Typical Cases of Annotators' Disagreement in Discourse Annotations in Prague Dependency Treebank

Šárka Zikánová, Lucie Mladová, Jiří Mírovský, Pavlína Jínová
Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské nám. 25, 118 00 Prague 1, Czech Republic
E-mail: {zikanova, mladova, mirovsky, jinova}@ufal.mff.cuni.cz

Abstract

In this paper, we present the first results of the parallel Czech discourse annotation in the Prague Dependency Treebank 2.0. Having established an annotation scenario for capturing semantic relations crossing the sentence boundary in a discourse, and having annotated the first sections of the treebank according to these guidelines, we report now on the results of the first evaluation of these manual annotations. We give an overview of the process of the annotation itself, which we believe is to a large degree language-independent and therefore accessible to any discourse researcher. Next, we describe the inter-annotator agreement measurement, and, most importantly, we classify and analyze the most common types of annotators' disagreement and propose solutions for the next phase of the annotation. The annotation is carried out on dependency trees (on the tectogrammatical layer), this approach is quite novel and it brings us some advantages when interpreting the syntactic structure of the discourse units.

1. Introduction

Current discourse annotation of Czech texts in the Prague Dependency Treebank (PDT) brought out some general questions which are common for annotation of higher and more complicated units in texts. Contrary to e.g. morphological tagging on the word layer, annotation of discourse, coreference, direct-speech-attribution etc. encounters two basic problems:

1) In classical linguistics, the system of "higher" language levels is usually less described than morphology and syntax. Therefore, there is a large amount of unclear cases at the beginning of annotation of these structures, where the very existence of a discourse/coreference/etc. relation in a single case is put under question.

2) The extent of the discourse units (discourse arguments) is formally almost non-predictable and it can depend on the annotators' understanding of the text.

Both these problems influence the measurement of inter-annotator agreement. In this paper, we present examples how we solve typical problematic structures in discourse annotation of Czech. Since our solutions should lead to a clear and reusable system, they are generally based on two principles: first, the nature of a "discourse relation" is getting more strict and gets into an opposition to other types of text relations (especially to coreference). Second, the syntactic structure of discourse arguments is taken into account.

The discourse annotation in PDT is linked to the previous layers of annotations, such as morphological analysis, syntactico-semantic analysis (so called tectogrammatics) including topic-focus articulation and some coreference types (Hajič et al., 2006). The aim of the discourse annotation is to indicate semantic relations crossing the sentence boundary (see Mladová et al., 2008). A discourse relation, be it expressed explicitly by means of a discourse connective, or implicitly, connects two "discourse arguments" (abstract objects, i.e. independent events, expressed mainly by independent clauses, cf. Asher, 1993). Every single discourse argument has a certain partial semantic feature, building together with its counterpart argument the whole of a discourse relation, e.g. the relation of reason links the Argument1 expressing the reason itself to the Argument2 expressing the fact.

In the first phase, the annotation is limited to the relations that are expressed by explicit discourse connectives (coordinating and subordinating conjunctions, particles, adverbs etc., cf. Prasad et al. 2007, 2008). This temporary formal restriction helps us understand better the character of discourse relations, set the annotation scenario clearly and train annotators. Implicit discourse relations will be annotated in further phases.

2. The process of discourse annotation in the Prague Dependency Treebank

PDT contains journalistic texts of all kinds, including e.g. sport results and television programs. For discourse annotation training, larger narrative texts (30 sentences and more) were selected, in which a higher occurrence of discourse relations can be assumed.

Annotators have at their disposal both plain text and the tectogrammatical analysis (tree structures). Annotation is carried out on the tectogrammatical trees (since we do not want to lose connection with the analyses of previous levels); however, its representation for annotators is very close to the plain text.

Annotators first search in the plain texts for possible discourse connectives and arguments of the connectives. Then they mark the assumed extent of discourse arguments on the tectogrammatical layer, link them with a discourse relation, and choose from the list of possible semantic types of the discourse relation. The appropriate discourse connective is also marked.

3. Evaluation of parallel annotations

In order to evaluate the inter-annotator agreement on selected text annotated by two or more annotators, we use F_1-measure for the agreement on arrows, types and connectives (their various combinations), and Cohen's κ (Cohen, 1960) for the agreement on types of arrows. By the agreement on arrows we mean agreement on the start and target nodes. Cohen's κ is used for measuring the agreement on types of those arrows where the annotators agreed on the start and target nodes. Tables 1 and 2 show results of three subsequent measurements (each performed on different data). The measurement #1 shows the average inter-annotator agreement between each of three annotators and an exemplar annotation. The measurements #2 and #3 show the agreement between two selected annotators.

Measurement	F_1 on arrows	F_1 on arrows and types	F_1 on arrows and connectives
Measurement #1 (74 sentences)	0.43	0.3	0.35
Measurement #2 (71 sentences)	0.44	0.39	0.44
Measurement #3 (68 sentences)	0.55	0.39	0.5

Table 1: The inter-annotator agreement on arrows, on arrows and types, and on arrows and connectives

Measurement	F_1 on arrows, types and connectives	Cohen's κ on types
Measurement #1 (74 sentences)	0.22	0.63
Measurement #2 (71 sentences)	0.39	0.81
Measurement #3 (68 sentences)	0.33	0.59

Table 2: The inter-annotator agreement on arrows, types and connectives, and on types

This first attempt at inter-annotator agreement confirms the general feelings that the taste of discourse annotation is more difficult than an annotation of lower levels in that it relies to a greater extent on individual annotators' interpretation of a broader context. If some of the restrictions are relaxed, the figures demonstrate a certain improvement, see below Sect. 4.3.

4. Cases of typical disagreement

The first evaluation of parallel annotations of selected texts brought up some interesting observations. Reflecting the results, we were able to distinguish several repeatedly occurring problematic issues in the annotations. The nature of these disagreements corresponds to the general problem of a formal description on such a high level of language, namely – the texts sometimes allow for different, equally relevant interpretations. So, as for the annotators, two general issues appeared to be difficult to decide: where the connective indeed connects two discourse-relevant text units, and, second, what is the exact extent of these units (arguments of the relation). These issues are closely analyzed in the sections 4.1 to 4.3, with real-data examples.

4.1 Semantic types of discourse relations

Contrary to our assumptions, a disagreement in the semantic type of the assigned relation is not so frequent. In other words, when annotators recognize presence of a discourse connective and determine the discourse arguments, all of them usually mark up the same type of the semantic relation.

4.2 Discourse and non-discourse relations: NPs and elided verbs

However, there was a relatively high degree of disagreement in the very recognition of a discourse relation in some typical cases. (The further examples represent the most common questions of the annotators.)

A trivial example is the fact that expressions acting as discourse connectives can be used in non-discourse contexts. Cf.
He took his hat ***and*** *went home.* (discourse-relevant coordination)
mother ***and*** *father* (discourse-irrelevant NP coordination)
The disagreement occurs when it is not clear whether the potential discourse connective refers to the whole sentence as an independent abstract object (discourse argument), or just to its complement, typically an NP. This ambiguity is common in sentences including verbs with a vague, general meaning (cf. 1; discourse connective is in bold).

(1)
[Arg1: Případ má několik problémových rovin.]
[Arg2: ***První*** *je fakt, že ačkoli uchazečka dosáhla při přijímacích zkouškách lepších výsledků než mužští uchazeči (bylo jich přijato 17 s horšími výsledky), nebyla ke studiu přijata právě a jenom proto, že je žena.]* (specification)

[Arg1: The case has several problematic points.]
[Arg2: ***The first*** *is the fact that although a female candidate succeeded in the entrance test better than male candidates (there were 17 accepted with worse results), she has not been accepted to study precisely and only because she is a woman.]* (specification)

According to one of the possible interpretations, the second sentence of the example (1) is a specification of the content of the first sentence. In this case, the relation is considered a discourse-relevant relation.
In another interpretation, the second sentence characterizes solely the NP *several problematic points*. Then the relation is not a matter of discourse analysis but

rather a relation of (one type of) coreference. Example (2) points to a similar situation.

(2)
*[Arg1: Při prohlídce střech Šternberského paláce si lze všimnout drobného, avšak charakteristického rozdílu mezi přístupem památkářů koncem 80. let a nyní**:COLON**] [Arg2: zatímco komíny staré sněmovny byly zbourány jako zbytečné a zůstala jen holá střecha, dělníci KDM mají přikázáno komíny všech čtyř objektů nejen ponechat, ale dokonce mírně přizdobit, aby tradiční kolorit malostranských střech časem nezmizel.]* (specification)

*[Arg1: When observing the roofs of the Sternberg Palace it is possible to note a small, but distinctive difference between the approaches of preservationists of late 80's and now**:COLON**] [Arg2: while chimneys of the old Parliament were demolished as functionless and only a clear roof was retained, the KDM workers are ordered not only to maintain chimneys of all the four objects, but even to decorate them slightly, so that the traditional local atmosphere of Lesser Town roofs does not eventually disappear.]* (specification)

In this case, the first argument involves either the whole clause, or just the NP *a small, but distinctive difference between the approaches of preservationists of late 80's and now*.
For a unification of annotation, we decided to consider the relations in these cases (1-2) as being coreferential rather than discourse-relevant.

In a similar vein, the existence of a discourse argument is often doubtful in structures with an elided verb in which a potential discourse connective occurs, cf. (3).

(3)
*Tato fakta svědčí i tom, že [Arg1: státní úředníci nemají dostatečný respekt,] [Arg2: možná **snad ani** představu o požadavcích Listiny základních práv a svobod]*. (gradation)

*These facts also suggest that [Arg1: state officials do not have enough respect,] [Arg2: perhaps not **even** an idea of the requirements of the Charter of Fundamental Rights and Freedoms]*. (gradation)

A question arises whether in (3) the connective connects independent abstract objects (*they have no respect and they have no idea*, cf. Figure 1 (3a)), or just parts dependent on the verb that are not discourse arguments (*they have no respect and no idea*, cf. Figure 2 (3b)).

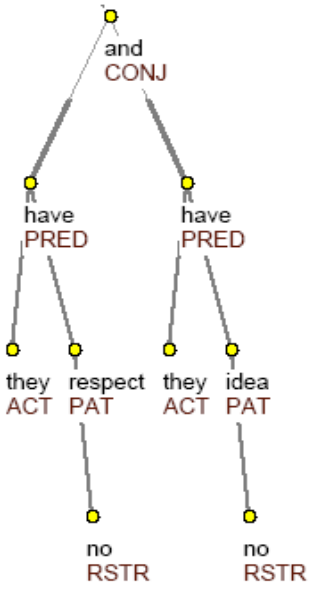

Figure 1: (3a) Tree representation for *They have no respect and they have no idea.*

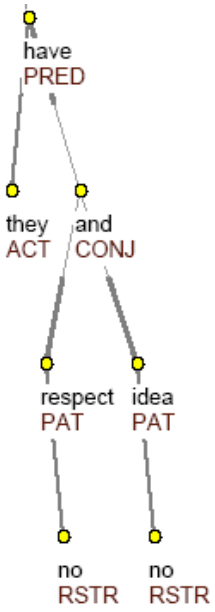

Figure 2: (3b) Tree representation for *They have no respect and no idea.*

In these cases, we consider both parts of the relation discourse arguments (separate clauses with an elided verb, as in 3a) if there is any modification that only applies to one of the coordination members and at the same time it is immediately dependent on the verb of this member (as *perhaps* in 3c). The modification (*perhaps*) constrains the two parts from being connected together into a simple coordination without an insertion of the elided verb (**no respect and perhaps no idea*). According to this criterion, the example (3) is understood to involve a discourse relation (cf. Figure 3 (3c)).

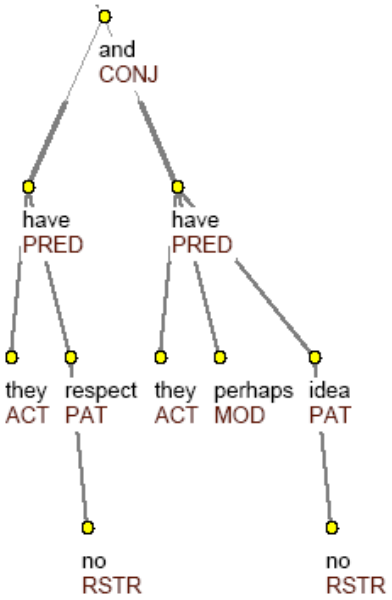

Figure 3: (3c) Tree representation for *They have no respect and* ***perhaps*** *no idea.*

Semantic difference between the two possible interpretations of structures like (1-3) may often not be crucial. Nevertheless, it is crucial to catch the same linguistic phenomena in the same way and to set clear borders of discourse annotation, in order to provide systematic and coherent linguistic data.

4.3 Extent (scope) of a discourse argument: Verbs of thinking and speaking

In some cases, there are no doubts about the existence of a discourse relation, but the extent (scope) of the discourse argument is arguable. Typically, there is annotators' disagreement in structures with governing verbs of thinking or speaking. Often it is not clear whether the discourse argument contains the governing verb or just the content of thought or speech (dictum), cf. (4).

(4)
[Arg1: Na tom, aby ve Šternberku ani v paláci Smiřických nevznikaly žádné příčky, trvají památkáři.]
[Arg2: Poslancům ***tudíž*** *nebude dopřáno žádné velké soukromí.]* (reason)

[Arg1: Preservationists insist that no partition walls will be built up neither in the Sternberg Palace nor in the Smiřický Palace.]
[Arg2: ***Therefore****, MP's will not enjoy great privacy.]* (reason)

In one of the annotators' interpretation of the discourse structure of (4), the governing verb is not included into the discourse argument:

[Arg1: No partition walls will be built up in the buildings.]
[Arg2: ***Therefore****, MP's will have no privacy.]*

In another solution the first argument is larger:
[Arg1: Preservationists insist that no partition walls will be built up in the buildings.]
[Arg2: ***Therefore****, MP's will have no privacy.]*

To ensure agreement, we recommended in these cases to take into consideration whether the meaning of the governing clause is substantial, i.e. whether it reflects an important operation to be carried out on the idea of the dependent clause. In (4) the governing verb is unambiguously a part of the discourse argument: it is necessary to know whether the preservationists *insist* on the idea, or for example, they *forbid* it.
To address this issue in the evaluation of the inter-annotator agreement, we have performed the same tests as before, this time allowing the annotators to disagree slightly either in the start or the target node of the arrows. By "slightly" we mean a difference of one level in the tree. For example, if node A is a parent of node B, then we consider arrows A→C and B→C to be in agreement, as well as arrows D→A and D→B. Tables 3 and 4 show results of the three measurements of the inter-annotator agreement, this time allowing for skipping one level at either the start or the target node.

Measurement	F_1 on arrows	F_1 on arrows and types	F_1 on arrows and connectives
Measurement #1 (74 sentences)	0.53	0.33	0.41
Measurement #2 (71 sentences)	0.5	0.44	0.5
Measurement #3 (68 sentences)	0.67	0.44	0.61

Table 3: The inter-annotator agreement on arrows, on arrows and types, and on arrows and connectives, with allowed skipping of one level either at the start or the target node

Measurement	F_1 on arrows, types and connectives	Cohen's κ on types
Measurement #1 (74 sentences)	0.24	0.56
Measurement #2 (71 sentences)	0.44	0.84
Measurement #3 (68 sentences)	0.39	0.53

Table 4: The inter-annotator agreement on arrows, types and connectives, and on types, with allowed skipping of one level either at the start or the target node

The numbers show improvement in agreement on arrows (about 10%) and on the combination with agreement on types and/or connectives (less than 5%), while Cohen's κ – measured solely on types – has either slightly improved or worsened, depending on the measurement (i.e. on the material tested).

5. Conclusion

As demonstrated by the results of parallel annotations, it is crucial at this moment to distinguish discourse relations from other types of relations within the sentence and in the text. At this stage, the research of discourse semantic relations and unification of discourse annotation is closely linked to syntactic analysis. Setting of annotation scenario can be only done consistently with regard to the syntactic construction. Likewise, it is necessary to determine the extent (scope) of discourse arguments in definable cases on the basis of syntactic structure.

In these tasks, the connection to the syntactico-semantic analysis of the tectogrammatical layer in the Prague Dependency Treebank appears as a rather convenient tool. It makes it possible to work with already established (and coherent) solutions of typical syntactic constructions, such as ellipses, coordinations etc.

6. Acknowledgements

The research reported in this contribution has been carried out under the grant projects of the Grant Agency of Czech Republic (GACR 405/09/0729, GACR 201/09/H057), the Center for Computational Linguistics CKL (LC536), the Czech Ministry of Education (MSM-0021620838, ME-10018), and the Grant Agency of Charles University in Prague (GAUK 103609).

7. References

Asher, N. (1993). *Reference to Abstract Objects in Discourse.* Dordrecht: Kluwer Academic Publishers.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), pp. 37--46.

Hajič, J. et al. (2006). *Prague Dependency Treebank 2.0.* Philadelphia: Linguistic Data Consortium.

Mladová, L., Zikánová, Š., Hajičová, E. (2008). From Sentence to Discourse: Building an Annotation Scheme for Discourse Based on Prague Dependency Treebank. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*. Marrakech, Morocco: European Language Resources Association, ISBN 2-9517408-4-0, pp. 1--7.

Prasad, R. et al. (2007). The Penn Discourse Treebank 2.0 Annotation Manual.

Prasad, R., Dinesh N., Lee A., Miltsakaki, E., Robaldo, L., Joshi, A., Webber, B. (2008). The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).*

http://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb-annotation-manual.pdf