ACL 2014

22-27 June, Baltimore

Constructing a Turkish-English Parallel TreeBank

Olcay Taner Yıldız†, Ercan Solak†, Onur Görgün†,†† and Razieh Ehsani†

† Işık University
†† Alcatel Lucent Teletaş
Istanbul, Turkey

FEYZİYE SCHOOLS FOUNDATION
IŞIK UNIVERSITY

Introduction

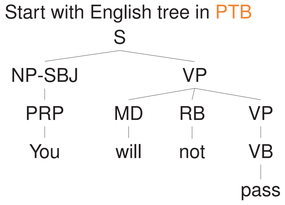
- Constructed a parallel constituency treebank
- Translated English trees from Penn Treebank
- 5K sentences
- Sentences have 15 tokens or less

Turkish

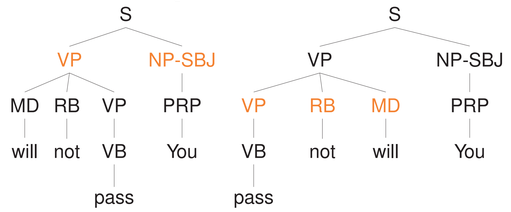
- Agglutinative
- (quite) Free word order
- Case markings indicate syntactic relations
- Morphemes have many allomorphs (vowel harmony)

Corpus construction strategy

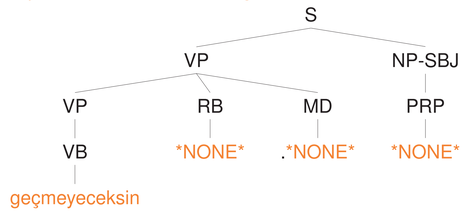
1. Start with English tree in PTB



2. Permute children to get Turkish word order

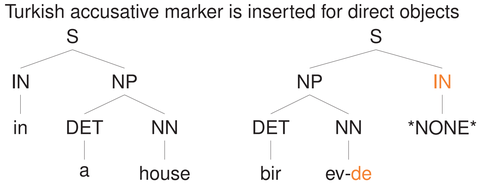


3. Replace leaves with Turkish glosses



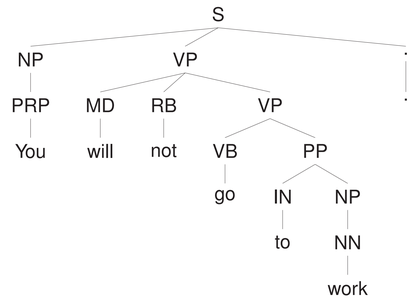
Prepositions

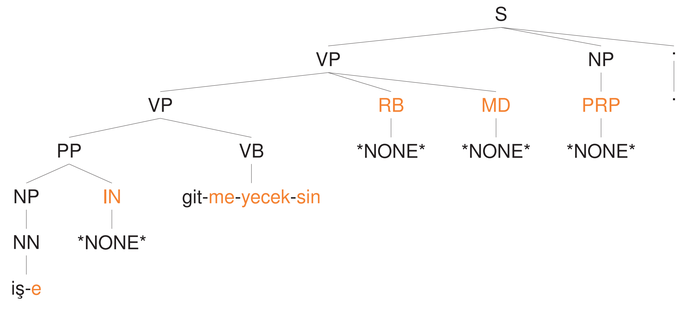
- Turkish has postpositional case markers
- PP's are permuted and prepositions become case markers
- Turkish accusative marker is inserted for direct objects



Ordering Rules and Heuristics

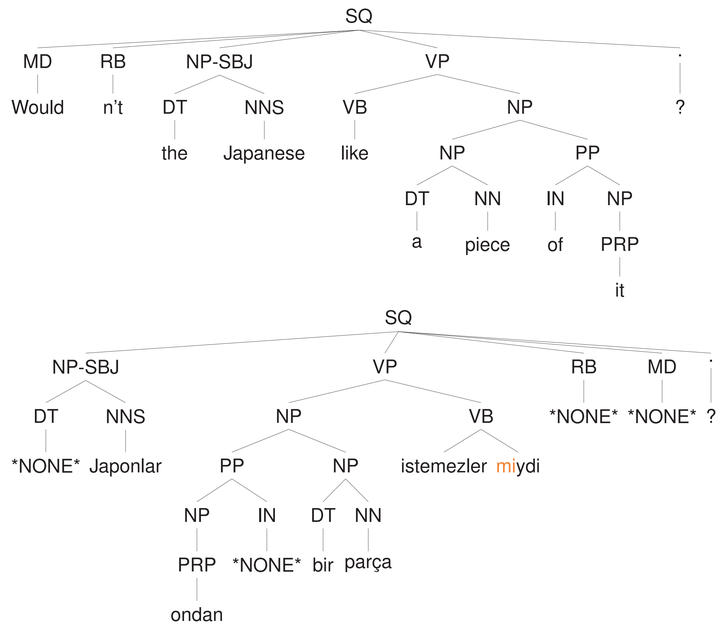
- English function words become morphemes
- Turkish morphotactics dictate the tree permutation order
- *NONE* replaces the leaf vacated by an English function word
- Personal pronoun becomes a morpheme





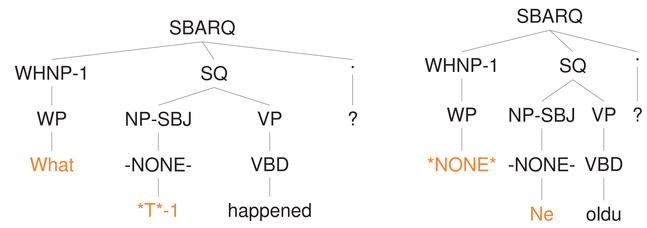
YES/NO Questions

- Turkish question morpheme-word mI is added
- Usually at the end or within the main word stem



WH- Questions

- In Turkish wh- questions, any constituent can be questioned by replacing it with question pronouns or adjectives
- PTB uses traces to indicate moved constituents in wh-questions
- Annotators insert question pronouns into leaves identified by traces

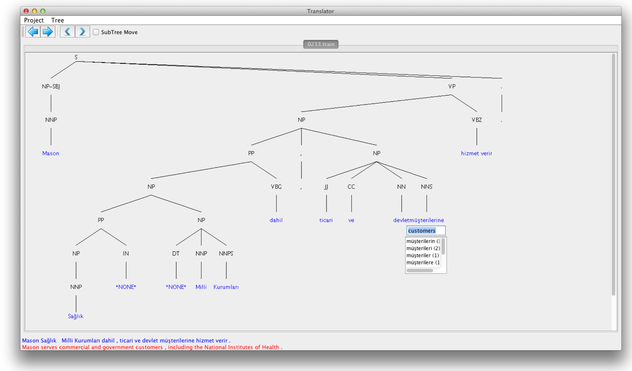


Replacement Heuristics

- 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.
  London → Londra
  United States → Birleşik Devletler
  but
  John → John
- Number agreement is a bit relaxed in Turkish
  çocuk-lar gel-di(ler).
- Tense ambiguity because of category and semantic differences.
  No perfect tense in Turkish, usually mapped to paste 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

email: {olcaytaner,ercan,razieh.ehsani}@isikun.edu.tr, onur.gorgun@alcatel-lucent.com