Prague Dependency Treebank Annotation Errors
A Preliminary Analysis

Vojtěch Kovář and Miloš Jakubíček

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
{xkovar3, xjakub}@fi.muni.cz

Abstract. This paper presents a basic analysis of syntactic annotation errors and inconsistencies in the Prague Dependency Treebank, the biggest corpus of Czech with manual syntactic annotation. The corpus is used for developing and testing of many syntactic analysers of Czech and the problems in the annotation have an essential impact on the evaluation of the quality of these parsers and the results of precision measurements. We identify some of the basic annotation problems and in some cases, we outline possible solutions.

Key words: error in text; annotation; Prague Dependency Treebank

1 Introduction

The Prague Dependency Treebank (PDT, [1]) forms the only big source of manually annotated Czech syntactic data. Currently, this corpus contains about two million tokens annotated in three layers – morphological, analytical and tectogrammatical – and it is of great use in the process of developing and testing syntactic analysers of Czech.

However, there is a large number of inconsistencies and errors in the data which makes using the corpus quite problematic and questionable. These flaws result from various reasons ranging from insufficient annotation guidelines and apparent mistakes to shortcomings of the annotation as it has been formalised. Furthermore, it is not clear what the percentage of the wrong annotation is and how it can affect the measurements that use the PDT as the gold standard data used for training and testing various algorithms.

In this paper, we present a preliminary analysis of errors and inconsistencies in a PDT sample. We describe some problems in annotation that were revealed during the work with this sample, try to figure out the sources of the particular problems and suggest possible solutions. We also estimate the overall percentage of error in our sample and, assuming that the sample is representative, in the whole corpus.

2 The Prague Dependency Treebank

The Prague Dependency Treebank has been developed according to the tradition of the Prague linguistic school – it uses the formalism of Functional Generative

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2009, pp. 101–108, 2009.

Description [2,3]. The annotation consists of three layers – morphological, analytical and tectogrammatical.

Within the scope of this paper, we are interested in the analytical layer only. This part of annotation contains the description of the dependency syntax in form of labeled dependency trees (acyclic oriented graphs over the input tokens). Furthermore, we will deal just with the structure of the trees, not the functional classification of the particular edges that is recorded as edge labels. This classification is not so critical and most parsers also do not label their outputs.

In the whole text, we refer to the current version of the corpus, PDT 2.0.

2.1 The Sample

For training and testing purposes, the PDT data is divided into 10 parts – 8 of them are provided for training, 2 are dedicated to testing.

For the purposes of this paper, we used the beginning of the first training set, *train-1* that was previously used by development of the *SET* parsing system [4]. Most of the examples come from the first 60 sentences of this testing set since these sentences were checked many times during the parser development and we know them very well.

3 Error Analysis

In this section, we present examples of errors, inconsistencies and other problems in our sample and briefly discuss various aspects of these problems.

3.1 Random Errors

The first group of problems we met during the parser development were apparent errors in the annotation. Such errors cannot be explained as inconsistencies or flaws of the annotation formalism, they are just random defects created by the human annotators. The existence of such random errors is unavoidable in all human annotated data and it must be presumed that anything done by humans, including the annotators, can and will be erroneous to some extent. However, every effort should be made to keep the number of annotator errors as low as possible.

As an example, we show the beginning of the sentence #00040 (see Figure 1). There are two apparent problems:

- The dependency *Většinu* ← *fax* (*Most* ← *fax*). There is no reason for this markup, the phrase *jako fax* (*as fax*) clearly belongs to the phrase *jako výkonnou kopírku* (*as an efficient copier*) – these two phrases should be joined in a coordination.
- The coordination in the top level of the tree. Previously mentioned coordination should be marked instead of this one and the whole structure should depend on the verb *používat* (*use*).

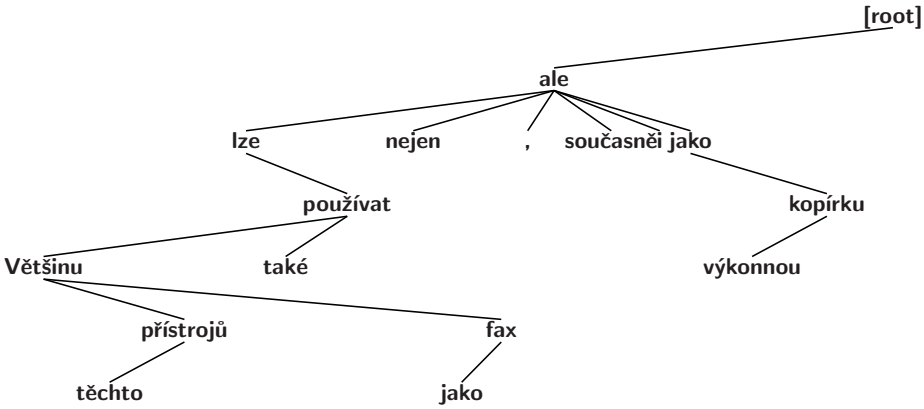

Fig. 1. The first part of the sentence #00040 as recorded in the PDT: *Většinu těchto přístrojů lze také používat nejen jako fax, ale současně i jako výkonnou kopírku...* (*Most of these devices can be used not only as fax but in the same time also as an effective copier...*)

We can see that such quite simple mistakes can globally change the structure of the tree (which might be also seen as a disadvantage of the dependency annotation formalism) and usage of sentences with such errors is very problematic in every possible application.

3.2 Inconsistencies

Inconsistencies in the annotation seem to be a bigger problem than random errors. They occur systematically in the corpus and are very common. Although an extensive manual for annotators is provided [5] to avoid these problems, there are still many language phenomena that are not described clearly enough or are even not described at all. The creativity of annotators then creates more annotation variants for a single phenomenon. According to our estimations, about 30 or 40 % sentences contain one of the phenomena that are marked inconsistently in some of the sentences.

Our first example of inconsistency shows annotation of punctuation in parts of item lists. Both sentences in Figures 2 and 3 contain an asterisk that has definitely the same meaning in both cases. However, the two annotations differ. No matter which of these variants is correct, in case of short sentences as in our two examples, the edge adjacent to the asterisk stands for 20 or 25 percent of sentence annotation.

The first example was rather technical and could be basically solved by some automatic or semi-automatic procedures. More serious inconsistencies can be found in annotation of frequent linguistic phenomena, e.g. passive verb forms. As illustrated in Figures 4, 5 and 6, this phenomenon is marked in various ways in the corpus.

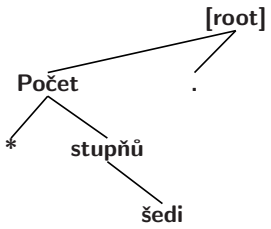

Fig. 2. The sentence #00048 as recorded in the PDT: * *Počet stupňů šedi* (* *Number of levels of grey*)

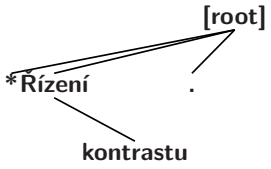

Fig. 3. The sentence #00053 as recorded in the PDT: * *Řízení kontrastu* (* *Contrast control*)

In Figure 4, the auxiliary verb *je* (*is*) is at the top of the phrase and the complement (*jaké*) depends on the participle form of the verb. In Figure 5, the auxiliary verb is still on the top of the phrase but the complements of the predicate depend on the auxiliary verb. Finally, in the third example (Figure 6), the participle is on the top of the phrase and all the rest including the auxiliary verb depends on the participle.

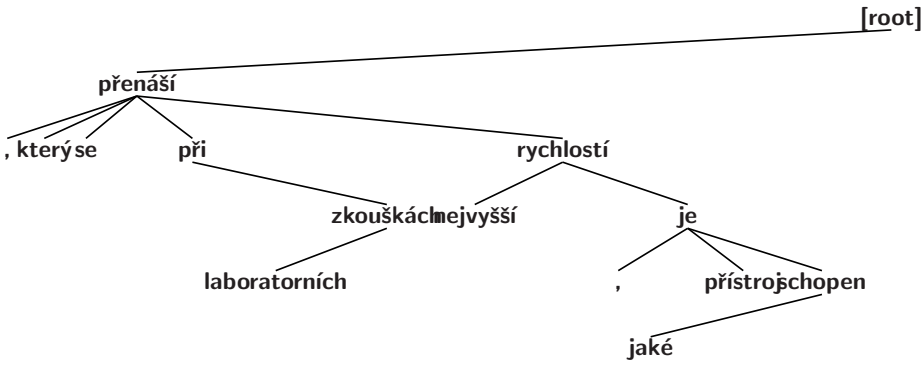

Fig. 4. Part of the he sentence #00012 as recorded in the PDT: *..., který se přenáší při laboratorních zkouškách nejvyšší rychlostí, jaké je přístroj schopen...* (*..., that is tranferred in laboratory tests with the highest speed that is the device able...*)

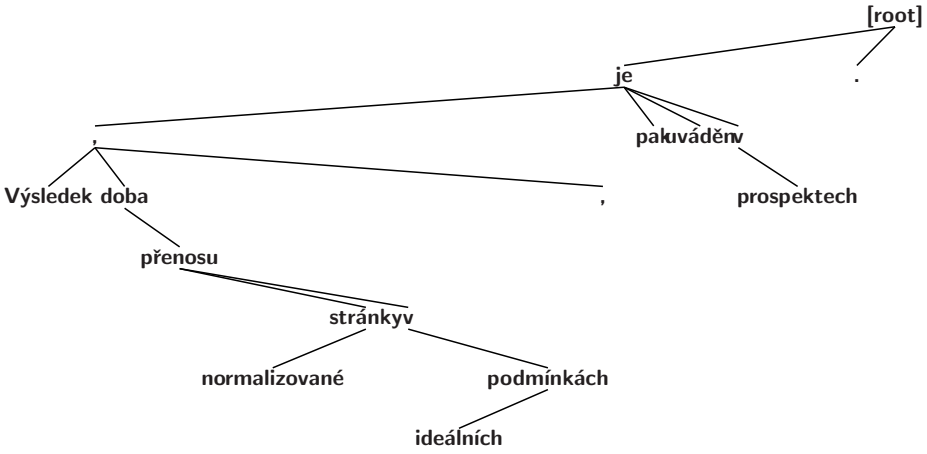

Fig. 5. The sentence #00013 as recorded in the PDT: *Výsledek, doba přenosu normalizované stránky v ideálních podmínkách, je pak uváděn v prospektech.* (*The result, the transfer time of the normalized page in ideal conditions, is then reported in brochures.*)

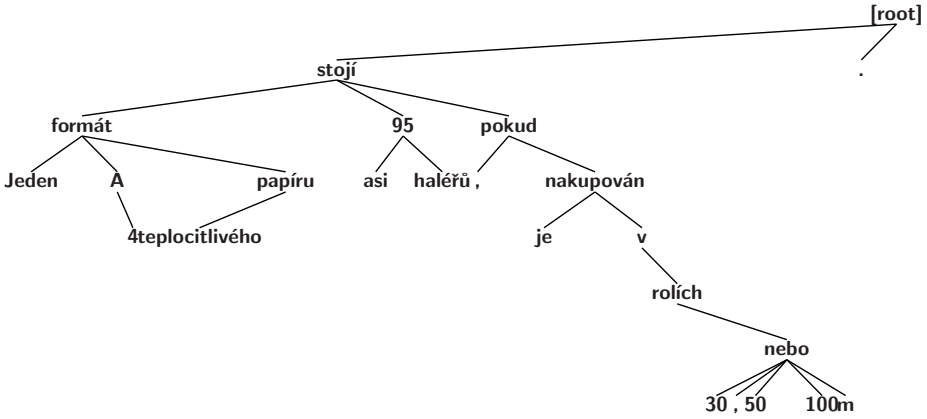

Fig. 6. The sentence #00028 as recorded in the PDT: *Jeden formát A 4 teplocitlivého papíru stojí asi 95 haléřů, pokud je nakupován v rolích 30, 50 nebo 100 m.* (*One page A 4 of the thermosensitive paper costs approximately 95 hellers, if it is bought in reels 30, 50 or 100 m.*)

This inconsistency in annotating the predicate structure is very painful since this structure determines the shape of the whole clause. For instance, in the process of parser developing and testing, developers (or training algorithms) have to search the most frequent annotation pattern for this case so that the parser has maximum precision against the data. However, they are doomed to fail when trained on such inconsistent data.

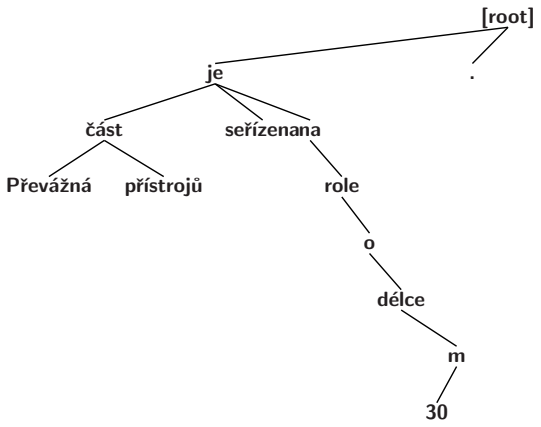

Fig. 7. The sentence #00030 as recorded in the PDT: *Převážná část přístrojů je seřízena na role o délce 30 m.* (*Most devices are adjusted for reels of length 30 m.*)

Another similar problem is annotation of phrases with numerals. Though they are well covered in the annotation manual that we previously mentioned, annotation of many sentences does not respect the instructions. An example is shown in Figure 7 (the "m" token should depend on the numeral in this case, according to the annotation manual).

3.3 The "Dirty" Cases

Although the problems showed above are the most remarkable ones, we are still far from a complete error list. In some cases, it is not clear if the particular annotation is a mistake or an intention.

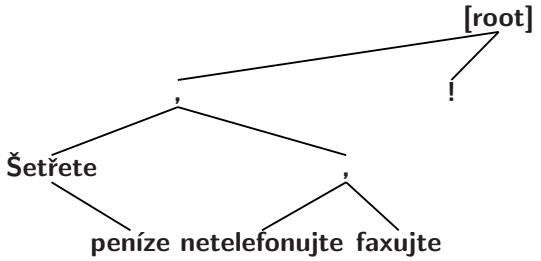

Fig. 8. The sentence #00004 as recorded in the PDT: *Šetřete peníze, netelefonujte, faxujte!* (*Save money, do not phone, fax!*)

This is the case shown in Figure 8 – it is not clear why the coordination is structured in this particular way. There might be a doubtful semantic hint

that members of the segment *netelefonujte, faxujte* (*do not phone, fax*) has closer relationship with each other than with the first phrase in the sentence, however, it is a question if any possible parser could reveal that hint. In our opinion, such cases should be annotated in the most straightforward way possible – as a flat coordination of three verbs. Unfortunately, the difference between the structured and the flat coordination markup in this case is more than 50 percent, which is to be taken as a flaw of the annotation formalism.

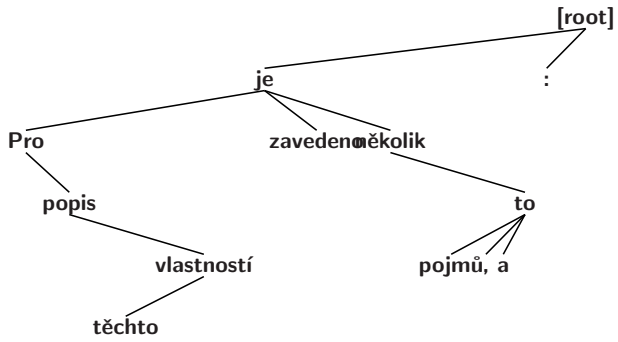

Fig. 9. The sentence #00047 as recorded in the PDT: *Pro popis těchto vlastností je zavedeno několik pojmů, a to:* (*For description of these properties, some terms are introduced, as follows:*)

In the last example (Figure 9), there is another strange annotation example in the end of the sentence – the phrase *pojmů, a to*. The structure of this phrase does not seem to have any rational basis, it just needs to fit into the dependency format *somehow*. It is even not clear, why these words should belong to one phrase.

4 On Parser Evaluation

As we have already mentioned above, even if big effort is made to eliminate annotator errors, some of them will still remain – and we dare to predict that the number of errors grows with the size of the data in a non-linear way. This raises a question whether treebanks represent a good way for measuring parser quality at all. Besides treebank consistency issues, there is more evidence which makes the use of treebanks for parser evaluation questionable: often the evaluation is significantly influenced by the different formalisms, annotations and last (but definitely not least!) by different linguistic insights and opinions. Finally, there have been recently proposed application-driven evaluation techniques for parsers (see e. g. [6]) which we believe to continue becoming more widely used in the parsing community. A detailed discussion of this topic is however outside of the scope of this paper.

5 Conclusion

We have described the main problems in the PDT annotation that we met during the process of parser development. We have presented examples of the selected problems and also showed the fact that in some cases, one simple mistake or inconsistency can lead to structural changes in the sentence annotation and significantly affect the results of parser tests and development. This can be considered as a negative feature of the dependency annotation formalism in general.

The total number of errors in the annotation in our sample is not clear because there is not a good characterization of what is an error. However, according to our estimations, the difference between the current state and the correctly and consistently annotated sentences would be 5 to 10 percent. This quite a big number may be one of the reasons why the current parsers of Czech are not so successful as parsers for English or German [7], although they have been intensively developed.

In the future, we want to perform more thorough critical analysis of the errors in the PDT corpus annotation and propose some automatic and semi-automatic methods leading to their elimination. We will also propose possible changes in the formalism and in the precision metrics used in the process of developing and testing syntactic parsers.

Acknowledgements This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536.

References

1. Hajič, J.: Building a syntactically annotated corpus: The Prague Dependency Treebank. In: Issues of Valency and Meaning, Prague, Karolinum (1998) 106–132.
2. Sgall, P.: Generativní popis jazyka a česká deklinace (Generative Description of the Language and the Czech Declension). Academia, Prague, Czech Republic (1967).
3. Sgall, P., Hajičová, E., Panevová, J.: The Meaning of the Sentence and Its Semantic and Pragmatic Aspects. Academia/Reidel Publishing Company, Prague, Czech Republic/Dordrecht, Netherlands (1986).
4. Kovář, V., Horák, A., Jakubíček, M.: Syntactic analysis as pattern matching: The SET parsing system. In: Proceedings of the 4th Language & Technology Conference, Poznań, Poland (2009).
5. Hajič, J., Panevová, J., Buráňová, E., Urešová, Z., Štěpánek, J., Pajas, P., Kárník, J.: Anotace na analytické rovině – Návod pro anotátory (2005) `http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/cz/a-layer`.
6. Miyao, Y., Sagae, K., Saetre, R., Matsuzaki, T., Tsujii, J.: Evaluating contributions of natural language parsers to protein-protein interaction extraction. Bioinformatics **25**(3) (2009) 394–400.
7. Baumann, S., Brinckmann, C., Hansen-Schirra, C., et al.: The MULI project: Annotation and analysis of information structure in German and English. In: Proceedings of the LREC 2004 Conference, Lisboa, Portugal (2004).