**1. Introduction**

- **Task:** to find correspondences between two tectogrammatical (deep-syntactic) trees that represent an English sentence and its Czech translation.
- **Motivation:** aligned tectogrammatical trees are needed for training tree-to-tree transfer models in our MT system.
- **Hypothesis:** tectogrammatical representations of Czech and English sentences are more similar compared to the similarity of the sentence surface shapes, thus higher agreement/precision in alignment should be achievable.

**2. Manually aligned data**

- 515 sentences (about 13,000 tokens) manually aligned on the word level, in parallel by two independent annotators.
- Three types of links distinguished: (a) sure links, (b) possible links, (c) phrasal links.
- The sentences were automatically parsed up to the tectogrammatical layer.
- Then the word alignment was transferred to the tectogrammatical trees in order to provide data for training and testing tectogrammatical aligners.

**3. Alignment algorithm**

**4. Function for scoring candidate node pairs**

- Based on a set of manually designed features - feature vector $f$
- Vector of feature weights
- Scalar product scoring function: $\text{score}(\text{cnode, enode}) = w \cdot f(\text{cnode, enode})$

**5. Evaluation**

- **inter-annotator agreement (f-measure)**
  - on aligning words: 82.1% (f-measure) on aligning t-nodes (i.e., after transferring the manual word alignment to t-trees): 94.7%.
  - performance of the automatic t-trees aligners (f-measure)
    - baseline: t-lemma sequences aligned by GIZA++: 82.6%.
    - alignment of t-trees by our feature-based aligner: 90.4% (10-fold cross validation).

**6. Conclusions**

- Inter-annotator agreement on aligning t-nodes (= content words) is considerably higher than the agreement on aligning all words of the original sentences.
- Our feature-based tectogrammatical aligner outperforms the GIZA++ baseline.