DESIGNING A NOOJ MODULE FOR TURKISH INFLECTIONAL ANALYSIS

AN EXAMPLE OF HIGHLY PRODUCTIVE MORPHOLOGY

Arianna Bisazza. FBK-Irst (Trento, Italy)
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
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] = kitaβım (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:
  
  \textit{one suffix $\iff$ one linguistic feature}

  
  \begin{align*}
  \text{sakin} & = \text{calm} \quad \text{(adj.)} \\
  \text{sakin}+\text{leş}- & = \text{to calm down} \quad \text{(v.int.)} \\
  \text{sakinleş}+\text{tir}- & = \text{to calm down so.} \quad \text{(v.tr.)} \\
  \text{sakinleştir}+\text{ebil}- & = \text{to be able to calm down so.} \quad \text{(v.)} \\
  \text{sakinleştirebil}+\text{ecek} & = \text{being(fut.) able to calm down so.} \quad \text{(n.)} \\
  \text{sakinleştirebilecek}+\text{im} & = \text{my being(fut.) able to calm down so.} \quad \text{(n.)} \\
  \text{sakinleştirebileceğim}+\text{i} & = \text{my being(fut.) able to calm down so.} \quad \text{(n.acc.)}
  \end{align*}

  
  “Seni \textit{sakinleştirebileceğimi sandım}”
  “I thought \textit{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 combinatorial) can be represented as a regular language (cf. Oflazer, 93)
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>b/NW + <E>/NW; |

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

| Kitap, N+FLX=endP+NW |
| Kitab, 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
Handling morphology:
Noun+NFVerbInflex.nom
Handling morphology: VerbInflex.nom
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

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Sizes</th>
<th>Unknown</th>
<th>Annotation</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Words</td>
<td>Types</td>
<td>#</td>
<td>%</td>
</tr>
<tr>
<td>UDHR</td>
<td>1626</td>
<td>720</td>
<td>22</td>
<td>3,05%</td>
</tr>
<tr>
<td>RevNAT O</td>
<td>69723</td>
<td>12932</td>
<td>411</td>
<td>3,18%</td>
</tr>
</tbody>
</table>
The module in action

“Seni sakinleştirebileceğimi sandım”
The module in action

<n+gen> <n+poss3s>

n sonunda selemnin biraktığı
bıraktığı mirasın bekçiliğini
şmaları sürekli olarak askeri
arınin gölgesinde kalmıştır.
geleşmeler, İttifak’ın siyasi
mirasın bekçiliğini
yapmanın ötesine
reformlarının gölgesinde
Bunun sebeplerinden
programlarının sonuçlarından
yapmanın ötesine geçmiştir. N
geçmiştir. Nitekim, bu kısa g
kalmıştır. Bunun sebeplerind
bazılı gayet açıktr. NATO
cık daha kolayca ölçülebilir

<n+gen> <wf>* <n+poss3s> (shortest
match)

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

<v+able+fut>

ör şey simbolik olmaktan öteye
r sürede NATO, Orta Avrupa’da
n dediği gibi, “Süratle tepki
lik Konseyi sahasında meydana
ak kuvvetleri hem kendilerini
geçemeyecektir
çıkabilecek
verebolecek
gelebilecek
savunabilecek
. Martin van Creveld Kudüs’tek
yüksek yoğunluluklu bir savaş i
, uzun mesafelerde konuşlarımdır
 teknolojik veya doğal bir afe
, 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
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.
Turkish is now supported by NooJ

The problem of inflected forms dictionary’s excessive size has been solved through NooJ formalisms and fonctionnalities, without need of external tools

Thanks for your attention…
Merci!
References

- Türkçe Sözlük, Türk Dil Kurumu Yayınları, 2005 (dictionary)
- K. Oflazer. Two-level description of Turkish Morphology. Proceedings of the Sixth Conference of EACL, 1993