Turkish and its Challenges for Language Processing

Kemal Oflazer

Carnegie Mellon University - Qatar
ko@cs.cmu.edu

Turkic Languages

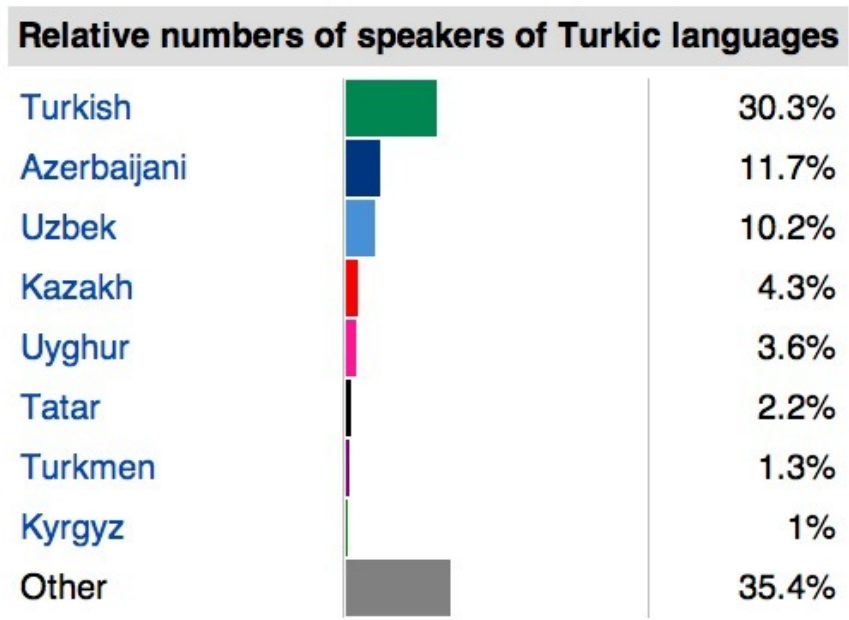
Turkic

Geographic distribution:	Originally from Western China to Siberia and Eastern Europe
Linguistic classification:	Altaic (controversial) • **Turkic**
Proto-language:	Proto-Turkic
Subdivisions:	Southwestern (Oghuz Turkic) Northwestern (Kipchak Turkic) Southeastern (Uyghur Turkic) Northeastern (Siberian Turkic) Oghur (Lir-, r-) Turkic "Arghu" (also subsumed under Oghuz)
ISO 639-5:	`trk`

Countries and autonomous subdivisions where a Turkic language has official status and/or is spoken by a majority

Image source: Wikipedia

- According to Wikipedia, Turkic languages are spoken as a native language by 165–200M people.



Turkic Languages

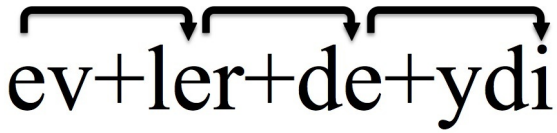
- Turkic (40)
  - Bolgar (1)
    - Chuvash [chv] (Russian Federation (Europe))
  - Eastern (7)
    - Ainu [aib] (China)
    - Chagatai [chg] (Turkmenistan)
    - Ili Turki [ili] (China)
    - Uyghur [uig] (China)
    - Uzbek, Northern [uzn] (Uzbekistan)
    - Uzbek, Southern [uzs] (Afghanistan)
    - Yugur, West [ybe] (China)
  - Northern (8)
    - Altai, Northern [atv] (Russian Federation (Asia))
    - Altai, Southern [alt] (Russian Federation (Asia))
    - Dolgan [dlg] (Russian Federation (Asia))
    - Karagas [kim] (Russian Federation (Asia))
    - Khakas [kjh] (Russian Federation (Asia))
    - Shor [cjs] (Russian Federation (Asia))
    - Tuva [tyv] (Russian Federation (Asia))
    - Yakut [sah] (Russian Federation (Asia))
  - Southern (12)
    - Azerbaijani (5)
      - Azerbaijani, North [azj] (Azerbaijan)
      - Azerbaijani, South [azb] (Iran)
      - Kashkay [qxq] (Iran)
      - Khalaj, Turkic [klj] (Iran)
      - Salchuq [slq] (Iran)
    - Turkish (4)
      - Balkan Gagauz Turkish [bgx] (Turkey (Europe))
      - Gagauz [gag] (Moldova)
      - Khorasani Turkish [kmz] (Iran)
      - Turkish [tur] (Turkey (Asia))
    - Turkmenian (1)
      - Turkmen [tuk] (Turkmenistan)
    - Crimean Tatar [crh] (Ukraine)
    - Salar [slr] (China)
  - Western (11)
    - Aralo-Caspian (4)
      - Karakalpak [kaa] (Uzbekistan)
      - Kazakh [kaz] (Kazakhstan)
      - Kyrgyz [kir] (Kyrgyzstan)
      - Nogai [nog] (Russian Federation (Europe))
    - Ponto-Caspian (4)
      - Balkar [krc] (Russian Federation (Europe))
      - Karaim [kdr] (Ukraine)
      - Krimchak [jct] (Ukraine)
      - Kumyk [kum] (Russian Federation (Europe))
    - Uralian (3)
      - Bashkort [bak] (Russian Federation (Europe))
      - Chulym [clw] (Russian Federation (Asia))
      - Tatar [tat] (Russian Federation (Europe))
  - Urum [uum] (Georgia)

Data Source: Ethnologue

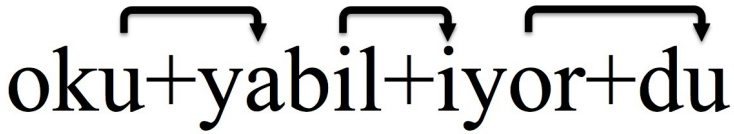
Turkic Languages - Characteristic Features

Phonology

- vowel harmony
- consonant assimilation



(they were in the houses)



((s)he was able to read)

Morphology

- Attach suffixes like “beads-on-a-string”
- No prefixes, no productive compounding
- Partial or full reduplication across words as a derivational process

Turkic Languages - Characteristic Features

Lexicon

- No noun classes or grammatical gender.

Word Order

- Subject – Object – Verb is the unmarked order.
- Based on the discourse context, any other order is usually possible.

- Some or all these features are shared with Mongolic, Tungusic, Korean and Japonic language families.

Sample Words Across Some Languages

sekiz (eight)

Türkiye Türkçesi	Azeri Türkçesi	Başkurt Türkçesi	Kazak Türkçesi	Kırgız Türkçesi
sekiz	säkkiz	higiz	segiz	segiz

Özbek Türkçesi	Tatar Türkçesi	Türkmen Türkçesi	Uygur Türkçesi	Rusça
säkkiz	sigiz	sekiz	säkkiz	vosem'

okumak (to read)

Türkiye Türkçesi	Azeri Türkçesi	Başkurt Türkçesi	Kazak Türkçesi	Kırgız Türkçesi
okumak	oḫumag	ukıv	okuv	okū

Özbek Türkçesi	Tatar Türkçesi	Türkmen Türkçesi	Uygur Türkçesi	Rusça
okımàk	uku	okamak	okımak	çitat'

cumhuriyet (republic)

Türkiye Türkçesi	Azeri Türkçesi	Başkurt Türkçesi	Kazak Türkçesi	Kırgız Türkçesi
cumhuriyet	respublika	respublika	respublika	respublika
		yömhöriyät		cumuriyat

Özbek Türkçesi	Tatar Türkçesi	Türkmen Türkçesi	Uygur Türkçesi	Rusça
cümhüriyät	respublika	respublika	respublika	respublika
	cömhöriyät		cumhuriyät	

Source: Turkish Language Institute (http://www.tdk.gov.tr/)

Turkish

- Lexicon heavily influenced by Arabic, Persian, Greek, Armenian, French, Italian, German, . . . , and recently English.
- Adopted Latin alphabet in 1928, literally overnight.
- Extensive “purification” of the lexicon in the 20th century,

My parents’ generation

Bir müsellesin mesahı sathiyesi zemini ile irtifaının zarbının nıfsına müsavidir.

My generation+

Bir üçgenin yüzey alanı tabanı ile yüksekliğinin çarpımının yarısına eşittir.

Turkish and NLP

Word Structure

- Pronunciation - Orthography mapping and its evolution
- Large number of very productive derivational morphemes
  - Essentially infinite word lexicon
  - Fixed size tag/feature encoding schemes do not work!
- Morphology and syntax interact in rather interesting ways.

Challenges

Pronunciation — Orthography Relation and its Evolution

- Morphological analysis really needs a TTS:
  - 2012'ye vs 2011'e:
    - No vowel to harmonize to in orthography
    - One needs to know how the pronunciation of the number ends.
  - 2/3'si, 2/3'ü, 15:00'te, 15:00'da
  - BAB'a vs AB'ye vs BBC'ye, vs BM'ye vs BM'e
- These are in general manageable by building a limited finite state model of how the pronunciation ends, as part of the analyzer.

Challenges

Pronunciation —- Orthography Relation and its Evolution

- The writer (usually of technical or news text) now implicitly assumes that the reader knows English, ... !
- Words are imported wholesale
  - with their orthography in their original language, but . . .
  - with suffixations based on their pronunciation in their original language!!!
  - Godot'yu . . .
  - serverlar ve clientlar
    - Worse server'lar ve client'lar
- For robust lexical processing, this needs to be handled.

Word Structure

- **ruhsatlandırılamamasındaki** - a word with 9 morphemes occuring once in a LM corpus.
- **ruhsat+lan+dır+ıl+ama+ma+sı+nda+ki**
- $\underbrace{\underbrace{\underbrace{\underbrace{\underbrace{\underbrace{\underbrace{\underbrace{\underbrace{\mathbf{ruhsat}}_{NOUN}\mathbf{+lan}}_{VERB}\mathbf{+dır}}_{VERB}\mathbf{+ıl}}_{VERB}\mathbf{+ama}}_{NOUN}\mathbf{+ma}}_{ADJ}\mathbf{+sı}}\mathbf{+nda}\mathbf{+ki}$
- You start with noun root and end up as an adjective after several derivations.
- *existing at the time of (it) not being able to acquire certification*

Word Structure

- But, in general things are saner!
- yapabileceksek
  - yap+abil+ecek+se+k
  - if we will be able to do (something)

- Average ≈ 3 morphemes/word (including the root)
  - But this is heavily skewed; high-frequency words usually have one morpheme!
- Average ≈ 2 morphological interpretations / word in running text.
  - But, 65% of words have one morphological interpretation.

Word	Morphemes	Ambiguity
bir	1	4
bu	1	2
da	1	1
için	1	4
de	1	2
çok	1	1
ile	1	2
en	1	2
daha	1	1
olarak	2	1
kadar	1	2
ama	1	3
gibi	1	1
olan	2	1
var	1	2
ne	1	2
sonra	1	2
ise	1	2
o	1	2
ilk	1	1

Word Structure

Productive Derivations

- Number of forms derivable from one root word

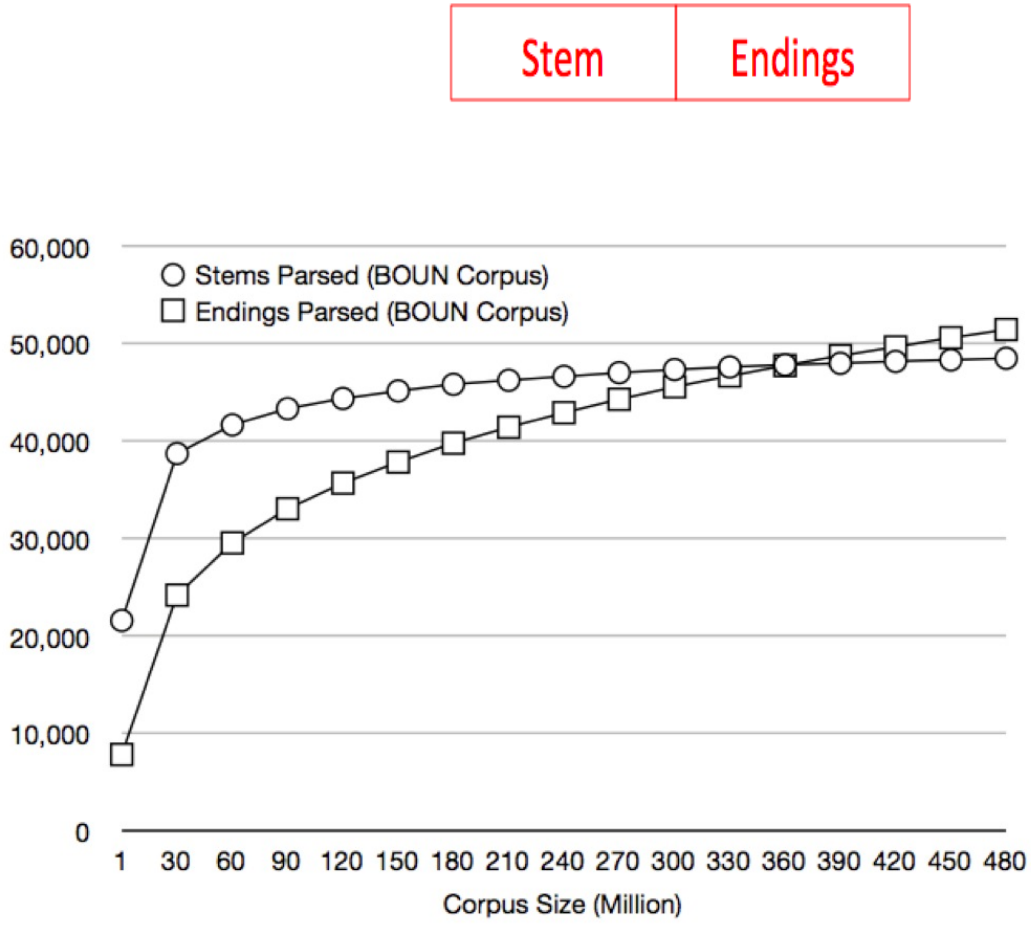
Root	# Derivations	# Words	Total
masa (Noun, (*table*))	0	112	112
	1	4,663	4,775
	2	49,640	54,415
	3	493,975	548,390
oku (Verb, (*read*))	0	702	702
	1	11,366	12,068
	2	112,877	124,945
	3	1,336,266	1,461,211

- Obviously not all make sense, but will be recognized as well-formed words

Word Structure

Some Statistics from BOUN News Corpus

- 4.1M unique words
- 5,539 new word forms were added going from 490M tokens to 491M tokens.
- Most frequent 50K words cover 89%.
- Most frequent 300K words cover 97%.
- 3.4M words appear less than 10 times
- 2.0M words appear once.



Haşim Sak, Tunga Güngör, and Murat Saraçlar: Resources for Turkish Morphological Processing. Language Resources and Evaluation,Vol. 45, No. 2, pp. 249–261, 2011

Challenges

Such a lexicon behaviour brings numerous challenges in

- Spelling correction,
- Tagset design,
- Language modeling,
- Syntactic modeling,
- Statistical Machine Translation

Challenges - Language Modelling

- Standard “word-based” language models have large out-of-vocabulary rates.

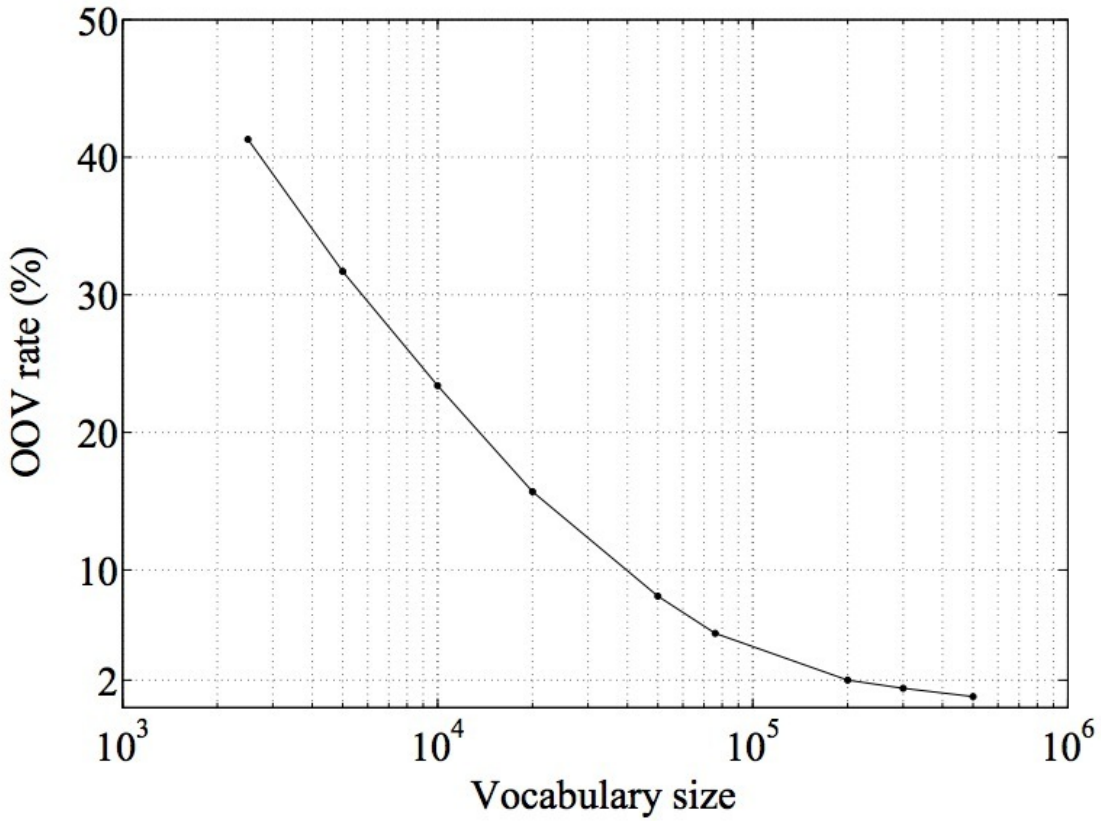


Figure 2.6. OOV rates for Turkish with different vocabulary sizes.

Language	**Vocab.**	**OOV**
English	60K	1%
Turkish	60K	8%
Finnish	69K	15%
Estonian	60K	10%
Hungarian	20K	15%
Czech	60K	8%

Ebru Arısoy, Statistical and discriminative language modeling for Turkish large vocabulary continuous speech recognition, PhD Thesis, Boğaiçi University, 2009

Challenges - Language Modelling

- Sublexical models provide much improved coverage.

Table 4.1. Results for different language modeling units (Real-Time Factor ≈ 1.5)

Recognition Units	Lexicon		UPW	n-gram	Coverage (per cent)		WER (per cent)	
	Size	AUL			Held-out	Test	Held-out	Test
Words	50 K	9.4	1.0	3-gram	92.7	91.9	29.9	29.4
	76 K	9.7	1.0	3-gram	94.9	94.6	27.7	27.0
	200 K	10.4	1.0	3-gram	98.0	98.0	25.5	24.1
	300 K	10.6	1.0	3-gram	98.7	98.6	25.1	23.9
	500 K	10.9	1.0	3-gram	99.1	99.2	**25.1**	**23.7**
Stem+endings	76 K	8.0	1.5	4-gram	99.7	99.6	24.1	23.2
	200 K	8.6	1.5	4-gram	99.8	99.8	**24.1**	**23.1**
Morphs (w/ WB morph)	50 K	7.0	2.4	5-gram	100	100	25.3	24.6
(w/o WB morph)	50 K	7.0	1.4	4-gram	100	100	24.7	23.9
(non-initials marked with "-")	76 K	6.7	1.4	4-gram	100	100	**24.1**	**22.9**

Ebru Arısoy, Statistical and discriminative language modeling for Turkish large vocabulary continuous speech recognition, PhD Thesis, Boğaiçi University, 2009

Word Order and Discourse

- More or less, anything goes, with minimal formal constraints.

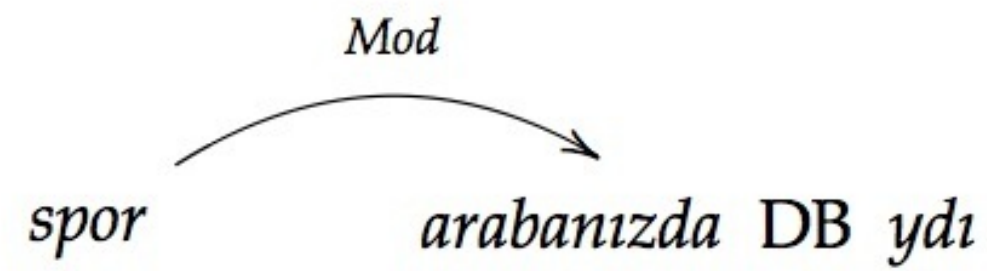
- Ekin Ayşe'yi gördü.
  - Ekin saw Ayşe.
- Ayşe'yi Ekin gördü.
  - It was Ekin who saw Ayşe.
- Gördü Ekin Ayşe'yi.
  - Ekin saw Ayşe (but was not really supposed to see her).
- Gördü Ayşe'yi Ekin.
  - Ekin saw Ayşe (and I was expecting that)
- Ekin gördü Ayşe'yi.
  - Ekin saw Ayşe (but someone else could also have seen her.)
- Ayşe'yi gördü Ekin.
  - Ekin saw Ayşe (but he could have seen someone else.)

- Formal grammar formalisms should be able to model word order and contextual background much more naturally.

Word Structure and Syntax

- Syntactic relations in Turkish are not between words but rather between Inflectional Groups
  - Chunks of inflectional morphemes separated by overt or covert derivational boundaries (DB).
  - $\underbrace{\textbf{ruhsat}}_{NOUN}\underbrace{\textbf{+lan}}_{VERB}\underbrace{\textbf{+dır}}_{VERB}\underbrace{\textbf{+ı1+ama}}_{VERB}\underbrace{\textbf{+ma+sı+nda}}_{NOUN}\underbrace{\textbf{+ki}}_{ADJ}$

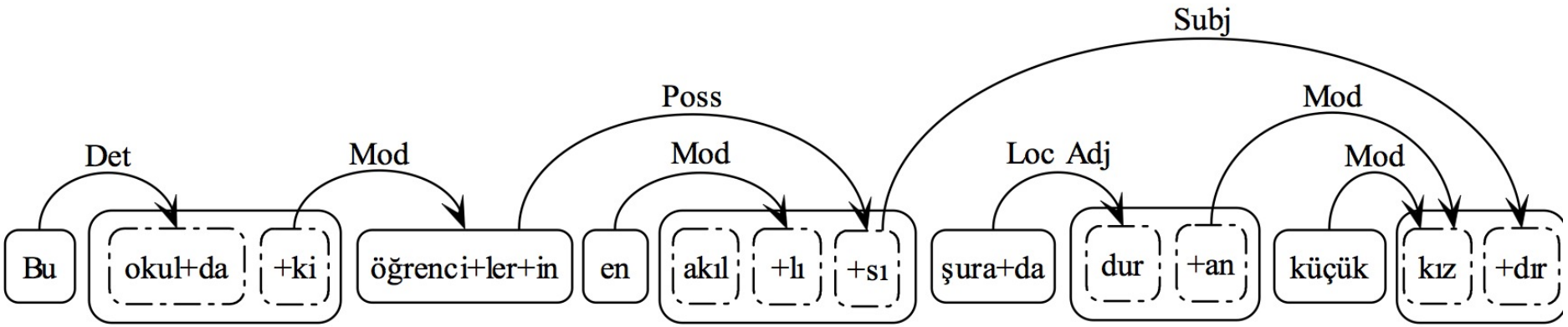
spor *arabanızdaydı*



sports *car-your-in DB it-was*

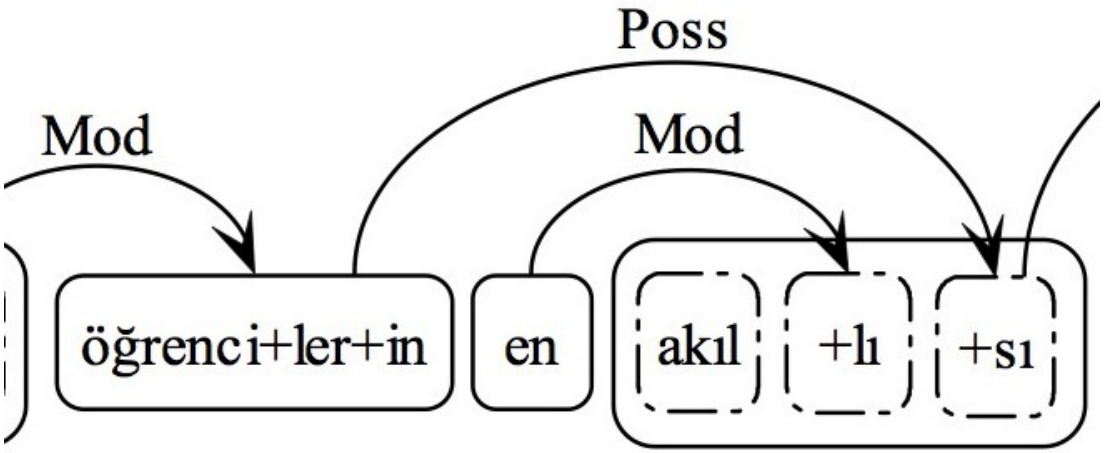
Word Structure and Syntax

- Different inflectional groups of a word can be involved in different syntactic relations.



Word Structure and Syntax

- Different inflectional groups of a word can be involved in different syntactic relations.



Word Structure

Derivations and Syntactic Relations

- Different inflectional groups of a word can be involved in different syntactic relations.
- Anonymous reviewer:
  - “You can’t do that! It violates the Lexical Integrity Principle.”
- Developer of the Syntactic Theory:
  - “Clearly, the principle needs to be revised!”

- The Turkish Dependency Treebank is encoded using such relations.
- Parsing accuracy should be based-on IG-to-IG relations, not word-to-word.

Challenges for Statistical Machine Translation

- How does English become Turkish?

if	we	will	be	able	to	make	...	become	strong
if	we	will	be	able	to	make	...	become	strong
...	strong	become	to	make	be	able	will	if	we
...	sağlam	+laş	+tır	+abil	+ecek	+se	+k		

... sağlamlaştırabileceksek

- BLEU will kill you if you get a single morpheme wrong!

Challenges for Statistical Machine Translation

- Make Turkish like English
  - Morphemes as words (Turkish)
  - I would not be able to do . . .
  - . . . yap +ama +yacak +tı +m
- Very long "sentences" $\Rightarrow$ alignment problems
  - 20 words $\Rightarrow\approx$ 60 morphemes.
- Decoder is responsible for both word order and morpheme order generation.
- Morphology frequently gets mangled.

Challenges for Statistical Machine Translation

- Make English like Turkish
  - Phrases as words (English)
  - Original English: . . . in their economic relations . . .
  - Original Turkish:. . . ekonomik ilişkilerinde . . .
  - Turkified English (:-)): . . . economic relation**+s+their+in** . . .
  - Preprocessed Turkish: . . . ekonomik ilişki**+lerinde** . . .
- Only align roots and assume the respective complex tags align.
- Much shorter English sentences, better alignment.
- Recall for English-side patterns are low during pre-processing.
  - Missing quite many phrasal patterns.
- There is now some work on hierarchical/syntax-based systems.

Nontechnical Challenges

- General lack of understanding/awareness of the technology.
- Lack of focused national initative.
- Everyone wants resources, yet not many are willing to contribute to building some.
- Not many natural producers of parallel texts involving Turkish.
- With very minor exceptions, no computational linguistics in other Turkic languages.

Now for the bright side

- Many useful resources and techniques have been developed over the last 2 decades.
  - Morphological analyzers, morphological disambiguators.
  - Numerous text corpora, speech corpora.
  - A modest dependency treebank of about 5500 sentences.
    - Used in CONLL Multilingual Dependency Parsing Competitions.
  - A dependency parser based on Nivre's MaltParser framework.

Now for the bright side

- Many useful resources and techniques have been developed over the last 2 decades.
  - A wide-coverage LFG parser based on ParGram framework.
  - Misc. Named Entity Recognizers and Gazetteers
  - A Turkish Discourse Bank.
  - A WordNet of about 15K synsets
  - Corpus of Spoken Turkish (in progress)
  - Turkish National Corpus (in progress)
- A respectable group of researchers working on Turkish language processing.
  - Many more needed given the number of speakers.

Thanks

- Questions?