Morphology and Lexicon-based Machine translation of Ottoman Turkish to Modern Turkish

Joomy Korkut

Motivation

- Ottoman Turkish is an extinct prestige dialect used in Turkey, from roughly 14th century until 1928.

- It was written in Perso-Arabic script, unlike Modern Turkish, which uses Latin script.

- A lot of money is spent on transcribing Ottoman Turkish texts and teaching this language to Turkish studies scholars.

Solution

- We present a rule-based algorithm to **translate** Ottoman Turkish text to Modern Turkish text.

- There is not enough data, transcribed or translated text between the two languages, so we **cannot** go for the statistical solution.

- Source and target languages are the **same language** with two different scripts, though the former has **very irregular** spelling.

Turkish Morphology

- Agglutinative language with vowel harmony

  form of the suffix *-ler, -lar* determined by the last preceding vowel
  ev + ***-ler*** = ***evler*** (houses)
  at + ***-lar*** = ***atlar*** (horses)

- Very low number of irregular conjugations/declensions
  (13 irregular verbs in aorist tense, only 4 in other tenses)

Turkish Morphology

- ***uygarlaştıramadıklarımızdanmışsınızcasına***
  ((behaving) as if you are among those whom we could not civilize)

- Let's replace the root with a synonym and decline again:
  medenileştiremediklerimizdenmişsinizcesine

- This should demonstrate how powerful vowel harmony is.

  This is so internalized for Turkish speakers that in some forms of poetry have "rhymes modulo vowel harmony".

Turkish Morphology

- ***uygarlaştıramadıklarımızdanmışsınızcasına***
  ((behaving) as if you are among those whom we could not civilize)

- In the Ottoman script, this would be written as such:

  اویغارلشدرہمہدقلرمزدنمشسکزجہسکہ

  roughly:
  avygarlşdrhmhdklrmzdnmşskzchskh

Turkish Morphology

- ***uygar -laş -tır(a) -ma -dık -lar -(ı)mız -dan -mış -sınız -casına***
  ((behaving) as if you are among those whom we could not civilize)

- In the Ottoman script, this would be written as such:

  اویغار لش دره مه دق لر مز دن مش سکز جه سکه

  roughly:
  avygar -lş -drh -mh -dk -lr -mz -dn -mş -skz -chskh

Challenges

- ## Orthographical ambiguity
  و corresponds to *v*, *o*, *ö*, *u* or *ü* depending on the context.
  ك corresponds to *k*, *g*, *ğ*, *y* or (nasal) *n* depending on the context.

- ## Missing vowels
  Most vowels, especially in suffixes, are omitted.

- ## Legacy spellings of loanwords
  Spellings of loanwords from Arabic and Persian are kept as original.

- ## Ambiguous word boundaries
  The letter ه, when it stands for *e* or *a* (but not *h*), doesn't attach to the left.
  This is achieved with space or the "zero width non-joiner" character.

Algorithm

1. Greedily parse suffixes from the end of the word and store them.

2. Once you find a presumptive root, look up in

    A. an Ottoman Turkish to Modern Turkish dictionary, which gives you a translated root directly.

    B. a Modern Turkish dictionary by generating a **regular expression** from the Ottoman spelling.

    - If no root is found, unparse the last suffix and look up again.

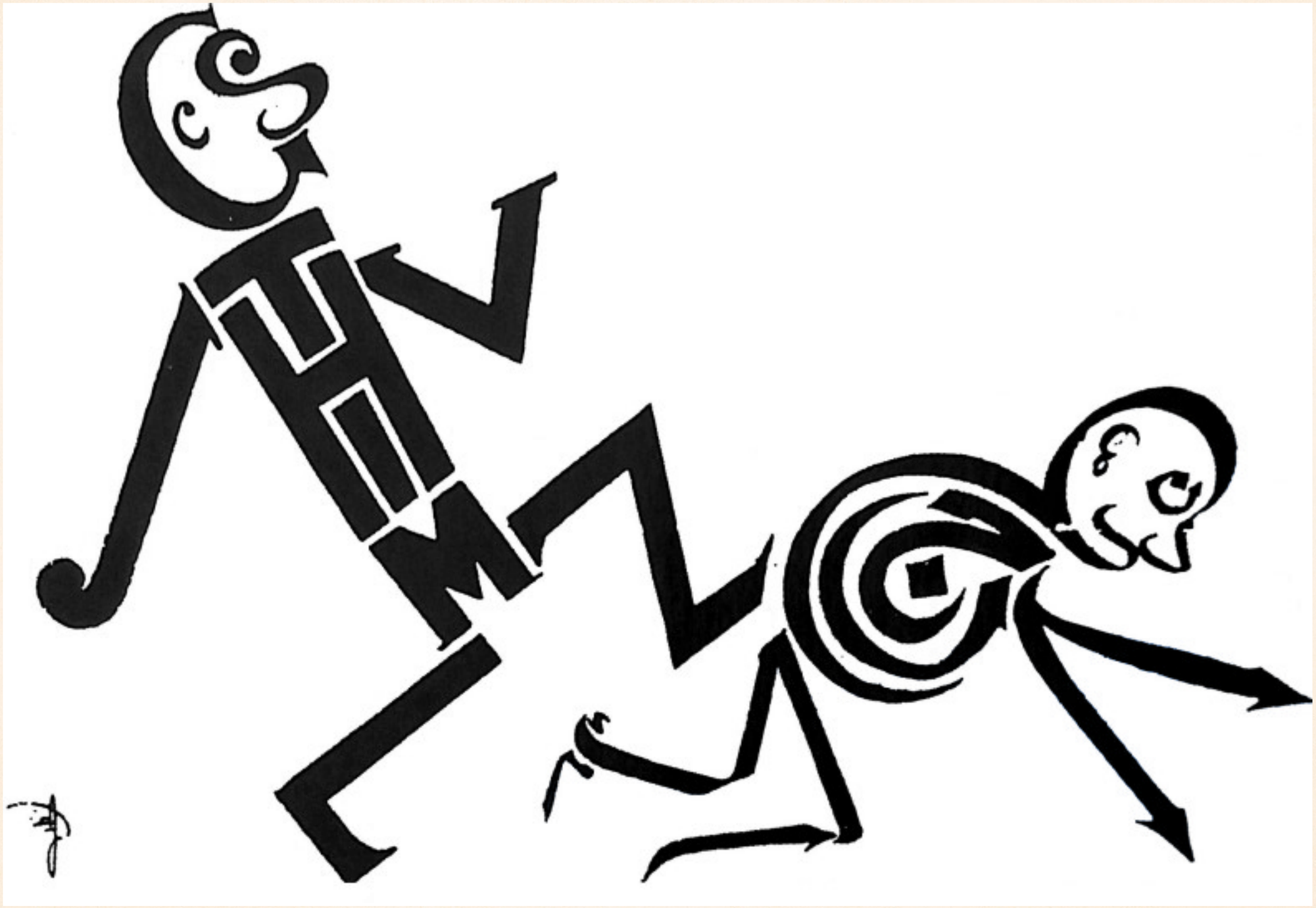
3. Add the translated suffixes to the translated root by conjugating and declining properly.

Regular Expression Generation

- For each Ottoman letter, we take into account what Modern Turkish letter it can stand for.

- Before and after each Ottoman letter, we consider whether there might be a missing vowel.

- For a word like گوز (Tur: *göz*, Eng: eye), we will generate the pattern `/^g(((a|e|i|ı|o|ö|u|ü)?v(a|e|i|ı|o|ö|u|ü)?)|o|ö|u|ü)z$/`



cartoon by Ramiz Gökçe, August 23rd 1928, Akbaba magazine