UZBEK AFFIX FINITE STATE MACHINE FOR STEMMING

Sharipov Maksud
Salaev Ulugbek
Yuldashov Ollabergan
Sobirov Jasur

Urgench State University

OUTLINE

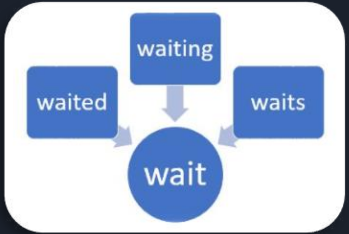
- Proposal
- Introduction
- Analysing
- Design
- Results
- Summary
- References

PROPOSAL

- Develop a morphological analysis tool for Uzbek stemming:
    - Morphotactic rules will be considered to achieve right stem
    - Finite State Machines (FSMs) will be designed
    - Avoid including lexicon
    - Perform morphologic analysis of a words from a large amount of text in high speed

Introduction

- In linguistic morphology and information retrieval, stemming is the process of reducing inflected or derived words to their word stem, base or root form
- *Stemming* is an algorithm that extract the morphological root of a word
- A **computer program** may be called a stemming program, stemming algorithm, or stemmer.

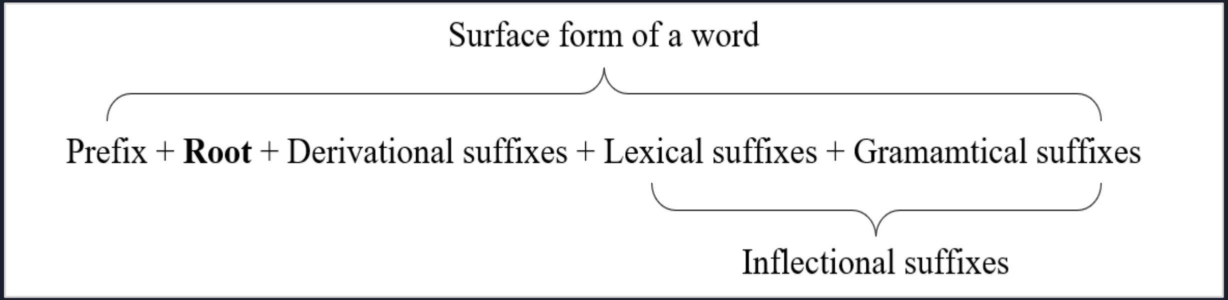


Source: http://mlwiki.org/index.php/Stemming

Introduction

Uzbek language spoken in general in Uzbekistan (and some other places in Central Asia). It is left to right written language. Uzbek language is an agglutinative, it is a language whose words are generated by adding affixes to the root forms. In such languages, given a word in its root form we can drive a new word by adding affix, and so on. Thus in many cases, a single Uzbek word may correspond to a many-word sentence or phrase in a non-agglutinative language. The language has a rich morphological structure.

Word form structure



Surface form of a word

Prefix + **Root** + Derivational suffixes + Lexical suffixes + Gramamtical suffixes

Inflectional suffixes

Source: Azim Hojiyev, O’zbek tili morfologiyasi, morfemikasi va so’z yasalishining nazariy masalalari. Toshkent-2010

ANALYSING

- In Uzbek, there are also suffixes that are written in the same form, but have different meanings. Herewith, there are suffixes that are formed by adding two or more suffixes side by side. For example, there are suffixes –*chi*, –*lik*, but there is also a monolithic form suffix –*chilik* according to the rule of separation of morphemes. When analyzing of the suffix –*chilik* can be a combination of 2 suffixes side by side ([–*chi*] +[–*lik*]) or mono form suffix (–*chilik*) depending to the content.
- *gul*+[–*chi*]+[–*lik*] gulchi (*eng* flower), gulchilik (*eng* floriculture)
- *dehqonchilik* (~~*dehqonchi*~~, it is incorrect word)

- There is also a suffix that is contain three suffixes side by side:

  [–*gar*]+[–*chi*]+[–*lik*], *odamgarchilik (eng* humanity*)*

ANALYSING (cont.)

- In the affix stemmer, while the lexicon is not used, it is planned to looking for a longer suffix (an undivided suffix) instead of one of the adjacent suffixes.

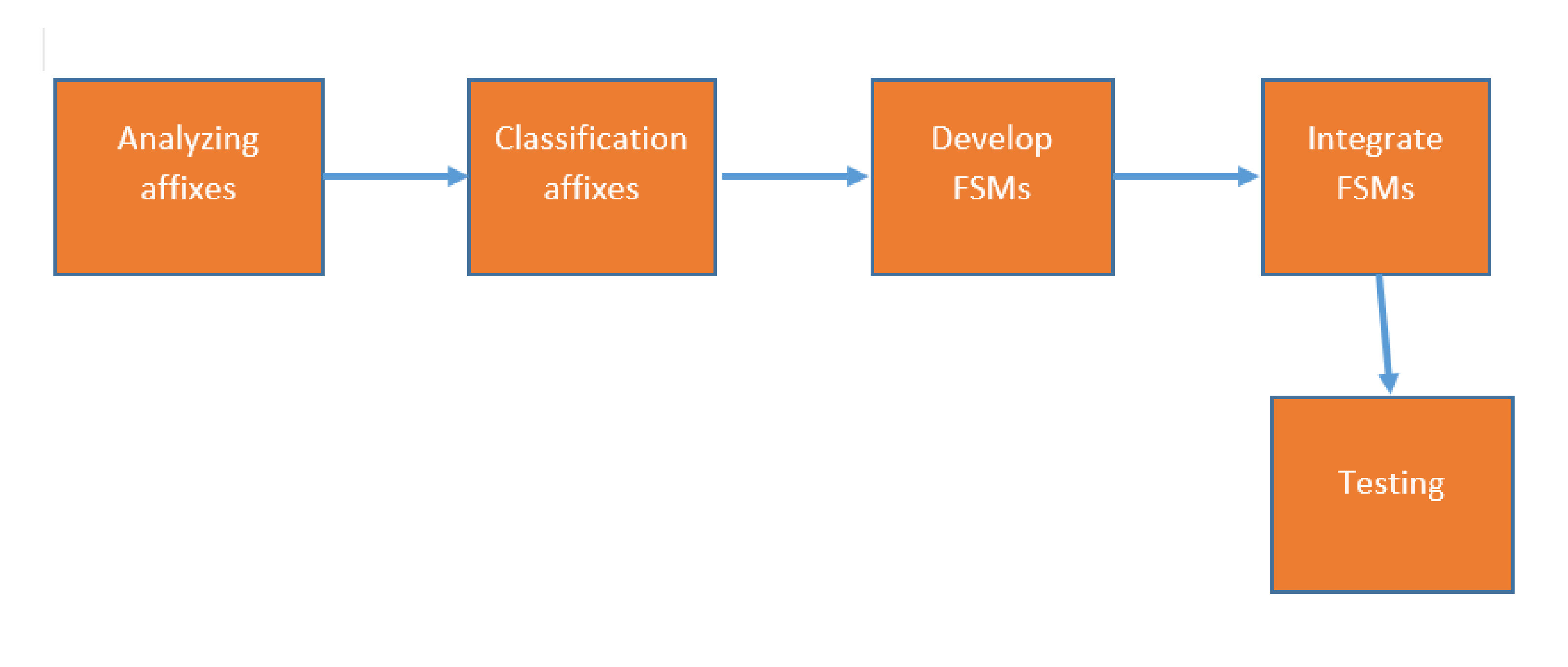
- In Uzbek, inflectional suffixes generally come after derivational suffixes. Some exceptional derivational suffixes that come after inflectional suffixes are: like o’ch/ir/gich (–gich - derivational suff., –ir - inflectional suff.).

- The suffix “–lar” besides of plural has a another meaning which is indicating his/her greeting to pointed person. Indeed, when the suffix used in a word, it formed by a different morphological structure:

dada/m/lar	root(dada) + Possessive (m) + Greeting (lar)	my father
olma/lar/im	root(olma) + Plural (lar) + Possessive(im)	my apples

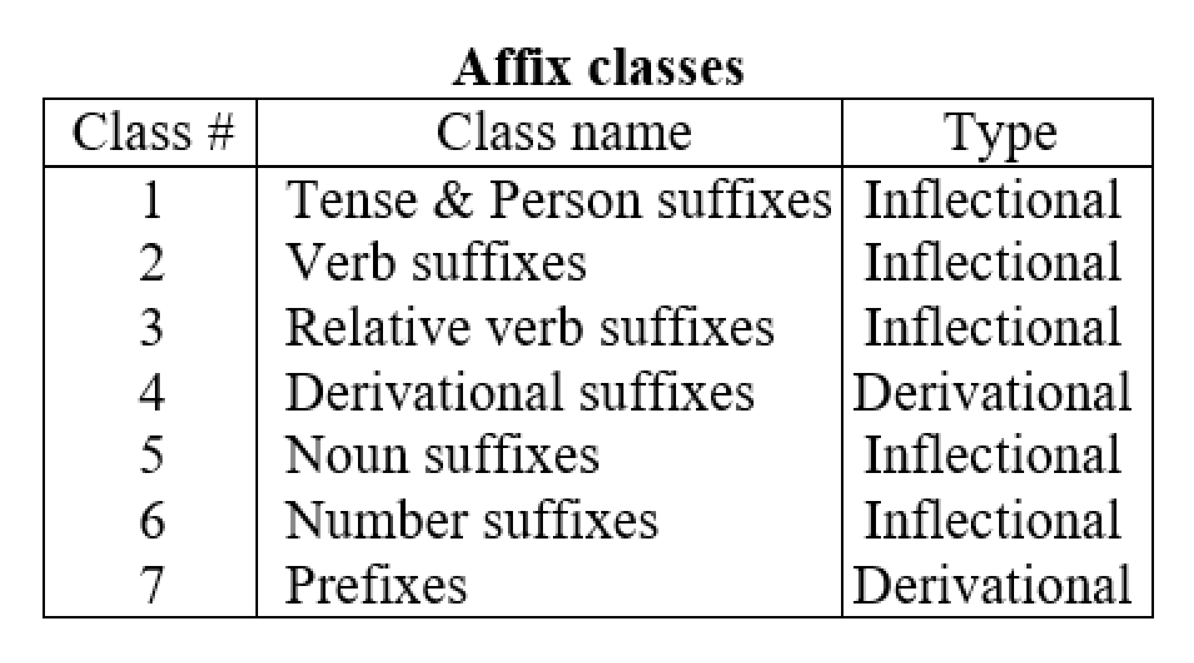
DESIGN

System design



METHODOLOGY

All affixes are classified into classes through their role and coming order.



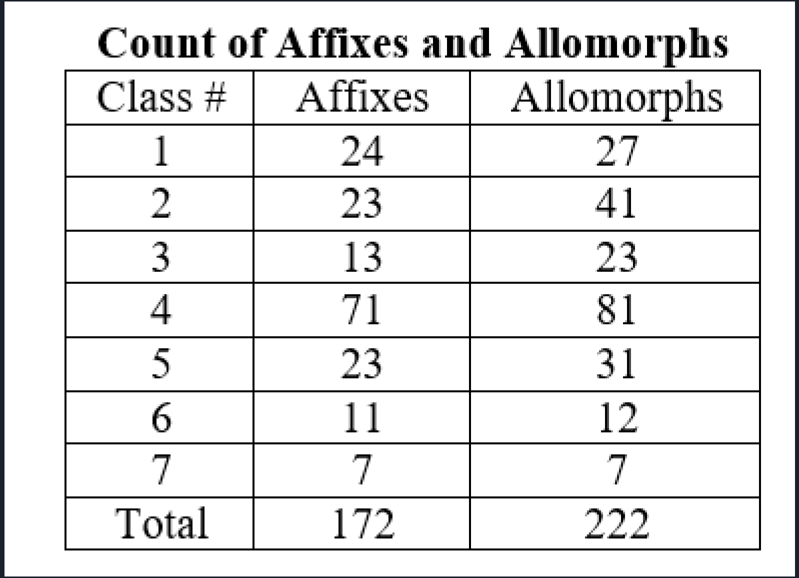
Affix classes

Class #	Class name	Type
1	Tense & Person suffixes	Inflectional
2	Verb suffixes	Inflectional
3	Relative verb suffixes	Inflectional
4	Derivational suffixes	Derivational
5	Noun suffixes	Inflectional
6	Number suffixes	Inflectional
7	Prefixes	Derivational

METHODOLOGY

An affix in Uzbek can have multiple allomorphs in order to provide sound harmony (as the phonological rules) in the word to which it is affixed.

- **G: g, k, q Y: a, y K: k, g**
  Q: k, g, g’, q T: t, d
- (): the letter between parentheses can be omitted



Count of Affixes and Allomorphs

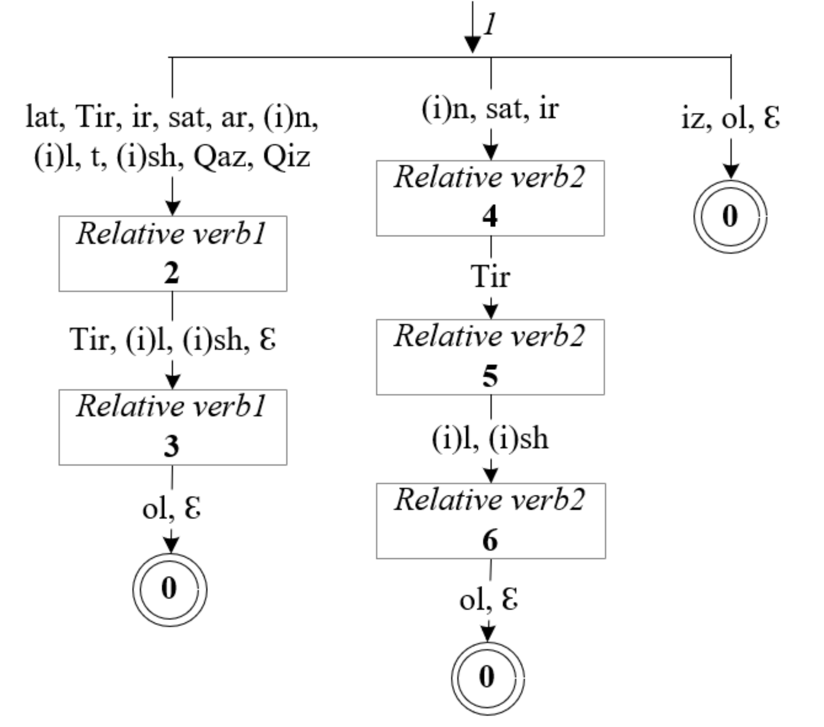
Class #	Affixes	Allomorphs
1	24	27
2	23	41
3	13	23
4	71	81
5	23	31
6	11	12
7	7	7
Total	172	222

The number of affixes of all derivational types is 77 and the number of with allomorphs is 87.
The number of all affixes of inflectional types is 95 and the number of with allomorphs is 135.

DESIGN

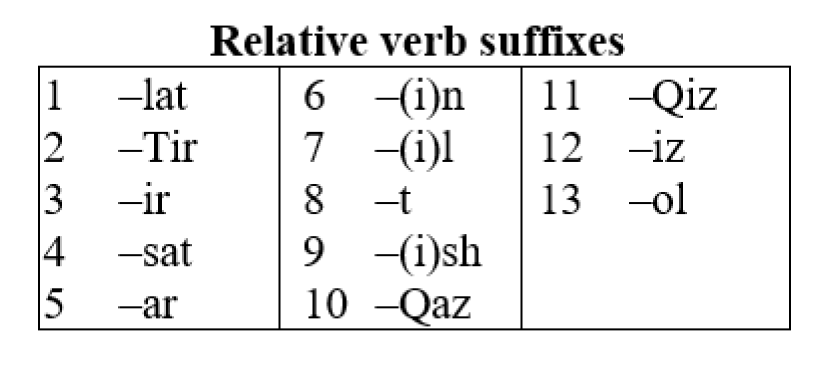
- Identyfing all affixes and perform morphologic analysis of a words from;
- Classified affixes into classes;
- Designed FSMs for each class which make a reversed order (right to left) analysis of a word. Performed following stages turn by turn to create FSMs:
    - Designing a left to right FSM;
    - Identification the affixes;
    - Inverting the left to right FSM and obtaining a non-deterministic finite state automaton (NFA);
    - Converting NFA to a deterministic finite automaton (DFA) and constructing the right to left FSM;

DESIGN



Relative verb suffixes left to right FSM

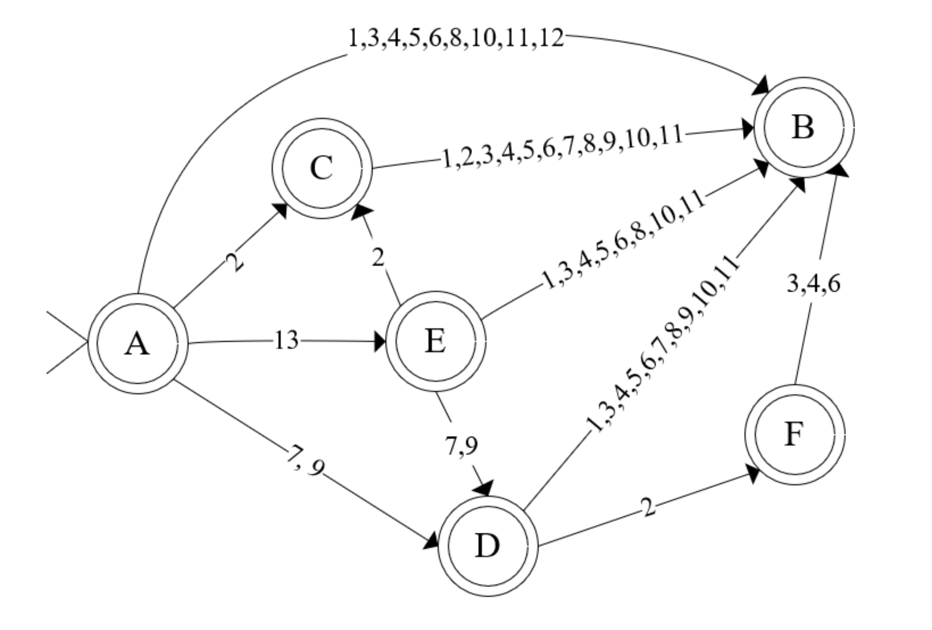
DESIGN



Relative verb suffixes

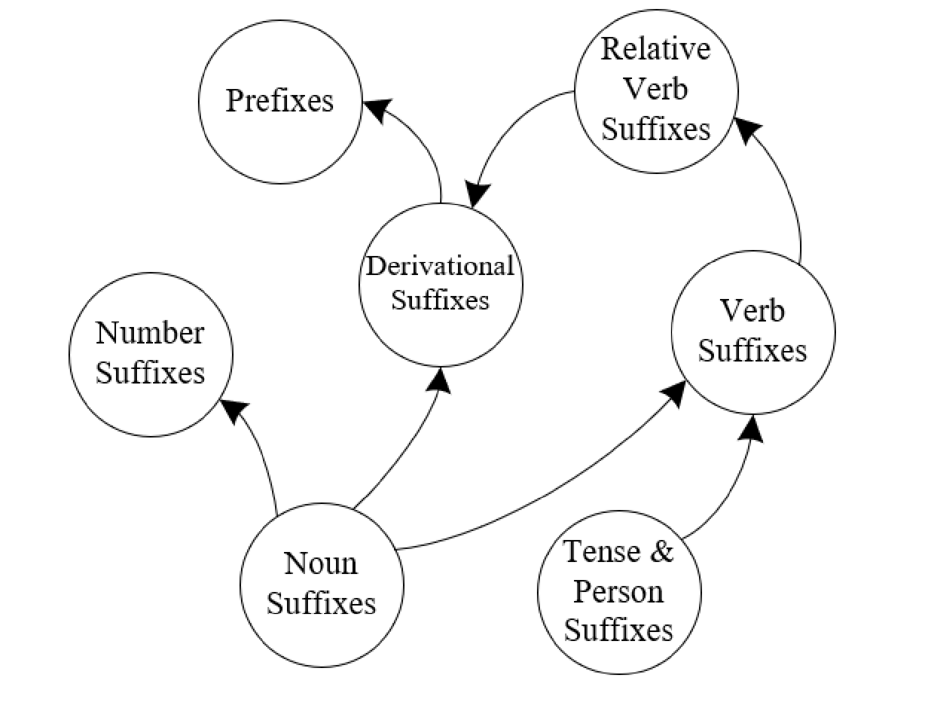
1	–lat	6	–(i)n	11	–Qiz
2	–Tir	7	–(i)l	12	–iz
3	–ir	8	–t	13	–ol
4	–sat	9	–(i)sh		
5	–ar	10	–Qaz		

Relative verb suffixes right to left FSM



DESIGN

Relationship of the FSMs



RESULT

The output taken after the analysis of the word "*bajartirilmayaptimi*" ("Is it not being performed?") by the main FSM.

Word, affix	Definition	Affix class
bajar	stem, verb	
–tir	relative verb suffix	Relative verb suffix
–il	relative verb suffix	Relative verb suffix
–ma	negative verb suffix	Verb suffix
–yap	continuous tense suffix	Tense & Person suffix
–ti	3nd single person suffix	Tense & Person suffix
–mi	question suffix	Tense & Person suffix

SUMMARY

- Examined all possibly affixes in Uzbek language and analyzed their morphotactic rules
- Classified the affixes into class respectively their role
- Created the FSMs for each class and main combining FSM
- Aviability moderating structure by affix addition and list exceptions in xml file
- Created the complete software for Uzbek stemming according the algorithm

REFERENCES

- Gülşen Eryiğit & Eşref Adalı, An affix striping morphological analyzer for Turkish, *Proceedings of the IASTED International Conference Artificial Intelligence and Applications*, Innsbruck, Austria, 2004, 299-304
- Gayrat Matlatipov and Zygmunt Vetulani. Representation of Uzbek Morphology in Prolog, Aspects of Natural Language Processing, 2009
- Ismailov A, Abdul Jalil M M, Abdullah Z and Abd Rahim N H 2016 A comparative study of stemming algorithms for use with the Uzbek language Computer and Information Sciences (ICCOINS) (Kuala Lumpur: IEEE) pp 7-12
- I I Bakaev and R I Bakaeva, Creation of a morphological analyzer based on finite-state techniques for the Uzbek language 2021 J. Phys.: Conf. Ser. 1791 012068
- I I Bakaev, T R Shafiev. Morphemic analysis of Uzbek nouns with Finite State Techniques. Journal of Physics: Conference Series. 1546 (2020) 012076 doi:10.1088/1742-6596/1546/1/012076
- Atadjanov J.A. Models of Morphological Analysis of Uzbek Words // Cybernetics and programming. — 2016. - № 6. - C.70-73. DOI: 10.7256/2306-4196.2016.6.20945.
- Abdurahim Mahmoud, Abdusalam Dawut, Peride Tursun and Askar Hamdulla. A Survey on the Methods for Uyghur Stemming. International Journal of Control and Automation. Vol. 9, No. 11 (2016), pp.143-158.
- http//dx.doi.org/10.14257/ijca.2016.9.11.13
- Mijit Ablimit, Tatsuya Kawahara, Akbar Pattar and Askar Hamdulla. Stem-Affix based Uyghur Morphological Analyzer. International Journal of Future Generation Communication and Networking. Vol. 9, No. 2 (2016), pp. 59-72. http://dx.doi.org/10.14257/ijfgcn.2016.9.2.07

Thank you!