Acquisition System for Arabic Noun Morphology

Saleem Abuleil   Khalid Alsamara   Martha Evens

Information System Department   Computer Science Department
Chicago State University   Illinois Institute of Technology
9501 S. King Drive, Chicago, IL 60628   10 West 31 Street, Chicago IL 60616
s_abuleil@hotmail.com kalsamara@hotmail.com   evens@iit.edu

Abstract

Many papers have discussed different aspects of Arabic verb morphology. Some of them used patterns; others used patterns and affixes. But very few have discussed Arabic noun morphology particularly for nouns that are not derived from verbs. In this paper we describe a learning system that can analyze Arabic nouns to produce their morphological information and their paradigms with respect to both gender and number using a rule base that uses suffix analysis as well as pattern analysis. The system utilizes user-feedback to classify the noun and identify the group that it belongs to.

1   Introduction

A morphology system is the backbone of a natural language processing system. No application in this field can survive without a good morphology system to support it. The Arabic language has its own features that are not found in other languages. That is why many researchers have worked in this area. Al-Fedaghi and Al-Anzi (1989) present an algorithm to generate the root and the pattern of a given Arabic word. The main concept in the algorithm is to locate the position of the root's letters in the pattern and examine the letters in the same position in a given word to see whether the tri graph forms a valid Arabic root or not.

Al-Shalabi (1998) developed a system that removes the longest possible prefix from the word where the three letters of the root must lie somewhere in the first four or five characters of the remainder. Then he generates some combinations and checks each one of them with all the roots in the file. Al-Shalabi reduced the processing, but he discussed this from point of view of verbs not nouns. Anne Roeck and Waleed Al-Fares (2000) developed a clustering algorithm for Arabic words sharing the same verbal root. They used root-based clusters to substitute for dictionaries in indexing for information retrieval. Beesley and Karttunen (2000) described a new technique for constructing finite-state transducers that involves reapplying a regular-expression compiler to its own output. They implemented the system in an algorithm called compile-replace. This technique has proved useful for handling non-concatenate phenomena, and they demonstrate it on Malay full-stem reduplication and Arabic stem inter-digitations.

Most verbs in the Arabic language follow clear rules that define their morphology and generate their paradigms. Those nouns that are not derived from roots do not seem to follow a similar set of well-defined rules. Instead there are groups showing family resemblances.

We believe that nouns in Arabic that are not derived from roots are governed not only by phonological rules but by lexical patterns that must be identified and stored for each noun. Like irregular verbs in English their forms are determined by history and etymology, not just phonology. Among many other examples, Pinker (1999) points to the survival of past forms became for become and overcame for overcome, modeled on came for come, while succumb, with the same sound pattern, has a regular past form succumbed. The same kinds of phenomena are especially apparent for proper nouns in Arabic derived from Indian and Persian names. Pinker uses examples like this, as well as emerging research in neurophysiology, to argue for the coexistence of phonological rules and lexical storage of English verb patterns.

We believe that further work in Arabic computational linguistics requires the development of a pattern bank for nouns. This paper describes the tool that we have built for this purpose. While the set of patterns for common nouns in Arabic may soon be established, newspapers and other dynamic sources of language will always contain new proper names, so we expect our tool to be a permanent part of our system, even though we may need it less often as time goes on.

2 Nouns in the Arabic Language

A noun in Arabic is a word that indicates a meaning by itself without being connected with the notion of time. There are two main kinds of noun: variable and invariable. Variable nouns have different forms for the singular, the dual, the plural, the diminutive, and the relative. Variable nouns are again divided into two kinds: inert and derived. The inert noun is not derived from another word, i.e. it does not refer to a verbal root. Inert nouns are divided into two kinds: concrete nouns (e.g., lion), and abstract nouns (e.g., love). Derived nouns are taken from another word (usually a verb) (e.g. office); they have a root to refer to. A derived noun is usually close to its root in meaning. It indicates, besides the meaning, the concrete thing that caused its formation (case of the agent-noun), or underwent its action (case of the patient-noun), or any other notions of time, place, or instrument. The following are the noun types:

A **genus noun** indicates what is common to every element of the genus without being specific to any one of them. It is the word naming a person, an animal, a thing or an idea. Example: رجل man   كتاب book

An **agent noun** is a derived noun indicating the actor of the verb or its behavior. It has several patterns according to its root. Example:

دارس   the person who studies

A **patient noun** is a derived noun indicating the person or thing that undergoes the action of the verb. Patient nouns have several patterns depending in the verbal root. Example:

مدروس   the thing that has been studied

An **instrument noun** is a noun indicating the tool of an action. Some instruments are derived; some are inert. Example: مفتاح key

An **adjective** is considered to be a type of noun in traditional Arabic grammar. It describes the state of the modified noun. Example: جميل beautiful   سيد Mr. محاضر Professor   كبير big

An **adverb** is a noun that is not derived and that indicates the place or the time of the action. Example:

شمال north   مدينة city   شهر Month

A **proper noun** is the name of a specific person, place, organization, thing, idea, event, date, time, or other entity. Some of them are solid (inert) nouns some of them are derived [Abuleil and Evens 1998].

3 Noun Classification

In this paper we focus on the following nouns: genus nouns, agent nouns, instrument nouns, adjectives, proper adjectives (adjectives derived from proper nouns), proper nouns, and adverbs. Some of these nouns are not derived from verbs and some are. All of these nouns use the same pattern when it comes to the dual form either for masculine or feminine, but there are many ways to form the plural noun. Some of the nouns have both masculine and feminine forms, some of them have just feminine forms and some have just masculine forms. A few nouns use the same format for both the plural and the dual (e.g. مدرسين teachers used for both dual and plural) For most nouns, when they end with the letter (ة), this indicates the feminine form of the noun, sometimes it does not, but it changes the meaning of the noun completely (e.g. مكتب office, مكتبة library). Sometimes the same consonant string with different vowels has different meanings (e.g. مدرسة school, مدرسة teacher). Nouns are not like verbs in the Arabic language, there is no clear rule to define the morphological information and generate the morphology paradigms for them. Instead each group of nouns follows its own pattern.

We have classified the nouns into 84 groups according to their patterns for singular, plural, masculine and feminine. We generated a method for each group to be used to find the morphological information and to form its paradigm. Very few of these groups have a

unique pattern for plural and singular; and most of them share the same pattern with other groups. Table 1 shows some examples of these groups and their patterns. The digit 9 stands for the letter "ayn [ع]", ' stands for "hamzh [ء]" and @ stands for "ta [ة]" since there is no corresponding letters in English for these letters.

Table 1. Pattern Classification

S-M	S-F	P-M	P-F
f9l	X	af9al	X
f9l	f9l@	af9la'	af9la'
X	f9l@	X	f9l
fa9l	fa9l@	f9al/f9l@	f9al/f9l@
f9al	X	X	af9l@
mf9l	X	mfa9l	X
fa9wl	X	fwa9el	X
mf9el	X	mfa9el	X
X	fa9l@	X	fa9lat
f9el	f9el@	f9la'	f9la'

S: Singular   F: Feminine
P: Plural   M: Masculine
X: not available

4   Acquisition System

The system reads the next noun in the text, isolates and analyzes the suffixes of the noun, generates its pattern, and uses either the Classified Noun Table, the Suffix/Pattern Analysis or the User-Feedback Module to find the group to which the noun belongs to identify the rules that applies to this group to generate all morphological paradigms with respect to the number and gender and updates the database. The system consists of several modules as shown in Figure 1.

4.1   Interface Module
This graphical user interface allows the user to interact with the system and handles the input/output. This module displays a main menu with two main options: collect nouns from documents and find morphological information.

4.2   Type-Finder Module
The main function of this module is to read the document and find the part of speech of the word: noun, verb, adjective, particle or proper

noun by running several tests: Database lookup, particle check, check on adjectives derived from proper nouns, parse of noun phrases and verb phrases, the affix check and the pattern check This module was built by Abuleil and Evens (1998, 2001). We use this module in our new system to find all nouns and extract them from the text.
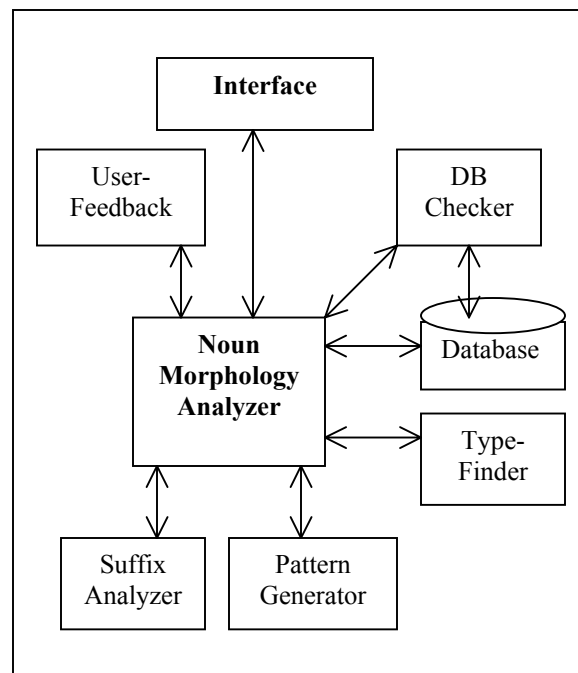

Figure 1. The Acquisition System

4.3   Database
The database includes a Classified Noun Table that contains each root noun (singular: masculine or feminine) and the number of the group to which the noun belongs. Each time the system identifies a new noun it adds its root to the Classified Noun Table**.**

4.4   Noun   Morphology   Analyzer Module
This is the core of the system, it calls different modules and performs different tasks to identify the noun and find its paradigm. First, it passes the noun to the suffix analyzer module to drop the suffix. Second, it passes it to the pattern generator module to find the pattern. Third, it analyzes the pattern to see whether it belongs to more than one group. It checks the Classified Nouns Table and then the suffix/pattern to

identify the group that the noun belongs to. If the system cannot identify the group then it calls the user-Feedback module to produce some questions to be answered by the user to reduce the number of alternatives to one. Finally, depending on the group the noun belongs to, it generates the morphological paradigms for number and gender and updates the database.

4.5   Suffix Analyzer Module
This module identifies the suffix, analyzes it and produces some lexical information about the noun like number and gender. First, it checks if any pronoun is concatenated with the noun. Second, it checks for a suffix indicating number. Third, it checks for a suffix indicating gender.

When the letter (ي) comes at the end of the noun there are two cases: it could be a part of the noun so we should not drop it, or it could be an extra letter as in relative nouns or when the pronoun is connected to the noun and it should be dropped in this case. When the noun ends with the letters (ين), most of the time it represents dual nouns but some times it represents both plural and dual nouns as in the following patterns: mfa9l, fa9l, mf9ull. Sometimes we have to check the pattern also to help in analyzing the suffix. We will handle these problems as special cases.

4.6   Pattern Generator Module
We have collected 62 different patterns used for both masculine and feminine, singular and plural after the suffix has been dropped see Appendix A. We used these patterns to generate a set of rules to build a finite-state diagram to be used to find the pattern for any noun. The input to this module is a noun after its suffix has been dropped in the previous step, the output is one or more patterns. If more than one pattern is found we validate the string by checking the pattern table.

The letter (م) and the letter (ا) at the beginning of the noun are sometimes the first characters of the noun, but sometimes they are separate words. We collected the nouns that begin with the letter (م) and the letter (ا) and saved them in a file to help us to distinguish between these two cases.

4.7   Database Checker Module
This module identifies any already classified noun or any noun derived from it. It gets the noun and its pattern from the noun morphology analyzer, finds all groups that contain the pattern, finds the singular noun (masculine or feminine) in each group and uses it to check the Classified Noun Table. If the noun exists it gets the group number to which it belongs and passes it to the Noun Morphology Analyzer to generate the results. For example the noun (ملاعب playground) has the pattern (mfa9l). This pattern appears in three different groups. See table 2.

Table 2. The Groups of the Noun "ملاعب"

Group#	Sing. Masc.	Sing Fem.	Plural Masc.	Plural Fem.
1	X	mf9l@	X	**Mfa9l**
2	Mf9l	X	X	**Mfa9l**
3	**mfa9l**	mf9l@	mf9lun/ mf9len	ملعبات

The nouns formed from these patterns have the following paradigms. See table 3.

Table 3. The Paradigms of the Noun "ملاعب"

Group#	Sing. Masc.	Sing Fem.	Plural Masc.	Plural Fem.
1	X	**ملعبة**	X	ملاعب
2	**ملعب**	X	X	ملاعب
3	**ملاعب**	ملعبة	ملعبون/ ملعبين	ملعبات

If the noun itself or any other noun derived from it has been previously classified we will find its noun root (singular noun) in the Classified Noun Table. The module will find the root (singular masculine) "ملعب" in the table and will get its group number "2" and pass it to Noun Morphology Analyzer to find the noun paradigms.

4.8   User-Feedback Module
This module gets all alternatives (groups) from the noun morphology analyzer module. It analyzes them and generates some questions to be answered by the user. It gets the answers, analyzes them and finds the group that the noun belongs to. The module asks questions like: Is the noun a singular? Is the noun a plural? Does

the noun have a masculine-singular format? Does the noun have a feminine-singular format?

Example:
Input: The noun (ملاعب playground)
Pattern: mfa9l
Number of groups that contain the pattern is 3.

Process:
Step #1: identify the groups

Group#	Sing. Masc.	Sing. Fem.	Plural Masc.	Plural Fem.
1	X	mf9l@	X	mfa9l
2	mf9l	X	X	mfa9l
3	mfa9l	mf9l@	mf9lun / mf9len	mf9lat

Step #2: Replace (X) with –1, given pattern with 1 and any thing else with 0.

Group#	Sing. Masc.	Sing. Fem.	Plural Masc.	Plural Fem.
1	-1	0	-1	1
2	0	-1	-1	1
3	1	0	0	0

Step #3: Add the one's in each column and subtract it from number of groups. Add the (-1's) in each column and subtract it from number of groups. Add the (0's) in each column.

Group#	Sing. Masc.	Sing. Fem.	Plural Masc.	Plural Fem.
1	-1	0	-1	1
2	0	-1	-1	1
3	1	0	0	0
A = Σ1's	**1**	**0**	**0**	**2**
B = Σ-1's	**1**	**1**	**2**	**0**
C = Σ 0's	**1**	**2**	**1**	**1**
A1 = #G – A	**2**	**3**	**3**	**1**
B1 = #G – B	**2**	**2**	**1**	**3**

From the table above we know that: the probability that the noun is singular masculine is 33.3% and the probability that it is a plural feminine is 66.6%.

Step #4: Pick the smallest value greater than 0 from the "A1" row and the "B1" row go from left to right and from top to bottom. Use the

column name to form questions. For the "A1" value use the following question: is the noun a [column name]? For the "B1" use the following question: does the noun have the [column name] format? Get the answer and drop invalid group(s).

Group#	Sing. Masc.	Sing. Fem.	Plural Masc.	Plural Fem.
1	-1	0	-1	1
2	0	-1	-1	1
A = Σ1's	**0**	**0**	**0**	**2**
B = Σ-1's	**1**	**1**	**2**	**0**
C = Σ 0's	**1**	**1**	**0**	**0**
A1 = #G – A	**2**	**2**	**2**	**0**
B1 = #G – B	**1**	**1**	**0**	**2**

Step #5: Repeat step 3 and step 4 until you end up with one group or all the values in both Row A1 and row B1 have the values either zero or the number of groups left.

Step #6: if more than one group is left from step #5 then find the largest value in the row "C" from left to right and ask the following question: which of the following [list all the options in that column] is the [column name] of the noun?

Group #	Sing. Masc.	Sing. Fem.	Plural Masc.	Plural Fem.
2	0	-1	-1	1
A = Σ1's	**0**	**0**	**0**	**1**
B = Σ-1's	**0**	**1**	**1**	**0**
C = Σ 0's	**1**	**0**	**0**	**0**
A1 = #G – A	**1**	**1**	**1**	**0**
B1 = #G – B	**1**	**0**	**0**	**1**

The questions the module generated from the previous example are:
Q1: *is the noun plural feminine?*
Answer: *yes*　// the system drops group#3
Q2: *does the noun have singular masculine format*?
Answer: No　// the system drops group#1

Result:
Group # 2: The noun (ملاعب playground) is a plural Feminine. The singular Masculine format is (ملعب), the singular Feminine format and plural masculine format are not available for this noun.

5   Examples

The following example shows how the system works. Assume that the input is the noun (مدربتهم their trainer), First the system calls the suffix analyzer module to drop the extra letter (pronoun: their) at the end (مدربت + هم), replace the letter (ت) with the letter (ة), generate the noun (مدربة trainer) and some lexical information about the noun.

Second, it passes the noun (مدربة trainer) to the pattern generator module to generate the pattern (mf9l@). Third, it checks the group table looking for this pattern (mf9l@). Fourth, if more that one group is found it uses the Database Checker Module to check the Classified Noun Table. Fifth, if the noun does not exist in the table, it calls the User-Feedback Module to analyze the groups (all alternatives) and asks the user some questions to assist in identifying the group see Table 4 and Table 5. The question that the module generated is:

Question: Does the noun have a masculine-singular format?
Answer: Yes
Result: drop group # 10 & group # 22

Table 4. First Cycle to Generate Question

Group #	Sing. Masc.	Sing Fem.	Plural Masc.	Plural Fem.
10	-1	1	0	-1
22	-1	1	0	-1
38	0	1	0	0
A = Σ1's	**0**	**3**	**0**	**0**
B = Σ-1's	**2**	**0**	**0**	**2**
C = Σ 0's	**1**	**0**	**3**	**1**
A1 = #G – A	**3**	**0**	**3**	**3**
B1 = #G – B	**1**	**3**	**3**	**1**

Table 5. Second Cycle to Generate Question

Group #	Sing. Masc.	Sing Fem.	Plural Masc.	Plural Fem.
38	0	1	0	0
A = Σ1's	**0**	**1**	**0**	**0**
B = Σ-1's	**0**	**0**	**0**	**0**
C = Σ 0's	**1**	**0**	**1**	**1**
A1 = #G – A	**1**	**0**	**1**	**1**
B1 = #G – B	**1**	**1**	**1**	**1**

Fifth, it generates the results: group#38 and updates the database. Table 6 shows system output for some input.

Table 6. System Output

Noun	مفاتيح keys	طائرة plane	صوتنا Our sound	كريمين generous
Suffix	----	----	نا	ين
Pattern	مفاعيل mfa9el	فاعلة fa9l@	فعل f9l	فعيل f9el
Group #	52	23	3	37
Result	Plural masc.	Singular Feminine	Singular feminine	Dual / plural masc.
Singular / Masc.	مفتاح	X	صوت	كريم
Singular / Fem.	X	طائرة	X	كريمة
Plural / Masc.	X	X	X	كرماء / كريمين
Plural / Fem.	مفاتيح	طائرات	اصوات	كريمات
Dual / Masc.	مفتاحين مفتاحان	X	صوتين صوتان	كريمين كريمان
Dual / Fem.	X	طائرتين طائرتان	X	كريمتين كريمتان

6   Results

To test our system we used nouns obtained from a corpus developed by Ahmad Hasnah based on text given to Illinois Institute of Technology, by the newspaper, *Al-Raya*, published in Qatar. We have tested each module in our system: the suffix analyzer modules, the pattern generator module, and the user-Feedback module. Table 7 shows the result of testing the system on 500 nouns.

Table 7. Suffix / Pattern / Noun Morphology Analyzer

	# correct	# incorrect	% correct	% incorrect
Suffix Analyzer	490	10	97%	3%
Pattern Analyzer	471	29	93%	8%
Noun Morph analyzer	451	49	90.2%	9.8%

As shown in Table 7 there were ten failure because of incorrect suffix analysis and 29 due

to missing patterns. These missing patterns have now been added. The suffix analysis problem is hard to correct because it arises from underlying ambiguities. If the noun has been classified previously the system does not have any problem to identify it and identify any noun derived from it.

The User-Feedback Module found most of the nouns that the Database Checker Module failed to identify. Table 8 shows a number of nouns identified by suffix/pattern, nouns identified by Database Checker Module and nouns identified by User-Feedback Modules. We believe that the more knowledge that the system gains and the more nouns that it adds to the Classified Noun Table the fewer questions have to be asked.

Table 8. Noun Classifier Methods

Nouns Identified by Database Checker	Nouns Identified by Suffix/ Pattern Analysis	Nouns Identified by User-Feedback Module
144	32	289
28.8%	7.1%	64.1%

7   Conclusion

We have built a learning system that utilizes user feedback to identify the nouns in the Arabic language, obtain their features and generate their paradigms with respect to number and gender. We tested the system on 500 nouns from newspaper text. The system identified 90.2% of them, 7.1% by just analyzing the suffix and the pattern of the noun, 28.8% by using the Database Checker Module and the Classified Noun Table and 64.1% by using User-Feedback Module. The system failed on 9.8% of the tested nouns.

References

Abuleil, S. and Evens, M., 1998. "Discovering Lexical Information by Tagging Arabic Newspaper Text", Workshop on Semitic Language Processing. COLING-ACL'98, University of Montreal, Montreal, PQ, Canada, Aug 16 1998, pp 1-7.

Abuleil, S. and Evens, M., 2002. Extracting an Arabic Lexicon from Arabic Newspaper Text. *Computers and the Humanities*, 36(2), pp. 191-221.

Al-Fedaghi, Sabah and Al-Anzi, Fawaz, 1989. "A New Algorithm to Generate Arabic Root-Pattern Forms". Proceedings of the 11th National Computer Conference, King Fahd University of Petroleum & Minerals, Dhahran, Saudi Arabia., pp 4-7.

Al-Shalabi, R. and Evens, M., 1998. "A Computational Morphology System for Arabic". Workshop on Semitic Language Processing. COLING-ACL'98, University of Montreal, Montreal, PQ, Canada, Aug 16 1998. pp. 66-72.

Beesley, K. and Karttunen, L., 2000. "Finite-State Non-Concatenative Morphotactics". Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics. Hong Kong, Oct 1-8, 2000. pp.191-198.

Hasnah, A., 1996. Full Text Processing and Retrieval: Weight Ranking, Text Structuring, and Passage Retrieval For Arabic Documents. Ph.D. Dissertation, Illinois Institute of Technology, Chicago, IL.

Roeck, A. and Al-Fares, W., 2000. "A Morphologically Sensitive Clustering Algorithm for Identifying Arabic Roots". Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics. Hong Kong, Oct 1-8, 2000. pp.199-206.

Appendix A. Patterns

Pattern	Used for	Example
f9l	sing – masc.	جمل
f9l	plural – masc.	جزر
f9l	plural – fem. / masc.	عرب
f9l	plural – fem.	صور
f9l	sing – masc.	ضوء
f9l@	sing. – fem.	صورة
mf9al	sing. masc.	مفتاح

Pattern	Used for	Example		Pattern	Used for	Example
f9l@	plural – masc.	قتلة		Mstf9l	sing. – masc.	مستخدم
aft9al	sing. – masc.	اختراع		mf9ll	sing. – masc.	مسلسل
anf9al	sing. – masc.	انفجار		Mstf9a	sing. fem.	مستشفى
astf9al	sing. - masc.	استثمار		mf9wl@	sing. – fem.	موسوعة
af9al	plural – fem.	اشجار		mf9el	sing. masc.	منديل
af9la'	plural – fem. / masc.	اغنياء		mfa9el	plural – fem.	مناديل
af9l@	plural – fem.	ادوية		mf9le@	sing. – fem.	مسرحية
af9el	sing. – masc.	ابريق		mfa9l	sing. – masc.	مقاتل
afa9el	plural – fem.	اباريق		mfa9l@	sing. – fem.	مظاهرة
f9lawat	plural – fem.	حمراوات		mf9wl	sing. – masc.	مشروع
fwa9l	plural – fem.	جوامع		mfa9el	plural – fem.	مشاريع
fwa9el	plural – fem.	موازين				
fe9al	sing- masc.	ميزان				
f9lan	plural – fem.	غزلان				
f9all	plural – fem.	قنابل				
tf9l@	plural – fem.	تكلفة				
f9wl@	plural –fem.	حكومة				
f9wl	sing. – masc.	عمود				
f9ll@	sing- fem.	قنبلة				
f9le@	sing. – fem.	جمعية				
f9le	sing.- masc.	صحفي				
f9el	sing – masc.	كريم				
f9el@	sing.- fem.	جزيرة				
f9al	sing.- masc.	مطار				
f9al	plural – fem.	جمال				
f9ale	plural – fem.	صحاري				
fa9l	sing. – masc.	عالم				
fa9l@	sing. – fem.	باخرة				
f9al@	sing. – fem.	خسارة				
f9al	plural – masc.	سجاد				
f9la'	plural – masc.	علماء				
f9la'	sing. – fem.	حمراء				
f9alel	plural – fem. / masc.	جماهير				
fa9wl	sing. masc.	صاروخ				
f9a'l	plural – fem.	حقائب				
tf9el	sing. – masc.	تمرين				
f9lwl	sing. – masc.	جمهور				
tfa9el	plural – fem.	تمارين				
fw9l@	sing. – fem.	جوهرة				
f9wal	sing. – masc.	عنوان				
f9awel	plural – fem.	عناوين				
mf9l@	sing. – fem.	منجرة				
mfa9l	plural – fem.	مناجر				
mf9l	sing. – masc.	مدرس				
mf9l@	sing. – fem.	مدرسة				
mf9l	sing. – masc.	مكتب				
mf9l@	sing. – fem.	مكتبة				
mft9l	sing. masc.	معتقل				