

**NooJ Conference, June 2009,
Tozeur**

DESIGNING A NOOJ MODULE FOR TURKISH INFLECTIONAL ANALYSIS

**AN EXAMPLE OF HIGHLY
PRODUCTIVE MORPHOLOGY**

Arianna Bisazza. FBK-Irst (Trento, Italy)

Outline

- Introduction
- Relevant features of Turkish
- Handling phonology
- Handling morphology
- The module in action
- TODOs and conclusions

Introduction

- No support for Turkish on NooJ platform so far
- Basic need: allow the user to perform linguistic searches on the text and write syntactic grammars => morphological analyzer
- By now focus on inflection (it is complex enough!) and leave derivation (easier to handle through the dictionary) to future work



Relevant features of Turkish

Relevant features of Turkish:

Phonology

- A few generic rules cause important variations in surface form (allomorphy) both of stems and suffixes :

vowel harmony

&

other phenomena...

Relevant features of Turkish: Phonology

Vowel harmony:

“given a syllable, determines which vowels can follow it in the same word”

Ex. Plural suffix [-lAr]: -ler/-lar

Türk + pl = Türkler

ev + pl = evler

Alman + pl = Almanlar

kuş + pl = kuşlar

A generic principle, concerns both stems and suffixes

Relevant features of Turkish: Phonology

Other phonological phenomena (some examples):

- Final silent/voiced consonant alternation (in stems)

Ex. **kitap+[-lm] = kitabım** (*my book*)

defter+[-lm] = defterim (*my notebook*)

- Inter-vowel “y” (in suffixes)

Ex. **kafa+[-A] = kafaya** (to the head)

kol+[-A] = kola (to the arm)

Relevant features of Turkish:

Morphology

Turkish is an agglutinative language:

- The vocabulary is built by a wide range of suffixes combinations
- Words can be very long and even correspond to whole English sentences

Relevant features of Turkish: Morphology

- Suffixation is compositional and virtually unlimited:

one suffix <=> one linguistic feature

sakin	= calm	(adj.)
sakin+leş-	= to calm down	(v.int.)
sakinleş+tir-	= to calm down so.	(v.tr.)
sakinleştir+ebil-	= to be able to calm down so. (v.)	
sakinleştirebil+ecek	= being(fut.) able to calm down so. (n.)	
sakinleştirebilecek+im	= my being(fut.) able to calm down so.(n.)	
sakinleştirebileceğim+i	= my being(fut.) able to calm down so.(n.acc.)	

“Seni sakinleştirebileceğimi sandım”

“I thought I could calm you down”

Relevant features of Turkish & NooJ

- Large morphologic production
 - > dictionary of inflected forms oversized!



Instead of compiling a huge dictionary we can use morphological grammars (*.nom*) to describe inflection and compute lemma & features of our corpus forms on the fly

Relevant features of Turkish & NooJ

...Why is this possible ?



- Word formation mechanisms are regular
- Suffix chains are easily decomposable
- Morphotactic (suffix combinatory) can be represented as a reg. language (cf. Oflazer, 93)


Relevant features of Turkish & NooJ

- Let's assume I have my morphological grammars ready... there's still something to handle: allomorphy.
- Instead of handling phonology & morphology in two passes, I tried to include all in one :
 - ▣ to be compatible with NooJ formalisms,
 - ▣ to decrease runtime of corpus analysis.



Handling phonology

Handling phonology

- Phonologic rules are generic principles of the language -> they apply to surface forms regardless to morphology
 - Thus, encoding phonologic variation together with morphotactic makes the grammars explode in complexity
- 
- Idea: make do with a limited power of expression, i.e. let the module recognize a *superset* of the correct inflected form of Turkish

Handling phonology: in the dictionary

- Stem allomorphy is handled in the dictionary of words used as bases for suffixation
(an automatically processed version of *TDK, 2005. Türkçe Sözlük, Türk Dil Kurumu Yayınları*)
- Phonological properties are encoded as inflectional paradigms => stem allomorphs generated once at dictionary compilation

DICT ENTRY (*tdk.dic*):

kitap,N+FLX=endP+NW

FLX RULE (*stemVariants.nof*):

endP = b/NW + <E>/NW;

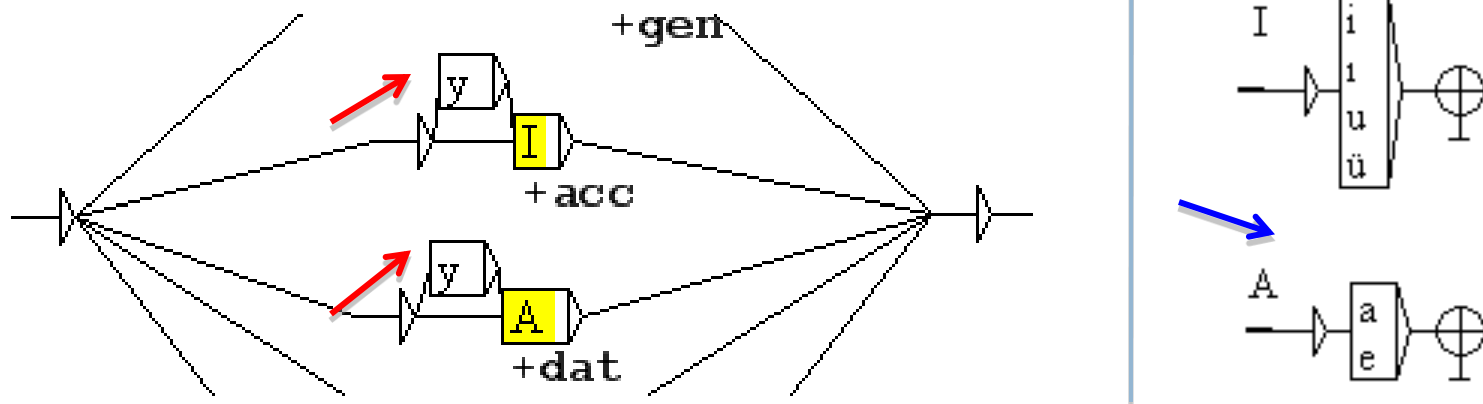
=> *DICT-FLX ENTRIES* (*tdk-flx.dic*):

kitap,N+FLX=endP+NW

kitap,N+FLX=endP+NW

Handling phonology: in the grammars

- Vowel harmony captured by vowel classes subgraphs...



- ...other variations by optional transitions



Handling morphology

Handling morphology

Inflectional morphology divided in two morphological grammars:

- Noun+NFVerbInflex.nom:
 - ▣ nouns,
 - ▣ nouns+copula,
 - ▣ non-finite verb forms
- VerbInflex.nom:
 - ▣ finite verb forms



The module in action

The module in action

- Dictionary of stems (*turkish_tdk.dic*): 45322 entries

=> 118581/349 states; 323 infos; recognizes 54347 forms

For the test:

- Corpus UDHR : *The Universal Declaration of Human Rights*

- Corpus RevNato : 35 articles of international politics published by NATO Review in 2005-2006

Corpus	Sizes		Unknown		Annotations	Time
	Words	Types	#	%	-	-
UDHR	1626	720	22	3,05%	1197	<2 s.
RevNATO	69723	12932	411	3,18%	20908	46 s.

The module in action

“Seni sakinleştirebileceğimi sandım”

example.not [Modified]

Characters
Tokens
Digrams
Annotations
Unknowns

Language is "Turkish (Turkey)(tr)".
Text Delimiter is: \n (NEWLINE)
Text contains 9 Text Units (TUs).
14 tokens including:
14 ...

seni sakinleştirebileceğimi sandım

34

sen,PRO+s+poss3s+case0

sen,PRO+s+acc

5

sakinleştir,V+compAcc+able+futAct+poss1s+acc

28

san,N+s+case0+CopPast+1s

san,V+compAcc+past+1s

Derivation

Inflection

The module in action

<N+gen> <N+poss3s>

n sonunda selefinin bıraktığı	mirasın bekçiliğini	yapmanın ötesine geçmiştir. N
bıraktığı mirasın bekçiliğini	yapmanın ötesine	geçmiştir. Nitekim, bu kısa g
şmaları sürekli olarak askeri	reformlarının gölgesinde	kalmıştır. Bunun sebeplerind
arının gölgesinde kalmıştır.	Bunun sebeplerinden	bazıları gayet açıktır. NATO
gelişmeler, İttifak'ın siyasi	programlarının sonuçlarından	çok daha kolayca ölçülebilir

<N+gen> <WF>* <N+poss3s> (*shortest*

matc

olan Yugoslavya) karşı savaş	açmasının gerekli olduğu konusunda
nlük hava kampanyası, ABD ile	Müttefiklerin askeri
akları tarafından kullanıldı;	Müttefiklerin sadece birkaç tanesi
sonra da AWACS uçaklarını ABD	şehirlerinin semalarında
'nün oluşturulması, stratejik	komutanlıkların yeniden düzenlenmesi

<V+able+fut>

r şey sembolik olmaktan öteye	geçemeyecektir	. Martin van Creveld Kudüs'tek
r sürede NATO, Orta Avrupa'da	çıkabilecek	yüksek yoğunluklu bir savaş i
n dediği gibi, "Süratle tepki	verebilecek	, uzun mesafelerde konuşlandır
lık Konseyi sahasında meydana	gelebilecek	teknolojik veya doğal bir afe
ak kuvvetleri hem kendilerini	savunabilecek	, hem de misyonu tehlikeye sok



TODOs and conclusions

TODOs and conclusions

- More tests, e.g. compare NooJ analysis with those of an existing morphological analyzer :
 - ▣ compute precision (are correct analysis there?)
 - ▣ compute noise (how many wrong analysis?)
- Deal with verbal inflection/derivational suffixes (passive, reflexive, causative...)
- Improve analysis of pronouns by writing a special grammar

TODOs and conclusions

- Run the grammars without constraints on the stem, with lower priority, to guess the lemma of unseen forms and gather candidate entries to enrich the dictionary

TODOs and conclusions

- Turkish is now supported by NooJ
- The problem of inflected forms dictionary's excessive size has been solved through NooJ formalisms and fonctionnalités, without need of external tools

Thanks for your attention...
Merci!

References

- *Türkçe Sözlük*, Türk Dil Kurumu Yayınları, 2005
(*dictionary*)
- A. Göksel and C. Kerslake. *Turkish: A Comprehensive Grammar*. Routledge, 2005
- K. Oflazer. *Two-level description of Turkish Morphology*. Proceedings of the Sixth Conference of EACL, 1993