A Corpus-Based Finite State Morphological Analyzer for Pashto

Fatima Tuz Zuhra and Mohammad Abid Khan
Department of Computer Science, University of Peshawar, Peshawar, Pakistan
fateeshah@yahoo.com, abid_khan1961@yahoo.com

Abstract

This paper provides details of the development of an inflectional morphological analyzer that can analyze different inflections of a Pashto verb, noun or adjective. The system is corpus-based. The developed system is capable to accept input in the form of a transliterated Pashto verbal, nominal or adjectival inflection; convert it to an Arabic-scripted Pashto equivalent; morphologically analyze the word and search and display all the sentences in the corpus, in which the word is used.

1. Introduction

Pashto is a morphologically rich language. There are countless applications of Natural Language Processing (NLP), one of which can be the development of a system that can provide all the morphological tags of a given word and search examples of the use of the word in a corpus of real life data. This work deals with the design and development of a similar application. The developed system can morphologically analyze as well as provide examples of the use of any verbal, nominal or adjectival inflection. These examples are searched from the Pashto corpus [1].

There can be several uses of the system, developed in this work. A linguist can use the system to morphologically analyze a particular word and see its daily life examples. Another and very important use of the system can be in the development of a part of speech (POS) tagger for Pashto language.

The rest of the paper is divided into the following sections. Section 2 provides a brief overview of the morphology of Pashto verbs, nouns and adjectives. Section 3 sheds light on the analysis of verbal, nominal and adjectival inflections. Section 4 is about the modeling and design of the morphological analyzer. In section 5, the implementation of the morphological analyzer is discussed. Section 5 provides details of the overall corpus-based morphological analyzer for Pashto.

2. A brief overview of Pashto morphology

It is important to provide a brief summary of the work, done by Pashto linguists, we studied before starting the computational work. They are Penzl [2], Khattak [3], Tegey and Robson [4], and Babrakzai [5]. The work of these linguists form the basis for the research work presented in this paper.

Khattak [3] identifies different facets, for which a Pashto verb inflects. He says, “The formal distinctions of the Pashto verb reflect a variety of categories: tense, aspect, mood and voice. Referring to the NPs in the subject or object position, the verb also inflects for person, number and gender.”

Khattak [3] further says that the morphology of the Pashto verb shows only two simple tenses: present and past. The future is expressed with the help of a model clitic *ba.*

Babrakzai [5] provides the basic structure of a Pashto verb, given below, where # indicates the potential positions for clitics.

Verb=[aspect # negative # stem + agreement #]

Babrakzai [5] provides the definition of agreement as follows:

“System of inflection that records a nominal’s inherent features (usually person, number, gender/ or case) on another category, generally a verb, adjective or a determiner”.

According to Tegey and Robson [4], agreement is indicated with personal endings, i.e. suffixes following the verb stem which show person and number.

The category of gender is restricted to the third person form of simple verbs and to the third person singular forms of the auxiliary [2] called copula verbs of 'to be' [6]. However, the category of gender is found in third person plural form of this auxiliary in Yousafzai dialect [7].

A Pashto noun inflects for gender, number and case [2]. Different Pashto grammarians [2, 8, 9] categorize the Pashto nouns into different masculine and feminine classes according to their final phonemes. Bellew [10] and others have also contributed significantly to the investigation about Pashto nouns. The Pashto adjectives have more or less the same inflectional properties and similar morphological behavior as those of Pashto nouns.

3. The analysis of verbal, nominal and adjectival inflections

Different verbal, nominal and adjectival inflections were manually extracted from about 30,000 words written Pashto data. These include over 2000 verbal, 2500 nominal and 1800 adjectival inflections. These inflections were decomposed into stems and affixes. This lengthy analysis phase revealed the personal suffixes for a Pashto verb given in table 1.

Table 1: Personal suffixes

Person	Suffix
First person singular (Present + Past)	-əm
First person plural (Present + Past)	-u
Second person singular (Present + Past)	-ee
Second person plural (Present + Past)	-əi
Third person singular and plural in present tense	-i
Third person masculine singular (Past)	-o
Third person masculine plural (Past)	-
Third person feminine singular (Past)	-a
Third person feminine plural (Past)	-ee

Various other verbal affixes, revealed in this analysis, are listed in table 2.

Table 2: Various affixes used in verb morphology

Morphological property	Affix
Perfective marking prefix	wə-
Past marking infix	-əl-
Passive participle suffix	-e
Perfect participle suffix	-e
Optative suffix	-e or -ɑy

The analysis of Pashto nominal inflections shows that the Pashto nouns have various types (classes), based on their ending phoneme. The Pashto nouns are classified in seven masculine and seven feminine classes. Each of these classes have a particular type of ending phoneme and the suffixation of each class is different from the other classes for reflecting the same facet. For example, the suffixes for direct plural formation of various masculine classes of nouns are given in table 3.

Table 3: Suffixes for various masculine classes of nouns

Noun class	Suffix
First masculine (animate)	-ɑn
First masculine (inanimate)	-una
Second masculine	-i (loud-stressed)
Third	-i (weak-stressed)
Fourth masculine (human)	-una
Fourth masculine (animal)	- ɑn
Fifth masculine	-gɑn or -wɑn
Sixth masculine	-una
Seventh masculine	-yɑn

There may be a chance that the direct plural forming suffix of two classes is the same, but in this case their other suffixes e.g. their vocative forming suffix will be different. Hence these are different classes.

The case of Pashto adjectives is similar to Pashto nouns, as revealed by the analysis of adjectival inflections. Based on the ending phonemes of Pashto adjectives, eight classes are defined [11].

4. Modeling and design of Pashto morphological analyzer

The morphological analyzer is modeled using Finite State Transducers (FSTs) as tools. FSTs combine lexicon and rules as said by Beesley and Karttunen [12]:

"An FST incorporates all the lexicon and rule information in a single network data structure, mapping directly between a language of underlying or "lexical" strings and a language of surface strings".

The rules devised in this research work are productive. Thus, more verbs, nouns and adjectives can be added to the system, without changing the rules.

After various affixes in the morphology were identified, the order in which these affixes are attached to the verbal, nominal or adjectival stem was determined. The determination of this order served as a

foundation for defining morphotactics for the Pashto verbal system. These morphotactics were then encoded in FSTs. In this section, some of these FSTs are presented. The glosses used in this discussion are given in table 4.

Table 4: The morphological tags

Word	Morphological Tag
Present	Pres
Past	Past
Perfective	Perf
Imperfective	Imperf
Imperative	Imp
Perfect Participle	PerfectPart
Optative	Opt
Passive Participle	Pass Part
Declarative	Dec
Subjunctive	Sub
First Person	F
Second Person	S
Third Person	T
Singular	Sg
Plural	Pl
Masculine	Mas
Feminine	Fem

The glosses used in nominal and adjectival FSTs are given in table 5.

Table 5: The words with their glosses

Word	Gloss	Word	Gloss
Adjective	Adj	Oblique case-II	OblII
Masculine	Mas	Vocative	Voc
Feminine	Fem	Singular	Sg
Direct	Dir	Plural	Pl
Oblique case-I	OblI		

A part of the verbal FST for modeling the present tense imperfective verbs is given in figure 1.

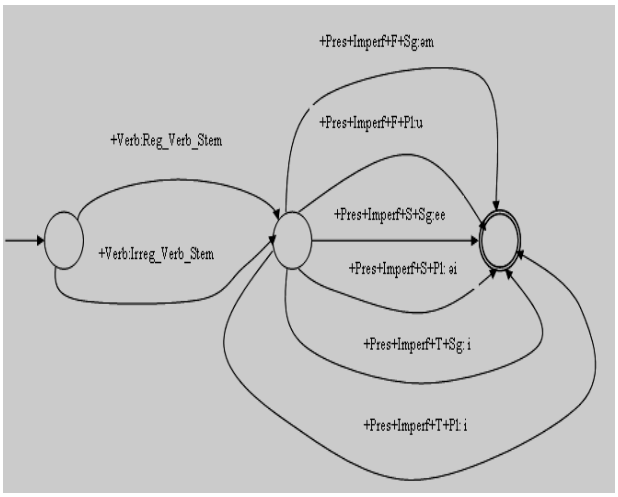

Figure 1: The present imperfective verbs

A part of the nouns' FST for modeling the second masculine class is provided in figure 2.

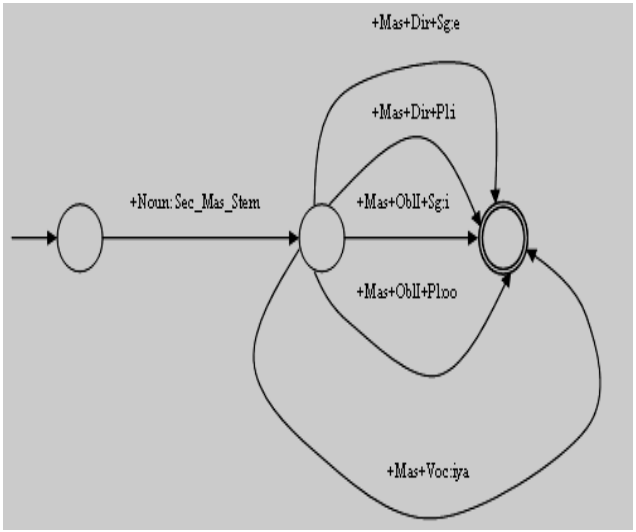

Figure 2: The second masculine class of nouns

Similarly, a part of the FST for the Pashto adjectives, which models the fifth class of adjectives, is given in figure 3.

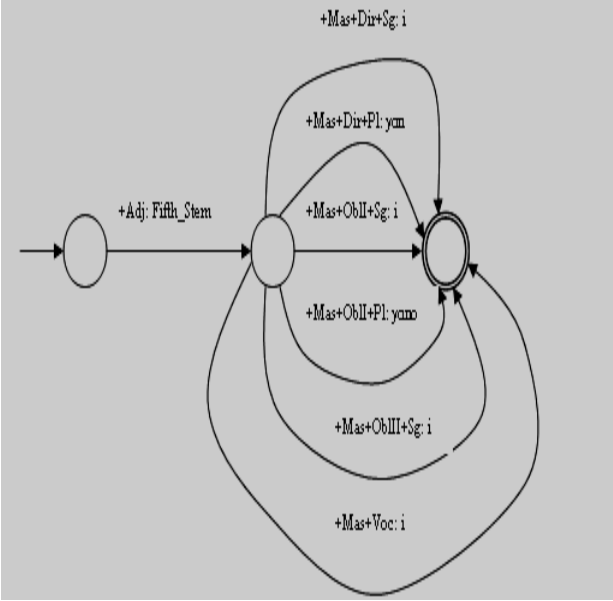

Figure 3: The masculine form of the fifth class of adjectives

These FSTs are ready to be implemented. The next section sheds light on the implementation of these FSTs.

5. Implementation of the morphological analyzer

The implementation details of the morphological analyzer are provided in this section. The FSTs, developed during the modeling and design phase, are implemented. For this implementation, four programming languages and tools are used, which are C# (in .NET framework), Xerox tools lexc and xfst, and Microsoft Access. A Romanized transliteration scheme, similar to that of Penzl [2], is used instead of actual Arabic script. Though, a great part of the transliteration symbols is adopted from [2], some symbols differ from that scheme. These differences are because of the diacritic symbols, used by Penzl, which are replaced by alternative keyboard symbols in this work because these diacritic symbols either are difficult to type or not available on keyboard. The symbols, used by Penzl, are shown in table 6 and the additions made to it in Table 7.

Table 6: Adopted transliteration symbols

Alphabet	Transliteration	Alphabet	Transliteration
ا	aa	ش	sh
ب	b	ښ	ss
پ	P	غ	gh
ت	T	ف	f
ټ	Tt	ق	q
ج	Dzh	ک	k
څ	Dz	ګ	g
چ	Tsh	ل	l
د	D	م	m
ډ	Dd	ن	n
ر	R	ڼ	nn
ړ	Rr	و	w
ز	Z	ی	y
ژ	Zh	ي	i
ږ	Zz	ې	ee
س	S	و	u

Table 7: Additional transliteration symbols

Alphabet	Transliteration	Alphabet	Transliteration
ؤ	Aw	ع	ah
و	Oo	ۀ	@
ح	h?	ى	@i
خ	X	ے	e
ذ	z?	ا	A?

All the FSTs are implemented in lexc, the binary files of its output were opened in xfst, and then saved in text files, where the lexical and corresponding surface strings were listed. These files were then read in the MS-Access database tables. One of these MS-Access tables is shown in figure 4.

PASHTONOUN : Table

Surface	Lexical
dzhmA?tuna	dzhmA?t+Noun+Mas+Dir+Pl
dzhmA?tuno	dzhmA?t+Noun+Mas+Obll+Pl
dzhmA?t	dzhmA?t+Noun+Mas+Dir+Sg
dzhmA?t	dzhmA?t+Noun+Mas+Obll+sg
ghwA?gA?nee	ghwA?+Noun+Fem+Dir+Pl
ghwA?gA?no	ghwA?+Noun+Fem+Obll+Pl
ghwA?	ghwA?+Noun+Fem+Dir+Sg
ghwA?	ghwA?+Noun+Fem+Obll+sg
ghwA?	ghwA?+Noun+Fem+Voc
hosoo	hos+Noun+Fem+Obll+Pl
hos@i	hos+Noun+Fem+Dir+Pl
hos@i	hos+Noun+Fem+Dir+Sg
hos@i	hos+Noun+Fem+Obll+sg
hos@i	hos+Noun+Fem+Voc
insA?na	insA?n+noun+Fem+Dir+Sg
insA?nA?n	insA?n+Noun+Mas+Dir+Pl
insA?nA?no	insA?n+Noun+Mas+Obll+Pl
insA?na	insA?n+Noun+Mas+Oblll+Sg
insA?na	insA?n+Noun+Mas+Voc
insA?n	insA?n+Noun+Mas+Dir+Sg
insA?n	insA?n+Noun+Mas+Obll+sg

Figure 4: The MS-Access nouns' table

Thus, a lexicon is obtained, with which all the rules of inflections of verbs, nouns and adjectives are incorporated. This lexicon contains various possible inflections of 200 root verbs, 250 root nouns and 140 root adjectives. This is the morphological analyzer for the verbal, nominal and adjectival inflectional system of Pashto.

6. The overall system

There are several components, designed and developed during this research work, in addition to the morphological analyzer. All these components are combined to develop the overall corpus-based finite state morphological analyzer for Pashto. All the components of the overall system are discussed briefly below.

The first component is a finite state morphological analyzer. This component analyzes any verbal, nominal or adjectival inflection morphologically subject to the condition that the part of speech, to be analyzed, is listed in the lexicon. This morphological analyzer is the result of the implementation of verbal, nominal and adjectival FSTs.

The second component of the system is a monitor corpus of written Pashto data [1]. This corpus currently contains Pashto data of 24,000 words and its size is increasing. This corpus is used for evaluating the results of the finite state morphological analyzer.

The third component is a Microsoft Access database. In this database, the output of the xfst is saved. This database contains a VERB, a NOUN and an ADJECTIVE table. All the surface forms and the corresponding lexical forms, obtained as an output of the implementation of FST, are stored in these tables.

The fourth component is an English-to-Pashto spelling transducer. This is one of the most wanted and most important components, designed and developed during this research work. This transducer can map from transliterated string to Arabic-scripted Pashto word.

All these components are integrated in a way, depicted in the flowchart in figure 5.

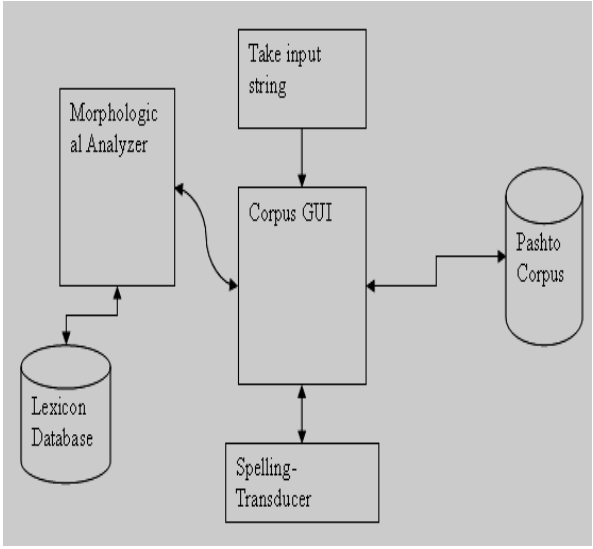

Figure 5: The flowchart of the whole system

By combining all the components, an application is developed that takes a transliterated Pashto verbal, nominal or adjectival inflection as input, convert it into an Arabic-scripted Pashto word, morphologically analyzes it, and provides all the sentences from the Pashto corpus, in which the input word is used. A sample interaction with this application, having a user-friendly interface, is given in figure 6.

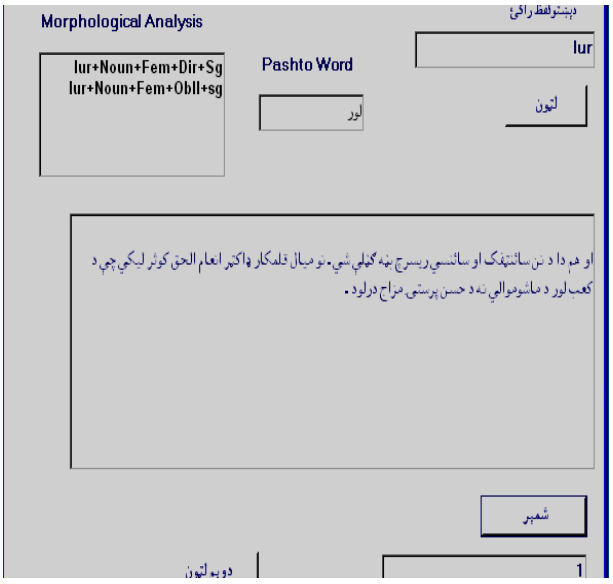

Figure 6: Sample interaction with the system

In the next section, the accuracy of the system is discussed and a brief error analysis is provided.

7. Error analysis

A total of 153 verbal, 200 nominal and 124 adjectival inflections were collected from written Pashto data and given to the system as input, out of which, 116 verbal, 118 nominal and 104 adjectives were correctly analyzed. Thus, the overall accuracy of the system is:

((116+118+104)/(153+200+124)) * 100 = 70.85%

The errors were analyzed. It was observed that 40% of the errors were because of limited lexicon size i.e. these were the verbal, nominal or adjectival inflections that the system could correctly analyze, only if the lexicon contained the stems of these inflections.

About 25% of the errors were because of the loan words, with inflections not according to Pashto morphology e.g. the Arabic inflection محققين muhəqqiqin (researchers).

10% of the errors were because of the variations in writing the same word by different authors. For example, some authors write وباسي wəbɑsi as اوباسي 'caused to get out'. The rest were miscellaneous types of errors such as the errors caused by compound words (derived words) and typographic errors by Pashto writers.

8. Conclusion

The research work, presented in this paper, deals with the analysis, design and implementation of a corpus-based finite state morphological analyzer for Pashto. This is the first morphological analyzer, developed for Pashto language. The accuracy of the morphological analyzer is about 71% that can be increased by increasing the size of the lexicon.

9. References

[1] M. A. Khan and F. T. Zuhra, "A General-Purpose Monitor Corpus of Written Pashto", in proc. *Conference on Corpus Linguistics*, Birmingham, July, 2007.

[2] H. Penzl, *A Grammar of Pashto*, University of Michigan, Ann Arbor, 1955.

[3] K. K. Khattak, *A Case Grammar Study of The Pashto Verb*, PhD thesis, Department of Phonetics and Linguistics School of Oriental and African Studies, Faculty of Arts, University of London, London, 1988.

[4] H. Tegey and B. Robson, *A Reference Grammar of Pashto*, Center for Applied Linguistics, Washignton, D. C, 1996.

[5] F. Babrakzai, *Topics in Pashto syntax*, PhD thesis, Linguistics department, University of Hawaii, Hawaii, 1999.

[6] F. T. Zuhra and H. Nauman, *The computational morphology of Pashto*, MSc thesis, Department of Computer Science, University of Peshawar, Peshawar, 2005.

[7] M. A. Khan and F. T. Zuhra, "A morphological analyzer for the past tense verbs in Pashto", in proc. *CLT07*, Bara Gali, Pakistan, 2007.

[8] M. A. Zyar, *Pashto Grammar*, Danish publishing association, Qissa Khwani Bazar Peshawar, 2003.

[9] S. Reshteen, *Pashto Grammar*, University Book Agency, Khyber Bazar Peshawar, 1994.

[10] H. W. Bellew, *Pashto Instructor – A Grammar of the Pukkhto Language*, Saeed Book Bank and Subscription Agency, Peshawar, 1986.

[11] F. T. Zuhra, *A corpus-based finite state morphological analyzer for Pashto,* MS (CS) thesis, Department of Computer Science, University of Peshawar, Peshawar, 2008.

[12] K. R. Beesley and L. Karttunen, *Finite State Morphology*, CSLI stdies in computational linguistics, 2003.