Analyzing the Most Common Errors in the Discourse Annotation of the Prague Dependency Treebank

Pavlína Jínová, Jiří Mírovský, Lucie Poláková
Charles University in Prague
Institute of Formal and Applied Linguistics
Malostranské nám. 25, Prague 1, Czech Republic
E-mail: `(jinova|mirovsky|polakova)@ufal.mff.cuni.cz`

Abstract

We present an analysis of the inter-annotator discrepancies of the Czech discourse annotation in the Prague Dependency Treebank 2.0. Having finished the annotation of the inter-sentential semantic discourse relations with explicit connectives in the treebank, we report now on the results of the evaluation of the parallel (double) annotations, which is an important step in the process of checking the quality of the data. After we shortly describe the annotation and the method of the inter-annotator agreement measurement, we present the results of the measurement and, most importantly, we classify and analyze the most common types of annotators' disagreement.

1 Discourse in the Prague Dependency Treebank 2.0

The Prague Dependency Treebank 2.0 (PDT; Hajič et al. [3]) is a manually annotated corpus of Czech journalistic texts, annotated on three layers of language description: morphological, analytical (the surface syntactic structure), and tectogrammatical (the deep syntactic structure) (Hajič et al. [2]). On the tectogrammatical layer, the data consist of almost 50,000 sentences.

Annotation of discourse structure in PDT uses a lexical approach, similarly to one of the Penn Discourse Treebank (PDTB 2.0, Prasad et al. [9]). It is based on identifying a discourse connective (an expression with text structuring function), which takes two text segments as its arguments and indicates a discourse meaning between them. The annotators mark three basic types of information: the connective (contrary to the Penn approach, there is no list of possible discourse connectives in advance), the two arguments of the connective (mostly clauses, sentences but sometimes also larger units, such as paragraphs) and the semantic type of the relation. At this stage of the project, we do not annotate implicit relations (relations without a connective).

A set of discourse semantic types was developed as a result of comparison of the sense hierarchy used in Penn (Miltsakaki et al. [5]) and the set of Prague tectogrammatical labels called functors (Mikulová et al. [4]).

Annotators had at their disposal both plain text and the tectogrammatical analysis (tree structures). The annotation was carried out on the tectogrammatical trees; a specialized annotation tool was developed for this purpose (Mírovský et al. [7]). The process of annotation spanned over two years (Mladová et al. [8]). Altogether in all PDT data, there have been 8,834 inter-sentential discourse arrows annotated.

Example I shows two sentences with a discourse relation of type *opposition* between them:

(I) *1. Čtyři ostrovní státy nabídly 260 vojáků.*
[Four island states offered 260 soldiers.]
2. Podle mluvčího Pentagonu jich ***ale*** *budou zapotřebí aspoň dva tisíce.*
*[****However****, according to the Pentagon spokesman, at least two thousand of them will be needed.]*

2 Evaluation of parallel annotations

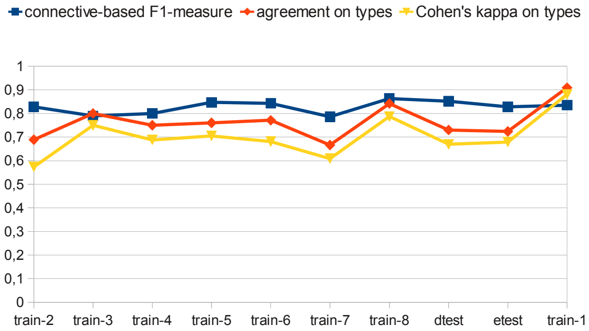
Several annotators annotated the data but (for obvious reasons of limited resources) each part of the data has only been annotated by one of them. Only 4% of the data (2,084 sentences) have been annotated in parallel by two annotators. We used the parallel (double) annotations for measuring the inter-annotator agreement, and for analyzing the most common errors, i.e. difficult parts of the annotation. Altogether, there have been 44 documents, 2,084 sentences and 33,987 words annotated in parallel.

To evaluate the inter-annotator agreement (IAA) on selected texts annotated in parallel by two annotators, we used the connective-based F1-measure (Mírovský et al. [6]), a simple ratio, and Cohen's κ (Cohen [1]). The connective based F1-measure was used for measuring the agreement on the recognition of discourse relations, a simple ratio and Cohen's κ were used for measuring the agreement on the type of relations recognized by both the annotators.[1]

In the connective-based measure, we consider the annotators to be in agreement on recognizing a discourse relation if two connectives they mark (each of the connectives marked by one of the annotators) have a non-empty intersection (technically, a connective is a set of tree nodes). For example, if one of the annotators marks two words *a proto [and therefore]* as a connective, and the other annotator only marks the (same) word *proto [therefore]*, we take it as agreement – they both recognized the presence of a discourse relation. (They still may disagree on its type.)

[1] In all our measurements, only inter-sentential discourse relations have been counted.

Graph 1 shows results of subsequent measurements of the agreement between the two most productive annotators during the two years of annotation. Each measurement was taken on approx. 200 sentences (3 to 5 documents).



Graph 1: The inter-annotator agreement in the subsequent measurements

The overall F1 measure on the recognition of discourse relations was 0.83, the agreement on types was 0.77, and Cohen's κ was 0.71. Altogether, one of the annotators marked 385 inter-sentential discourse relations, the other one marked 315 inter-sentential discourse relations.

The simple ratio agreement on types (0.77 on all parallel data) is the closest measure to the way of measuring the inter-annotator agreement used on subsenses in the annotation of discourse relations in the Penn Discourse Treebank 2.0, reported in Prasad et al. [9]. Their agreement was 0.8.

Table 1 shows a contingency table of the agreement on the four major semantic classes, counted on the cases where the annotators recognized the same discourse relation. The simple ratio agreement on the four semantic classes is 0.89, Cohen's κ is 0.82. The agreement on this general level in the Penn Discourse Treebank 2.0 was 0.94 (Prasad et al. [9]).

	contr	**contin**	**expans**	**tempor**	**total**
contr	137	2	5	1	145
contin	1	49	5		55
expans	4	8	60	3	75
tempor		1	1	7	9
total	142	60	71	11	284

Table 1: A contingency table of the agreement on the four discourse super types (semantic classes)

3 Analysis of the discrepancies

In our measurements and analyses of the IAA, we observe two main types of disagreement: (1) disagreement in identifying the connective, i.e. a situation when one annotator recognized a connective that the other did not, and (2) disagreement in the semantic type of the relation, i.e. a situation when both the annotators recognized a relation anchored by the same connective but they did not characterize this relation in the same way semantically.

3.1 Disagreement in identifying the discourse connective

The analysis of the cases of disagreement in the discourse connective identification shows that it is in the vast majority a mistake of one of the annotators (80 cases in sum). A situation where different interpretations of both annotators are correct only occurs once in our parallel data.

10% of these disagreement cases are rather of a technical than of a linguistic nature: both annotators marked the discourse relation but one of them forgot to add the connective. 15% represent the cases in which one of the annotators overlooked a typical connective – and with it also the relation itself, 75% stand for cases in which the connective is represented by an expression that in Czech also has a non-connective function. This ambiguity probably contributes to the fact that these expressions are more easily left out when reading the text than the typical connectives are. This applies especially for so-called rhematizers, i.e. particles with a rheme signalling function (e.g. *také [also]*, *jenom [only]*, for details on rhematizers see Mikulová et al. [4]).

3.2 Disagreement in the semantic type of a discourse relation

Two main types of disagreement can be distinguished in determining the semantic type of a discourse relation (approximately 60 cases in sum):

First, there are cases of disagreement that are clearly a mistake of one of the annotators: the context does not allow for the asserted interpretation. It represents 26% of the total amount of disagreement and we see their main source in misunderstanding the text.

Second, some cases of disagreement in the semantic type cannot be considered mistakes of the annotators. 7% arise as a consequence of the fact that some of the relations seemed to be defined very clearly but in the course of the annotation they proved to be quite difficult to distinguish from one another in a complicated real-data situation (e.g. a case of relation of *explication* vs. relations of *reason-result* and *specification*).

The remaining cases of disagreement in the semantic type can be divided into (i) agreement and (ii) disagreement within the four major semantic classes. (i) represents situations where each annotator assigned a different type of a relation **within one** of the four basic semantic classes (which are *temporal relations*, *contingency*, *contrast* and *expansion*) and both these interpretations are equally correct (approx. 26% of all cases of disagreement in the semantic type). This situation is typical especially for relations from the *contrast* group. This type of disagreement does not need to be treated as a complete disagreement: the IAA measurement method in the Penn Discourse Treebank 2.0 considers such cases as agreement on the higher level in the sense hierarchy (cf. Table 1).

The remaining type of disagreement are cases where two semantic types **from different** major semantic classes have been used for interpretation of one relation (approx. 41% of all cases of the disagreement in type). These cases of disagreement arise directly from contexts that allow both interpretations. Often, some semantically rather vague connectives contribute to those cases. It is illustrated by example (II). The relation between the sentences in the example text was interpreted as *reason-result* (one argument is the reason of the other one) and also as *conjunction* (the second argument adds new information to the first one). Both interpretations are possible here.

(II) *Za nabídku by se nemusel stydět ani Don Carleone - nebylo možné jí odolat.* ***A tak*** *do roka a do dne dostalo práci 440 shanonských občanů a do pěti let jich bylo už desetkrát tolik.*

Not even Don Carleone would have to be ashamed of that offer - it was impossible to resist. ***And so*** *440 people of Shannon got a job within a year and a day, and within five years, they were already ten times as many.*

This type of disagreement between a relation from the *contingency* group (e.g. *reason-result*) and a relation from the *expansion* group (e.g. *conjunction, equivalence*) is the most frequent disagreement across different major semantic classes. This situation, in our opinion, follows from a certain grade of vagueness of some journalistic formulations – we are allowed to treat the text sequences both causally and as a simple build-up of the previous context.

4 Conclusion

We presented an evaluation and analysis of disagreements in the annotation of the inter-sentential discourse relations with an explicit connective in the Prague Dependency Treebank. The results show that agreeing on an existence of a discourse relation via a surface-present discourse connective is a manageable task, whereas the annotation of the semantic type of the relation depends heavily on the interpretation of the text. The comparison of our annotation with the annotation of Penn Discourse Treebank 2.0 showed that on two levels of granularity of discourse types (senses), the inter-annotator agreement was roughly the same in these two projects.

Almost all cases of disagreement in identifying a connective have to be interpreted as a mistake of an annotator. A majority of them originates from the fact that some of these expressions in Czech have also other functions than the one of a discourse connective. As for the disagreement in the semantic type, annotators' mistakes are not so frequent. This type of disagreement arises from (1) semantic closeness of some relation types, (2) semantic ambiguity/vagueness of some contexts, and (3) in the case of the relation of *explication* also from the complex nature of the relation.

Acknowledgments

We gratefully acknowledge support from the Grant Agency of the Czech Republic (grants P406/12/0658 and P406/2010/0875) and the Ministry of Education, Youth and Sports in the Czech Republic, program KONTAKT (ME10018) and the LINDAT-Clarin project (LM2010013).

References

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

[2] Hajič J., Vidová Hladká B., Böhmová A., Hajičová E. (2000) The Prague Dependency Treebank: A Three-Level Annotation Scenario. In: *Treebanks: Building and Using Syntactically Annotated Corpora (ed. Anne Abeille)*, Kluver Academic Publishers

[3] Hajič J., Panevová J., Hajičová E., Sgall P., Pajas P., Štěpánek J., Havelka J., Mikulová M., Žabokrtský Z., Ševčíková-Razímová M. (2006) Prague Dependency Treebank 2.0. *Software prototype, Linguistic Data Consortium,* Philadelphia, PA, USA, ISBN 1-58563-370-4, www.ldc.upenn.edu

[4] Mikulová M. et al. (2005) Annotation on the tectogrammatical layer in the Prague Dependency Treebank. *Annotation manual.* Available from http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/pdf/t-man-en.pdf

[5] Miltsakaki E., Robaldo L., Lee A., Joshi A. (2008) Sense annotation in the Penn Discourse Treebank. In: *Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science, Vol 4919,* pp. 275–286

[6] Mírovský J., Mladová L., Zikánová Š. (2010) Connective-Based Measuring of the Inter-Annotator Agreement in the Annotation of Discourse in PDT. In: *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010),* Beijing, China, pp. 775–781

[7] Mírovský J., Mladová L., Žabokrtský Z.: Annotation Tool for Discourse in PDT. In: *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, Tsinghua University Press, Beijing, China, pp. 9–12

[8] 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

[9] 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*, CD-ROM