DAAD
Deutscher Akademischer Austausch Dienst
Servicio Alemán de Intercambio Académico

CLE
Unlocking Information for Human Development

Automatic Derivation of Nouns from Adjectives in Pashto

Tariq Naeem
Muhammad Abid Khan

DAAD
Deutscher Akademischer Austausch Dienst
Servicio Alemán de Intercambio Académico

CLE
Unlocking Information for Human Development

Outline

- Introduction
- In the World
- Problem Statement
- Data Collection / Analysis
- Modeling of Data
- Implementation
- Error Analysis
- Conclusion

DAAD
Deutscher Akademischer Austausch Dienst
Servicio Alemán de Intercambio Académico

CLE
Unlocking Information for Human Development

Introduction

- Derivations, in computational linguistics, are also called lexeme formation and word formation.
- Derivational affixes change completely the grammatical category of the root words to which they are connected. For e.g.

Adjectives	Nouns
Wicked	Wickedness
Dark	Darkness
Foolish	Foolishness

DAAD
Deutscher Akademischer Austausch Dienst
Servicio Alemán de Intercambio Académico

CLE
Unlocking Information for Human Development

In the World..

- There are quite a lot of Natural Language Processing (NLP) applications.
- Morphological analysis is a pre-imperative in different areas of NLP, for instance:
  - Lemmatizers
  - Parsers
  - Stemmers
  - Spell checkers
  - Part of Speech taggers
  - Speech Recognition systems

DAAD
Deutscher Akademischer Austausch Dienst
Servicio Alemán de Intercambio Académico

CLE
Unlocking Information for Human Development

Contd..

- We looked in to the work of several other researchers in the field of derivation in other languages.

1. Paumari verbs as derivational additions depicting particular affixes.
2. Lexemes in Malay was inspected to find the association between morphologically stems which can adapt to derivational morphology in pairs

DAAD
Deutscher Akademischer Austausch Dienst
Servicio Alemán de Intercambio Académico

CLE
Unlocking Information for Human Development

Problem Statement

- Does class changing derivational morphological analysis exist in Pashto language?

- Under what circumstances, words in Pashto language change their grammatical category from adjectives to nouns?

DAAD
Deutscher Akademischer Austausch Dienst
Servicio Alemán de Intercambio Académico

CLE
Unlocking Information for Human Development

Contd..

3. Arabic language showed that part of words are deduced from stems by insertion of affixes.
4. Turkish grammar was examined for dealing with productive derivational strategies.
5. Hindi original root words were examined to find patterns arranged by understanding the properties of the affixes.

DAAD
Deutscher Akademischer Austausch Dienst
Servicio Alemán de Intercambio Académico

CLE
Unlocking Information for Human Development

Data Collection

- Developing corpus of Pashto language was the first and foremost activity.

- Data was collected from different areas, like, news, memos, letters, research articles, books, fiction, sports and magazines, making it a representative corpus.

DAAD
Deutscher Akademischer Austausch Dienst
Servicio Alemán de Intercambio Académico

CLE
Unlocking Information for Human Development

Data Analysis

- Pashto corpus was created using the corpus improvement tool XML Aware Indexing and Retrieval Architecture (XAIRA).
- Tagging of Pashto text was done in Extensible Markup Language (XML)
- XAIRA tagged files were used in SQL Server 2012 tables.

DAAD
Deutscher Akademischer Austausch Dienst
Servicio Alemán de Intercambio Académico

CLE
Unlocking Information for Human Development

Contd..

Table: Derivational stems and suffixes

DESKTOP-HU8H0NP....bo.ExtractedData

	Root	GramClass1	Affix	RootAffix	GramClass2
	تور	Adjective	والي	توروالي	Noun
	كلک	Adjective	والي	كلكوالي	Noun
	ژوندیّ	Adjective	ون	ژوندون	Noun
	نښتي	Adjective	ون	نښتون	Noun
	بيل	Adjective	تون	بيلتون	Noun
	ځاي	Adjective	تون	ځايتون	Noun
	مټين	Adjective	توب	مټينتوب	Noun
	مور	Adjective	تيا	مورتيا	Noun

- The table has entry against every grammatical class.
- The derivations were broke down into stems and affixes

DAAD
Deutscher Akademischer Austausch Dienst
Servicio Alemán de Intercambio Académico

CLE
Unlocking Information for Human Development

Modeling of Data

- We model our corpus data into Finite State Transducers (FSTs).
- We used FSTs because they are mathematically derived models which efficiently compute many useful NLP functions and weighted transitions on strings.
- They are efficient and more effective.

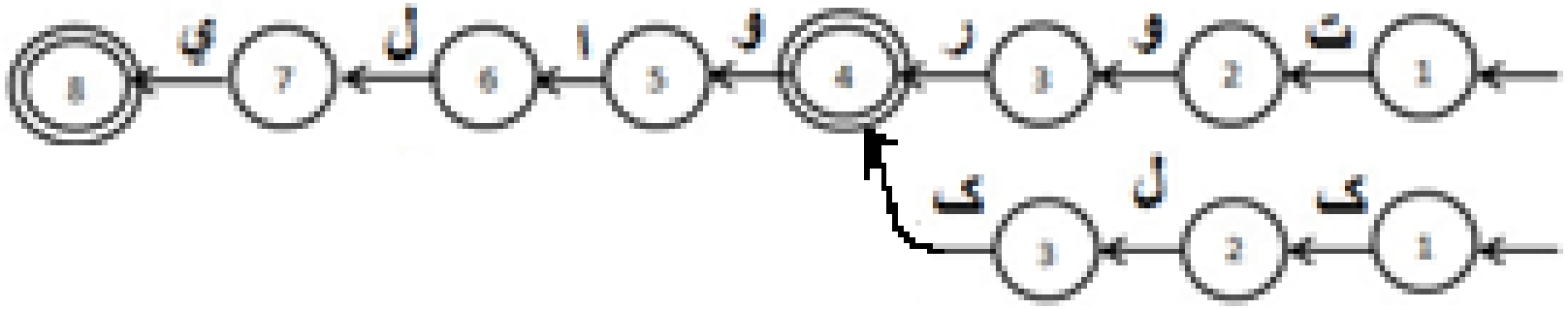
DAAD
Deutscher Akademischer Austausch Dienst
Servicio Alemán de Intercambio Académico

CLE
Unlocking Information for Human Development

Contd..

The eight (8) unique rules defined were modeled by FSTs.

1. The adjective تور (Black) to a noun تورواليي (Blackness) and adjective کلک (hard) to a noun کلک والي (hardness). It is shown as:

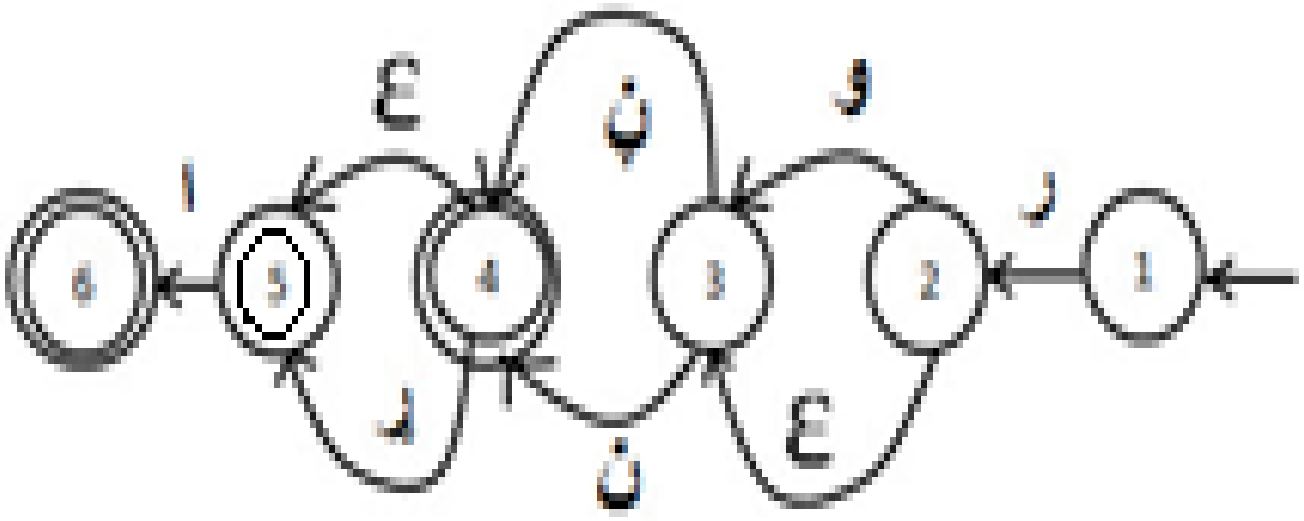


DAAD
Deutscher Akademischer Austausch Dienst
Servicio Alemán de Intercambio Académico

CLE
Unlocking Information for Human Development

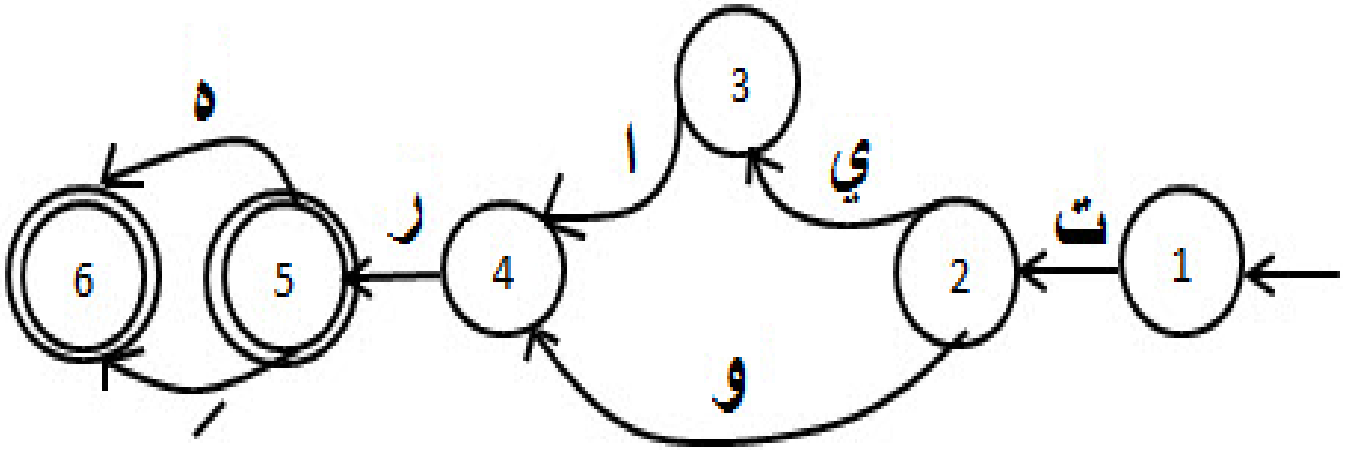
Contd..

2. رُونړ**[ronrr]** or رُونِ**[ronrr]** ‘bright’, رّنړا**[rarra]** or رّنا **[rarra]** ‘brightness’.



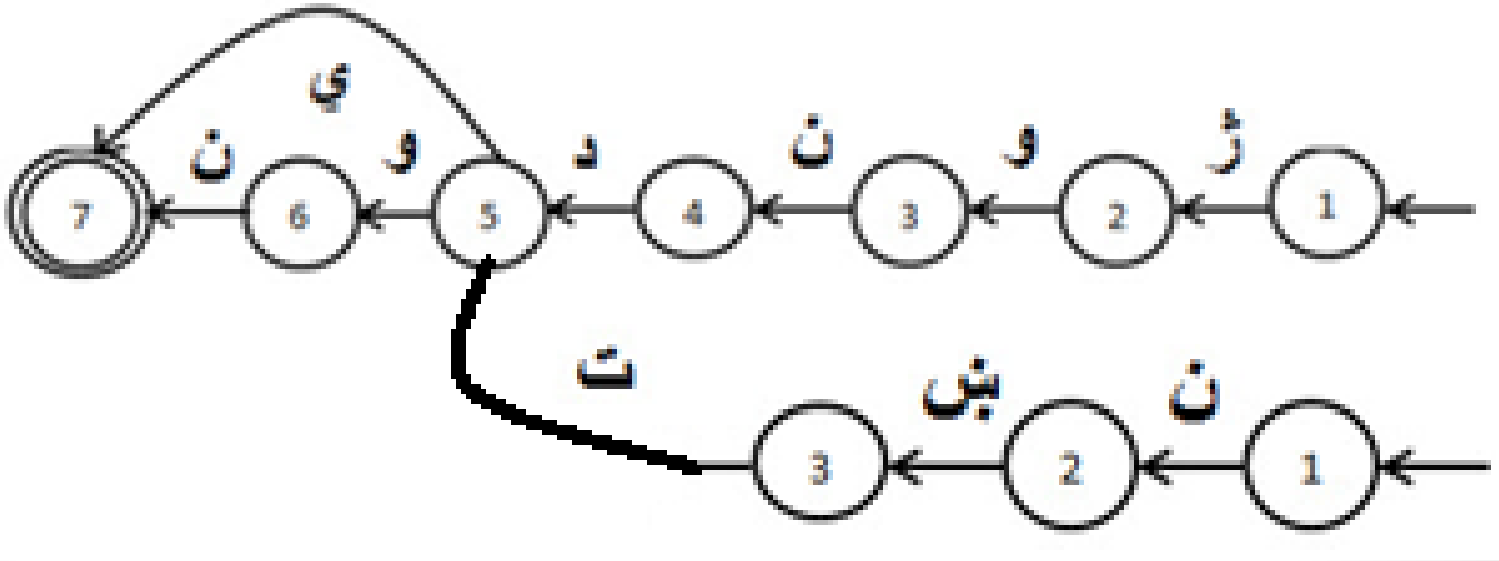
Contd..

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



Contd..

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

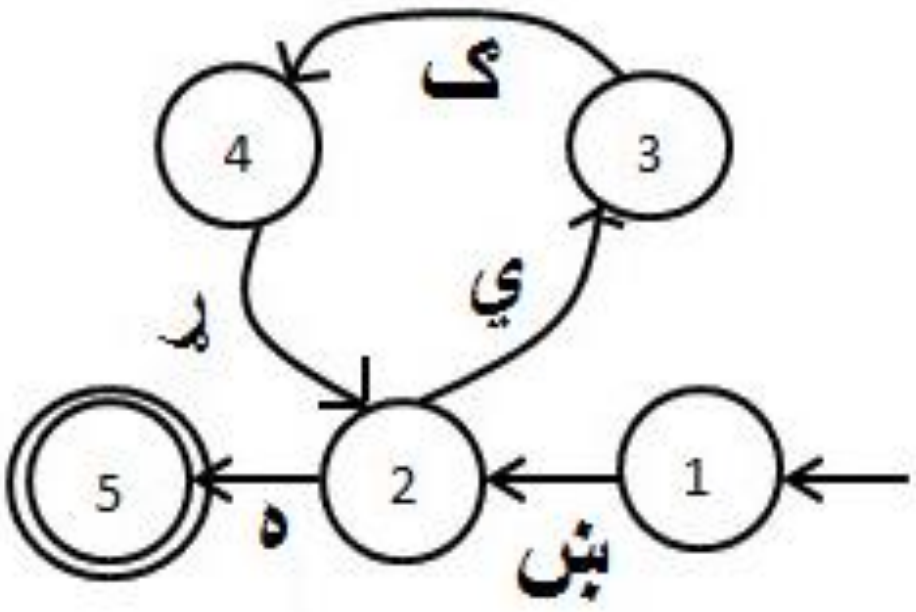


DAAD
Deutscher Akademischer Austausch Dienst
Servicio Alemán de Intercambio Académico

CLE
Unlocking Information for Human Development

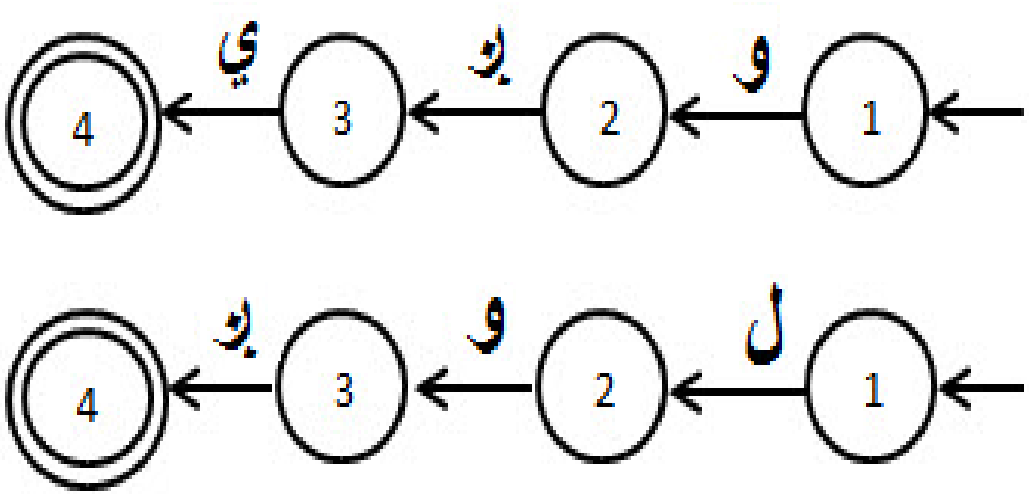
Contd..

5. بنه [khah] ‘good’, بنیګړه [khaegarrah] ‘goodness’.



Contd..

6. وږي **[wagey]** ‘hungry’ or لوږَ **[lawaga]** ‘hunger’

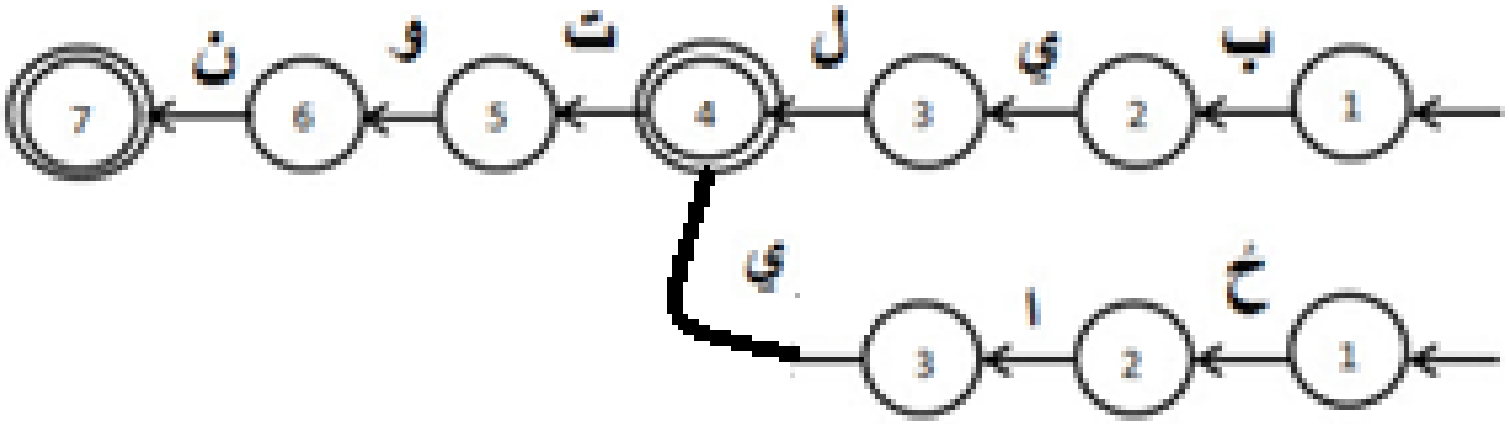


DAAD
Deutscher Akademischer Austausch Dienst
Servicio Alemán de Intercambio Académico

CLE
Unlocking Information for Human Development

Contd..

7. بيل [bael] ‘separate’, بيلتون [baeltoon] ‘separation’; ځاي [zaey] ‘a place’, ځايتون [zaeytoon] ‘a dwelling place’, ‘a home’, ‘a birthplace’;

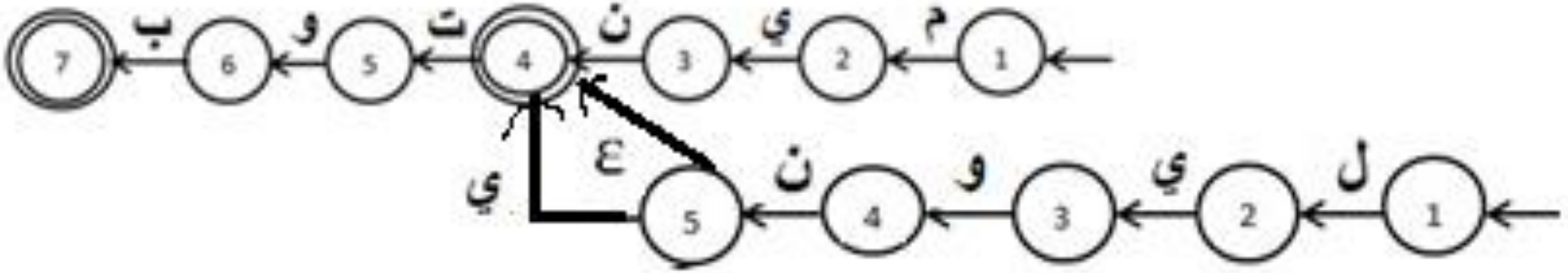


DAAD
Deutscher Akademischer Austausch Dienst
Servicio Alemán de Intercambio Académico

CLE
Unlocking Information for Human Development

Contd..

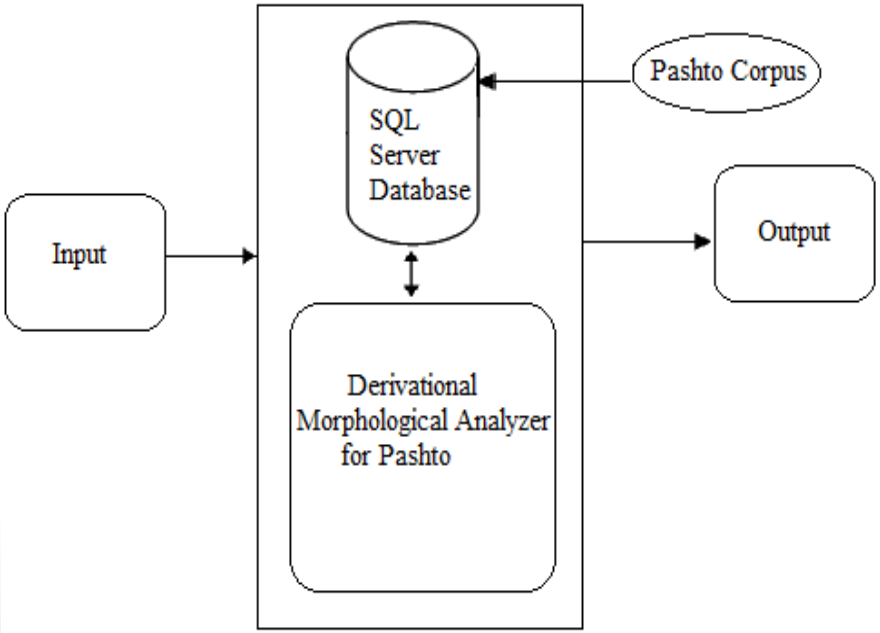
8. مَين [maen] ‘affectionate’ مَينتوب [maentoob] ‘affection’, ‘love’, ليوني [lewane] ‘mad’, ليونتوب [lewantoob] ‘madness’;

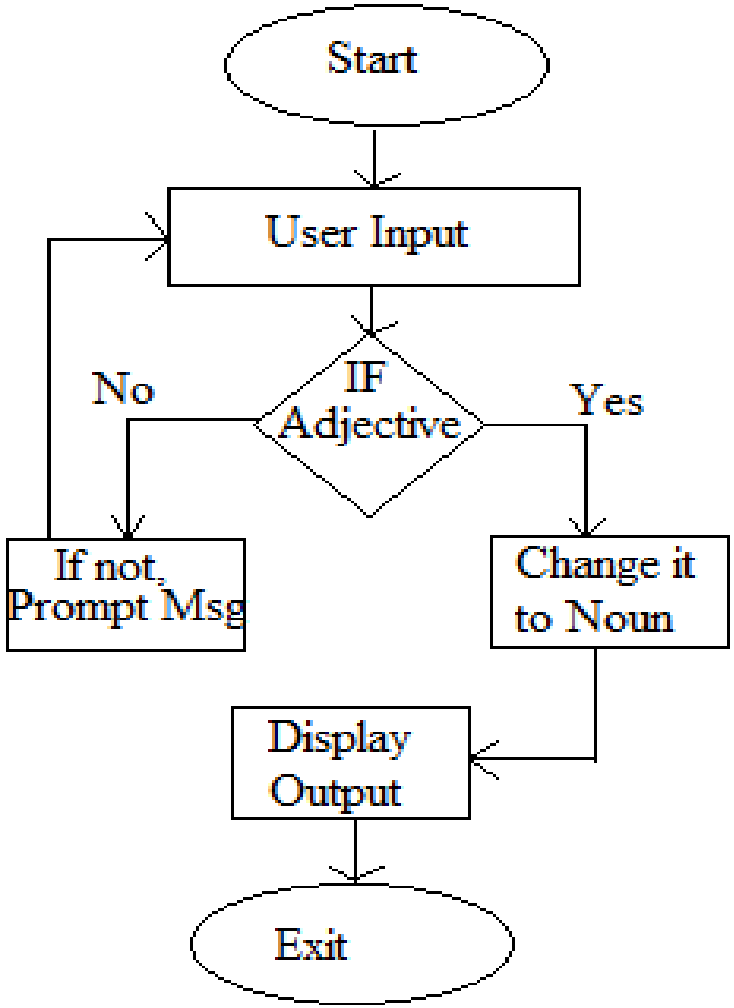


DAAD
Deutscher Akademischer Austausch Dienst
Servicio Alemán de Intercambio Académico

CLE
Unlocking Information for Human Development

Design and Flowchart of System





DAAD
Deutscher Akademischer Austausch Dienst
Servicio Alemán de Intercambio Académico

CLE
Unlocking Information for Human Development

Implementation

- We used four (4) programming languages/tools.

1. We used XEROX LEXC compiler to draw finite state diagrams from the lexicon of adjectives and nouns in Pashto.
2. After finite state diagrams, we used XEROX XFST to compile the .txt files generated by LEXC.
3. MS SQL Server was used for corpus and
4. MS CSharp was used as a front end to fetch data from the database.

DAAD
Deutscher Akademischer Austausch Dienst
Servicio Alemán de Intercambio Académico

CLE
Unlocking Information for Human Development

Error Analysis

- 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%**

DAAD
Deutscher Akademischer Austausch Dienst
Servicio Alemán de Intercambio Académico

CLE
Unlocking Information for Human Development

Conclusion

- The collation feature for Pashto language is not available in older versions of MS SQL Server.
- It was made available in MS SQL Server 2012 to store Pashto text
- Derivation is sensitive because a slight change of ◌َ **[zabbar],** ◌ِ**[zeir]** and ◌ُ [**paish]** can be disastrous.

DAAD
Deutscher Akademischer Austausch Dienst
Servicio Alemán de Intercambio Académico

CLE
Unlocking Information for Human Development

Future Work

- 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.

DAAD
Deutscher Akademischer Austausch Dienst
Servicio Alemán de Intercambio Académico

CLE
Unlocking Information for Human Development

UNIVERSITY OF PESHAWAR

Questions please.