An All-Words Sense Annotated Turkish Corpus

Sinan Akçakaya Olcay Taner Yıldız

Department of Computer Engineering, Işık University, İstanbul, Turkey

ICNLSP 2018

Outline

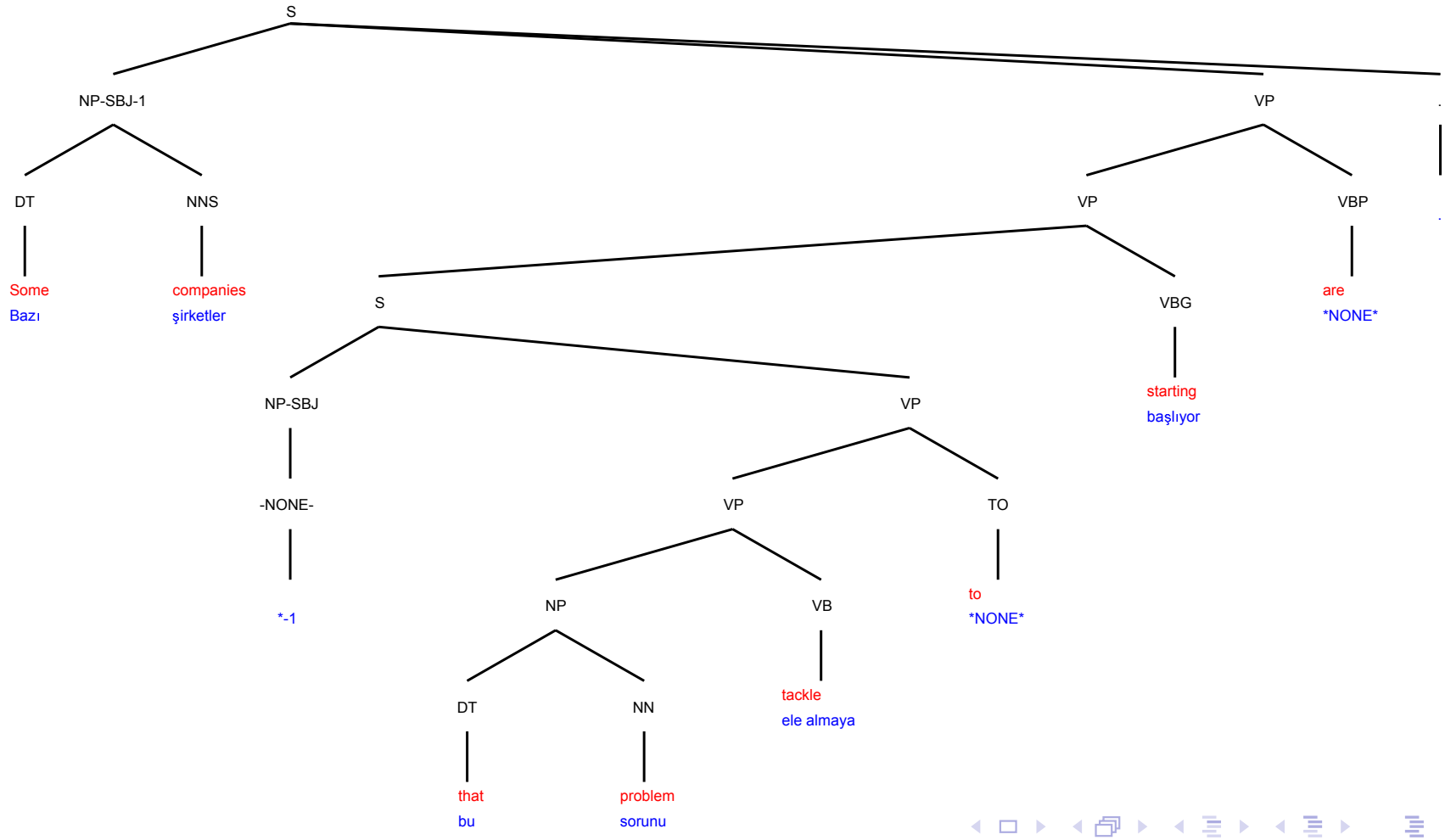
Word Sense Disambiguation

- Word Sense Disambiguation (WSD) is a historical task.
- An acute need for these kinds of resources, especially for non-English languages.
- We present a sense tagged Turkish corpus, which has been built on the preceding parallel treebank construction.

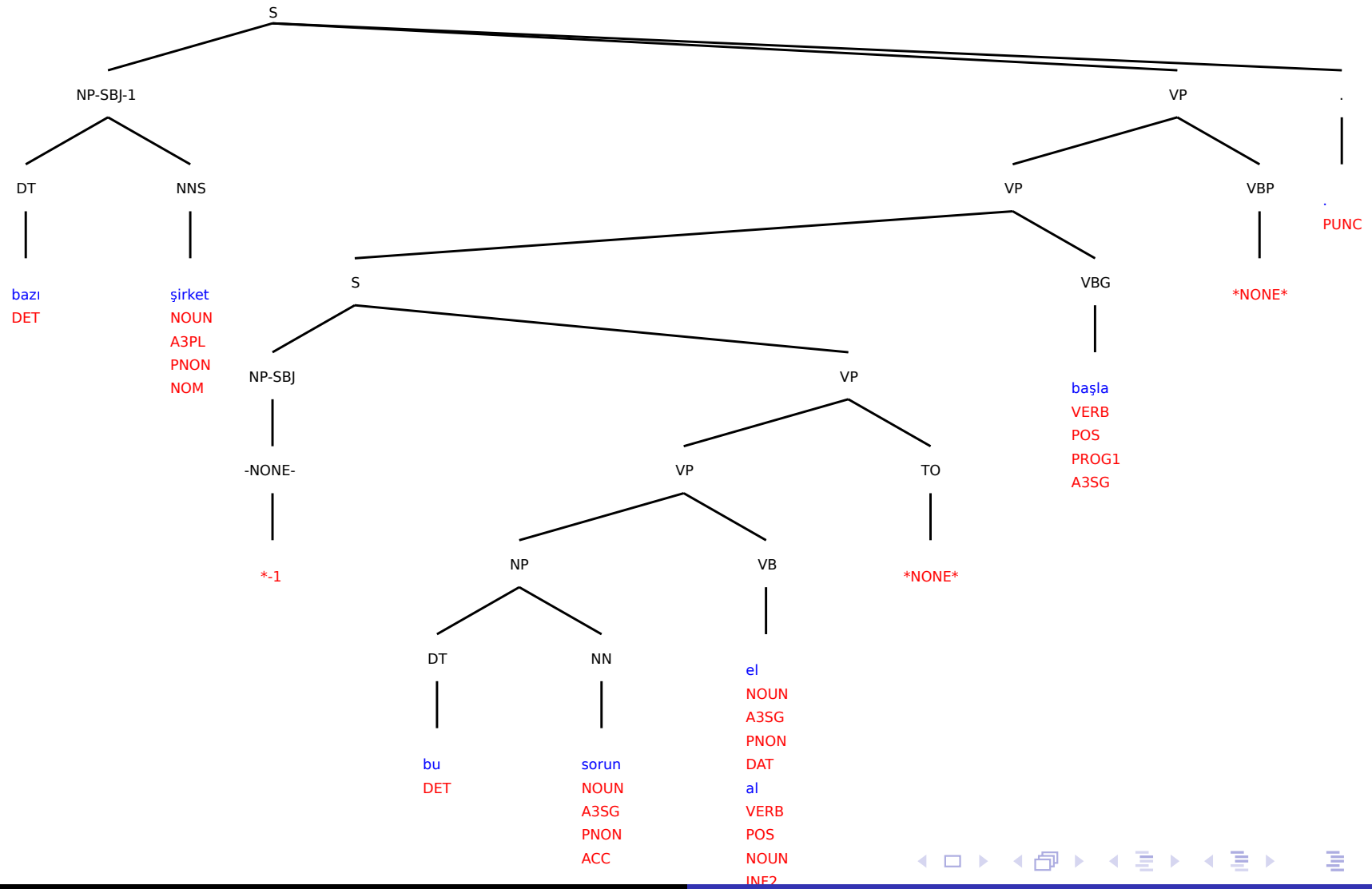
Sense Annotated Corpora

- English
  - SemCor (23,000 different lemmas of 234,000 instances)
  - Line-hard-serve corpus (4,000 sense-tagged examples)
  - DSO corpus (192,800 occurrences of 191 English words)
  - Open Mind Word Expert (90,000 instances of 288 nouns)
  - SENSEVAL competitions
- Turkish
  - Turkish Lexical Sample Task (5,385 samples of 26 highly ambiguous words)
  - Turkish Lexical Sample Dataset (15 words with at least 100 samples)

Translation: "Some companies are starting to tackle that problem" and its Turkish translated form



Morphological Analysis



Sense Inventory

```
<SYNSET>
<ID>TUR10-0066140</ID>
<LITERAL>baba
<SENSE>1</SENSE>
</LITERAL>
<POS>n</POS>
<DEF>...</DEF>
<EXAMPLE>...</EXAMPLE>
</SYNSET>
```

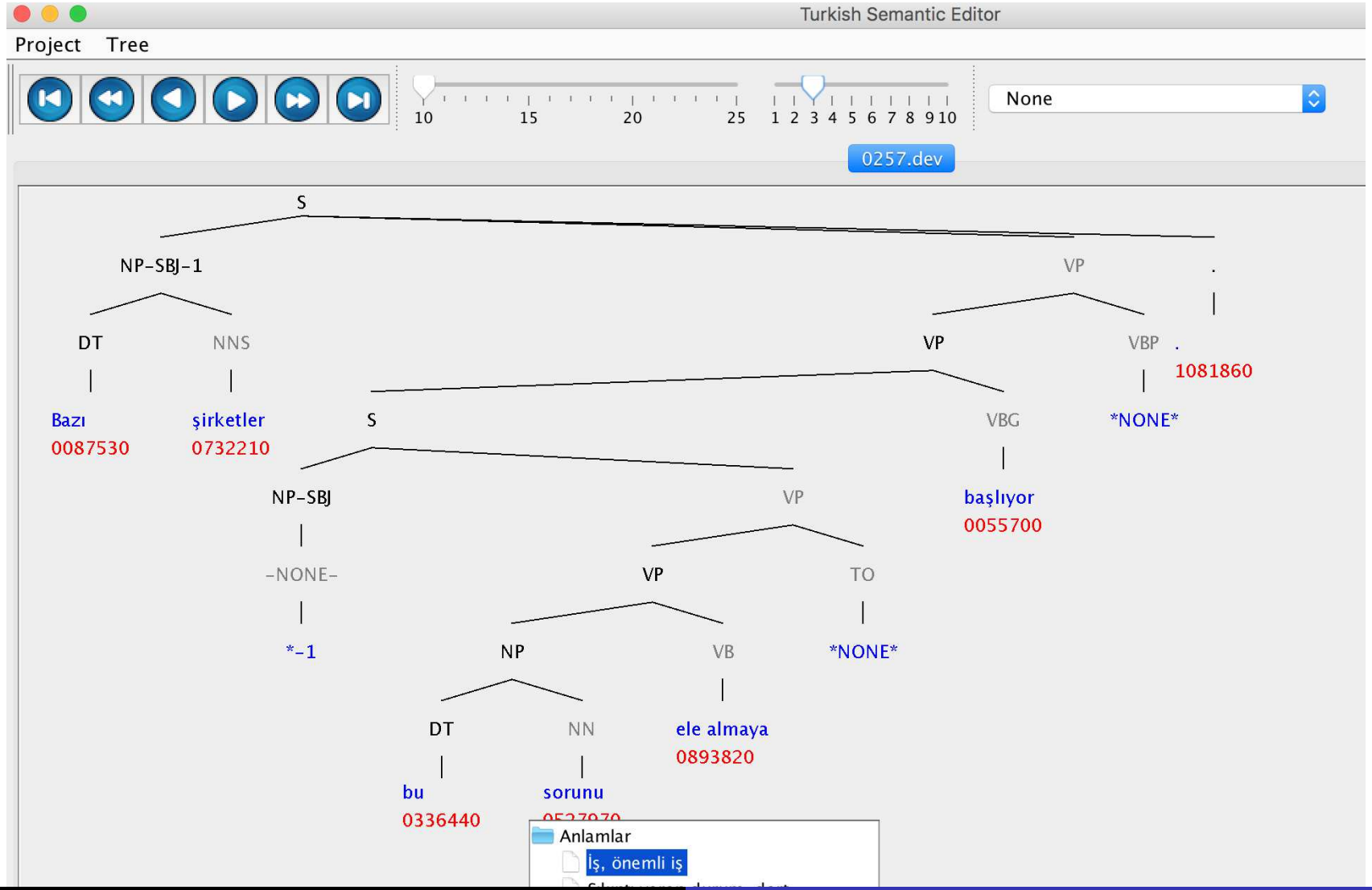
Extracting candidate senses from the dictionary

- Simple Senses (Example: sorunu)
  sorun + NOUN + A3SG + PNON + ACC (the problem)
  sorun + NOUN + A3SG + P3SG + NOM (her/his problem)
  soru + NOUN + A3SG + P2SG + ACC (your question)
- Senses for derivational suffixes (Example: sessizlik)

ses+NOUN+A3SG+PNON+NOM	voice
^DB+ADJ+WITHOUT	silent
^DB+NOUN+NESS	quite, silence

- Senses for collocations (Example: iki yüzlü)
  İki yüzlü ile dostluk kurma (Do not make friends with a hypocrite)
  Available senses: "iki", "yüz", "yüzlü", "iki yüzlü'

Annotation Tool



Data format for Tree Leaves

- Initial Penn-Treebank Data Format

  ```
  (NN problem)
  ```

- Modified Data Format

  ```
  (NN {turkish=sorunu}
  {english=problem}
  {morphAnalysis=sorun+NOUN+A3SG+PNON+ACC}
  {metaMorphemes=sorun+yH}
  {semantics=TUR10-0703650})
  ```

Comparison of Turkish sense annotated corpora

Corpus	Example	Lemma	Coverage	Syntactic Parse
Semeval-2007	5,385	26	Lexical sample	Available
TLSD	3,616	35	Lexical sample	Unavailable
Our corpus	83,474	7,524	All-words	Available

Distribution of word types

Word Type	Sample Size	Distinct Sample Size
Noun	33,320	5,882
Verb	11,981	747
Adjective	9,591	739
Adverb	3,704	227
Number, Range	4,460	
Punctuation	14,372	
Pronoun, etc.	6,046	
Total	83,474	7,595

10 most-frequent words and their rankings

Word	# of Occurence	# of Senses
olmak (be, happen)	1,072	25
etmek (auxiliary verb)	765	9
dolar (dollar)	622	1
hisse (share, stock)	616	3
bay (mr.)	481	3
şirket (company, firm)	391	1
satmak (sell)	383	5
milyon (million)	380	2
yapmak (do, make)	375	20
demek (say, tell)	339	15

Baseline results over 10-fold cross-validation

- 25.04% accuracy for the random baseline and 61.26% for the most frequent sense (MFS) baseline.
- In Turkish, TLSD has 29.15% accuracy for the MFS baseline.
- In English, SENSEVAL-2 and SENSEVAL-3 have 57% and 60.9% for the MFS baseline.

Review

- Our experience on manual tagging of TDK senses in a Turkish-English parallel treebank.
- During the process, same software infrastructure have been used with the previous tasks.
- Creation of this dataset will offer the first all-words sense annotated corpus in Turkish.

Advantages over other annotated corpora of Turkish

- Word coverage is much more extensive, since all words are labeled.
- Root forms and morphological structure are on hand.
- Syntactic features available in the parsed sentences makes it possible to acquire more information about the words.

Questions?