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

Marie Mikulová, Jan Štěpánek

Charles University in Prague, MFF ÚFAL

Performance

The annotators book the time spent working to a web form.

Average speed: 9.2 sentences per hour.

No one gets slower over time.

The fastest annotator is at the same time the one with lowest error rate and highest agreement with others.

Used to determine the wages.

Prague Czech-English Dependency Treebank

Penn Treebank texts (WSJ part) translated to Czech.

Stand-off annotation principle: 4 layers of annotation.

Tectogrammatical layer: 39 possible attributes per node (8.42 used in average).

Error rate

Numbers from different periods can not be compared directly (different data, checking procedures, rules etc.), but the rank of the annotators remains similar.

How to judge the quality of the annotation?

Inter-annotator agreement

No “gold standard” data → all pairs of annotators compared.

Non-trivial task: deleted and added nodes complicate the alignment.

Baseline: output of a parser. The agreement between any two annotators is always higher then the agreement between the baseline and the worst annotator.

PCEDT (Czech part)

Annotation tool: Tree editor TrEd.

Format: PML (Prague Markup Language).

Annotation scheme: similar to Prague Dependency Treebank 2.0.

1st phase of annotation: 4 years, 4-7 annotators.

49208 sentences, over 1 million tokens. Almost 10% annotated in parallel.