Morphological Analysis of Tajik
Notes and Preliminary Results

Gulshan Dovudov and Vít Baisa

Natural Language Processing Centre, Faculty of Informatics
Masaryk University, Brno, Czech Republic
387176@mail.muni.cz, xbaisa@fi.muni.cz

Abstract. In this article we describe state of art of morphological analysis of Tajik language. At first we comment retrieval of prefixes and postfixes. Then we introduce an algorithm for semi-automatic morphological analysis of one-root Tajik words. The algorithm works with a database of roots, prefixes and suffixes and in the case a new root or a new affix is found the algorithm adds it into the database on the basis of manual analysis.

Key words: Tajik language, morphological analysis, prefix, root, postfix, suffix, affix

1 Introduction

The Tajik language belongs to the Iranian group of languages which is a part of the extensive Indo-European language family. With its grammatical system, Tajik language belongs to the group of languages of analytic type. A rich system of inflectional case forms existed in the ancient Iranian languages but it is fully lost in present Tajik. Case forms in Tajik are expressed by purely syntactical means: prepositional and postpositional construction, *izafet* combination and word order. A category of gender is also almost lost despite it existed in the ancient Iranian languages too. In Tajik, only verbs have a developed system of syntactical and analytical forms.

Generally, every Tajik word can be segmented into three parts – *morphemes*: *prefix*, *root* (the lexical kernel of a word bearing its basic semantic value) and *postfix* (*suffix* + *ending*). That is why they can be expressed as one of four models: R, Pr $\oplus$ R, Pr $\oplus$ R $\oplus$ Ps and R $\oplus$ Ps (Root, PRefix and PoStfix). Table 1 shows some examples of Tajik words segmented into morphemes.

2 Definitions

Let us mention some definitions applicable to Tajik language:
Root – the main part of a word. Root is the mandatory part of every word.
Affix – an auxiliary part of a word added to the root which serves to word formation and expressing of grammatical meanings. Affixes form words only in conjunction with roots. Affixes alone do not bear any lexical meaning.

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2010, pp. 21–27, 2010.

Table 1. Examples of Tajik words segmented into morphemes

Structure	Tajik word	In latin	Translation
R	китоб	kitob	book
R	кор	kor	work
R	зиёд	ziyod	many
R	ист	ist	stand
R ⊕ Ps	китоб-ча	kitob-cha	litle book
R ⊕ Ps	кор-гар	kor-gar	worker
R ⊕ Ps	соя-бон	soya-bon	tilt
R ⊕ Ps	шарм-гин	sharm-gin	timid
Pr ⊕ R	но-умед	no-umed	despair
Pr ⊕ R	бар-зиёд	bar-ziyod	excessive
Pr ⊕ R	на-рав	na-rav	don't go
Pr ⊕ R ⊕ Ps	бар-омад-ан	bar-omad-an	to ascend
Pr ⊕ R ⊕ Ps	(на-ме)-рав-и	(na-me)-rav-i	you do not go
Pr ⊕ R ⊕ Ps	(на-ме-фур)-омад-ам	(na-me-fur)-omad-am	I do not descend

Prefix – a morpheme standing before the root and changing its lexical or grammatical meaning. Prefixes are divided into two groups: simple and compound. Compound prefixes (disyllabic and trisyllabic) are formed by concatenating of appropriate number of simple prefixes.
Postfix – a part of a word which follows directly a root, consisting of suffixes and endings. Postfixes as well as prefixes are divided into two groups, but a compound postfix can consist of 2–8 simple suffixes.
Suffix – a kind of affix morpheme which follows a root and comes before an ending.
Base – a part of a word that remains after the cutting-off ending. A base may be only root or root with affixes.

3 Database of Morphemes

3.1 Postfixes

In this paper, we assume that every input word is correct. I.e. there are no errors in spelling. The procedure of morphological analysis is based on previously prepared fixed database of morphemes – roots, prefixes and postfixes.

Morphological analysis of a word equals to segmenting that word into three mentioned components.

Database of postfixes of Tajik literary language was expanded step by step based on iterative processing of representative texts (see Section 6). As a result, database of 2,533 suffixes with their frequencies of occurrence was made.

Table 2 shows the frequency of postfixes of different level of complexity. The level represents number of simple postfixes in compound postfix. 0 level of complexity means that there is no postfix in a word.

Table 2. List of postfixes with frequencies

L	Count	Frequency
0	0	46.89650
1	113	39.25153
2	755	11.12421
3	1,017	2.35906
4	540	0.35571
5	86	0.01142
6	17	0.00129
7	3	0.00019
8	2	0.00006

3.2 Prefixes

Database of prefixes was created combinatorially and elaborated by statistical method.

We have complete list of simple (one-syllable) prefixes at our disposal: ба (ba-), бар (bar-), бе (be-), би (bi-), бо (bo-), боз (boz-), бу (bu-), во (vo-), дар (dar-), ма (ma-), ме (me-), на (na-), но (no-), то (to-), фар (far-), фур (fur-), ҳам (ham-), ҳаме (hame-) and ҳар (har-).

These 19 prefixes represent all simple prefixes. Since any compound prefix may be created as a concatenation of two or three simple prefixes we can generate all double and triple prefixes by permutations. There are 342 (19*18) double and 5,814 (19*18*17) triple possible prefixes. It is obvious that simple prefixes may not repeat in compound prefixes.

These hypothetical prefixes were checked semi-automatically and the result was list of 19 simple, 39 double and 8 triple real prefixes.

Table 3 provides a list of all currently known prefixes ordered by their frequency. Frequencies and counts for both prefixes and suffixes were derived from representative texts (see Section 6).

3.3 Coverage of the Database

At the moment it is difficult to even estimate the coverage of our database of morphemes.

Processing of about 1,700,000 words yielded 66 prefixes, 26,479 roots and 2,533 postfixes. After processing of other texts (about 1,140,000 words) we obtained only 2 new prefixes, 4,443 new roots and 360 new postfixes. It is about 4.5% of new prefixes, about 16.77% of new roots and 14.21% of new postfixes.

Since these new morphemes have very low frequency we cam assume that the coverage is considerably high.

Table 3. List of prefixes with frequencies

#	Prefix	Freq.	#	Prefix	Freq.	#	Prefix	Freq.
1	ме (me)	48.356	23	барна (barna)	0.048	45	бозname (bozname)	0.003
2	на (na)	11.421	24	воме (vome)	0.048	46	меби (mebi)	0.002
3	бе (be)	7.113	25	бозме (bozme)	0.031	47	нафур (nafur)	0.002
4	ҳам(ham)	6.913	26	мефар (mefar)	0.029	48	вонаме (voname)	0.002
5	бар (bar)	6.369	27	наби (nabi)	0.024	49	намефур (namefur)	0.002
6	наме (name)	5.033	28	барнаме (barname)	0.021	50	ҳаме (hame)	0.002
7	но (no)	3.568	29	намедар (namedar)	0.020	51	бебоз (beboz)	0.001
8	бо (bo)	2.639	30	дарме (darme)	0.015	52	бифар (bifar)	0.001
9	би (bi)	2.616	31	ноба (noba)	0.013	53	бубар (bubar)	0.001
10	ҳар (har)	1.128	32	бино (bino)	0.009	54	мена (mena)	0.001
11	ба (ba)	1.055	33	баҳам (baham)	0.008	55	ноби (nobi)	0.001
12	дар (dar)	0.681	34	надар (nadar)	0.008	56	нодар (nodar)	0.001
13	боз (boz)	0.675	35	бано (bano)	0.006	57	ноҳам (noham)	0.001
14	ма (ma)	0.386	36	дарбар (darbar)	0.006	58	ҳамебар (hamebar)	0.001
15	во (vo)	0.377	37	дарна (darna)	0.006	59	намефар (namefar)	0.001
16	мебар (mebar)	0.362	38	барҳам (barham)	0.005	60	фар (far)	0.001
17	барме (barme)	0.316	39	бозна (bozna)	0.005	61	беба (beba)	0.001
18	бу (bu)	0.267	40	вона (vona)	0.004	62	бозма (bozma)	<0.001
19	то (to)	0.124	41	мефур (mefur)	0.004	63	вома (voma)	<0.001
20	медар (medar)	0.101	42	нафар (nafar)	0.004	64	дарма (darma)	<0.001
21	набар (nabar)	0.073	43	дарname (darname)	0.004	65	фур (fur)	<0.001
22	Намебар (namebar)	0.058	44	дарҳам (darham)	0.003	66	барма (barma)	<0.001

4 Semi-Automatic Morphological Analysis

Quality of semi-automatic morphological analysis of a word strongly depends on the database of morphemes. An output of the analysis is either a segmentation of a word into three parts (Pr, R and Ps), or information that the word can not be segmented into known morphemes. It is quite clear that a negative result is a consequence of incompleteness of our database.

For this reason it seems natural to expand the database by adding new morphemes manually identified by an expert during morphological analysis, see Section 5.

The algorithm for semi-automatic morphological analysis for Tajik words is depicted in the form of flowchart on Figure 1 on the next page.

We have all Tajik words consisting of one or two letters in our database and therefore the analysis will process only words with strictly more than two characters. If the analyser gets one- or two-character word it immediately outputs result, i.e. root.

Morphological analysis of a word consists of the following steps. Block 1 represents recognition of a prefix. Since Tajik prefixes contain at least two letters, we pick two letters from the beginning of input. Then we select all prefixes from database which start with these two letters.

If none of these selected prefixes is contained in the word it is natural to assume that the word begins with the root. If the prefix is identified it is removed from the word, and the remaining fragment of the word is analyzed in block 2.

The process in block 2 is similar to the previous block. If at least one root is found then it is removed from the word and, again, the remaining fragment goes to block 3. Here it is compared with postfixes from the database. If at least

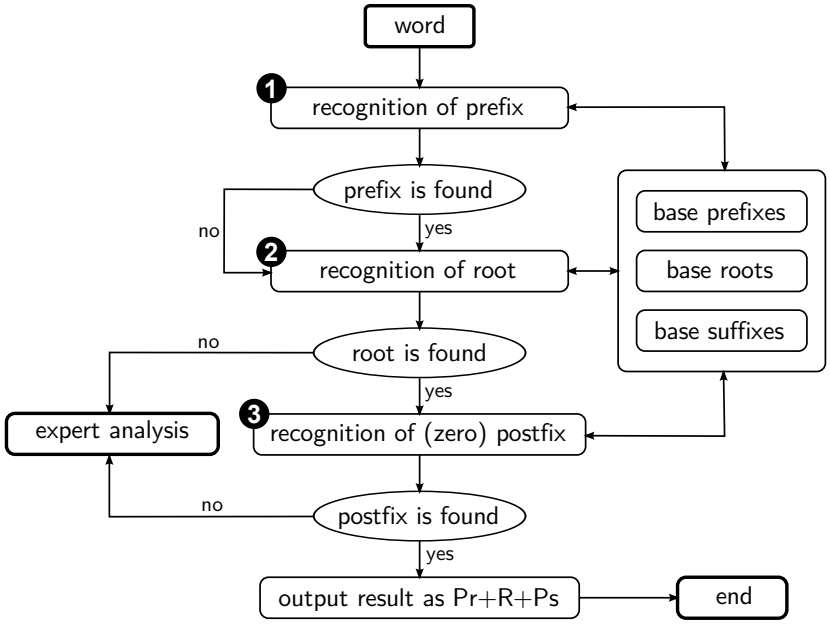

Fig. 1. Algorithm for morphological analysis

one postfix is found the process is completed successfully and the input word is represented as a concatenation of three morphemes – prefix, root and postfix.

It is quite clear that an analysis of some words may result in only root or root with either prefix or postfix.

It is also obvious that morphological analysis of some words can not find any possible segmentation into morphemes and since we expect correctly spelled word, this phenomenon is caused by the fact that the database do not contain corresponding morphemes. In such cases the word is sent to manual analysis.

5 Linguistic Analysis

If morphological analysis of a word fails, the word has to be analysed manually. A language expert segments the word into morphemes. The algorithm then determines whether any of these morphemes are already in the database. If not, the new morphemes are added.

6 Data Resources Description

Representative sample of about 8,000 pages (about 4,000,000 words) was used for text processing described above. Texts were taken from literary works, newspaper articles and from professional literature in Tajik language. For more details, see Table 4 on the following page.

Table 4. Texts used for text processing

No.	Author	Document	Pages
1	Abu Ali Ibn Sino	AL-Konun	200
2	Abulkosim Firdavsi	Shohnoma	200
3	Sadriddin Ayni	Yoddoshtho	280
		Yatim	220
		Kahramoni khalki tojik – Temurmalik	150
4	Bobojon Gafurov	Tojikon	200
5	Sotim Ulugzoda	Piri hakimoni mashrikzamin	150
6	Nazirjon Tursunov	Ta'rikhi tojikon	400
7	Muhammadjon Shakuri	Panturkizm va sarnavishti khalki tojik	346
		Khuroson ast injo	360
		Sadri Bukhoro	187
8	F. Muhammadiev	Kulliyot	100
9	L. Sherali	Namunai ash'or	300
10	Jalol Ikromi	Asarhoi muntakhab	100
11	Abdumalik Bahori	Bozgasht	100
		Sohili Murod	100
12	Rahim Jalil	Odamoni Jovid	100
13	M.Ganiev	MS'Word	50
14	Hakim Rahimi	Oila va oiladori	150
		Farhangi zaboni tojiki	150
15	newspapers	Jumhuriyat	270
		Sugd	280
		Sadoi mardum	200
		Charkhi gardun	400
		Lochin	192
		Mash'al	181
		Nohid	161
		Salomat Boshed	1,040
		Sukhani Khalk	1,309
		Khirman	100

7 Future Work

Our goal is to extend the database of morphemes by processing other literary texts gathered from electronic books, newspapers, internet etc.

We will also put all available documents together and make a corpus of Tajik with more than 5,000,000 millions of tokens.

Extensive work on the morphological analyser should also lead to development of a spell checker, POS tagger and an algorithm for morphological disambiguation.

With these tools we will be able to annotate data in the corpus automatically.

All these goals should hopefully end with high-quality data sources of Tajik language.

8 Tajik Language Processing – State of Art

There are few works connected with Tajik language processing. Besides localisation of some software and creating Tajik keyboard layout [9] it is necessary to mention Russian-Tajik and Tajik-Russian dictionary (Usmanov and Soliev) and text-to-speech synthesizer [10,11] based on syllables.

Acknowledgements

This work has been partly supported by Erasmus Mundus Action II lot 9: Partnerships with Third Country higher education institutions and scholarships for mobility, and by the Ministry of Education of CR within the Center of basic research LC536.

References

1. Rastorgueva, V. S. *A brief sketch of the grammar of the Tajik language (supplement to the Tajik-Russian Dictionary)*. Moscow: State Publishing House of Foreign and National Dictionaries, 1954, pp. 529–570.
2. Buzurgzoda, L., Niyazmuhammedov, B. *Grammar of the Tajik language. Part 1: Phonetics and Morphology*. Stalinabad. 1944, 112 p.
3. Active Tajik literary language. Volume 1: Lexicology, phonetics and morphology. Dushanbe: Irfon, 1973. pp. 452.
4. Rustamov, S. *Derivation of nouns in the modern Tajik literature language*. Dushanbe, 1972, pp. 90.
5. Amonova, F. R. *Noun affixal derivation in modern Persian and Tajik languages*. Dushanbe, 1982, pp. 55.
6. Usmanov, Z. D., Dovudov, G. M. *On forming the prefix base to the literary Tajik.* Reports of the Academy of Sciences of the Republic of Tajikistan, no. 6, 2009, pp. 431–436.
7. Usmanov, Z. D., Soliev, O. M., Dovudov, G. M. *On a set of postfixes of Tajik literature language.* Reports of the Academy of Sciences of the Republic of Tajikistan, no. 2. 2010, pp. 99–103.
8. Usmanov, Z. D., Dovudov, G. M. *On statistical regularities of tajik morpheme basis*. Reports of the Academy of Sciences of the Republic of Tajikistan, no. 3. 2010, pp. 188–191.
9. Soliev, O. M. *Mathematical model of optimal keyboard layout and its applications*. Ph.D. thesis. `http://www.mitas.tj/dissov/kandidat/avtoreferat/o_soliev.pdf`
10. Khudoiberdiev, K. A. *Tajik Text-to-Speech Synthesizer*. Ph.D. thesis. `http://www.tajik-tts.narod.ru/`
11. Khudoiberdiev, K. A. *Complex of Program Synthesis Tajik* `http://www.mitas.tj/dissov/kandidat/avtoreferat/khudoiberdiev.pdf`