Connective-Based Measuring of the Inter-Annotator Agreement in the Annotation of Discourse in PDT

Jiří Mírovský, Lucie Mladová, Šárka Zikánová
Charles University in Prague
Institute of Formal and applied Linguistics
{mirovsky,mladova,zikanova}@ufal.mff.cuni.cz

Abstract

We present several ways of measuring the inter-annotator agreement in the ongoing annotation of semantic inter-sentential discourse relations in the Prague Dependency Treebank (PDT). Two ways have been employed to overcome limitations of measuring the agreement on the exact location of the start/end points of the relations. Both methods – skipping one tree level in the start/end nodes, and the connective-based measure – are focused on a recognition of the existence and of the type of the relations, rather than on fixing the exact positions of the start/end points of the connecting arrows.

1 Introduction

1.1 Prague Dependency Treebank 2.0

The Prague Dependency Treebank 2.0 (PDT 2.0; Hajič et al., 2006) is a manually annotated corpus of Czech. It belongs to the most complex and elaborate linguistically annotated treebanks in the world. The texts are annotated on three layers of language description: morphological, analytical (which expresses the surface syntactic structure), and tectogrammatical (which expresses the deep syntactic structure). On the tectogrammatical layer, the data consist of almost 50 thousand sentences.

For the upcoming release of PDT, many additional features are planned, coming as results of several projects. Annotation of semantic inter-sentential discourse relations is one of the planned additions.

To ensure the highest possible quality of the annotated data, it would be best if several annotators annotated the whole data in parallel. After solving discrepancies in the annotations of the individual annotators, we would get a high-quality annotation. This approach is sometimes employed, but most of the times, the available resources prohibit it (which is also the case of the discourse annotation project). Manual annotation of data is a very expensive and time consuming task. To overcome the restriction of limited resources, each part of the data is annotated by one annotator only, with the exception of a small overlap for studying and measuring the inter-annotator (dis-)agreement.

1.2 Inter-Annotator Agreement in Computational Linguistics

Measuring the inter-annotator agreement has long been studied (not only) in computational linguistics. It is a complex field of research and different domains require different approaches.

Classical measures *recall*, *precision* and *F-measure* offer the most straightforward and intuitively interpretable results. Since they do take into account neither the contribution of chance in agreement, nor different importance of different types of disagreement, etc., other more or less elaborate coefficients for measuring the inter-annotator agreement have been developed. Cohen's κ (Cohen, 1960) is suitable for classification tasks and tries to measure the agreement "above chance". Krippendorff's α (Krippendorff, 1980) can be used if we need to distinguish various levels of disagreement. Rebecca Passonneau (2004) offered a solution for measuring agreement between sets of elements (like words in coreferential chains). Variants of these coefficients can be used for measuring agreement among more than two annotators. A comprehensive overview of methods for measuring the inter-annotator agreement in various areas of

Coling 2008: Poster Volume, pages 775–781,
Beijing, August 2010

computational linguistics was given in Artstein and Poesio (2008).

For measuring the inter-annotator agreement in the annotation of semantic inter-sentential discourse relations in PDT, we have chosen two measures. The relations do not form natural chains (unlike e.g. textual and grammatical coreference) and a simple *F_1-measure* is well suited for the agreement on existence of the relations. For the agreement on types of the relations, which is a typical classification task, we use Cohen's κ.

Our research has then been focused not on "how to measure" the agreement (which coefficient to use), but rather on "what to measure" (which phenomena), which is the topic of this paper.

2 Annotated Phenomena

Since the Prague Dependency Treebank 2.0 already contains three layers of linguistic annotation, two of which (the analytical layer – surface syntax, and the tectogrammatical layer – underlying syntax and semantics) are tree representations, we took advantage of these existing analyses and carry out the annotation of discourse phenomena directly on the trees (the tectogrammatical layer). It means that we capture the discourse relation between any two (sub)trees in the document by drawing a link (an arrow) between the highest nodes in the (sub)trees, see Figure 1.

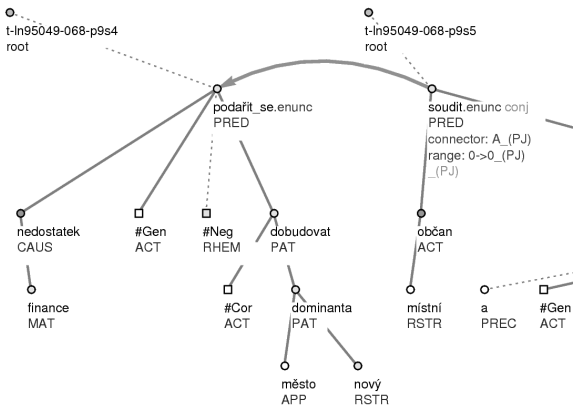

Figure 1. A discourse arrow between two nodes represents a discourse relation between two trees – subtrees of the nodes.

Discourse relations we annotate are in principle semantic relations that apply between two abstract objects (Asher, 1993) (i.e. discourse units or text spans) and help make the text a coherent whole. These relations are often signaled by the presence of a discourse connective, i.e. expressions as *"ale"*, *ačkoliv"*, *"tedy"*, *"ovšem"* (in English *"but"*, *"although"*, *"then"*, *"however"* etc. In the first phase of the project, we only annotate relations (link the (sub)trees) where such a connective is present.

Every relation gets assigned two important attributes: first, the discourse connective that anchors the relation, and, second, the semantic type of the relation. For assigning semantic relations in the discourse, we developed a set of 22 discourse-semantic tags (Mladová et al., 2009). It is inspired partly by the set of semantic labels used for the annotation of the tectogrammatical layer in PDT 2.0, relations within the sentence (the tectogrammatical syntactico-semantic labels called functors, Mikulová et al., 2005) – since some of the semantic relations apply also intra-sententially, like causal or contrastive relations; and partly by the set of semantic tags in the Penn Discourse Treebank 2.0 (Prasad et al., 2008), a discourse annotation project for English with similar aims.

Hence, there are three important issues for the inter-annotator measurement on the discourse level of annotation in PDT: the agreement on the start and target nodes of the discourse relation (and so the extent of the discourse arguments), the agreement on the discourse connective assigned to the relation, and, last but not least, the agreement on the semantic type of the relation.

3 Measuring the Inter-Annotator Agreement in the Annotation of Discourse in PDT 2.0

3.1 Simple (Strict) Approach

The basic method we use for measuring the inter-annotator agreement requires a perfect match in the start and end points of the relations. We calculate *recall* and *precision* between the two annotators. Since these measures are not symmetric in respect to the annotators, we use their combination – *F_1-measure* – which is symmetric. At each node, we compare target

nodes of the discourse relations created by the two annotators. We consider two relations to be in agreement strictly only if they share both the start node and the target node.

A second number we measure is an agreement on the relation and the type. For considering two relations to be in agreement, we require that they share their start and target nodes, and also have attached the same type.

Similarly, we measure an agreement on the relation and the connective, and an agreement on the relation, the type and the connective.

Attaching a type to a relation can be understood as a classification task. We calculate two numbers – simple ratio agreement and Cohen's κ – on the types attached to those relations where the annotators agreed on the start and the target nodes. Cohen's κ shows the level of agreement on the types above chance.

For completeness, we also calculate simple ratio agreement on the connectives attached to those relations the annotators agreed on.

Table 1 shows results of these measurements on two hundred sentences annotated in parallel by two annotators.[1]

measure	value
F_1-measure on relations	0.43
F_1-measure on relations + types	0.34
F_1-measure on relations + connectives	0.41
F_1-measure on rel. + types + connect.	0.32
agreement on types	0.8
agreement on connectives	0.95
Cohen's κ on types	0.74

Table 1. The inter-annotator agreement for a strict match.

3.2 Skipping a Tree Level

Requiring a perfect agreement on the start node and the target node of the discourse relations turns out to be too strict for a fair evaluation of the inter-annotator agreement. It often happens that the annotators recognize the same discourse relation in the data but they disagree either in the start node or the target node of the relation.

In Zikánová et al. (2010), we elaborate on typical cases of this type of disagreement and show that in many times, the difference in the start node or the target node is only one level in the tree. We have also shown that these disagreements usually depend on a subtle and not crucial difference in the interpretation of the text.

Figure 2 shows an example of a disagreement caused by a one-level difference in the target node of a relation. The two trees (a cut of them) represent these two sentences:

"Vím, že se nás Rusů bojíte, že nás nemáte rádi, že námi trochu pohrdáte. Ale Rusko není jenom Žirinovskij, Rusko není jenom vraždění v Čečensku."

(In English: *"I know that you are afraid of us Russians, that you dislike us, that you despise us a little. But Russia is not only Zhirinovsky, Russia is not only murdering in Chechnya."*)

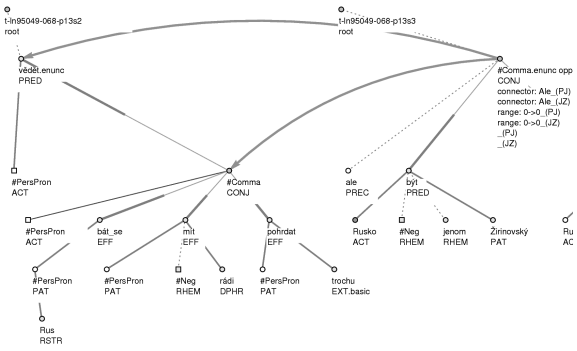

Figure 2. Disagreement in the target node.

Both annotators recognized the discourse relation between the two sentences, both selected the same type (opposition), and both marked the same connective (*"Ale"*, in English *"But"*). The disagreement in the target node is caused by the fact that one annotator has connected the second sentence with *"knowing that something is going on"*, while the other has connected it directly with the expression *"something is going on"*.

We have shown in Zikánová et al. (2010) that allowing for skipping one tree level either at the start node or the target node of the relations leads to an improvement in the inter-annotator

[1] The annotators did not know which part of the data will be used for the measurement. The agreement was measured on 200 sentences (6 documents). PDT 2.0 contains data from three sources. The proportion of the sentences selected for the measurement reflected the total proportion of these data sources in the whole treebank.

agreement (F_1-*measure* on the relations) of about 10%. To be exact, by allowing to skip one tree level we mean: 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. Table 2 shows present results of this type of measurement, performed on the same data as Table 1.

measure	value
F_1-measure on relations	0.54
F_1-measure on relations + types	0.43
F_1-measure on relations + connectives	0.49
F_1-measure on rel. + types + connect.	0.39
agreement on types	0.8
agreement on connectives	0.92
Cohen's κ on types	0.73

Table 2. The inter-annotator agreement with one-level skipping.

The results seem to be consistent, since the improvement here is similar to the previously published test. The F_1-*measure* on the relations improved from 0.43 to 0.54. On the other hand (and also consistently with the previous test), simple ratio agreement on types (or connectives) and Cohen's κ on types, all measured on those arrows the annotators agreed on, do not change (more or less) after skipping one level is allowed. For these three measures, skipping one level only adds more data to evaluate and does not change conditions of the evaluation.

3.3 Connective-Based Approach

Further studies of discrepancies in parallel annotations show that skipping one level does not cover all "less severe" cases of disagreement.

Figure 3 presents an example of a disagreement in the start node of a relation with a two-level distance between the nodes. The two trees (a cut of them) represent these two sentences:

"Racionální kalkulace vlastníků nájemních bytů je proto povede k jedinému závěru: jakékoliv investice do oprav a modernizace nájemního bytového fondu jsou a budou ztrátové. Proto je další chátrání nájemních domů neodvratné."

(In English: *A rational calculation of the owners of the apartments will lead them to the only conclusion: any investment in repairs and renovation of the rental housing resources is and will be loss-making. Therefore, further dilapidation of the apartment buildings is inevitable."*)

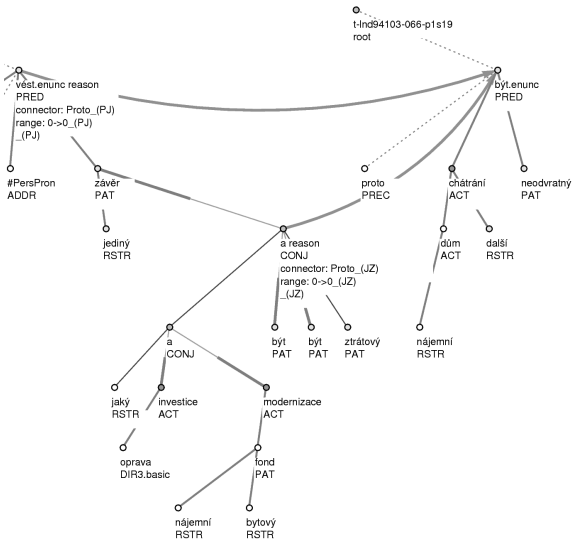

Figure 3. Two-level disagreement in the start nodes

The difference between the annotators is that one of them started the relation at the phrase *"will lead to the only conclusion: any investment ... is and will be ..."*, while the other started the relation directly at the phrase *"any investment ... is and will be ..."*.

However, both the annotators admittedly recognized the existence of the discourse relation, they also selected the same type (reason), and marked the same connective (*"Proto"*, in English *"Therefore"*).

Figure 4 shows an example of a disagreement caused by a different selection of nodes and by the opposite direction of the arrows. The trees represent these sentences: *"To je jasné, že bych byl radši, kdyby tady dosud stál zámek a ne tohle monstrum. Ale proč o tom stále uvažovat?"*

(In English: *It is clear that I would prefer if there still was a castle here and not this monster. But why keep thinking about it forever?"*)

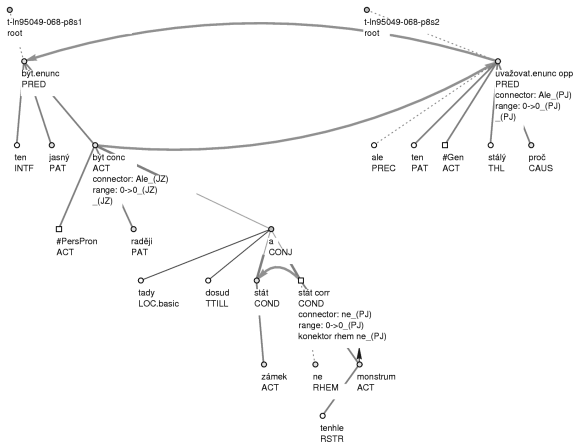

Figure 4. Disagreement in the nodes and in the direction of the arrows.

This time, both annotators recognized a presence of a discourse relation and marked the same connective (*"Ale"*, in English *"But"*). They did not agree on the start/end nodes and on the type of the relation (opposition vs. concession).

Figure 5 shows another type of "slight" disagreement. This time, the annotators agreed on everything but the range of the relation. They agreed both on the type (reason) and the connective (*"tak"*, in English *"Thus"*). The three trees (again a cut of them) represent these three sentences:

"Podle šéfa kanceláře představenstva a. s. Škoda Zdeňka Lavičky jsou však v říjnu schopny fungovat prakticky všechny závody bez vážnějšího omezení. To je v rozporu s tvrzením vedení koncernu z minulého týdne, ve kterém škodovácký management tvrdil, že se odstávka dotkne většiny provozů a závodů Škody Plzeň, která má v současnosti 28000 zaměstnanců. Vzniká tak podezření, že se vedení koncernu snažilo vyvinout tlak na vládu a donutit ji k zaplacení dluhů."

(In English: *"According to Zdeněk Lavička, the chief of the board of directors of Škoda corp., virtually all factories are able to operate in October without serious limitations. It contradicts the statement of the syndicate administration from the last week, in which the management of Škoda claimed that the downtime would affect most of the plants and factories of Škoda Plzeň, which presently has 28,000 employees. Thus a suspicion arises that the syndicate administration tried to exert pressure on the government and force it to pay the debts."*)

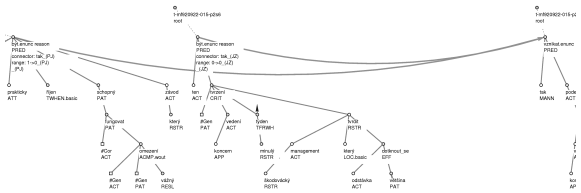

Figure 5. Disagreement in the range of the discourse relation.

The difference between the annotators is in the range of the start part of the arrows. One of the annotators marked the two first sentences as a start point of the relation, while the other marked the second sentence as the start point only. They agreed on the target point of the relation being the third sentence.

Inspired by these examples, we designed another – a connective-based – measure for evaluating the inter-annotator agreement of the discourse relations. It seems that although the annotators sometimes fail to mark the same start/target nodes, or to select the same type or the same range of the relations, they usually agree on the connective. This idea is also supported by high levels of the simple ratio agreement on connectives measured on relations the annotators agreed on from Tables 1 and 2 (0.95 and 0.91). These numbers show that once the annotators agree on a relation, they almost always agree also on the connective.[2]

The connective-based measure considers the annotators to be in agreement on recognizing a discourse relation if they agree on recognizing the same connective (please note that we only annotate discourse relations with explicitly expressed connectives).

Table 3 shows results of the evaluation of the inter-annotator agreement, performed using the connective-based measure, on the same data as Tables 1 and 2.

[2] This is only an interpretation of the numbers, not a description of the annotation process; in fact, the annotators usually first find a connective and then search for the arguments of the discourse relation.

measure	value
F_1-measure on relations	0.86
F_1-measure on relations + types	0.56
F_1-measure on rel. + start/end nodes	0.43
F_1-measure on rel. + types + nodes	0.34
agreement on types	0.65
agreement on start/end nodes	0.50
Cohen's κ on types	0.56

Table 3. The inter-annotator agreement evaluated with the connective-based measure.

This time (compared with Tables 1 and 2, i.e. the simple strict measure and the one-level skipping measure), the agreement (*F_1-measure*) on relations is much higher – 0.86 (vs. 0.43 and 0.54). On the other hand, simple ratio agreement (and Cohen's κ) measured on relations recognized by both annotators are lower than in Tables 1 and 2. Although the annotators might have recognized the same discourse relation, a (possibly small) difference in the interpretation of the text caused sometimes not only a disagreement in the positions of the start/end nodes, but also in the type of the relation.

The simple ratio agreement on types from Table 3 (0.65) is probably the closest measure to the way of measuring the inter-annotator agreement on subtypes in the annotation of discourse relations in the Penn Discourse Treebank 2.0, reported in Prasad et al. (2008). Their agreement was 0.8.

4 Conclusion

We have presented several ways of measuring the inter-annotator agreement in the project of annotating the semantic inter-sentential discourse relations with explicitly expressed connectives in the Prague Dependency Treebank. We have shown examples from parallel annotations that substantiate the importance of the alternative approaches to the evaluation of the agreement.

Skipping a tree level in the start node or the end node of the relations helps to recognize factual agreement in some cases where the strict approach detects disagreement. We have shown that it is still too strict and that there are cases which we would like to classify as agreement but the measure does not recognize them.

The connective-based measure seems to be the closest one to what we would like to consider a criterion of agreement. It disregards the actual nodes that are connected with a discourse relation, and even disregards the direction of the relation. In this sense, it is the most benevolent of the three measures.

It does not mean that the simple strict measure or skipping a tree level are inferior or obsolete ways of measuring the agreement. All the measures focus on different aspects of the agreement and they are all important in the process of annotating the corpus, studying the parallel annotations and improving the annotation instructions. We may agree on the fact that on this level of language description, it is very hard to achieve perfect agreement (Lee at al., 2006), yet we should never cease the effort to further specify and clarify the ways of annotation, in order to catch the same linguistic phenomena in the same way, and thus provide systematic and coherent linguistic data.

Acknowledgments

We gratefully acknowledge support from the Czech Ministry of Education (grant MSM-0021620838), the Grant Agency of the Czech Republic (grants 405/09/0729 and P406/2010/0875), the Czech Science Foundation (grant 201/09/H057), and the Grant Agency of Charles University in Prague (GAUK 103609).

References

Artstein R. and M. Poesio. 2008. *Inter-coder agreement for computational linguistics.* Computational Linguistics 34/4, pp. 555–596.

Asher, N. 1993. *Reference to Abstract Objects in Discourse.* Kluwer Academic Publishers

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

Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M., Žabokrtský, Z., and M. Ševčíková-Razímová. 2006. *Prague Dependency Treebank 2.0.* CD-ROM, LDC2006T01, Linguistic Data Consortium, Philadelphia, USA.

Krippendorff, K. 1980. *Content Analysis: An Introduction to Its Methodology*. Chapter 12. Sage, Beverly Hills, CA, USA.

Lee, A., Prasad, R., Joshi, A., Dinesh, N., and B. Weber. 2006. *Complexity of dependencies in discourse: Are dependencies in discourse more complex than in syntax?* J. Hajič and J. Nivre, (eds.). Proceedings of the 5th Workshop on Treebanks and Linguistic Theories (TLT 2006). Prague, Czech Republic, pp. 79–90.

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

Mladová L., Zikánová, Š., Bedřichová, Z., and E. Hajičová. 2009. *Towards a Discourse Corpus of Czech.* Proceedings of the Corpus Linguistics Conference, Liverpool, Great Britain, in press (online proceedings: http://ucrel.lancs.ac.uk/publications/cl2009/).

Passonneau, R. 2004. *Computing Reliability for Coreference.* Proceedings of LREC, vol. 4, Lisbon, Portugal, pp. 1503–1506.

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

Zikánová Š., Mladová L., Mírovský J., and P. Jínová. 2010. *Typical Cases of Annotators' Disagreement in Discourse Annotations in Prague Dependency Treebank.* LREC 2010, Malta, in press.