Annotation of Selected Non-dependency Relations in a Dependency Treebank

Kristýna Čermáková, Lucie Mladová, Eva Fučíková, Kateřina Veselá

Charles University in Prague
Institute of Formal and Applied Linguistics

E-mail: {cermakova, mladova, fucikova, vesela}@ufal.mff.cuni.cz

Abstract

The following paper has two aims. First, it introduces a procedure of a manual annotation of selected linguistic phenomena across a large-scale dependency treebank of English. The method was designed to provide higher consistency of annotated data, and so higher credibility of the treebank. Second, the first expert task completed by means of this method is being described – the annotation of rhematizers and discourse connectives and their modifiers, i.e. annotation of some non-dependency relations in a dependency approach.

1 Motivation

Disagreements between annotators' judgments in corpora annotation are a disturbing factor in the machine training. Nevertheless, they represent an important indicator of functionality of the used approach; they localize theory deficiencies, and are also useful for finding interesting language phenomena.

Prior to the first release of the Prague English Dependency Treebank (PEDT) [2], a set of control scripts was established to increase consistency of the annotated data. Since the PEDT adopted its annotation scheme and guidelines from its Czech "mother treebank" – the Prague Dependency Treebank (PDT) [7], we expected some of the repeated disagreements to be caused by different treatment of language-specific phenomena in the two corpora, and therefore to be solved by adding new, language-specific annotation rules. Hence, the most frequent disagreements in PEDT were identified, analyzed, and rectified through additional manual annotation. The control scripts revealed that the divergences had been partly a matter of capturing conventions; partly they were difficult linguistic issues that required deeper linguistic knowledge. The former was handled by introducing automatic changes [8], and the latter was solved by the development of a new annotation method, the so called expert annotation. In this paper, we offer the description and evaluation of this specific manual annotation procedure.

2 Prague English Dependency Treebank

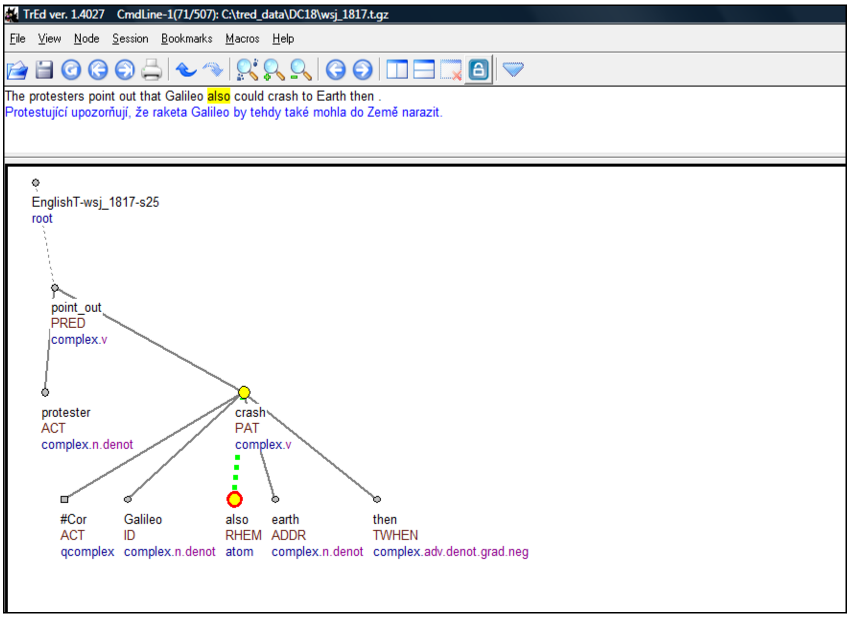
The Prague English Dependency Treebank (PEDT) represents the English-language part of the Prague Czech-English Dependency Treebank (PCEDT, [1]). PCEDT is a parallel corpus developed by Czech linguists primarily for the purpose of experiments in machine translation with a special emphasis on dependency-based (structural) translation. PCEDT and therefore also PEDT are based on the long-standing Praguian linguistic tradition and the Functional Generative Description of language (FGD) [6], adapted for the current computational linguistics research needs. The first publicly released version PEDT 1.0 [2] comprises annotation of approximately 25% of all approximately 49 000 sentences from Penn Treebank III – the Wall Street Journal section.

Wall Street Journal texts in PEDT are manually annotated on the tectogrammatical layer which represents underlying syntactic structure and captures semantic relations. The tectogrammatical layer annotation comprises dependency structure in the form of a dependency tree, including semantic labeling (the so called functors), valency annotation, and some coreference relations. Currently, routine annotation proceeds (approximately 50% of the data have already been annotated), and annotation principles are refined. The goal of the project is to annotate the entire PTB III – WSJ. Simultaneously with the annotation of English data, Czech translations are annotated giving rise to the parallel PCEDT.

3 Annotation of Specific Phenomena

The standard manual annotation of the tectogrammatical layer in PEDT proceeds since late 2006 on the dependency-converted tree structures. The division of the data into the original WSJ sections is preserved, with each annotator receiving one section at a time to be examined tree by tree. The inter-annotator agreement is measured regularly on a subset of simultaneously annotated data. For the two main attributes in the treebank, i.e. structure and functor, the agreement ranges from to 81 to 91% (functor) and 90 to 96% (structure) with a slight rising tendency.

A specific procedure was proposed to solve especially problematic syntactic issues, such as the treatment of some of the non-dependency edges described further in Section 4 of this paper. First, the nature of the problem was identified; second, the PEDT was scanned and sentences with the occurrence of the problematic phenomenon (problematic lexemes, phrases or functors) were selected and located into filelists. Particular questionable parts (nodes of the trees) within the filelists were highlighted in the annotation tool (see Figure 1) [4], and finally, “expert” annotators trained for the given linguistic task examined the filelists across all corpus sections. Where possible, correct analysis was pre-annotated automatically, for instance the correct lemmatization of typically unambiguous multiword expressions such as *as a result*, *for example* or *in other words*.



The protesters point out that Galileo also could crash to Earth then .
Protestující upozorňují, že raketa Galileo by tehdy také mohla do Země narazit.

Figure 1: A highlighted node for the expert annotation in the tree editor TrEd

4 Non-dependency Relations in PEDT

Based on the tradition of FGD, PEDT is a dependency corpus. However, it is able to convey various non-dependency relations by means of non-dependency edges which are established to represent primarily parataxis and some other specific relations. In our specific phenomena annotation, we focused on the problematic non-dependency edges that are used to capture proposition modifiers (expressions with a lesser degree of integration into the syntactic structure, such as modal and attitude markers, focusing and additive expressions, etc.). In PEDT, these are roughly reflected by nodes with the semantic role of:

- expressions referring to preceding contexts (PREC),
- rhematizers (RHEM),
- conjunction modifiers (CM),
- and marginally attitude (ATT).

Nodes with the functor (semantic label) PREC function as discourse connectives. Basic forms of these linking expressions are adverbials (*consequently*), particles (*yet*), some prepositional phrases (*in addition*), and paratactic connectives (*therefore*). From the point of view of traditional English grammars [5], they are partly homonymous with verbal complements, most often with temporal and spatial ones. If they assist in connecting paratactically conjoined elements, they are usually assigned the functor for conjunction modifiers (CM).

The nodes with the functor RHEM function as rhematizers, i.e. expressions whose function is to signal the topic-focus articulation categories in a sentence, namely the communicatively most important categories – the focus and contrastive topic [3], and their scope is indicated by a non-dependency edge. Rhematizers, e.g. *even, just, solely, exactly, precisely, only, alone, merely, simply, especially, particularly, in particular* resemble adverbials but they differ from them by their ability to modify not only a verb and adjective but also a syntactic noun. Additive expressions such as *too, also, again, equally, similarly, likewise, as well, in addition* rank sometimes among rhematizers although their function is primarily connective.

The first specific phenomena annotation completed throughout the PEDT is the annotation of the described semantic groups, mainly PREC and RHEM.

5 Rhematizers and Discourse Connectives in PEDT

In this section, we describe some of the repeatedly occurring problems concerning non-dependency edges from the linguistic point of view. Within the group examined, there is a huge functional homonymy among various uses of the same lexical unit. Sometimes, only a larger context and its analysis from the point of view of topic-focus articulation and/or discourse structure are needed to interpret the function of a particle or adverbial correctly. However, in certain cases, even with substantial background knowledge, an unambiguous solution is not to be found. For our annotation it was crucial to distinguish between the cases in which the given expression had its original adverbial meaning (i.e. there was a proper dependency relation), and the cases in which it functioned otherwise (i.e. as a node with a non-dependency edge).

5.1 Rhematizers and Extent Adjuncts

One of the most disputed problems is the homonymy of expressions with the semantic component of extent in their meaning. If they express the extent or degree, they are interpreted as extent adjuncts, e.g. (1a). But, if the interpretation allows also the "*primarily*" meaning, as possibly in (1b), it can be treated both as a syntactic member with the semantic role of extent or as a focalizing element.

(1a) *He has cancelled numerous campaign appointments and was largely inaccessible to the media until the stock story broke.*

(1b) *The enormous inflation over the past 30 years was largely due to monetary policy.*

Another problematic group is represented by "typical" rhematizers, e.g. (2a). However, it should be noted that if such "typical" rhematizers modify a numeral or another quantitative expression, e.g. (2b) they act precisely like regular extent modifiers without a rhematizing function, cf. (2c):

(2a) *We invited only friends.*

(2b) *We invited only five friends.*

(2c) *We invited exactly five friends.*

In (1b) and (2b), both interpretations are correct depending on the point of view of the analysis. Yet, for the purpose of semantic labeling in PEDT, a uniform rule for such cases had to be established. Therefore, (1b) was treated as a rhematizer and (2b) as an adjunct of extent.

5.2 Rhematizers and Discourse Connectives

Rhematizers relate to a smaller or larger part of a clause, i.e. they can have a narrower or wider scope. Rhematizers with a narrow scope are easy to recognize, the only problematic usage of rhematizers in English is their position right before the verb, e.g. (3a), and (3b). With regard to the preceding context, annotators are able to distinguish between a narrow scope, e.g. (3a) and a wide scope of the rhematizer, e.g. (3b). Hence, only a larger context can help determine the scope of the rhematizer, and recognize whether the rhematizer relates to elements that precede and/or follow in the surface word order.

Further, if the rhematizer has an additive meaning and stands before the verb as in (3b), it can coincide with the function of a discourse connective. Discourse connectives, unlike rhematizers, relate always to two arguments, they connect two text spans. When a sentence-initial *also* is separated by a comma, it is treated always as a discourse connective (3c). The difference between (3b) and (3c) is considered formal, not semantic. Therefore, *also* in (3b) can be treated both as a rhematizer (RHEM) and a discourse connective (PREC) with equal validity. The type of sentences as in (3b) caused major problems in the expert annotation, also because of its high frequency in the treebank texts.

(3a) *Crude oil for November delivery edged up by 16 cents a barrel to $ 20.75 a barrel. Heating oil prices also rose.*

(3b) *The complex restructuring transforms London-based WCRS from primarily a creator of advertising into one of Europe's largest buyers of advertising time and space. It also creates a newly merged world-wide ad agency controlled by Eurocom.*

(3c) *The company said a drop in activity in the powerboat industry reduced sales volume at its two marine-related operations. Also, the company said its commercial products operation failed to meet forecasts.*

6 Inter-annotator Agreement

Three annotators who also have long-term experience with standard annotation of PEDT were trained for the specific task, and they annotated approx. 3000 problematic structures each. 515 structures were annotated by all three of them as a set of data for IAA measurement. The measurement itself is derived from the basic IAA measurement script for standard annotations [8], and it proceeds roughly as follows: Within the set of nodes either highlighted or touched by any of the three trained annotators, agreement regarding four attributes was computed between each pair of annotators. The attributes are the following: **structure** (parent node), **functor**, **tectogrammatical lemma** of the node, and **a/aux.rf**, i.e. links to the lower layer of surface syntax. The results are summed up in the Table 1.

Attribute/ Annotator pair	Structure	Functor	T-lemma	A/aux.rf
A x B	91.2%	91.1%	98.9%	96.2%
B x C	90.6%	89.6%	98.9%	92.7%
C x A	92.1%	89.3%	99.1%	93.6%

Table 1: IAA Measurement for specific phenomena annotation in the PEDT

In terms of expert annotation as such, the results can be considered satisfactory. They show that the method applied supported a unified approach to the phenomenon of focusing and additive expressions. The results are slightly higher for the same two annotator pairs compared to their recent results in standard, much more complex annotation. They also prove that the most difficult task for annotators (both in standard and specific phenomena annotation) is the agreement in functor. To sum up, before the implementation of the expert annotation, the most frequent instances of inter-annotator disagreement were caused by the lack of precise guidelines which allowed more interpretations. In the expert annotation, such disagreements occur extremely rarely.

7 Conclusion and Future Work

Our specific annotation proved to be an effective solution of the annotation of complicated phenomena. It eliminates mistakes, and simultaneously does not inhibit the standard annotation from proceeding. We were able to refine more than 9000 sentences, which is approx. 19% of the treebank. As we expected, the problematic issues examined such as the distinction between the rhematizing, discourse linking and simply adverbial function of homonyms or defining the scope of a rhematizer are too complex to be currently successfully resolved merely by automatic annotation.

The new annotation method demonstrated a significant difference between English and Czech not only in terms of standard word order principles (as is generally known) but also in terms of rhematizer positioning principles. On the one hand, English word order is largely determined by grammatical principles and as such it displays less flexibility than Czech word order. On the other hand, rhematizer positioning in English is far more flexible than it is in Czech. The data concerned in the specialized annotation show that English (unlike Czech) is able to place rhematizing expressions on the border between the topic and focus notwithstanding their distance (in the linear surface word order) from the focus proper, i.e. the informationally most weighted element. Anyway, it was not the surface position but the scope of pre-verbal rhematizers that was most problematic issue even for trained annotators.

Based on the positive results of the fist specific phenomena annotation, another run is in preparation which will focus on the annotation of complicated comparative structures.

8 Acknowledgements

The research this paper is based upon was supported by the following grants: GAUK 103609, GA201/09/H057, GA405/09/0729, MŠMT ČR LC536, and EU – Companions.

References

[1] Čmejrek, Martin, Jan Cuřín, Jiří Havelka, Jan Hajič, and Vladislav Kuboň. 2005. Prague Czech-English Dependency Treebank. In *EAMT 2005 Conference Proceedings*, p. 73–78.

[2] Hajič, Jan et al. 2009. *Prague English Dependency Treebank 1.0*, Software or data, Institute of Formal and Applied Linguistics, Charles University in Prague.

[3] Hajičová, Eva, Barbara Partee Hall, and Petr Sgall. 1998. *Topic-focus articulation, tripartite structures, and semantic content.* Boston: Kluwer Academic Publishers.

[4] Pajas Petr, Štěpánek Jan. 2008. Recent Advances in a Feature-Rich Framework for Treebank Annotation. In *The 22nd International Conference on Computational Linguistics – Proceedings of the Conference*, Manchester, p. 673–680.

[5] Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, Jan Svartvik. 2004. *A Comprehensive Grammar of the English Language.* London: Longman.

[6] Sgall, Petr, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Prague: Academia.

[7] Šindlerová, Jana, Lucie Mladová, Josef Toman, Silvie Cinková. 2007. *An Application of the PDT-Scheme to a Parallel Treebank.* In: NEALT Proceedings Series, Vol. 1, Proceedings of the 6th International Workshop on Treebanks and Linguistic Theories, Bergen, Norway, p. 163-174.

[8] Toman, Josef. 2009. *Automatická anotace angličtiny na tektogramatické rovině (Automatic Annotation of English on the Tectogrammatical Layer*). Master’s thesis. Charles University in Prague.