The Turkish Treebank

Kemal Oflazer
Sabancı University / CMU

Turkish

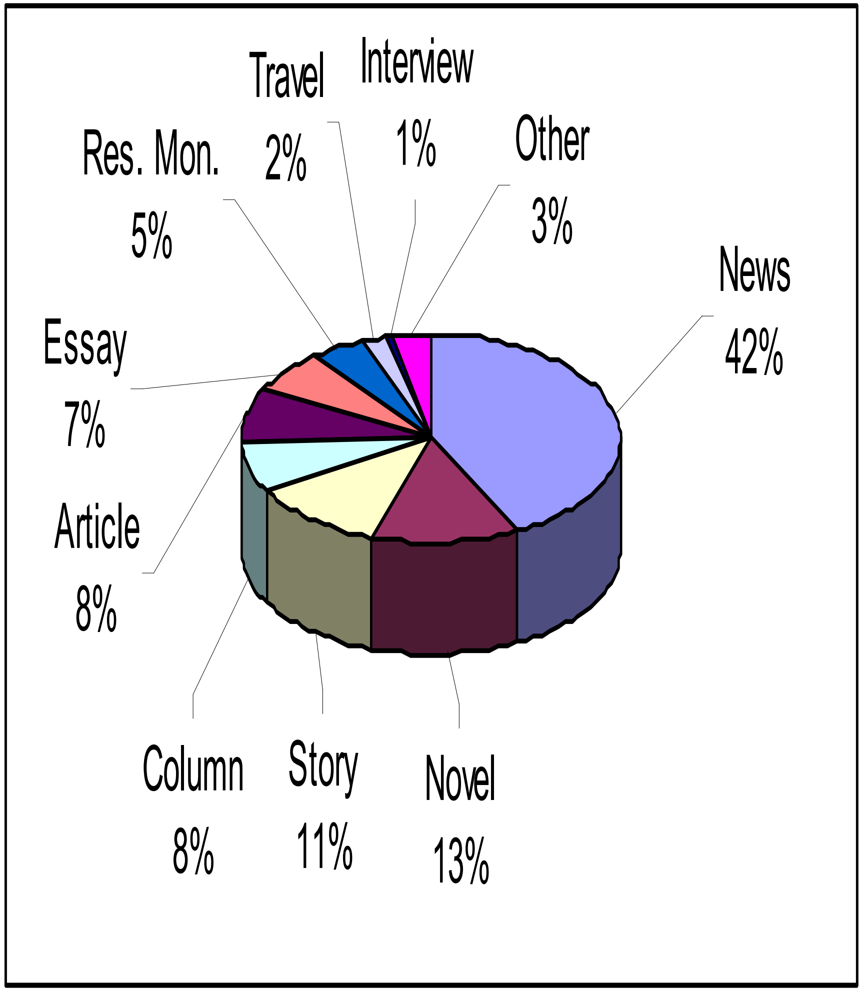
- Complex agglutinative morphology with very productive inflectional and derivational processes
  - Word forms consist of morphemes concatenated to a root morpheme or to other morphemes
  - Practically infinite vocabulary

- Free constituent order
  - Unmarked constituent order is SOV
  - All orders are possible with minimal formal constraints.

- Syntactic relations in the treebank are between sublexical units called Inflectional Groups

The Turkish Treebank

- About 5600 sentences of 10 genres of post-1990 written Turkish text extracted from METU Turkish Corpus
  - full morphological annotation
    - derivations explicitly marked
  - surface dependency relations



Lexical Annotation

- A word is represented sequence of inflectional groups (IGs) of the form

  Lemma+Infl$_1$^DB+Infl$_2$^DB+…^DB+Infl$_n$

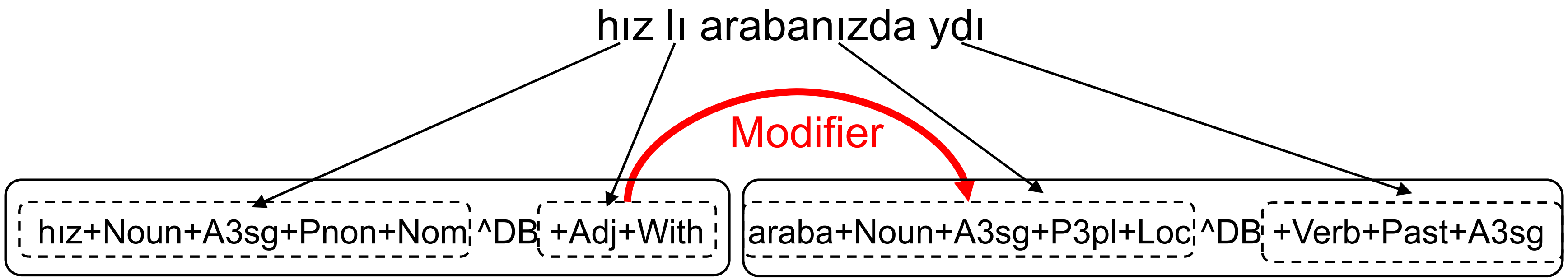
- Encoded using an extensive set of
  - Inflectional, derivational and morpho-semantic features

- **iyileştiriliyorken**
  - **(literally) while it is being caused to become good**
  - **while it is being improved**
- **iyi+Adj ^DB+Verb+Become^DB+Verb+Caus ^DB+Verb+Pass+Pos+Pres^DB+Adverb+While**

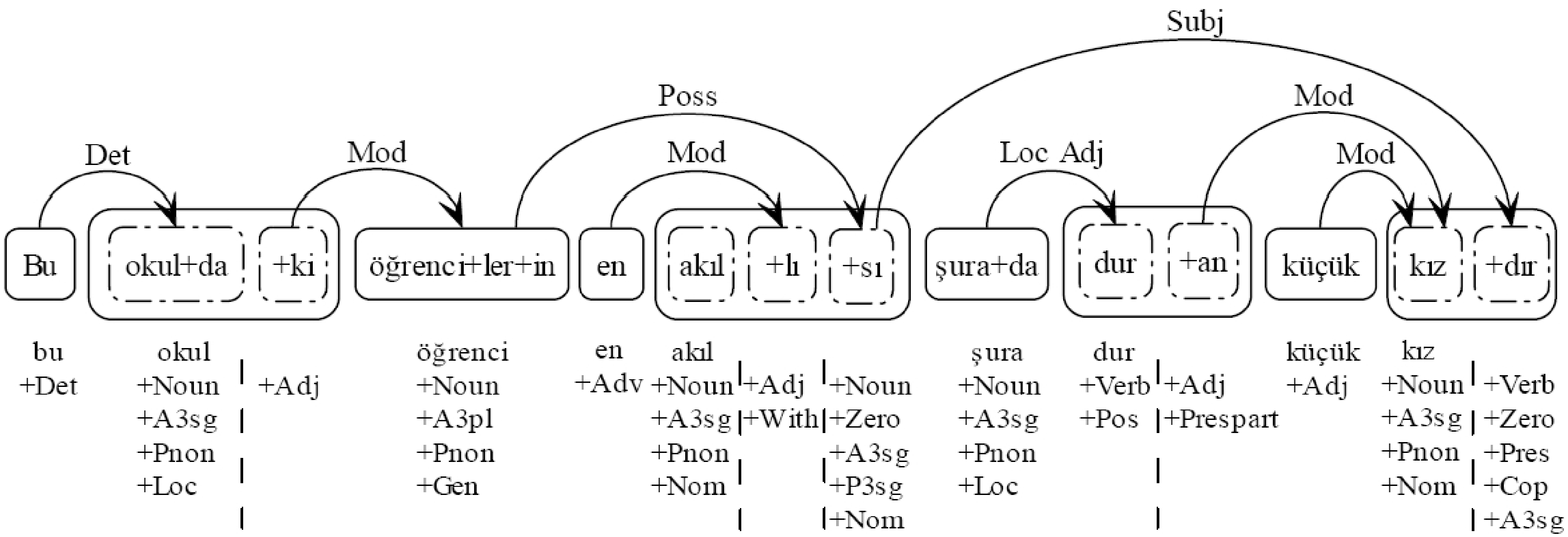
Syntactic Annotation

- Surface dependency relations are between inflectional groups
  - About 20 relation types
- Links emanate from the last IG of the dependent and land on some IG of the head IG

hızlı arabanızdaydı
(it was in your fast/speedy car)



Syntactic Annotation



bu +Det | okul +Noun +A3sg +Pnon +Loc | +Adj | öğrenci +Noun +A3pl +Pnon +Gen | en +Adv | akıl +Noun +A3sg +Pnon +Nom | +Adj +With | +Noun +Zero +A3sg +P3sg +Nom | şura +Noun +A3sg +Pnon +Loc | dur +Verb +Pos | +Adj +Prespart | küçük +Adj | kız +Noun +A3sg +Pnon +Nom | +Verb +Zero +Pres +Cop +A3sg

This school-at+that-is student-s-' most intelligence+with+of there stand+ing little girl+is

The most intelligent of the students in this school is the little girl standing there.

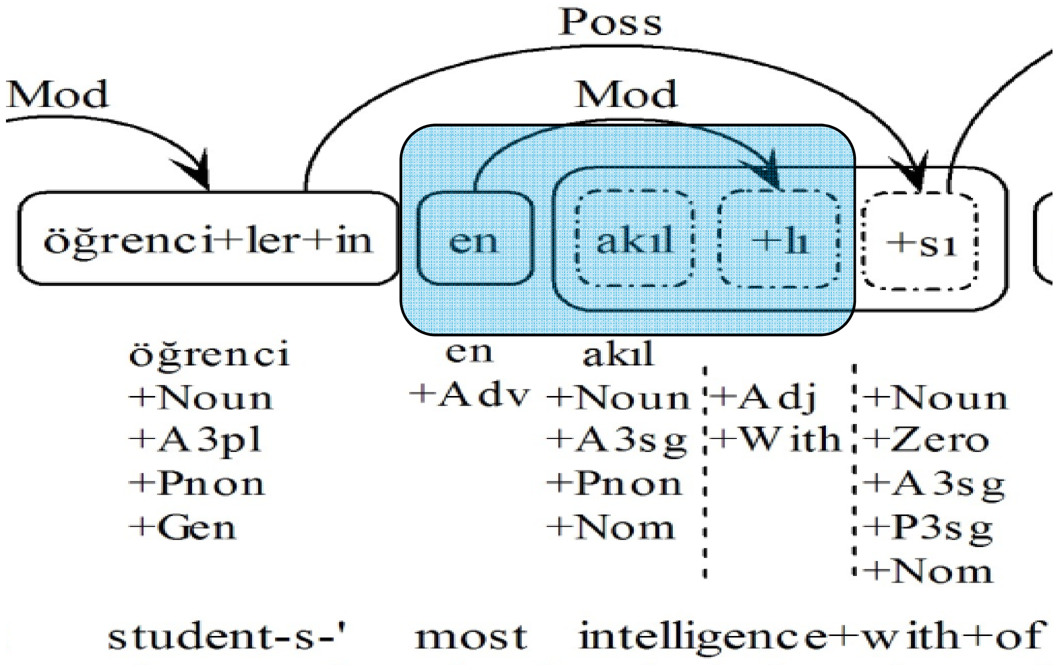
Figure 1

Dependency links in an example Turkish sentence.

+'s indicate morpheme boundaries. The rounded rectangles show words while IGs within words that have more than one IG are indicated by the dashed rounded rectangles. The inflectional features of each IG as produced by the morphological analyzer are listed below the IG.

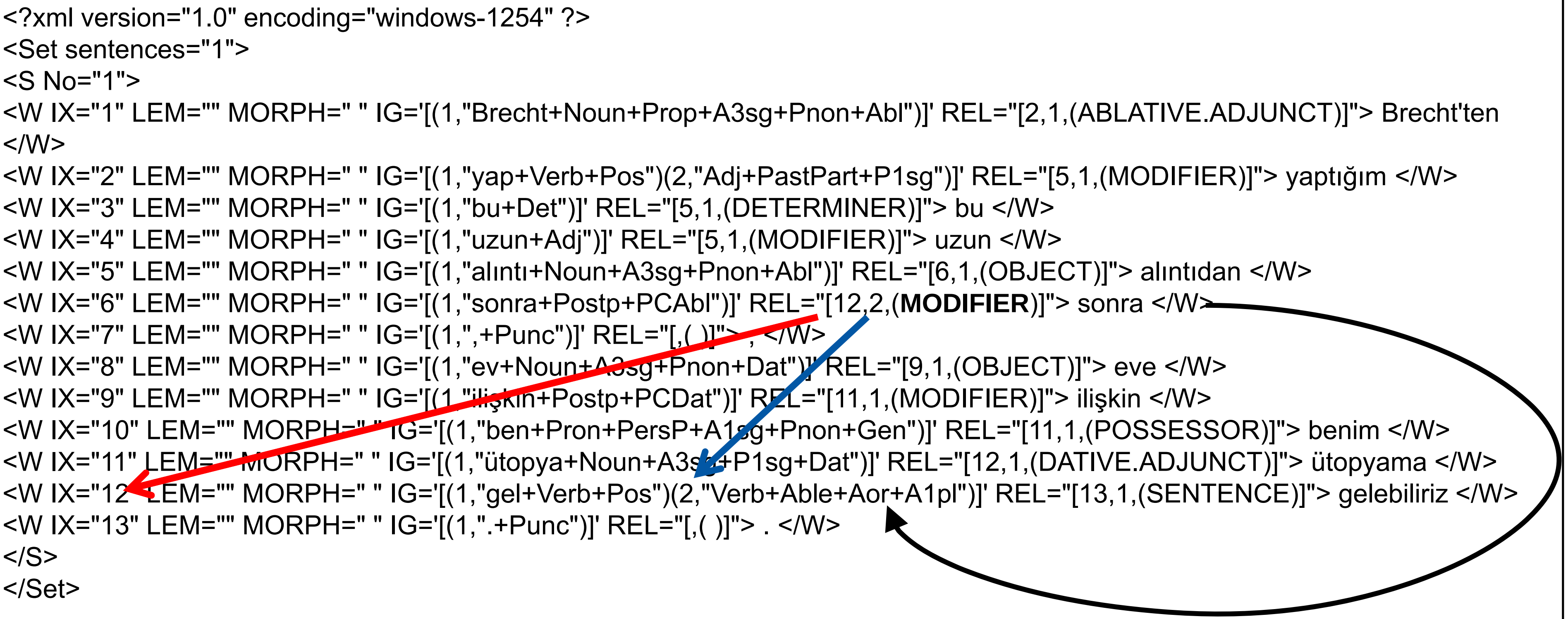
Syntactic Annotation

- The intensifier adverbial *en* (most) modifies the intermediate derived adjective *akıl+lı* (with intelligence/intelligent)



Syntactic Annotation

- Each sentence is encoded using XML



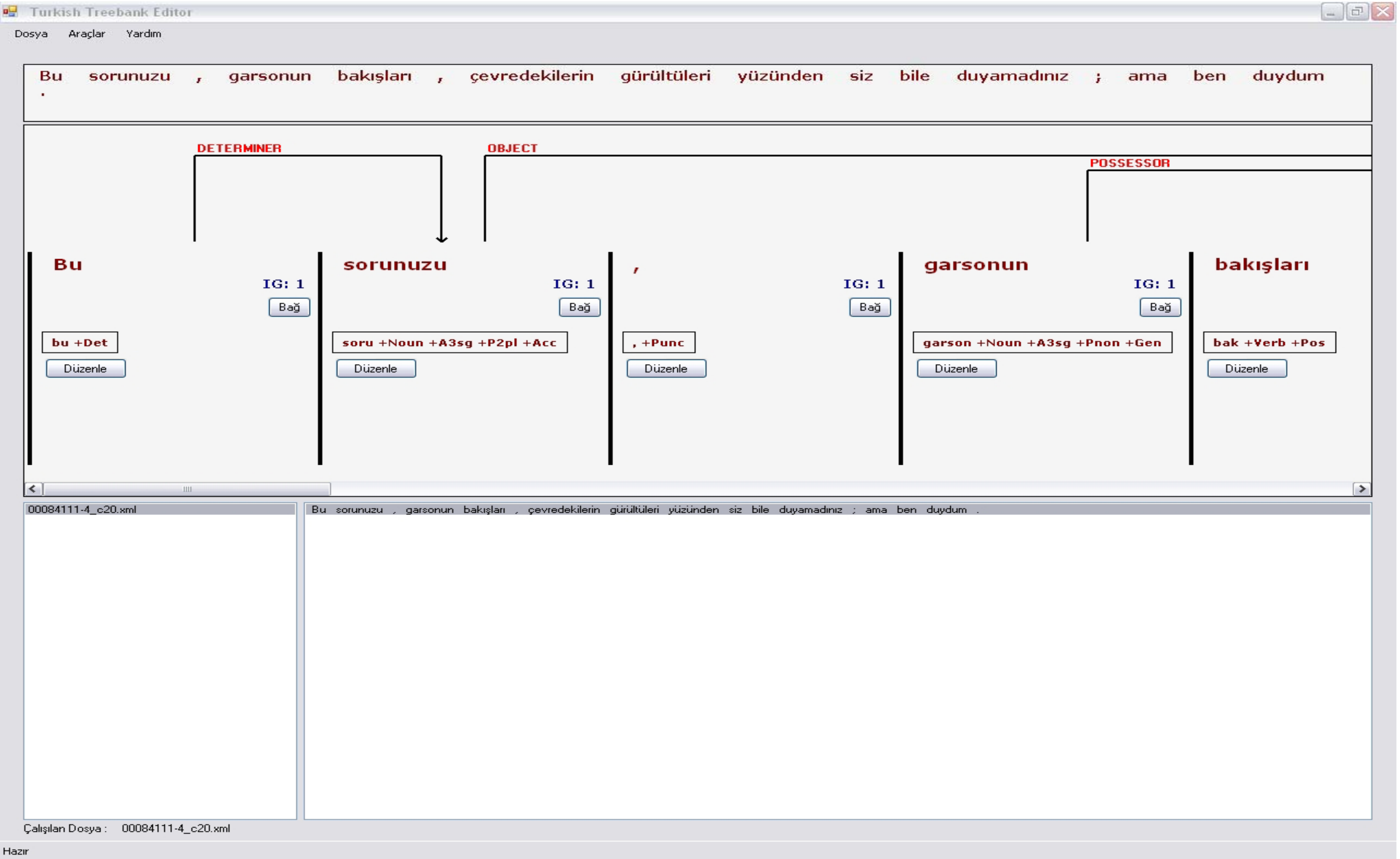
```
<?xml version="1.0" encoding="windows-1254" ?>
<Set sentences="1">
<S No="1">
<W IX="1" LEM="" MORPH=" " IG='[(1,"Brecht+Noun+Prop+A3sg+Pnon+Abl")]' REL="[2,1,(ABLATIVE.ADJUNCT)]"> Brecht'ten
</W>
<W IX="2" LEM="" MORPH=" " IG='[(1,"yap+Verb+Pos")(2,"Adj+PastPart+P1sg")]' REL="[5,1,(MODIFIER)]"> yaptığım </W>
<W IX="3" LEM="" MORPH=" " IG='[(1,"bu+Det")]' REL="[5,1,(DETERMINER)]"> bu </W>
<W IX="4" LEM="" MORPH=" " IG='[(1,"uzun+Adj")]' REL="[5,1,(MODIFIER)]"> uzun </W>
<W IX="5" LEM="" MORPH=" " IG='[(1,"alıntı+Noun+A3sg+Pnon+Abl")]' REL="[6,1,(OBJECT)]"> alıntıdan </W>
<W IX="6" LEM="" MORPH=" " IG='[(1,"sonra+Postp+PCAbl")]' REL="[12,2,(MODIFIER)]"> sonra </W>
<W IX="7" LEM="" MORPH=" " IG='[(1,",+Punc")]' REL="[,( )]"> , </W>
<W IX="8" LEM="" MORPH=" " IG='[(1,"ev+Noun+A3sg+Pnon+Dat")]' REL="[9,1,(OBJECT)]"> eve </W>
<W IX="9" LEM="" MORPH=" " IG='[(1,"ilişkin+Postp+PCDat")]' REL="[11,1,(MODIFIER)]"> ilişkin </W>
<W IX="10" LEM="" MORPH=" " IG='[(1,"ben+Pron+PersP+A1sg+Pnon+Gen")]' REL="[11,1,(POSSESSOR)]"> benim </W>
<W IX="11" LEM="" MORPH=" " IG='[(1,"ütopya+Noun+A3sg+P1sg+Dat")]' REL="[12,1,(DATIVE.ADJUNCT)]"> ütopyama </W>
<W IX="12" LEM="" MORPH=" " IG='[(1,"gel+Verb+Pos")(2,"Verb+Able+Aor+A1pl")]' REL="[13,1,(SENTENCE)]"> gelebiliriz </W>
<W IX="13" LEM="" MORPH=" " IG='[(1,".+Punc")]' REL="[,( )]"> . </W>
</S>
</Set>
```

Tools

- Wide-coverage morphological analyzer + unknown word processor
- Collocation/multi-word construct processor
- Heuristic rule-based link suggestions
- GUI-based annotation tools

- Now, we can use a parser directly to create a tree for validation or fix by an annotator

Tools



Bu sorunuzu , garsonun bakışları , çevredekilerin gürültüleri yüzünden siz bile duyamadınız ; ama ben duydum .

bu +Det

soru +Noun +A3sg +P2pl +Acc

, +Punc

garson +Noun +A3sg +Pnon +Gen

bak +Verb +Pos

00084111-4_c20.xml

Bu sorunuzu , garsonun bakışları , çevredekilerin gürültüleri yüzünden siz bile duyamadınız ; ama ben duydum .

Çalışılan Dosya : 00084111-4_c20.xml

How was it exploited?

- Mostly based on a user poll 3 years after release
  - CONLL 2006 and CONLL 2007 Multilingual Dependency Parsing Competitions
  - Ph.D. Thesis on Turkish Dependency Parsing
  - Automatic Induction of a CCG Grammar for Turkish
  - Syntactic Tools for Turkish Text Watermarking
  - Turkish Subcategorization Frame Acquisition
  - Turkish Semantic Role Labeling
  - As a reference for a planned annotation project for poly-synthetic American Indian languages.
- Treebank is available at http://www.ii.metu.edu.tr/~corpus/treebank.html

Problems

- Inconsistencies
- Speed
- Bugs in tools crept into annotations
- Punctuation
- Original plan was to have semantic roles in addition to the syntactic roles

Future

- Clean-up
- Increase size
  - A new test set had to be created for CONLL 2007
  - 300 sentences annotated/checked in about 30 hours.
- Semantic roles

Acknowledgements

- TÜBİTAK – Turkish NSF for initial funding
- Nart Atalay, Bilge Say
- Sabine Buchholz, Gülten Cebiroğlu, Ruken Çakıcı, Deniz Yüret