Constructing a Turkish-English Parallel TreeBank
Olcay Taner Yildiz, Ercan Solak, Onur Gorgun, and Razieh Ehsani

Introduction
- Constructed a parallel constituency treebank
- Translated English trees from Penn Treebank
- 5k sentences
- Sentences have 15 tokens or less

Turkish
- Agglutinative
- (quite) Free word order
- Case markings indicate syntactic relations
- Morphemes have many allomorphs (vowel harmony)

Corpus construction strategy
1. Start with English tree in PTB
2. Permute children to get Turkish word order
3. Replace leaves with Turkish glosses

Ordering Rules and Heuristics
- English function words become morphemes
- Turkish morphotactics dictate the tree permutation order
- "NONE" replaces the leaf vacated by an English function word
- Personal pronoun becomes a morpheme

YES/NO Questions
- Turkish question morpheme-word mI is added
- Usually at the end or within the main word stem

Replacement Heuristics
- There is no the in Turkish. Depending on the context we have the -> "NONE", bu (this), o (that).
- Proper nouns are translated if there is a gloss.
- "*NONE*" replaces the leaf vacated by an English function word
- Turkish question morpheme-word mI is added
- Number agreement is a bit relaxed in Turkish
- Tense ambiguity because of category and semantic differences.
- No perfect tense in Turkish, usually mapped to past tense
- Annotators choose the closest tense

Tools
- Visualize the trees at each step
- Easy permutation and replacement
- Helps the annotators with online analysis
- Search a pattern in the treebank

Next steps
- Larger corpus, translate the whole of PTB
- Morphological analysis and movement of morphemes into "NONE" leaves
- Tree modifications after the basic translation, addition of levels
- Use in SMT