DEVELOPING A SIMPLIFIED MORPHOLOGICAL ANALYZER FOR ARABIC PRONOMINAL SYSTEM

Mohammad Mahyoob

Assistant Professor, Department of Languages and Translation,
Alula Campus, Taibah University, Madinah, KSA

ABSTRACT

This paper proposes an improved morphological analyser for Arabic pronominal system using finite state method. The main advantage of the finite state method is very flexible, powerful and efficient. The most important results about FSAs, relates the class of languages generated by finite state automaton to certain closure properties. This result makes the theory of finite-state automata a very versatile and descriptive framework. The main contribution of this work is the full analysis and the representation of morphological analysis of all the inflections of pronoun forms in Arabic. In this paper we build a finite state network for the inflectional forms of the root words, restricted to all the inflections and grammatical properties of generating the dependent and independent forms of pronouns in Arabic language. The results show high score of accuracy in the output with all the needed linguistic features and the evaluation process of output is conducted using f-score test and the achievement is at the rate of 80% to 83%. The results from the study also provide the evidence that Arabic has strong concatenative word formations.

KEYWORDS

Computational Morphology, Arabic pronominal system, Arabic Morphological Analyzer, finite state automaton

1. INTRODUCTION

Like other language, Arabic language has a regular and irregular pattern in word formations, but the forms of regularity are more than irregularity. Studying the regular forms in Arabic nouns and verbs proved that all the inflections of Arabic subject-verb agreement are regular. And the study of Arabic dual formation of nouns are all regular as well as most of plural masculine and feminine formation [1]. There are some limited irregular forms for the plural feminine and masculine and that is called broken plural in Arabic. The study provides instances for Arabic independent pronouns formation. Unlike English, Arabic has a very rich inflection system. There are different inflections for gender, person, number and case even if there is an abstract representation for the first person singular perfective verbs in Arabic, but it is understood that there is a covert pronoun reference. However, for each other inflection, there are distinctive affixes which are added to the noun or verbs. There has been much work on the Morphological generation and analysis for the lexical forms of languages. The handling of morphological rich inflectional languages is still problematic for natural language processing [2].

For many reasons; Arabic language forms a real challenge for NLP experts, as the absence of diacritic marks is the most prominent challenge, the second one is the consonantal root and infixing [3]. However, Arabic has more regular dual and plural forms than irregular [4] . Arabic has a concatenative morphology in all the aspects of forming dual nouns and most of plural nouns. The irregularity of Arabic is considered for some plural forms which is called broken plural [5]. Like English, Arabic has some irregular plural forms.

DOI: 10.5121/ijnlc.2020.9202

There are some characteristics for inflectional morphology which can be represented by the following [6]:

a. Systematic
Adding an affix to a stem has the same grammatical or semantic effect for all stems

b. Productive
New addition to the language automatically conform to the rules for affixation

c. Preservative

The broad grammatical category of the word is not altered in the inflectional process
The current study contributes to our knowledge by addressing the concatenative forms of Arabic pronominal system. This paper also studies the processing of all pronoun forms in Arabic language using finite state machine.

2. RELATED WORKS

Several studies on standard Arabic morphological analysers have been done and published. These studies dealt with nouns, verbs and adjectives. To our knowledge, no real work has been done specifically on Arabic inflected forms of pronouns from a computationally morphological perspective.

The first manual works for Arabic morphological analysers were done by Buckwalter [7]; Graff et al. [8]; Habash and Rambow, [9]. Next, a morphological analyser was developed for Egyptian Arabic using annotated data by Habash et al. [10]. A lot of development for the previous work using hand crafted rules. Then a new morphological analyser was developed based on an annotated corpus by Eskander et al. [11]. They provided the morpho-syntactic values for the lexemes along with inflected stem. A morphological analyser for standard Arabic nouns and verbs using finite state machines were developed by Mahyoob and Algaraady, [12] [13]. A new version appeared with new technique for a Moroccan Arabic and a Sanaani Yemeni Arabic morphological analyser by (Al-Shargi et al., [14] Recently Abuata and Al-Omari [15]. developed a rule-based morphological analyser to segment affixes and clitics in GLF text. Most of these versions have their advantages and disadvantages, we cannot name one of these as a comprehensive Arabic morphological analyser. From a theoretical perspective; Salem's work discussed the distribution of reduced and unreduced pronouns in standard Arabic using generative linguistic framework proposed by Halle and Marantz 1993. The study investigated a theoretical analysis of phonological relationship between reduced and unreduced pronouns [16]. Another study was done by Alqarni [17]; he discussed the pronominal system in Arabic on the light of Halle and Marantz's distributed morphology framework as well. He investigated the gender, number, person and case distribution in Standard Arabic. Real and deep work is needed to cover all the Arabic inflections forms for all the lexemes in the language. To our knowledge, this study can be considered as the first morphological analyser specialized for Arabic pronominal system using deterministic finite state automaton. Hahn's study investigates the analysis of pronominal arguments, conjuncts and bound pronouns. The study shows the violation of the null representation in the subject position for the pronominal argument and compares the null conjuncts and bound pronoun in standard Arabic [18].

3. PRONOUNS IN ARABIC

There are different numbers of pronouns in Arabic. The independent and dependent pronouns. The independent pronouns are not joined to any word and appear as a separate word in subject position and nominative case. Unlike English, Arabic has three types of number; singular, dual and plural as illustrated in table 1 below.

Table1. Subject Pronoun in Arabic

Person	Gender	Plurality	pronoun
First person	Masculine / Feminine	Singular	Ana (I)
	Masculine / Feminine	Dual/ plural	Nahnu (we)
Second person	Masculine	Singular	Anta(You)
		Dual	Antuma(you)
		Plural	Antum(you)
	Feminine	Singular	Anti(you)
		Dual	Antuma(you)
		Plural	Antuna(you)
Third person	Masculine	Singular	Hwa (he)
		Dual	Humaa(they)
		Plural	Hum(they)
	Feminine	Singular	Hiya(she)
		Dual	Humaa(they)
		Plural	Hunna(they)

There are two types of dependent pronouns in Arabic; possessive pronoun and objective pronoun. These pronouns can come after nouns, verbs and particles as shown in the table 2 below:

Table 2. Object Pronouns in Arabic

Person	Gender	Plurality	pronoun
First person	Masculine / Feminine	Singular	-i, -ni (my, me)
	Masculine / Feminine	Dual/ plural	-na (ours, us)
Second person	Masculine	Singular	-ka(your, you)
		Dual	-kuma(your, you)
		Plural	-kum (your, you)
	Feminine	Singular	-ki (your, you)
		Dual	-kuma (your, you)
		Plural	-kun (your, you)
Third person	Masculine	Singular	-hu (his, him)
		Dual	-huma(Their, them)
		Plural	-hum(Their, them)
	Feminine	Singular	-ha(hers, her)
		Dual	-humaa(their, them)
		Plural	-hunna(their, them)

4. FINITE STATE AUTOMATON (FSA)

On the first interpretation, an FSA can be seen as simply an oriented graph with labels on each arc. Finite State Automaton (FSA) a finite-state automaton A is a 5-tuple (Σ, Q, δ, q_0, F) where Σ is a finite set called the alphabet, Q is a finite set of states, i ∈ Q is the initial state, F ⊆ Q is the set of final states and E ⊆Q x (Σ U{∈}) x Q is the set of edges [19]

The automaton takes a finite sequence of 0s and 1s as input. For each state, there is a transition arrow leading to a next state for both 0 and 1. A DFA jumps deterministically from a state to another by following the transition arrow. For example, if the automaton is currently in state S0 and current input symbol is 1 then it deterministically jumps to state S1. A DFA has a start state (denoted graphically by an arrow coming in from nowhere) where computations begin, and a set of accept states (denoted graphically by a double circle) which helps define when a computation is successful [20].

A deterministic finite automaton is a 5-tuple, (Q, Σ, δ, q_0, F), consisting of

- (Q) a finite set of states
- (Σ) a finite set of input symbols called the alphabet
- δ a transition function ($\delta : Q \times \Sigma \rightarrow Q$)
- q_0 a start state ($q_0 \in Q$)
- F a set of accept states ($F \subseteq Q$)

The machine starts in the start state q_0, the machine will transit from state to state with the data according to the transition function δ. Finally, the machine accepts data if the last input of this data causes the machine to halt in one of the accepting states. Otherwise, it is said that the automaton rejects the string [20].

5. THE FRAMEWORK

The framework of this paper is as the following:

The input string, the lexicon, and the morphological rules; these are all represented in the finite machine. The rules are applied with corresponds to the application of finite state automaton. The identification of strings is processed in the lexicon and all the morphological information encoded in the lexicon is represented naturally in the finites state machines. Every word in the dictionary is represented in the automaton. The morphological analysis of the input word could also be represented in the finite state automaton. The framework of the study can be observed in the following figure:

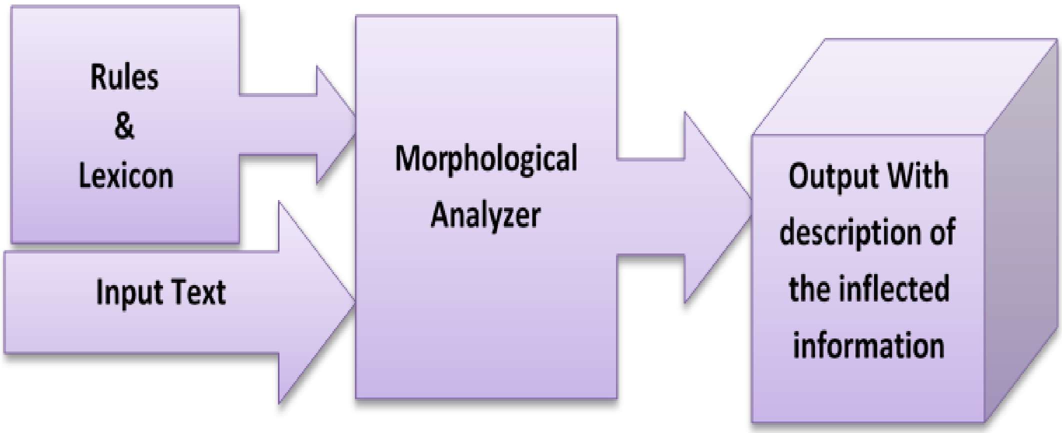


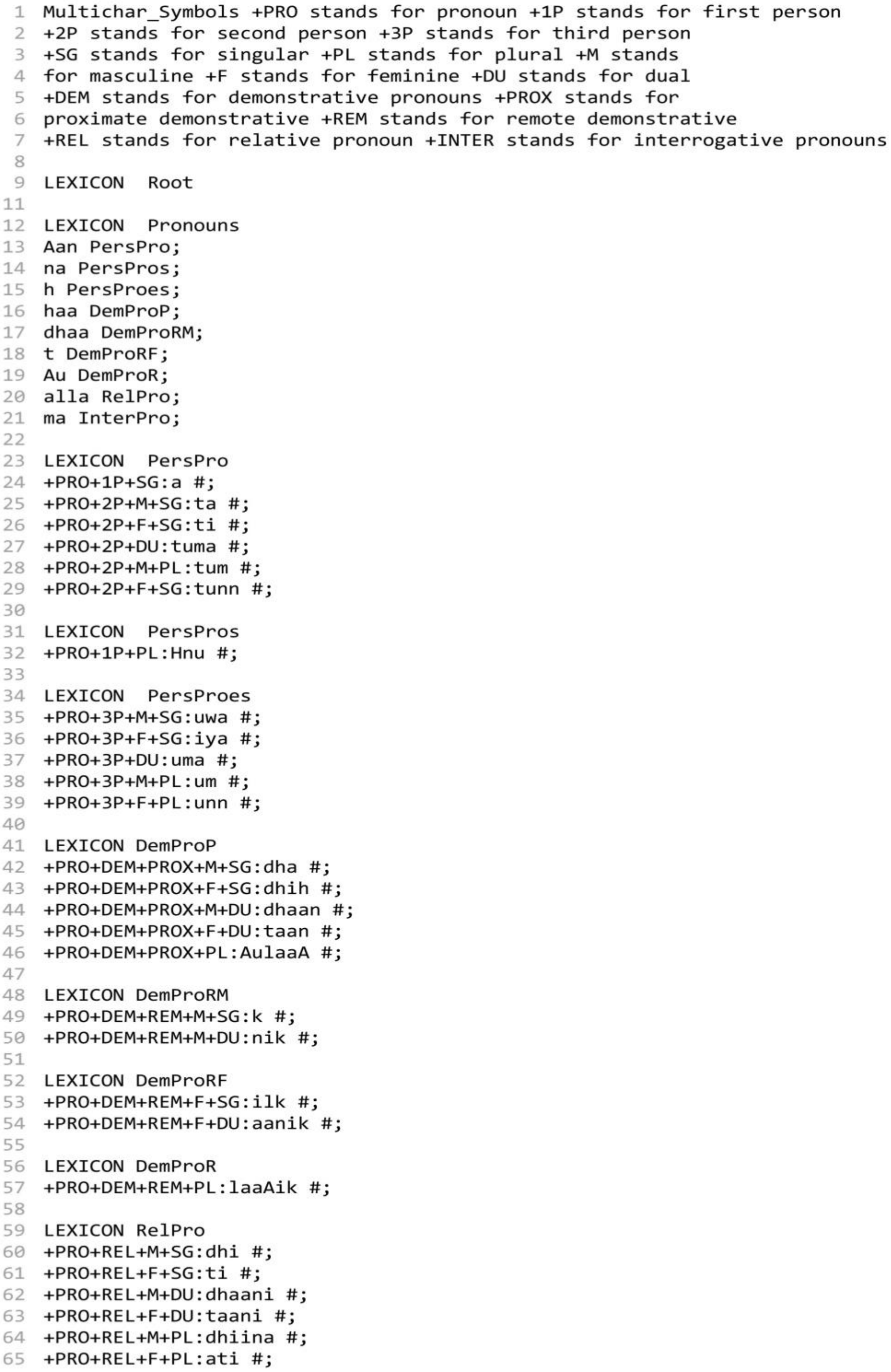
Figure 1. Model of the study

6. EXPERIMENT

To enable the tool to handle all the inflected forms of the word we developed a lexicon with all the possible inflection of the word. Most of the pronouns in Arabic are concatenation morphemes and that could make the high accuracy in our system output.

6.1. The lexicon of our work

Designing a computational lexicon for the morphological analyses as well as evaluating the output of this work presupposes a predefined morphological independent specification of the output morphological representation. The morphological representation is composed of a set of morphological description and a set of rules and principles. The development of this lexicon is done semi-automatically using finite state techniques which could be used to encode the morphological analysis with the application of these certain rules as shown in figure 2.



```
Multichar_Symbols +PRO stands for pronoun +1P stands for first person
+2P stands for second person +3P stands for third person
+SG stands for singular +PL stands for plural +M stands
for masculine +F stands for feminine +DU stands for dual
+DEM stands for demonstrative pronouns +PROX stands for
proximate demonstrative +REM stands for remote demonstrative
+REL stands for relative pronoun +INTER stands for interrogative pronouns

LEXICON  Root

LEXICON  Pronouns
Aan PersPro;
na PersPros;
h PersProes;
haa DemProP;
dhaa DemProRM;
t DemProRF;
Au DemProR;
alla RelPro;
ma InterPro;

LEXICON  PersPro
+PRO+1P+SG:a #;
+PRO+2P+M+SG:ta #;
+PRO+2P+F+SG:ti #;
+PRO+2P+DU:tuma #;
+PRO+2P+M+PL:tum #;
+PRO+2P+F+SG:tunn #;

LEXICON  PersPros
+PRO+1P+PL:Hnu #;

LEXICON  PersProes
+PRO+3P+M+SG:uwa #;
+PRO+3P+F+SG:iya #;
+PRO+3P+DU:uma #;
+PRO+3P+M+PL:um #;
+PRO+3P+F+PL:unn #;

LEXICON DemProP
+PRO+DEM+PROX+M+SG:dha #;
+PRO+DEM+PROX+F+SG:dhih #;
+PRO+DEM+PROX+M+DU:dhaan #;
+PRO+DEM+PROX+F+DU:taan #;
+PRO+DEM+PROX+PL:AulaaA #;

LEXICON DemProRM
+PRO+DEM+REM+M+SG:k #;
+PRO+DEM+REM+M+DU:nik #;

LEXICON DemProRF
+PRO+DEM+REM+F+SG:ilk #;
+PRO+DEM+REM+F+DU:aanik #;

LEXICON DemProR
+PRO+DEM+REM+PL:laaAik #;

LEXICON RelPro
+PRO+REL+M+SG:dhi #;
+PRO+REL+F+SG:ti #;
+PRO+REL+M+DU:dhaani #;
+PRO+REL+F+DU:taani #;
+PRO+REL+M+PL:dhiina #;
+PRO+REL+F+PL:ati #;
```

Figure 2. Sample of the developed lexicon

The above graph shows the development of the lexicon for this study. It contains the root, different inflection forms and the rules which lead to the correct output. These are some of the input strings which are processed by the automata generating the output strings which are discussed in the coming sections.

7. Results, Discussion and Analysis

The following figure indicates that by using finite state machine for processing the language data is a real representation of the analysis for the different forms of pronouns in Arabic. This is a significant improvement, because the achievable results of the data using FSM scores a very high regularity in Arabic independent pronouns. R represents the root and S represents the different gender, person and numbers inflections.

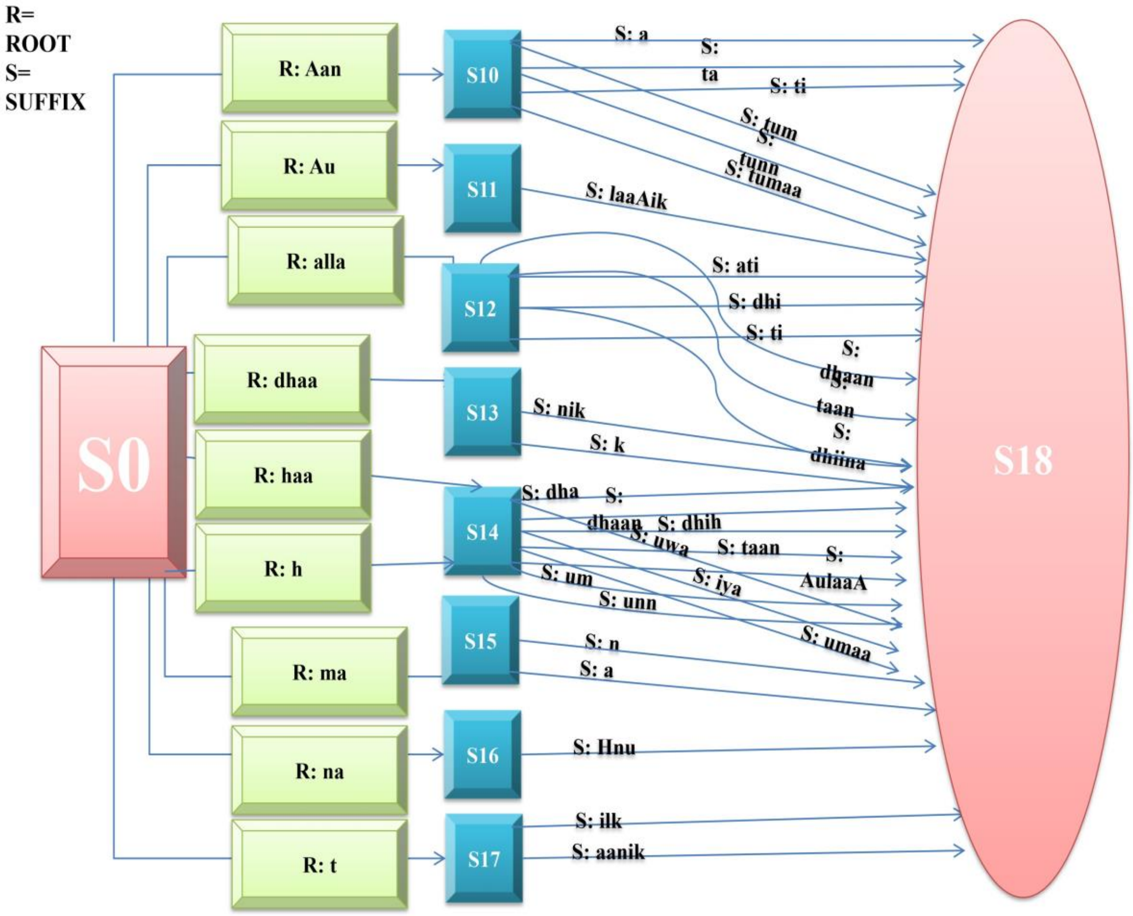


Figure 3. Representsation of the data in finite state machine

The central task of a morphological analyser is to assign a morphological description into input lexemes.

Figure 3 shows the following:

a. Nineteen states
b. Fifty-one transition (arcs)
The initial state is S0, the final state is S18.
Σ**:** {(Roots: Aan, Au, alla, dhaa, haa, h, ma, na,t;), (suffixes: a, ta, ti, tum, tunn, tumaa, laaAik, ati, dhi, ti, haan, taan, dhiina, nik, k,dha, dhaan, dhih, uwa, taan, iya, aulaaA, um, unn, umaa, n, a, hnu, ilk, aanik)}

7.1. Used Symbols

The symbols which are used in this study and represented in lexicon and FSM are as follow: +PRO stands for pronoun, +1P stands for first person, +2P stands for second person, +3P stands for third person, +SG stands for singular +PL stands for plural, +M stands for masculine, +F stands for feminine, +DU stands for dual, +DEM stands for demonstrative pronouns, +PROX stands for proximate demonstrative, +REM stands for remote demonstrative, +REL stands for relative pronoun, +INTER stands for interrogative pronouns. More details about the inflections used in the graph 1 are clearly indicatec in the lexicon.

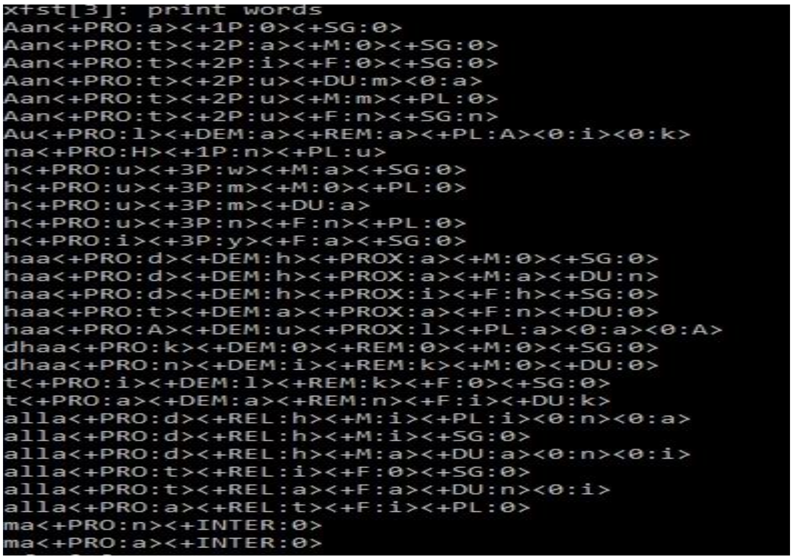
7.2. Transition function matrix

In table 3 what we observe is that the transition of each arc and the form it indicates.

Table 3. Transition function matrix

From	To	Output
0	1,2,3,4,5,6,7,8,9	pronoun (root variation)
1,2,3,4,5,6,7,8,9	10,11,12,13,14,15,16,17	Pronouns (root variation)
10,11,12,13, 14,15,16,17	18	*All the inflections of different types of pronouns*

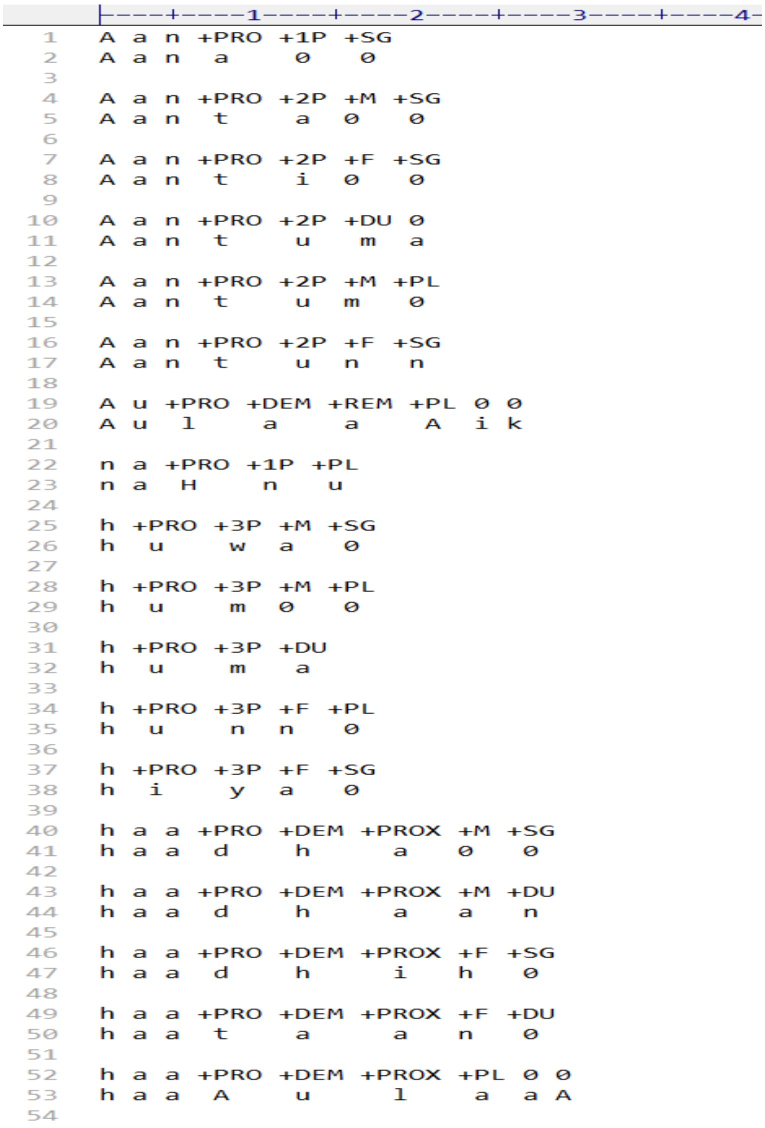
The graphs 1 illustrates the generated morphological information for each entry using finite state machine. All the forms of pronouns are presented in the output along with the linguistic information of each form. It presents the method of adding the features and the alteration from one form to another by its inflections such as singular, dual plural, masculine, feminine, etc. The graph presents the change of one pronoun type to another; for example, relative pronouns, demonstrative pronouns, etc.



```
xfst[3]: print words
Aan<+PRO:a><+1P:0><+SG:0>
Aan<+PRO:t><+2P:a><+M:0><+SG:0>
Aan<+PRO:t><+2P:i><+F:0><+SG:0>
Aan<+PRO:t><+2P:u><+DU:m><0:a>
Aan<+PRO:t><+2P:u><+M:m><+PL:0>
Aan<+PRO:t><+2P:u><+F:n><+SG:n>
Au<+PRO:l><+DEM:a><+REM:a><+PL:A><0:i><0:k>
na<+PRO:H><+1P:n><+PL:u>
h<+PRO:u><+3P:w><+M:a><+SG:0>
h<+PRO:u><+3P:m><+M:0><+PL:0>
h<+PRO:u><+3P:m><+DU:a>
h<+PRO:u><+3P:n><+F:n><+PL:0>
h<+PRO:i><+3P:y><+F:a><+SG:0>
haa<+PRO:d><+DEM:h><+PROX:a><+M:0><+SG:0>
haa<+PRO:d><+DEM:h><+PROX:a><+M:a><+DU:n>
haa<+PRO:d><+DEM:h><+PROX:i><+F:h><+SG:0>
haa<+PRO:t><+DEM:a><+PROX:a><+F:n><+DU:0>
haa<+PRO:A><+DEM:u><+PROX:l><+PL:a><0:a><0:A>
dhaa<+PRO:k><+DEM:0><+REM:0><+M:0><+SG:0>
dhaa<+PRO:n><+DEM:i><+REM:k><+M:0><+DU:0>
t<+PRO:i><+DEM:l><+REM:k><+F:0><+SG:0>
t<+PRO:a><+DEM:a><+REM:n><+F:i><+DU:k>
alla<+PRO:d><+REL:h><+M:i><+PL:i><0:n><0:a>
alla<+PRO:d><+REL:h><+M:i><+SG:0>
alla<+PRO:d><+REL:h><+M:a><+DU:a><0:n><0:i>
alla<+PRO:t><+REL:i><+F:0><+SG:0>
alla<+PRO:t><+REL:a><+F:a><+DU:n><0:i>
alla<+PRO:a><+REL:t><+F:i><+PL:0>
ma<+PRO:n><+INTER:0>
ma<+PRO:a><+INTER:0>
```

Graph 1. Generated forms with morphological information

The output wordlist with the different forms and all the morphological information is shown in the following figure. The network of finite state shows the output with different inflections of the pronoun’s forms such as first person singular, second person dual, etc. gender is represented as masculine and feminine. Most of the linguistics features are added in the lexicon and the tool is processing and generating the different forms with all these features. The output of this process is observed and appeared in figure 4 below. Two lines for each word, the first is showing the morphological analysis and the second provides the form in the language.



```
A a n +PRO +1P +SG
A a n  a    0   0

A a n +PRO +2P +M +SG
A a n  t    a  0   0

A a n +PRO +2P +F +SG
A a n  t    i  0   0

A a n +PRO +2P +DU 0
A a n  t    u   m  a

A a n +PRO +2P +M +PL
A a n  t    u  m   0

A a n +PRO +2P +F +SG
A a n  t    u  n   n

A u +PRO +DEM +REM +PL 0 0
A u  l    a    a    A  i k

n a +PRO +1P +PL
n a  H    n   u

h +PRO +3P +M +SG
h  u    w  a   0

h +PRO +3P +M +PL
h  u    m  0   0

h +PRO +3P +DU
h  u    m   a

h +PRO +3P +F +PL
h  u    n  n   0

h +PRO +3P +F +SG
h  i    y  a   0

h a a +PRO +DEM +PROX +M +SG
h a a  d    h     a   0   0

h a a +PRO +DEM +PROX +M +DU
h a a  d    h     a   a   n

h a a +PRO +DEM +PROX +F +SG
h a a  d    h     i   h   0

h a a +PRO +DEM +PROX +F +DU
h a a  t    a     a   n   0

h a a +PRO +DEM +PROX +PL 0 0
h a a  A    u     l    a  a A
```

Figure 4. Output sample wordlist

The quantitative information is extracted with regard to various linguistic elements. It is extremely useful not only in building and formulating rules but also testing them. The obtained results of this study are represented in the following table.

Table 4. obtained results

Data	1400	700
Correct output	98.00	99.00
Generated forms	30	30
F-score	80.43%	83.17%

What we find in the above table is the percentage of F-score 80.43%. The concatenative process of pronoun inflected forms indicated that Arabic has a regular inflected system in Nouns, Verbs and Pronouns as well.

8. CONCLUSION

The present study was designed to develop and evaluate a morphological analyser for Arabic pronouns forms. Studies on morphological parsing with the focus on a particular linguistic structure such pronouns in Arabic can lead to better understanding of the problem involved. The results indicated that the output is effective and evaluated with a high degree of success. Most of the needed information for the word analysis was presented in wordlist of the tool used in this study. This study adds all the necessary information and values to the lexical concatenating pronoun according to their patterns.

9. LIMITATION AND FUTURE WORK

Despite the developed Morphological Analyser for Arabic pronominal System, there is a need for a comprehensive investigation of all inflections and derivations related to pronouns in all cases of nominative, accusative, genitive, etc. This work has open up several questions that need of further investigation to establish whether the verb, noun and adjectives in Arabic can be regularly derived and inflected. Further experimental investigations are needed to estimate the level of the concatenative and non-concatenative formation in Arabic lexical categories.

REFERENCES

[1] Cavalli-Sforza, V., Soudi, A. and Mitamura, T., 2000, April. Arabic morphology generation using a concatenative strategy. In Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference (pp. 86-93). Association for Computational Linguistics.

[2] Beesley, K. R. 1998. 'Arabic morphology using only finite-state operations'. Proceedings of the Workshop on Computational Approaches to Semitic Languages, Montreal, Quebec.

[3] McCarthy, J. (1991). 'A Prosodic theory of nonconcatenative Morphology' Linguistic Inquiry,12 (3): 373-418.

[4] Haywood, J.A. and Nahmad, H.M.1965. A New Arabic Grammar of the written Language.London: Lund.

[5] Qassem, S., & Mahyoob, M. (2015). Semi-Automatic annotation of Arabic Corpus: a Morphosyntactic study (Doctoral dissertation, Aligarh Muslim University).

[6] Ritchie, G. D., Russell, G. J., Black, A. W., & Pulman, S. G. (1992). Computational morphology: practical mechanisms for the English lexicon. MIT press.

[7] Tim Buckwalter. 2004. Buckwalter Arabic Morphological Analyzer Version 2.0. LDC catalog number LDC2004L02, ISBN 1-58563-324-0.

[8] David Graff, Mohamed Maamouri, Basma Bouziri, Sondos Krouna, Seth Kulick, and Tim Buckwalter. 2009. Standard Arabic Morphological Analyzer (SAMA) Version 3.1. Linguistic Data Consortium LDC2009E73.

[9] Nizar Habash and Owen Rambow. 2006. MAGEAD: A Morphological Analyzer and Generator for the Arabic Dialects. In Proceedings of ACL, pages 681– 688, Sydney, Australia.

[10] Habash, R. Eskander, and A. Hawwari. 2012a. A Morphological Analyzer for Egyptian Arabic. In NAACL-HLT 2012 Workshop on Computational Morphology and Phonology (SIGMORPHON2012), pages 1–9

[11] Ramy Eskander, Nizar Habash, and Owen Rambow. 2013. Automatic Extraction of Morphological Lexicons from Morphologically Annotated Corpora. In Proceedings of tenth Conference on Empirical Methods in Natural Language Processing.

[12] M. Mahyoob(2018). Deterministic Finite State Automaton of Arabic Verb System: A Morphological Study, international journal of Computational Linguistics, Computer Science Journals, Malaysia.

[13] Mohammad Mahyoob & Jeehaan Algaraady. 2018. Towards Developing a Morphological Analyser for Arabic Noun Forms. International journal of Linguistics and computational Applications.

[14] Faisal Al-Shargi, Aidan Kaplan, Ramy Eskander, Nizar Habash, and Owen Rambow. 2016. A Morphologically Annotated Corpus and a Morphological Analyzer for Moroccan and Sanaani Yemeni Arabic. In Proceedings of the International Conference on Language Resources and Evaluation (LREC), Portoroz, Slovenia

[15] Belal Abuata and Asma Al-Omari. 2015. A rule-based stemmer for Arabic Gulf Dialect. Journal of King Saud University - Computer and Information Sciences, 27(2):104 – 112.

[16] Albuhayri, S. 2013. The Pronominal system in Standard Arabic: strong, clitic and affixal pronouns (MA thesis). Arizona State University.

[17] M Alqarni, 2020 Pronominal System in Standard Arabic: A Distributed Morphology Analysis International Journal of Arabic-English Studies,

[18] Hahn, M., 2011. Null conjuncts and bound pronouns in Arabic. In Proceedings of the 18th international conference on Head-Driven Phrase Structure Grammar, university of washington (pp. 60-80).

[19] Roche, E., & Schabes, Y. (Eds.). (1997). Finite-state language processing. MIT press.

[20] Beesley, K. R. and L. Karttunen 2003. Finite State Morphology. Stanford, Calif., Csli.