Morphological Analysis of Tajik – Notes and Preliminary Results

Gulshan Dovudov, Vít Baisa

Faculty of Informatics
Masaryk University
Brno

RASLAN 2010, Karlova studánka

Introduction

- Tajik language
- four models: R, Pr $\oplus$ R, Pr $\oplus$ R $\oplus$ Ps, R $\oplus$ Ps

Examples of segmentations

structure	Tajik word	in latin	translation	root	word
R	китоб	kitob	book	noun	noun
R	сурх	surkh	red	adj	adj
R	ист	ist	stand	verb	verb
Pr⊕R	но-умед	no-umed	despair	noun	adj
Pr⊕R	бар-зиёд	bar-ziyod	excessive	adj	adj
Pr⊕R	на-рав	na-rav	don't go	verb	verb
Pr⊕R⊕Ps	то-мактаб-ӣ	tomaktabi	preschool	noun	adj
Pr⊕R⊕Ps	на-ме-рав-ем	name-rav-em	We do not go	verb	verb
Pr⊕R⊕Ps	на-ме-фур-омад-ам	na-me-fur-omad-am	I do not descend	verb	verb
R⊕Ps	ҳафт-а	haft-a	week	num	noun
R⊕Ps	китоб-ча	kitob-cha	litle book	noun	noun
R⊕Ps	сурх-ҳо	surkh-ho	reds	adj	noun

Database of morphemes – Postfixes

- iterative processing of representative texts
- 2,533 suffixes with their frequencies

Postfixes – Examples

L	count	freq.	examples	in latin
0	0	46.89650	-	-
1	113	39.25153	ҳак, ҳо, ӣ, а, гар, гӣ, ум	hak, ho, i, a, gar, gi,um
2	755	11.12421	ҳо-и, ванд-он,зор-аш	ho-i, vand-on, zor-ash
3	1,017	2.35906	ча-тоб-ро,тар-ин-ам	cha-tob-ro, tar-in-am
4	540	0.35571	иш-манд-он-и	ish-mand-on-i
5	86	0.01142	ум-ин-аш-он-ро	um-in-ash-on-ro
6	17	0.00129	и-ят-нок-тар-ин-и	i-yat-nok-tar-in-i
7	3	0.00019	и-ят-нок-и-аш-он-ро	i-yat-nok-i-ash-on-ro
8	2	0.00006	и-ят-нок-тар-ин-ат-он-ро	i-yat-nok-tar-in-at-on-ro

Prefixes

- generating list of all possible prefixes:
  - simple
  - double
  - triple
- limiting it by statistical research
- resulted in 19 simple, 39 double and 8 triple

Prefixes – Examples

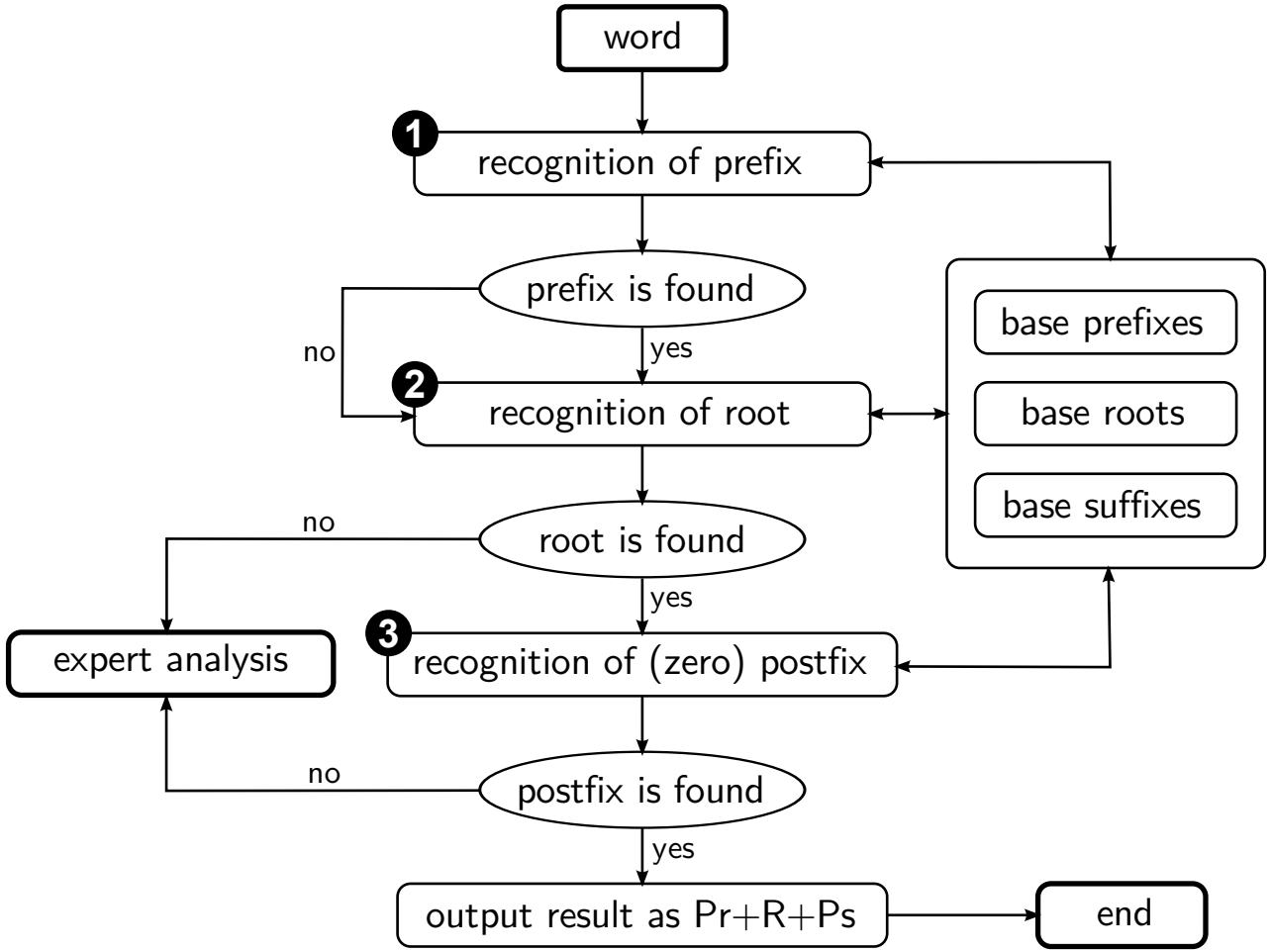
L	examples	in latin
1	ба-, би-, бо-, во-, дар-, ма-	ba-, bi-, bo-, vo-, dar-, ma-
2	ноба-, ноби-, нодар-, ноҳам-	noba-, nobi-, nodar-, noham-
3	барнаме-, дарнаме-, намефур	barname-, darname-, namefur

Coverage of the Database

- the first phase: processing of 1,700,000 words
  - 66 prefixes
  - 26,479 roots
  - 2,533 postfixes
- the second phase: other 1,140,000 words
  - 2 new prefixes (4.5 %)
  - 4,443 new roots (16.77 %)
  - 360 new postfixes (14.21 %)

Semiautomatic Morphological Analysis

- depends on quality of the database of morphemes
- output: segmentation of a word into parts (R, Ps, Pr) or info that the word can not be segmented into known morphemes
- $\rightarrow$ expert manual analysis


word
1
recognition of prefix
prefix is found
no
yes
2
recognition of root
root is found
no
yes
3
recognition of (zero) postfix
postfix is found
no
yes
expert analysis
output result as Pr+R+Ps
end
base prefixes
base roots
base suffixes

Data Resource Description

- 8,000 pages and 4,000,000 words
- literary works, newspaper articles and professional literature

Future works

- POS tagger and morphological disambiguation
- further extending of the database
- corpus of Tajik with at least 5,000,000 tokens
- spell checker in hunspell format (for OO)

Tajik Language Processing: State of Art

- localisation of some software
- keyboard layout
- Russian-Tajik, Tajik-Russian dictionary
- syllable-based text-to-speech synthesizer