Ways of Evaluation of the Annotators in Building the Prague Czech-English Dependency Treebank

Marie Mikulová, Jan Štěpánek
Charles University in Prague
MFF ÚFAL
{mikulova,stepanek}@ufal.mff.cuni.cz

Abstract

In this paper, we present several ways to measure and evaluate the annotation and annotators, proposed and used during the building of the Czech part of the Prague Czech-English Dependency Treebank. At first, the basic principles of the treebank annotation project are introduced (division to three layers: morphological, analytical and tectogrammatical). The main part of the paper describes in detail one of the important phases of the annotation process: three ways of evaluation of the annotators - inter-annotator agreement, error rate and performance. The measuring of the inter-annotator agreement is complicated by the fact that the data contain added and deleted nodes, making the alignment between annotations non-trivial. The error rate is measured by a set of automatic checking procedures that guard the validity of some invariants in the data. The performance of the annotators is measured by a booking web application. All three measures are later compared and related to each other.

1. Introduction

The annotation of a corpus is always a complex task, especially if the corpus is large and the added linguistic information is rich. A system for evaluation of the annotation and annotators should by an integral part of any annotation project. An example of such a large complex corpus is the Prague Czech-English Dependency Treebank (PCEDT).

The PCEDT is planned to be a corpus of (deeply) syntactically annotated parallel texts (in English and Czech) intended chiefly for machine translation experiments. The texts for the PCEDT (its first version was described in Čmejrek et al., 2004) were taken from the Penn Treebank (Marcus et al., 1993). For the Czech part of the PCEDT, the English texts were translated into Czech. As a base of the process of creation of the corpus (hierarchical system of annotation layers, annotation rules) we use the already accomplished Prague Dependency Treebank (PDT) 2.0 (Hajič et al., 2006). Similarly to the PDT 2.0, written sentences in the PCEDT are represented on three layers: morphological layer (lemmas, tags, morphological categories), analytical layer (surface structure, dependencies, analytical functions) and tectogrammatical layer.

The tectogrammatical layer (Mikulová et al., 2006) contains all the information that is encoded in the structure of a sentence and its lexical items: deep, semantic-syntactic structure, functions of its parts, "deep" grammatical information, coreference and topic-focus articulation including deep word order. Every sentence is represented by a tectogrammatical tree. A node of the tree either represents a semantic unit present in the surface shape of the sentence (an autosemantic word with its function words like prepositions, subordinating conjunctions, auxiliary verbs) or it is a newly established node that has no counterpart on the surface - in case of ellipsis.

We adhere to the stand-off annotation principle: the layers of annotation are separated from the input data and from one another, they are interlinked by references leading always from the hierarchically higher layers to the lower ones. For example, there are two references to the analytical layer from a tectogrammatical node representing a prepositional group: one pointing at the preposition and one at the noun.

The tectogrammatical annotation scheme is complex (39 different attributes, 8.42 attributes filled on average for a node; the annotation manual has more than 1000 pages) and takes a long time (in one hour, one annotator can annotate 9.2 sentences in average in the first phase of the annotation), and then it is important to measure the quality of the annotations and the annotators.

While organizing the annotation of the PCEDT (especially its Czech part, which is the main concern of this article), we will prop ourselves upon multifarious experiences (both positive and negative) gained from the production of the PDT 2.0. In the PCEDT project the quality of the work of a particular annotator is judged by several ways:

- the annotation agreement between annotators is measured,
- the output of the automatic checking procedures tells us how often an annotator makes mistakes compared to the others,
- the annotators book the time they spend annotating; it allows later to evaluate their performance and the relation of the efficiency to the error rate.

In the next sections we describe these ways to measure and evaluate the annotation and the annotators.

Overall	K	94,08%			
	Ma	94,01%			
	A	93,83%			
	O	93,78%			
	Z	84,58%			
Structure	A	88,62%	is_dsp_root	K	95,86%
	Ma	88,60%		A	95,83%
	O	87,92%		Ma	95,75%
	K	87,88%		O	95,72%

	Z	69,28%		Z	89,72%
a/aux.rf	K	93,82%	is_generated	K	96,24%
	Ma	93,58%		A	96,05%
	A	93,55%		Ma	96,03%
	O	93,53%		O	96,02%
	Z	82,45%		Z	90,27%
a/lex.rf	K	96,26%	is_member	K	94,72%
	Ma	96,12%		A	94,70%
	A	96,00%		Ma	94,50%
	O	95,90%		O	94,25%
	Z	89,67%		Z	85,47%
annot_comment	K	96,52%	is_parenthesis	Ma	95,42%
	Ma	96,40%		K	95,40%
	A	96,30%		O	95,27%
	O	96,27%		A	95,15%
	Z	90,43%		Z	88,72%
compl.rf	K	96,32%	is_state	K	96,50%
	Ma	96,22%		Ma	96,25%
	A	96,12%		O	96,13%
	O	96,03%		A	96,13%
	Z	90,18%		Z	90,35%
functor	K	85,70%	t_lemma	K	93,76%
	Ma	85,67%		Ma	93,60%
	O	85,57%		O	92,70%
	A	85,13%		A	92,42%
	Z	66,80%		Z	81,60%

Table 1: Inter-annotator agreement

2. The annotation agreement between annotators

The basic way how to evaluate an annotation is to measure the inter-annotator agreement. However, the structure to be compared is very complex. The algorithm aligning two tectogrammatical trees built upon the same analytical tree is complex accordingly: the nodes have to be aligned one to one in both the compared files. Since annotators can delete nodes as well as add new ones, not all the nodes must be aligned in every tree pair. Various heuristics are used to align nodes that differ in some attribute values, details are described in (Klimeš, 2006). Once the trees are aligned node to node, we just compare the values of all the attributes of all the aligned nodes. To evaluate the structural agreement, we treat the identifier of a node's parent as a new attribute of the node. Complex attributes (lists, structures etc.) need further manipulation in order to be compared. For example, identifiers of linked analytical nodes have to be sorted; for annotator's comment, we only compare its type, because the text can vary even when having the same meaning.

Since there is no "golden" annotation, we just measure the agreement of all the pairs of annotators (see Table 1, data from December 2007 - from the beginning of the process; average value is shown for every attribute, and average value over all the attributes and structure is presented as "Overall"). As a baseline, we use the output of an automatic procedure with which the annotators start their work (marked "Z" in the table). Note that the agreement among annotators is always higher than the agreement between any annotator and the baseline.

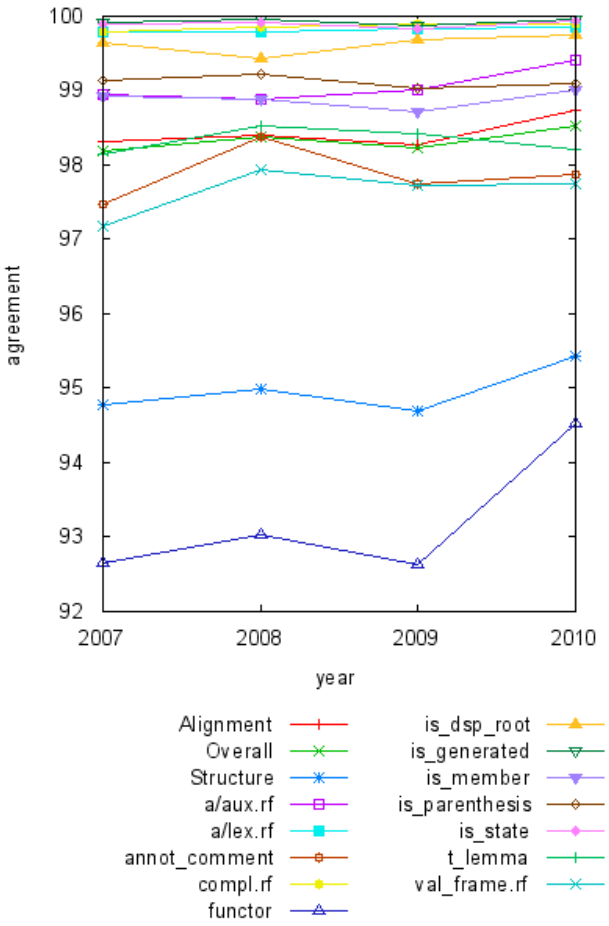

Figure 1: Inter-annotator agreement over four years.

The inter-annotator agreement was satisfactory over the whole process. In Figure 1, we can see that the two main figures (functor and structure) improved (probably due to changes in the annotation rules), while the other figures rather fluctuated. The attributes with a lower difference between baseline and the annotators (about 5%, i.e. is_state, is_generated, is_dsp_root, compl.rf, annot_comment, and a/lex.rf – see Table 1) tend to contain more errors, or have too vague annotation rules. Another reason of the lower difference between the baseline and the annotators is also the rareness of the phenomena described by the attributes (direct speech: is_dsp_root, formulation of state: is_state) that causes the annotators to forget to mark their occurrence. It is probably worth considering to annotate these phenomena separately by an annotator that can concentrate fully on them.

The annotator that agrees most with all the others ("K") is at the same time the annotator that makes the least errors and submits the most sentences (see next sections).

3. Error rate

Using the list of errors generated by the automatic checking procedures (we describe the system for annotation quality checking in Mikulová and Štěpánek, 2009) we can count how often the annotators make errors (only those errors the procedures can detect, of course): the number of errors an annotator made is divided by the number of sentences or nodes he annotated. Table 2 shows the comparison of the error rate for 4 annotators in December 2007 (at the beginning of the process) and current numbers for 7 annotators from July 2009. The numbers from different periods cannot be compared directly because since the beginning there have been more than 30 new checking procedures, which means the current list of errors is longer. On the other hand, the rank of the annotators can be compared.

The table shows that our current best annotator ("K") had approximately 30 errors per 100 sentences and 1.62 errors per 100 nodes. His error rate has not got worse over the two years and he remains the best annotator. The table further shows that the differences in error rate between annotators can be great and that all the annotators keep their positions: no one gets markedly better nor worse – annotator "Ma" is the only annotator whose error rate gets worse over time, but not a lot, the main reason being probably the personality of the annotator. The comparison of veteran annotators and the new ones that have been annotating only for a short time is also interesting: it shows that knack, practice, and experience lead to quality of the annotation.

	December 2007	July 2009	
Who	Errors per 100 sentences	Errors per 100 sentences	Errors per 100 nodes
K	29.7851	1.5103	0.0806
O	39.6699	4.0331	0.2067
Ma	61.4087	8.4670	0.4533
A	63.2318	6.3583	0.3265
L	-	15.0668	0.8010
Mi	-	16.2241	0.8460
J	-	19.0476	1.0971

Table 2: Error rate

4. Performance of the annotators

In the annotation process, even the time spent working by the annotators is measured. The annotators book the time to a web form. For each month the web application counts the annotators' performance over the month and the over-all performance. Performance of the annotators shows us how the annotation process is difficult and time-consuming, how much time a particular annotator or an average annotator spends on one sentence. According to these facts we can estimate how much time the annotation process of the whole planned data volume will take, if the annotation schema does not change. Based on these estimates, we can decide whether it is useful or necessary to extend the annotation team (with regard to financial capacity, of course) or whether it is suitable or urgent to modify and simplify the annotation schema to speed the process up.

The data are important among others to determine the wages; on the basis of the data we tariff a sentence (annotators are being paid monthly according to the number of sentences they have annotated).

Table 3 shows performance of the annotators in October 2008 and June 2009, Table 4 shows the over-all performance. Monitoring the performance illustrates the differences between annotators, but also the fluctuation of each particular annotator. We can also observe the inverse proportionality of the performance and error rate (see section 3): the more efficient an annotator is (he annotates more data), the less errors he makes. This seems to go against the "more haste less speed" principle, nevertheless, regular and frequent annotation of a particular volume of the data appears to be far more essential for a lower error rate and also higher efficiency, because it involves repeated confrontation with annotation rules. The annotator can master the rules more easily, he works faster and makes less errors because of ignorance (there are still errors of inadvertence, but their numbers are not so high and do not change so much). An annotator that works irregularly, just once in a longer period, has not cultivated his intuition, practice, experience, and therefore he works more slowly and produces more errors.

Who	Hours	Sentences	Sentences per hour	Minutes per sentence
A	114.25	963	8.4289	7.1184
I	827.00	7006	8.4716	7.0825
J	105.70	1001	9.4702	6.3357
K	107.00	1430	13.3645	4.4895
L	266.41	1716	6.4412	9.3150
Ma	78.00	615	7.8846	7.6098
Mi	169.98	1655	9.7364	6.1624
O	289.02	3211	11.1100	5.4006

Table 4: Over-all performance of the annotators

5. Conclusion

In the article, we have presented some organizational aspects of building of a large syntactical treebank. We described three ways to measure and evaluate the annotation and annotators.

We believe that all the three methods described here (inter-annotator agreement, measuring of the error rate, and performance of the annotators) are important for the annotation process and evaluation of the annotators. No annotation can be considered high-quality without measuring the inter-annotator agreement. However,

especially in cases of complex and long-term annotation tasks, it is appropriate, if not unavoidable, to append also further measurements of quality of the annotation and annotators.

We believe that being at the very end of the PCEDT project with more than 95 % of the data already annotated our proposals are sufficiently backed by our experience and practice.

Acknowledgement

The research reported in this paper was supported by the Ministry of Education, Youth and Sport of the Czech Republic (within the project LC 536), by the Grant Agency of the Czech Republic (within the project P406/2010/0875 and 406/10/P193) and by the EuromatrixPlus project sponsored by the European Commission (FP7-ICT-2007-3-231720 of the EU, 7E09003 of the Czech Republic).

	October 2008				June 2009			
Who	Hours	Sentences	Sentences per hour	Minutes per sentence	Hours	Sentences	Sentences per hour	Minutes per sentence
A	18.50	147	7.946	7.551	-	-	-	-
I	100.50	742	7.383	8.127	101.50	1229	12.108	4.955
J	11.50	97	8.435	7.113	2.00	28	14.000	4.286
K	33.00	418	12.667	4.737	23.50	332	14.128	4.247
L	46.00	143	3.109	19.301	27.88	365	13.092	4.583
Ma	40.00	310	7.750	7.742	-	-	-	-
Mi	17.85	142	7.955	7.542	24.91	358	14.372	4.175
O	37.81	403	10.659	5.629	56.65	632	11.156	5.378

Table 3: Performance of the annotators

References

Hajič J. et al. (2006). *The Prague Dependency Treebank 2.0*. CD-ROM. Linguistics Data Consortium Cat. No. LDC2006T01. Philadelphia, PA, USA. URL: http://ldc.upenn.edu, http://ufal.mff.cuni.cz/pdt2.0.

Klimeš V. (2006). *Analytical and Tectogrammatical Analysis of a Natural Language*. PhD Thesis. MFF UK, Prague.

Marcus M., Santorini B. and Marcinkiewicz M. A. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, (19):313–330.

Mikulová M. et al. (2006). *Annotation on the tectogrammatical level in Prague Dependency Treebank*. Annotation manual. Technical report ÚFAL TR-2006-30. MFF UK, Prague.

Mikulová M. and Štěpánek J. (2009). Annotation Quality Checking and Its Implications for Design of Treebank (in Building the Prague Czech-English Dependency Treebank). In *Proceedings of the Eight International Workshop on Treebanks and Linguistic Theories*. 4-5, Milan, Italy.