Automatic Derivation of Nouns from Adjectives in Pashto

Tariq Naeem
naeemtarik@gmail.com
Mohammad Abid Khan
mabid@upesh.edu.pk
Department of Computer Science, University of Peshawar, Peshawar

Abstract

The paper explains the development of class changing derivational morphological analyzer of Pashto language. This can examine various derivations of Pashto nouns from adjectives using a corpus. The lexemes (Pashto words) are tested through finite state machines and are able to accept input in the form of Pashto adjectival derivation. The input is given in Arabic-scripted Pashto equivalent form. The derivational morphological analyzer displays all the occurrences of lexemes (words) and sentences by a simple search in the corpus.

1. Introduction

There are quite a lot of Natural Language Processing (NLP) applications. Morphological analysis is a pre-imperative in different areas of NLP e.g. are lemmatizers, stemmers and spell checkers in dictionaries/corpora. This paper addresses the development of an application that can perform the derivational analysis of the inputted word showing its root along its morphological units and features (also called morpho-analysis) by retrieving all the sentences of the use of the word from Pashto corpus.

A significant usage of the system, prepared in this paper, can be in the improvement of Part of Speech (POS) tagging program and stemmers in Pashto language.

Derivations, in computational linguistics, are also called lexeme formation and word formation. According to [1] derivational affixes change completely the grammatical category of the root words to which they are connected. Derivation is divided into two parts i.e. class-maintaining derivation and class-changing derivation. This paper explains the class changing aspect of derivation in Pashto.

The rest of the work is devised in following manner. Section 2 covers the related work of class changing derivations. Section 3 explains the existing of class changing derivation of nouns from adjectives in Pashto language. Section 4 describes the transformation of those linguistics patterns, discussed in section 3, in FSTs. Discussions and results are drawn in section 5. Section 6 concludes the complete study. Limitations and future work of the proposed system is discussed in section 7.

2. Related Work

[2] Investigated Paumari verbs as derivational additions which assisted in different ways. The author depicted the particular affixes utilized as a part of Paumari language with eminent relevance.

[3] Inspected the association between morphologically referred lexemes in Malay which can adapt to derivational morphology in pairs. Their tests demonstrate translation from Malay into English. These authors determined that new lexemes can be shaped by three morphological procedures i.e. Affixation, Compounding (the shaping of new lexemes by placing two or more lexemes together) and Reduplication (also called word repetition) [1]. Inflection and derivation are because of the sub-classifications of affixations yet they are in contrast to each other [2].

As per [4], morphological generation is a critical matter for producing derivative frames of word from semantic representation. Arabic language, of Semite peoples, has frequent derivational appearances. The research work of [4] demonstrates that great part of the Arabic vocabulary is comprised of words which are deduced from stems by insertion of prefixes, infixes and suffixes. In initial stage, these authors perceive prefixes and suffixes. Then at that point utilizing affixes limitation, they filter the mixed up relationship of affixes to perceive the radical word (root word). Their outcomes show whether this

'radical word' (new lexeme) has a place in Arabic language [4]. The work of these authors influences the derivation of words from the root words thereby changing its class. These authors demonstrated the linguistic constructs of word categorization and morpho-syntactic enactment on a written work. As a result, the new word gives a whole different meaning and changes its grammatical family [4] [3].

The work of [5] examines the role of sub-lexical units as a response for dealing with productive derivational strategies, in the building of a lexical utilitarian grammar for Turkish. Such sub-lexical units make it possible to reveal the internal structure of words with various derivations to the grammar rules consistently [5]. In conclusion, this reminds more minimal and sensible guidelines. Further, the semantics of the findings can similarly be purposely seen in grouping.

The procedures of framing new words from existing words through the affixation of morphemes or by removal of affixes where the resulted new lexeme is different from its previous original root goes to a different grammatical category. [6] Trained a derivational analyzer for Hindi over a past existing inflectional analyzer. In the proposed methodology, the authors determined words in Hindi language which were examined to get the derivational affixes. Then the patterns were arranged by understanding the properties of the affixes.

[7] Proposes morphological analysis using word and paradigm (Prefix-Suffix) model using a corpus. These authors evidenced the parser precision and effective use of memory and the corpus.

[8] Give a clear engineering model of a basic and precise framework for developing a morphological analyzer with finite state transducers (FST) using Telugu noun forms. These authors have indicated the regular expressions and unicode for their data format. Rule based methodology was utilized for constructing the morphological analyzer for Tamil language. The machine learning techniques were utilized for carrying out morphological analysis for Tamil

[9]. These researchers have trained separate modules for verb and noun. These authors have segregated their morphological analyzer in three levels. The first module, to begin with, is the pre-processing level. The input is changed over into a section of a sequence of units for handling by the morphological analyzer. Second module segments the linguistic units called morphemes. As per the morpheme boundary, preprocessed words are fragmented into smaller chunks. In third and last module is distinguishing morphemes. The segmented morphemes are given to the developed analyzer; it then predicts the linguistic classification to the segmented morphemes.

A blended methodology (of previous methodologies) for building a morphological analyzer for Malayalam language is used by [10]. These authors combined the methodologies of root (word) recognizer and addition suffix stripping approach. They enforced a lexical dictionary for better outcome and fast processing.

[11] Have done a relative study on Malayalam language using distinctive methodologies e.g. Brute force technique, Root driven strategy and Suffix stripping. By this hybrid approach, these authors took the advantage of both paradigms and had separate groups of classes whose morphophonemic patterns are the same. Their Malayalam morph-analyzer assists in automatic spelling and sentence structure checking, NLU (Natural Language Understanding), Speech synthesis, POS (Part of speech) taggers and parsers.

Before commencing this computational study, the work done by Pashto linguists was studied. They are Roberts [12], Tegey [13], Ziyar [14], Tegey and Robson [15], Babrakzai [16] and Rishteen [17]. The work of these language specialists frame the premise for the exploration work exhibited in this paper.

1. Class Changing Derivatives from Adjectives to Nouns in Pashto

The work of [1] and [2] can also be implemented for Pashto language. The affixes (prefixes and suffixes) can also be found in Pashto. According to [18], derivation can take place from adjectives to abstract nouns as follows:

Example 3.1

وږي [wagey]'hungry' or لوږَ [lawaga]'hunger'.
لوږه تنده پر غالبه شوه یک باره
[Lowagah tandah prey ghaliba showa yak barah]
په صورت ور پاتِ نه شه طاقت توان
[pah sorat wer paeti nah sha taqat twan]

"Hunger and thirst all at once overpowered him.
In his body no power or strength remained." [18]

In the above example, it is clear that the suffix ي [ye] of وږي is removed and prefixل [laam] is added to the root instead to change the word from adjective to noun to make it a new lexeme.
The word لوږَ or لوږه is same, the word is sometimes written with َ [zabar] and sometimes ه. This is a poetic stanza and we know that poets molds words the way they want and that's why it is sometime very hard to understand what they want to say.

Example 3.2

تږي [tagey] ‘thirsty’, تنده [tandah] or تندَ [tandah] ‘thirst’.

لوړه تنده نه شته دّ قانع په قناعت کښي

[lowagah tandah nah shtah da qanae pa qina’at ke]

دا کیمیا چه زده کا په خرقه کښي اُمرا وي

[da chemia che zda ka pah kharqah kae omurah we]

“In the contentment of the contented man, there is neither hunger nor Thirst;

And they who acquire this alchemy will be noble, tho’ clad in rags.” [18]

The تنده [tandah] ‘thirst’ is formed by dropping two letters in the تږي [tagey] ‘thirsty’ i.e. (ږ and ي)and three other letters i.e. (د , ن and ه) are affixed with ت [tey] to make it a noun from adjective.

Example 3.3

رُونړ [ronrr] or رُونڼ [ronrr] ‘bright’, رَنړا [rarra] or رَنڼا [rarra] ‘brightness’.

په رنړا ئي دَ چا کار نه پوره کیږي

[pah rarrae ked a cha kaar nah porah kege]

دّ آسمان برق و بریښنا ده دا دنیا

[da asman barq o braekhna dah da dunya]

“By the Light of it the business of this life cannot be perfected;

For this is as the lightening and the light of the sky.”[18]

Sometimes the lexeme takes ئِي **[aey]**, as in the following example:

لکه نمره په جهان وُ خیژي رنړائي شي

[lakah namra pah jahan aokhejey rarrae she]

دم قدم هسي زنده کاندِ اخلاص

[dam qadam hasse zinda kande ikhlaas]

“As when the sun riseth on the world, *Light* and *Brightness* cometh,

So doth friendship and affection give life to both breath and footstep?” [18]

In both examples discussed above و of the word رُونړ **[ronrr]** is bumped off and ا **[alif]** is appended with the root word in order to change the lexeme from adjective to noun.

Example 3.4

تور [tor] ‘dark’ or ‘black’ تّیاره [tiyarah] or تیارَ **[tiyara]** ‘darkness’ or ‘blackness’.

کل جهان توره تیاره شه له هغه ګرد و غبار

[kul jahan tora tiyarah sha lah hagha garrd o ghubaar]

آسمان رعد بریښیده تکه شمسیران

[asmaan ra’ad brikheda takah shamsheraan]

The whole world filled with *Darkness* from this dust and vapour;

In the heavens thunder and lightning flushed as from swords.”[18]

The infix و in the lexeme تور **[tor]** is removed and یا is infixed to change it to a new lexeme تیارَ **[tiyara]** i.e. from adjective to noun.

Example 3.5

یون ګران په لار دي بوالهوس ته

[yoon Gran pah laar de boalhoos tah]

مرد هغه ګنړه چه بنیګړي که بنا

[mard hagha garrah che khaegarre keh bina]

“Journey on this road is difficult to the fickle and capricious:

Consider him a man who layeth the foundation of *Goodness*”. [18]

In the above illustration, ګړ and ي is infixed with the lexeme بنه to make it noun from adjective.

The whole of the nouns of the above classes mentioned from 1 to 5 are feminine. The following i.e. 6 to 8 are all masculine, with the exception of those words formed by affixing تیا **[tiya]**, ستیا **[stia]**, ګلوي [galwe]and ولي **[wali]**, are feminine.

Example 3.6

مخ دِ سپین لکه آفتاب وُه

[makh de speen lakah aftab woh]

تر آفتابَ ئي لا تاب وُه

[ter aftaba ye la taaba woh]

ولي اوس دا هسي تور شه

[wali aos da hassi toor sha]

په تورواليِ لکه سکور شه

[pah toorwaali lakah skoor sha]

“Thy countenance was white like unto the sun- yea!
It was brighter than the orb of the day:
But now, alas! It is become so black,
That it’s *Blackness* is like unto charcoal”. [18]

It is justified from above examples that the lexeme تور **[toor]** is an adjective by suffixing والي **[waali]** to تور **[toor]** makes تور والي **[toorwaali]** which is a noun.

Example 3.7

کله ما وته امید دَ خپل ژوندوُن شي

[kalah ma wata umeed da khpal zowandoon she]

په هجران به ئي ژوندوُن راته زبون شي
[pah hijraan ba ye zowandoon rata zaboon she]
"When shall I entertain hope for my own *E*xistence?
Since separated from her, *Life* itself to me is infamous". [18]

The ئ **[ye]** of the word ژوندئ **[zowande]** is removed and ون **[woon]** is affixed instead to change the word into a new lexeme.

Example 3.8

ناګاه ويښه شوه له خوبّ
[na gah wekha showa lah khoob]
ز ړه ئی ډک له مَيَن توبَ
[zrra ye ddak la maentoob]
کښيناسته نګاه ئی وُ کړ
[kaenaasta nigah ye okarr]
يار ئی نه ليد آه ئی وُ کړ
[yaar ye na lid aah ye okarr]
"Suddenly she awoke from her slumbers,
Her heart filled with *Love* and *Affection*.
She sat up and gazed around, but signed
For she beheld not her beloved one." [18]

In this example, second stanza, the word توبَ **[toob]** is suffixed with مَيَن **[maen]** which is an adjective to make a new lexeme مَيَن توبَ **[maentoob]**, a noun.

خداي دِ نه کا ندِ بيلتون دَ دوهٴ يارانو
[khudae de nah kande baeltoon da dowao yaraano]
په بيلتون عاشق په روغ صورت بيمار دي
[pah baeltoon aashiq pah rough sorat bemaar de]
"God forbid that *Separation* should be caused between two lovers;
For in *Separation* the lover, though healthy in body, is sick at heart." [18]
In the above example, ټون [toon] is suffixed with the word بيل [bael] to make it a noun i.e. بيلتون [baeltoon]
چه په ديدن مِو ړتيا نه شوه
[che pah dedan de morrtiya na showah]
اوس د يار غمو کړي مو ړ
[os de yaar ghamo karre moorr]
"Whereas from her presence thou didst not acquire *Satiety*,
Grief on her account has now *Satiated* thee" [18]

In the second stanza of the above example ړ مو **[]** is an adjective but by putting the suffix تيا with مو ړ makes ړتيا مو which is a noun.

4. Modeling and Finite State Transducers

Finite state machines (FSM) also known as finite state transducers (FST) efficiently compute many useful Natural Language Processing (NLP) functions and weighted transitions on strings. In many fields of NLP, FSTs are applied as a core technology in developing spell checkers, parsers, POS taggers, speech recognition systems, morphological analyzers, information retrieval systems and lexical analyzers [19]. This work encouraged the modeling and design of FSTs. Using FSTs, the natural language modeling is efficient and more effective because they are mathematically derived models, by Chomsky, and are well-understood [20].

a. وږي **[wagey]** 'hungry' or لوږَ **[lawaga]** 'hunger'

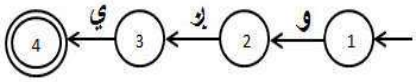

Figure 1: The modeling of adjective وږي 'hungry'

The strings recognized in this finite automata is لوږَ **[lawaga]**'hunger' with zabar and لوږه **[lawagah]** 'hunger' with ه **[hy]** as it ends at state 5. Both are nouns.

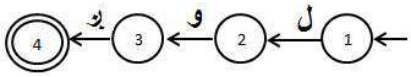

Figure 2: The FST of noun لوږَ'hunger'

The above automaton, its derivative is وږي [wagey] 'hungry' is form when the finite automaton starts from the initial state and ends straight at the final state 4 is an adjective. In this finite automaton we have 6 states.

b. تږي **[tagey]** 'thirsty',تنده **[tandah]** or تندّ **[tandah]** 'thirst'.

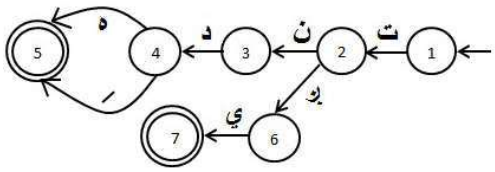

Figure 3: Automaton of adjective changed to noun

In this finite automaton, we have 7 states. Starting from state 1, marked as initial state and rests at state 5 and 7 as they are the final states. This automaton forms three strings. First, if we start with state 1 and move along with transition to state 5, two strings are identified which has the same semantics. The strings are تندّ**[tandah]** 'thirst' with zabar and تنده **[tandah]**

‘thirst’ with ه **[hy]** are nouns. Its derivatives are formed by following the path from state 1, 2, 6, and 7 i.e. تږي **[tagey]** ‘thirsty’ is an adjective.

c. رُونړ **[ronrr]** or رُونڼ **[ronrr]** ‘bright’, رّنړا **[rarra]** or رّنڼا **[rarra]** ‘brightness’.

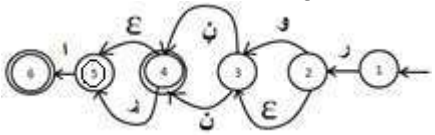

Figure 4: Automaton of both adjective رُونڼ ‘bright’ and noun رّنړا [rarra] ‘brightness’

The above automaton accepts variety of words. It has total 6 states with 3 marked as final and 1 as initial.

The first word which is recognized by this automaton is رُونڼ **[ronrr]** ‘bright’ if we follow state 1, 2 and 3. You can get رُونړ **[ronrr]** ‘bright’ by following the path from 1, 2, 3, 5 and 6. The first derivative of the two nouns are رّنړا **[rarra]** ‘brightness’ is recognized by state 1, 2, 3, 4, 5 and 6. Furthermore, if you go from the transition marking state 1, 2, 3, 4 and 5 will get the string رّنڼا **[rarra]** ‘brightness’.

d. تور [tor] ‘dark’ or ‘black’ تّیاره [tiyarah] or تیارَ **[tiyara]** ‘darkness’ or ‘blackness’.

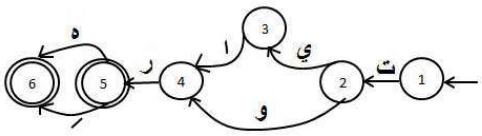

Figure 5: The FST of adjective تور ‘dark’ and noun تّیاره ‘darkness’

The next string and its derivative can be recognized by visiting the state 1, 2, 4 and 5 and its derivative, Princeton, New Jersey.

can be extracted by the following two paths from state 1 to state 6 taking the alternative path یَ to state 3 will lead you to تیارَ **[tiyara]** ‘darkness’ or ‘blackness’ and the string تّیاره [tiyarah] ‘darkness’ or ‘blackness’.

e. ښه [khah] ‘good’, ښیګړه [khaegarrah] ‘goodness’.

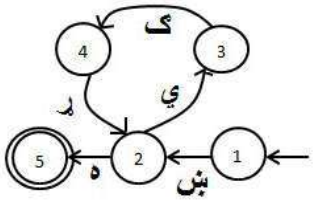

Figure 6: FST of adjective ښه ‘good’ and noun ښیګړه ‘goodness’

The strings ښه [khah] ‘good’ and ښیګړه [khaegarrah] ‘goodness’ is recognized in a similar and simple manner. The word ښه [khah] ‘good’ is recognized by visiting state 1, 2 and 3 only follow the alternative path from the initial state.

f. تور **[toor] ‘black’,** توروالي **[toorwaali] ‘blackness’;** کلک **[klak] ‘hard’,** کلک والي **[klakwaali] ‘hardness’**

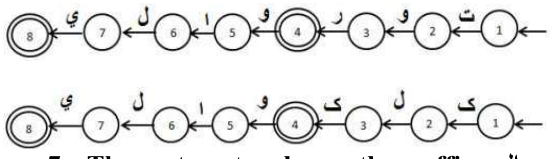

Figure 7: The automata shows the suffix والي **[waali] with adjectives to make it a noun**

The automata of توروالي **[toorwaali] ‘blackness’** and کلک والي **[klakwaali] ‘hardness’** is more or less the same. The have the same number of states with the same states as final states. The string توروالي **[toorwaali] ‘blackness’** is recognized as follow. The string تور **[toor] ‘black’** is recognized by starting at the initial state and ends at state 4. And its derivative توروالي **[toorwaali] ‘blackness’** take all the way long from initial to the latter final state i.e. state 8.

Starting at state 1 and stops at state 4 give us the string کلک **[klak] ‘hard’**. If we go straight from state 1 to state 8 we get the adjective کلک والي **[klakwaali] ‘hardness’**.

g. ژوندیّ **[zowande] ‘alive’ or ‘existing’,** ژوندون **[zwande] ‘life’, ‘existence’;** نښتي **[nkhate] ‘captive’, ‘prison’,** نښتون **[nakhtoon] ‘captivity’, ‘imprisonment’.**

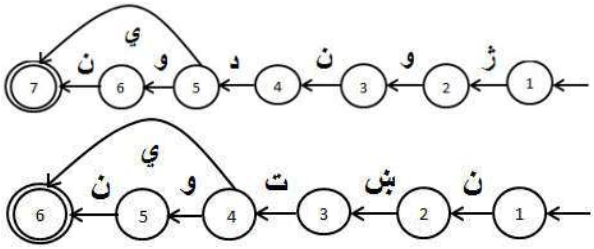

Figure 8: Automata depicts adjectives with suffix ون [oon] converting it to nouns

The finite automata of ژوندون **[zwandon] ‘life’** and نښتون **[nakhatoon] ‘captivity’, ‘imprisonment’** have the same number of states. The string ژوندون **[zwandon]** ‘life’ is produced by starting at the initial state i.e. state 1 and to states 2, 3, 4, 5, 6 and finally stopping in state 7, marked as the final state. Its adjective ژوندیّ **[zowande]** ‘alive’ or ‘existing’ takes all the way long from initial state 1 to states 2,

3, 4, 5 and then jumping straight to state 7 on the input symbol يّ.

Similarly, the other word نښتون **[nakhatoon]** starting at state 1 and to states 2, 3, 4, 5 and finally stopping at the final state 6, reading all the input symbols generates a noun. The string نښتي **[nkhate] 'captive', 'prison'** is an adjective and can be produced if we start crawling from state 1 and to states 2, 3, 4 and finally to state 6.

h. بيل [bael] 'separate', بيلتون [baeltoon] 'separation'; ځاي [zaey] 'a place', ځايتون [zaeytoon] 'a dwelling place', 'a home', 'a birthplace'; مَين [maen] 'affectionate' مَينتوب [maentoob] 'affection', 'love', ليوني [lewane] 'mad', ليونتوب [lewantoob] 'madness'; مړ [morr] 'satiated', مورتيا [morrtiya] 'satiety'; خمسور [khamsoor] 'impudent', خمسور تيا [khamsoortiya] 'impudence', 'familiarity'.

The finite automata with the suffix تون and توب are drawn. The strings can be recognized in the following manner.

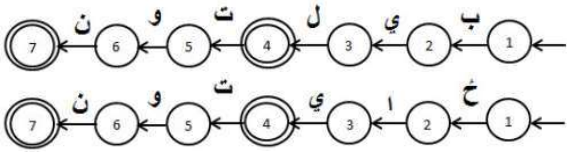

Figure 9: FSTs indicates the suffix تون [toon] with adjectives converting it to nouns

Starting at initial state and visiting only 2, 3, and 4 will give you the string بيل **[bael] 'separate'** but if you move along till state 7, you'll get بيلتون **[baeltoon] 'separation'** which is the derivative of the first.

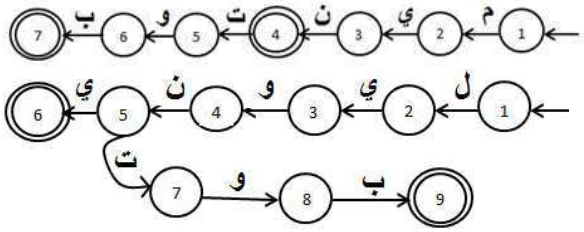

Figure 10: Automaton drawn shows adjectives changed into noun by the suffix توب [toob]

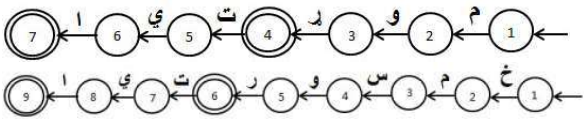

Figure 11: FSTs depicts the suffix تيا [tiya] with adjectives to convert it into nouns

5. Results and Discussions

In this section, the implementation of derivational morphological analyzer is discussed. The FSTs, created during the modeling stage, are executed and ensured.

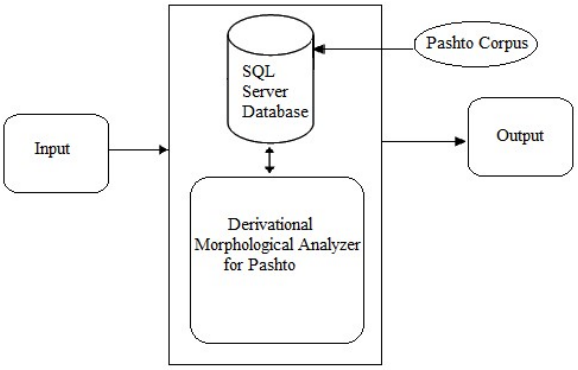

Figure 12: Entire framework overview

5.1 Pashto Corpus and its use in proposed work

Developing a corpus of Pashto language was the first and foremost activity to be used in proposed work. A Pashto corpus was created utilizing the corpus improvement tool XML Aware Indexing and Retrieval Architecture (XAIRA). Tagging of Pashto text was done in Extensible Markup Language (XML). The corpus is composed of Pashto textual material. Data was collected from different areas, like, news, memos, letters, research articles, books, fiction, sports and magazines, making it a representative corpus. Then, these XAIRA tagged files were used in Microsoft Structured Query Language (SQL) Server Management Studio 2012 tables. Each entry in database tables has sorting rules based on alphabet or language and comparison styles using collation feature of Microsoft SQL Server. The 'nvarchar data type' was used to insert Pasto text in SQL tables. The window locale collation to insert Pashto text is "Pashto_100_BIN". Collation determined for unicode data specify rules for text entered in columns. The corpus as of now contains 0.5 million words containing lexemes with its grammatical class. The corpus is utilized for considering the derivational morphological arrangement of Pashto. The derivations were broke down into stems and affixes. An example is displayed given in table 1.

Table 1 Derivational Stems and Suffixes

DESKTOP-HU8H0NP....bo.ExtractedData

	Root	GramClass1	Affix	RootAffix	GramClass2
	تور	Adjective	والي	توروالي	Noun
	کلک	Adjective	والي	کلکوالي	Noun
	ژوندێ	Adjective	ون	ژوندون	Noun
	نښتي	Adjective	ون	نښتون	Noun
	بیل	Adjective	تون	بیلتون	Noun
	ځاي	Adjective	تون	ځایتون	Noun
	ټین	Adjective	توب	ټینتوب	Noun
	مور	Adjective	تیا	مورتیا	Noun

The assessment of derivations demonstrates that the nouns in Pashto language have several types based on suffixes. Each of these examples illustrated in Example 3.1 to example 3.8 has an exceptional type of morpheme unlike from the other families for depicting the equivalent aspect.

5.2 Implementation of derivational morphological analyzer

The automata were implemented which were formulated during the modeling and design stage. Four programming languages and tools were used for implementation. First, the XEROX LEXC tool was used. LEXC (Lexicon Compiler), is a writing apparatus for making vocabularies and lexical transducers. The LEXC tool was used to draw finite state diagrams from the lexicon of adjectives and nouns in Pashto.

A universal application XEROX, (XEROX Finite State Transducer) XFST is very useful for processing FSTs. Compilation from .txt files are done efficiently by XFST reading from binary files. XFST gives numerous approaches to get data around a system and to review and adjust its structure. The input of XFST, in our proposed study, was the output generated by LEXC compiler.

```
LEXICON Root
  مور        Adjective;
  خمسور      Adjective;

LEXICON Adjective
NounSuffix;
#;

LEXICON NounSuffix
  تیا #;
    #;
```

Figure 13: Sample lexicon processed by XEROX's LEXC

The sample of lexicon in figure 13 is the representation of lexicon developed and stored in notepad with utf8-mode for reading Pashto text. The LEXC compiler starts evaluating the Root lexicon first when compiled by XEROX's LEXC compiler. Then, it trails through other lexicons by suffixation rules.

```
Starting in utf8-mode.
lexc> compile-source lexc_CCD.txt
opening "lexc_CCD.txt"
Opening 'lexc_CCD.txt'...
Root...2, Adjective...2, NounSuffix...2
Building lexicon...Minimizing...Done!
SOURCE: 2.5 Kb. 11 states, 11 arcs, 4 paths.

lexc>
```

Figure 14: LEXC interface

Figure 14 shows the compilation of lexicon displayed in figure 12. The source in LEXC compiler depicts a total of 11 states, 11 arcs and 4 paths transducer. The lexicon of the finite state transducers, thus developed, contains 621 nouns and 953 adjectives. This information is carried and processed by XEROX XFST tool. The XFST generated output file which was forwarded and stored in Microsoft SQL Server. Finally, the Microsoft Visual Studio CSharp dot net framework was used as a front end to work on the corpus stored in Microsoft SQL Server.

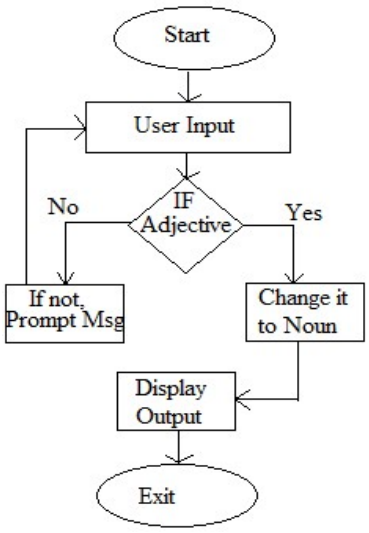

Figure 15: Flowchart of the system

The flowchart of figure 15 depicts the complete working of derivational morphological analyzer converting adjectives to nouns. The screen shot of sample interaction with the proposed system is given in figure 16:

Figure 16: The developed derivational morphological analyzer

The above morphological analyzer analyzes derivation from adjectives to nouns. In corpus, the sentence containing the query word retrieves all the sentences and displays them in the left pane. Similarly, the morphological analysis is done in the search query.

5.3 Error Analysis

The accuracy of the system was measured by error analysis was performed. The output was recorded and matched manually with the system generated output. A sample of 174 nouns and 169 adjectives were collected from the written Pashto data and given to the system as input. Out of this input 147 nouns and 142 adjective words were accurately examined. Consequently, the complete accuracy of the overall system is:

((147+142) / (174+169)) * 100 = 84.25%

Due to limited lexicon size, the system could not search and analyze the root word with its derivational affixes in Pashto corpus.

15.75% of the errors were because of typographic variation and borrowed words from other languages in Pashto.

6. Conclusion

In this paper, the derivational properties of Pashto language are discussed. There are two kinds of derivations that exist in Pashto i.e. Class changing and class maintain derivation.

Being the first derivational morphological analyzer created for Pashto, it achieved 84% overall accuracy and it can be further improved by using big corpus. A lot of work has been done to build a representative corpus in Microsoft SQL Server 2012. The collation feature, not available in older versions, was used to store the corpus in its original form.

Due to complexities involved in derivation, this area is very sensitive for making new words to enrich Pashto because a slight change of ◌ً **[zabbar],** ◌ٍ [**zeir]** and ◌ٔ [**paish]** can be disastrous.

7. Limitations and Future Work

The current state-of-the-art derivational morphological analyzer changes adjectives to nouns. This paper discussed class changing derivation of Pashto. Further work can be done on class maintaining derivatives in Pashto language. Also, the corpus size can be increase to achieve higher accuracy for derivational rules.

References

[1] M. Aronoff and K. Fudeman, "What is morphology?", Fundamentals of linguistics 1, Blackwell Publishing, Malden, MA, **2005**, 1-3

[2] Chapman S., Paumari Derivational Affixes *Associacao Internacional de Linguistica SIL Brasil* **2008**, 9, 10.

[3] M.A Malik, "An approach to the study of linguistics" New Kitab Mahal, Lahore, **2010**, 116-118

[4] P. Nakov; H.T. Ng. Translating from morphologically complex languages: a paraphrase-based approach *HLT' 11: Proceedings of the 49th Annual Meeting of the ACL* **2011**, 1299.

[5] H.G. Raverty, "A grammar of the Pashto or Afghan language" Bombay, Calcutta, **2007**, 9-12

[6] Buckwalter T.; Arabic Morphological Analyzer version 2.0 LDC **2004**, 11.

[7] K.G. Mkanganwi, Shona (derivational) Morphology: Observation in Search of a Theory. *Zambezia* **2002**, 175,176

[8] Cetinoglu Ozlem; Oflazer Kemal, Morphology-syntax interface for Turkish ACL **2006**, 153,160, 10.3115/1220175.1220195

[9] Kanuparthi N.; Inumella A.; Sharma D.M., Hindi derivational morphology analyzer, ACL **2012**, 12

[10] Uma Maheshwar Rao G; Ambar Kulkarni P.; Christopher Mala, The study effect of length in morphological segmentation of agglutinative languages *ACL*, **2012**, 18, 19.

[11] D.L Sneha and K. Bharadwaja, "A novel approach for morphing Telugu Noun forms using finite state transducers" IJERT, **2013**, 2(7), 550.

[12] V.P Abeera, S. Aparna, R.U Rekha, K. Anand, Dhanalakshmi, Soman and Rajendran, "Morphological Analyzer for Malayalam using Machine Learning" ICDEM'10, **2010**, 253

[13] Vinod P M; Jayan V; Bhadran V K, Implementation of Malayalam morphological analyzer based on hybrid approach *ACL* **2012**, 310

[14] Jisha P.; Jayan; Rajeev; Rajendran, Morphological Analyzer for Malayalam – A comparison of different approaches *IJCSIT* **2009**, *2*. 156,157

[15] Roberts Taylor, Clitics and Agreement **2000**, 17-23, Massachusetts Institute of Technology (MIT), Linguistics

[16] Tegey Habibullah, The grammar of clitics: Evidence from Pashto (Afghani) and other languages **1977**, University of Illinois
[17] Ziyar Mujawar Ahmad, Pashto grammar, **2005**, 83-99, vol-3, Danish Publishers branch association
[18] Tegey Habibullah; Robson Barbara, A reference Grammar of Pashto **1996**, 46-88, Center for Applied Linguistics, Washington D.C
[19] Babrakzai F., Topics in Pashto syntax, PhD thesis, Linguistics department, University of Hawaii, Hawaii, **1999**
[20] Rishteen SaqeedUllah, Pashto grammar, **2001**, 389-392, University Book Agency publisher
[21] C.E. Biddulph, Afghan Poetry of the 17th Century: being selected from the poem of Khushal Khan Khattak Denzil lbbetson Atlantic publishers, London **2006**, 31.
[22] Beesley Kenneth R.; Karttunen Lauri, Finite State Morphology, **2003**, 501-505 Stanford, CA: CSLI Publications
[23] Chomsky Noam, On certain formal properties of grammars, **1959**, 141-166, Massachusetts Institute of Technology, Massachusetts and The institute for advanced study