Turkish morphology in WebLicht

Çağrı Çöltekin

University of Tübingen
Seminar für Sprachwissenschaft

SFCM 2015

Turkish NLP pipeline in WebLicht

- Tokenization
- Morphological analysis
- Morphological disambiguation
- Dependency parsing

Turkish NLP pipeline in WebLicht

- Tokenization
- Morphological analysis
- Morphological disambiguation
- Dependency parsing

This short talk is only about some of the challenges in Turkish NLP because of the morphological complexity.

The classical example

İstanbul-lu-laș-tır-a-ma-yabil-ecek-ler-imiz-den-miș-siniz

‘You were (evidentially) one of those who we may not be able to convert to an Istanbulite’

Productive derivational morphology

İstanbul-lu-laş-tır-a-ma-yabil-ecek-ler-imiz-den-miş-siniz

-lu makes adjectives/nouns from nouns

- *İstanbul-lu* 'someone from Istanbul'
- *Stuttgart-lı* 'someone from Stuttgart'

-laş makes verbs from adjectives/nouns, with the meaning 'to become ...'

- *İstanbul-lu-laş-* 'to become an Istanbulite'
- *diktatör-leş-* 'to become a dictator'

Productive derivational morphology

İstanbul-lu-laş-tır-a-ma-yabil-ecek-ler-imiz-den-miş-siniz

-lu makes adjectives/nouns from nouns

- *İstanbul-lu* ‘someone from Istanbul’
- *Stuttgart-lı* ‘someone from Stuttgart’

-laş makes verbs from adjectives/nouns, with the meaning ‘to become ...’

- *İstanbul-lu-laş-* ‘to become an Istanbulite’
- *diktatör-leş-* ‘to become a dictator’

Some challenges:

- A lexicon of all derived words is not feasible
- Ambiguity: the same suffix may have both lexicalized and productive usage
- Some suffixes repeat (*göz-lük-lük* ‘place for eye glasses’, *göz-lük-çü-lük* ‘profession of making or selling eye glasses’) :

Voice suffixes

İstanbul-lu-laş-tır-a-ma-yabil-ecek-ler-imiz-den-miş-siniz

-tır is the causative marker

- *İstanbul-lu-laş-tır* ‘to cause someone to become an Istanbulite’
- *oku-t-tur-...* ‘...to cause someone to cause someone to read’

- Passive suffix may also repeat twice

Voice suffixes

İstanbul-lu-laş-tır-a-ma-yabil-ecek-ler-imiz-den-miş-siniz

-tır is the causative marker

- *İstanbul-lu-laş-tır* ‘to cause someone to become an Istanbulite’
- *oku-t-tur-...* ‘...to cause someone to cause someone to read’

- Passive suffix may also repeat twice

- Theoretically unbounded number of suffixes
- Even if the number is limited, representation as a typical feature is problematic
- Ambiguity: some multiple forms are for emphasis, not for double causation

Other verbal inflections

İstanbul-lu-laş-tır-a-ma-yabil-ecek-ler-imiz-den-miş-siniz

-a/-(y)abil indicate ability/possibility, -ma is the negative marker

- *İstanbul-...-a-ma-* 'not to be able to cause someone to become an Istanbulite'
- *İstanbul-...-a-ma-yabil-* 'may not be able to cause someone to become an Istanbulite'

Other verbal inflections

İstanbul-lu-laş-tır-a-ma-yabil-ecek-ler-imiz-den-miş-siniz

-a/-(y)abil indicate ability/possibility, -ma is the negative marker

- *İstanbul-…-a-ma-* ‘not to be able to cause someone to become an Istanbulite’
- *İstanbul-…-a-ma-yabil-* ‘may not be able to cause someone to become an Istanbulite’

- Nothing new, repetition and ambiguity
- A finite verb may have about 10 inflectional suffixes marking voice, tense, aspect, modality and person/number

Subordination

İstanbul-lu-laş-tır-a-ma-yabil-ecek-ler-imiz-den-miş-siniz

-ecek makes a subordinate clause

- *İstanbul-...-ecek* 'someone who may not possibly be converted to an Istanbulite'
- Now the word acts like a noun (referring to a person)

-ler is the plural marker

-imiz (normally) marks the possessor (first person plural)

- *ev-imiz* 'our house'
- but, here it marks the subject of the subordinate clause

-den marks for ablative case

- *İstanbul-...-ecek* 'of those we may not be able to converted an Istanbulite'

Subordination

İstanbul-lu-laş-tır-a-ma-yabil-ecek-ler-imiz-den-miş-siniz

-ecek makes a subordinate clause

- *İstanbul-...-ecek* ‘someone who may not possibly be converted to an Istanbulite’
- Now the word acts like a noun (referring to a person)

-ler is the plural marker

-imiz (normally) marks the possessor (first person plural)

- *ev-imiz* ‘our house’
- but, here it marks the subject of the subordinate clause

-den marks for ablative case

- *İstanbul-...-ecek* ‘of those we may not be able to converted an Istanbulite’

- We have two POS tags with inflections, the verb of the subordinate clause and the resulting noun
- Features may conflict: the verb has `Person=1` while the noun has `Person=3`

Copular suffixes

İstanbul-lu-laş-tır-a-ma-yabil-ecek-ler-imiz-den-miş-siniz

-(y)miş	marks for past tense and evidentiality, copula part '(y)' is dropped because of the phonological context
-siniz	marks for first person plural

Copular suffixes

İstanbul-lu-laş-tır-a-ma-yabil-ecek-ler-imiz-den-miş-siniz

-(y)miş	marks for past tense and evidentiality, copula part '(y)' is dropped because of the phonological context
-siniz	marks for first person plural

- Now we have three POS tags, two of them are predicates
- The predicates have different feature values, different subjects

Copular suffixes

İstanbul-lu-laş-tır-a-ma-yabil-ecek-ler-imiz-den-miş-siniz

-(y)miş marks for past tense and evidentiality, copula part ‘(y)’ is dropped because of the phonological context

-siniz marks for first person plural

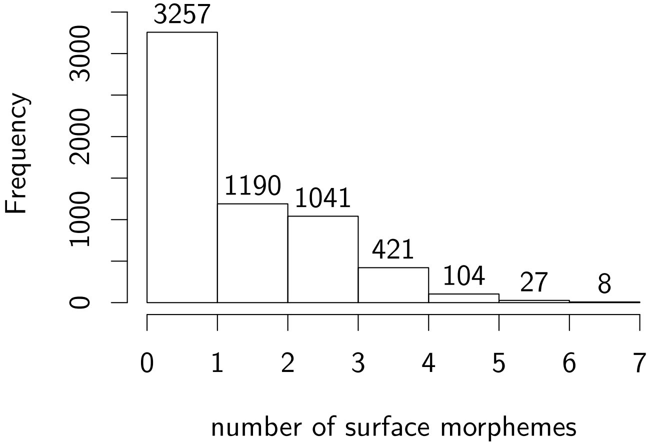
- Now we have three POS tags, two of them are predicates
- The predicates have different feature values, different subjects

⟨İstanbul-lu-laş-tır-a-ma-yabil⟩⟨-ecek-ler-imiz-den⟩⟨miş-siniz⟩

Summary

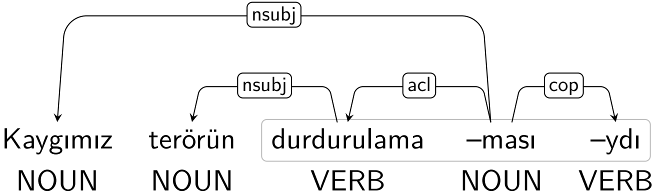
- Theoretically unbounded, repeated suffixes
- Large number of tags means sparsity for machine learning methods
- Multiple POS tags, multiple syntactic units in a single word
  - Multiple/conflicting feature values
  - Parts of a word may participate in different syntactic relations
  - Tokenization (for syntax) depends on morphological analysis/disambiguation
- Ambiguity
- Free word order

Morphological complexity in the real world



*Counts over a corpus of approx. 6K hand-annotated tokens, excl. punctuation.

An example dependency analysis



‘Our worry was (the fact) that terror could not be stopped’