) texts.

17 http://ufal.mff.cuni.cz/pdtsl

Acknowledgement

We gratefully acknowledge the support of the Czech Ministry of Education through the grant No. MSM-0021620838 and ME 838 and the Grant Agency of Charles University in Prague through the grant No. GAUK 52408/2008.

We wish to thank Jan Hajič, whose comments stimulated us to make our paper better. We are grateful to Petr Pajas for presenting a wide dependency tree.

References

Cieri, Ch., D. Miller and K. Walker (2004). The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text. In *Proceedings of the 4th LREC*, Lisbon, Portugal, 69-71.

Godfrey, J. J., E. C. Holliman and J. McDaniel (1992). SWITCHBOARD: Telephone speech corpus for research and development, IEEE ICASSP, 517-520.

Fitzgerald, E. (2009). *Reconstructing spontaneous speech.* PhD thesis, Baltimore, Maryland.

Fitzgerald, E. and F. Jelinek (2008). Linguistic resources for reconstructing spontaneous speech text. In *LREC Proceedings*, Marrakesh, Morocco, 1–8.

Greenbaum, S. (ed.) (1996). *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press.

Hajič, J. and B. Hladká (1997). Tagging of inflective languages: a comparison. In *Proceedings of ANLP'97*, Washington, DC, 136--143.

Hajič, J. et al. (2006). *The Prague Dependency Treebank 2.0,* (Linguistic Data Consortium, Philadelphia, PA, USA), Cat. No. LDC2006T01.

Hana, J., D. Zeman, J. Hajič, H. Hanová, B. Hladká and E. Jeřábek (2005). *Manual for Morphological Annotation*. TR-2005-27, Ústav formální a aplikované lingvistiky, MFF UK.

Hoekstra, H., M. Moortgat, I. Schuurman and T. van der Wouden (2001). Syntactic Annotation for the Spoken Dutch Corpus Project. In Daelemans, W.; Simaan, K.; Veenstra. J.; Zavrel, J. (eds.): *Computational Linguistics in the Netherlands* 2000. Amsterdam/New York, Rodopi, 73-87.

Chomsky, N. (1993). *Lectures on Government and Binding: The Pisa Lectures.* Holland: Foris Publications, 1981. Reprint. 7th Edition. Berlin and New York: Mouton de Gruyter.

Králík, J. and L. Uhlířová (2007). The Czech Academic Corpus (CAC), its history and presence, In *Journal of quantitative linguistics*. 14 (2-3): 265-285.

Mikulová, M. (2008). *Rekonstrukce standardizovaného textu z mluvené řeči v Pražském závislostním korpusu mluvené češtiny. Manuál pro anotátory.* TR-2008-38, Institute of Formal and Applied Linguistics, MFF UK.

Müllerová, O. (1994). *Mluvený text a jeho syntaktická výstavba.* Academia, Praha.

Nilsson, J., J. Hall and J. Nivre (2005). MAMBA meets TIGER: Reconstructing a Treebank from Antiquity. In *Proceedings of NODALIDA 2005 Special Session on Treebanks for Spoken and Discourse*, Copenhagen Studies in Language 32, Joensuu, Finland, 119-132

Pajas, P. and J. Štěpánek (2005). *A Generic XML-based Format for Structured Linguistic Annotation and its Application to the Prague Dependency Treebank 2.0*. TR-2005-29, Institute of Formal and Applied Linguistics, MFF UK.

Ribarov, K., A. Bémová and B. Hladká (2006). When a statistically oriented parser was more efficient than a linguist: A case of treebank conversion, In *Prague Bulletin of Mathematical Linguistics*, 1 (86):21-38.

Sgall, P., E. Hajičová and J. Panevová (1986). *The meaning of the sentence in its semantic and pragmatic aspects*, ed. by J. Mey. Reidel, Dordrecht; Academia, Praha.

Shriberg, E. (1994). *Preliminaries to a Theory of Speech Disfluencies*. PhD thesis, University of California, Berkeley.

Svartvik, J. and R. Quirk (1980). *A Corpus of English Conversation*. Lund.

Šmilauer, V. (1972). *Nauka o českém jazyku*. Praha.

Šmilauer. V. (1969). *Novočeská skladba.* Státní pedagogické nakladatelství. Praha.