



# Arabic Stemming

# Agenda

- Text Normalization Tasks:
  - Stemming
  - Lemmatization
- Arabic Stemming and Stemmers.
- How to build Arabic stemmer.

```
>>> porter= nltk.PorterStemmer()  
>>> stemming=[porter.stem(w) for w in tokens]  
>>> for w in stemming:  
        print w
```

حكى  
بعضهم  
قال  
:  
كنت  
في  
سفر  
فضللت  
عن  
الطريق  
'  
فرأيت  
بيتاً

```
>>> lancaster= nltk.LancasterStemmer()
>>> stemming=[lancaster.stem(w) for w in tokens]
>>> for w in stemming:
    print w + '-'
```

حكى- بعضهم- قال- :- كنت- في- سفر- فضلت- عن- الطريق- ، - فرأيت- بيتاً- في- الفلاة  
 . -فأتيته- فإذا- به- أعرابيّة- ، - فلما- رأتنى- قالت- من- تكون- ؟- قلت- ضيف- ، -  
 قالت- أهلاً- ومرحباً- بالضيف- ، - انزل- على- الرحب- والسعة- . - قال- فنزلت- فقدمت-  
 لي- طعاماً- فأكلت- ، - وماءً- فشربت- ، - فبينما- أنا- على- ذلك- إذ- أقبل- صاحب-  
 البيت- . - فقال- من- هذا- ؟- فقالت- ضيف- . - فقال- لا- أهلاً- ولا- مرحباً- ، - ما- لنا-  
 وللضيف- ، - فلما- سمعت- كلامه- ركبت- من- ساعتى- وسرت- ، - فلما- كان- من- الغد-  
 رأيت- بيتاً- في- الفلاة- فقصدته- فإذا- فيه- أعرابيّة- فلما- رأتنى- قالت- من- تكون-  
 ؟- قلت- ضيف- . - قالت- لا- أهلاً- ولا- مرحباً- بالضيف- ، - ما- لنا- وللضيف- ، - فبين-  
 ما- هي- تكلمنى- إذ- أقبل- صاحب- البيت- فلما- رآنى- قال- من- هذا- ؟- قالت- ضيف-  
 . -

```
>>> wl=nltk.WordNetLemmatizer()
>>> rr=[wl.lemmatize(w) for w in tokens]
>>> for word in rr:
    print word + '-'
```

حكى- بعضهم- قال- :- كنت- في- سفر- فضلت- عن- الطريق- ، - فرأيت- بيتاً- في- الفلاة  
 . -فأتيته- فإذا- به- أعرابيّة- ، - فلما- رأتنى- قالت- من- تكون- ؟- قلت- ضيف- ، -  
 قالت- أهلاً- ومرحباً- بالضيف- ، - انزل- على- الرحب- والسعة- . - قال- فنزلت- فقدمت-  
 لي- طعاماً- فأكلت- ، - وماءً- فشربت- ، - فبينما- أنا- على- ذلك- إذ- أقبل- صاحب-  
 البيت- . - فقال- من- هذا- ؟- فقالت- ضيف- . - فقال- لا- أهلاً- ولا- مرحباً- ، - ما- لنا-  
 وللضيف- ، - فلما- سمعت- كلامه- ركبت- من- ساعتى- وسرت- ، - فلما- كان- من- الغد-  
 رأيت- بيتاً- في- الفلاة- فقصدته- فإذا- فيه- أعرابيّة- فلما- رأتنى- قالت- من- تكون-  
 ؟- قلت- ضيف- . - قالت- لا- أهلاً- ولا- مرحباً- بالضيف- ، - ما- لنا- وللضيف- ، - فبين-  
 ما- هي- تكلمنى- إذ- أقبل- صاحب- البيت- فلما- رآنى- قال- من- هذا- ؟- قالت- ضيف-  
 . -

# Arabic Stemming

- What is stemming?

Stemming is the process of removing any affixes from words and reducing these words or return to their roots. For example, stemming the English word computing produces the root compute. This is the same root produced by the word computation.

- How is the stemmer implemented?

The first thing the stemmer does is remove the longest suffix and the longest prefix. It then matches the remaining word with the verbal and noun patterns, to extract the root.

# Arabic Stemming

- What are stemming types?

## 1- Light Stemmer

Ex: الطالب <=== طالب

## 2- Root-Based Stemmer

Ex: الطالب ===> طلب

# Arabic Stemmers

<http://www.nltk.org/api/nltk.stem.html>

- Assem's Arabic Light Stemmer ( BETA )

<https://arabicstemmer.com/>

- ISRIStemmer.

<http://www.nltk.org/api/nltk.stem.html>

- Tashaphyne Arabic Light Stemmer:

<https://pypi.python.org/pypi/Tashaphyne/>

- Khoja Arabic Stemmer (root-based stemmer)

<http://zeus.cs.pacificu.edu/shereen/research.htm>

- FARASA stemmer

<http://qatsdemo.cloudapp.net/farasa/register.html>



# Assem's Arabic Light Stemmer ( BETA )

## Description

Welcome to the Arabic Light Stemming Algorithm made for [Snowball](#), it's fast and can be generated in many programming languages (through Snowball).

## Demo

Type some Arabic text and press "Stem!" button or "File" to read from a local ".txt" file

مكتبة لمعالجة الكلمات العربية وتجذيعها

STEM!

FILE

Stats

words: 5

stems: 5

ratio: 1.2

مكتب معالج كلم عرب تجذيع

[/ Arabic Language Technologies](#) / [ALT Server](#) / [Demos](#) / Farasa

Please enter your text:

أدخل النص المراد معالجته:

المدرسة  
الطالب  
يفكرون

Please note that there are some limitations to try the Dependency Parser:

- The demo is confined to process only three sentences per request. each sentence shouldn't exceed 20 words.
- The length of the text to be processed should be within 400 characters.

Lemmatization أصول الكلمات



Process text معالجة النص

Clear text مسح النص

Text length 21 عدد أحرف النص

مدرسة . طالب . فكر

C:\Users\Amira\Anaconda2\Lib\site-packages\Tashaphyne\module2.py

Project

DLLs

Doc

envs

etc

include

info

Lib

Library

libs

man

Menu

nlTK\_data

pkgs

projects

ANLP\_Section1.py

Stemming\_section.py

StopWordRemove.py

Scripts

share

sip

tcl

Tools

nonadmin

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

```
word=u"الغُرَائِيَّةُ"
ArListem=ArabicLightStemmer();
stem=ArListem.lightStem(word);
print stem

print ArListem.get_unvocalized();
word=u'أَفْتَكَاتِبَانِي'
stem=ArListem.lightStem(word);
print ArListem.get_stem();
print ArListem.get_right();

ArListem=ArabicLightStemmer();
word=u'فَتْمَبَرِين'
stem=ArListem.segment(word);
print str(ArListem.get_affix_list()).decode("unicode-escape");
print ArListem.get_segment_list();
```

python(44) x python(45) x python(46) x python(47) x python(48) x python(49) x python(50) x python(51) x python(52) x python(53) x python(54)

Python 2.7.12 |Anaconda 4.1.1 (64-bit)| (default, Jun 29 2016, 11:07:13) [MSC v.1500 64 bit (AMD64)] on win32

In[2]: runfile('C:/Users/Amira/Anaconda2/Lib/site-packages/Tashaphyne/module2.py', wdir='C:/Users/Amira/Anaconda2/Lib/site-packages/Tasha

عرب

العربية

كاتب

7

[[{'prefix': u'ف', 'root': u'مَبر', 'suffix': u'ين', 'stem': u'تَمبر'}, {'prefix': u'فت', 'root': u'مَبر', 'suffix': u'ين', 'stem': u'مَبر'},

set([(1, 5), (2, 5), (0, 7)])

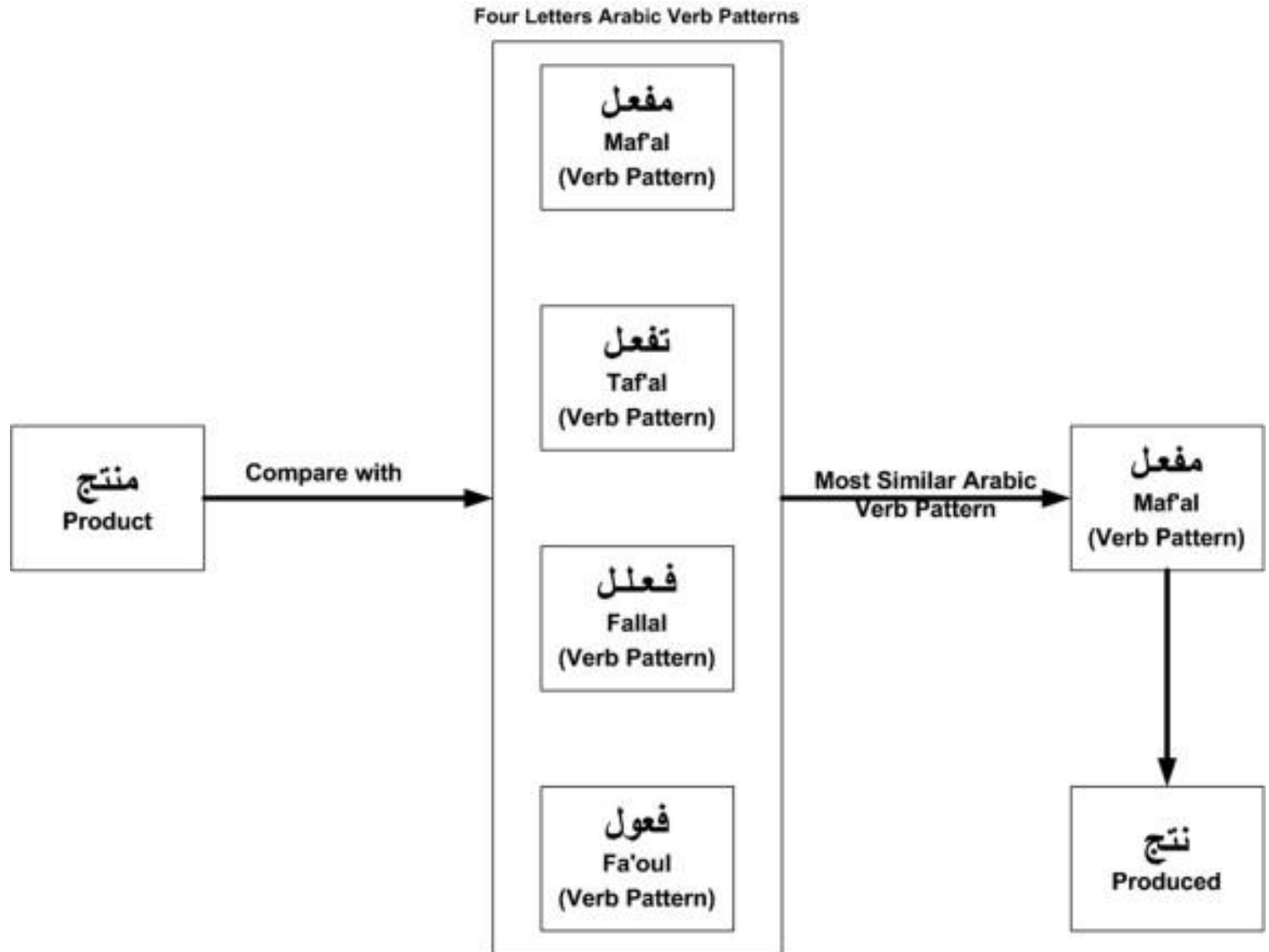
# Arabic root-based stemmer algorithm

**Input:** A text file that contains the Arabic words

**Output:** Arabic Triliteral Verb/Verbs

1. Remove Arabic prefix(es) from each word
2. Normalize 3 shapes of (Alif, "أ", "إ", "إِ") to (Bare Alif, "ا")
3. Remove suffix(es) from each word
4. Determine word length after removing affixes (prefixes and suffixes)
5. Identify Arabic patterns having same lengths to word length in step 4.
6. Compare each pattern identified in step 5 with extracted word from step 3
7. Select the closest pattern:
  - a. Choose the pattern from the set of Arabic patterns having same lengths to word length which has the highest number of common Arabic letters with the Arabic word extracted from step 3.
  - b. Determine the pattern which has the largest matching corresponding letters with the generated word from step 3 which is considered as the right pattern, where the corresponding Arabic letters within the extracted word from step 3 will not be compared with three Arabic letters (Faa', "ف"), (Ayn, "ع"), (Laam, "ل") within the pattern under consideration.
8. Eliminate all matched letters in step 7. The Arabic letters of the Arabic word extracted from step 3 which corresponds to the Arabic letters (Faa', "ف"), (Ayn, "ع"), and (Laam, "ل") in the selected pattern (found in step 7.a) are selected to constitute the extracted Arabic root.
9. Refine the extracted Arabic root by converting some of the Arabic letters.

# Arabic root-based stemmer algorithm



# Khoja stemmer algorithm

Algorithm: Khoja Root-Based Stemming Algorithm

Purpose: Stemming Arabic Words

Input:

- Dataset
- Stop-word list
- Assets and patterns files

Output: Stemmed Dataset

Procedure:

1. Replace initial َ , ِ , ِ with .
2. Stop-words removal.
3. Remove punctuation, non-letters and diacritics.
4. Remove definite **articles** from the beginning of the word.
5. Remove the letter (ﺀ) from the beginning of the word and (ﻪ) from the end of the word.
6. Remove prefixes and suffixes
7. Comparing the resulting word to patterns stored in the dictionary, if the resulting root is meaningless the original word is returned without changes.

**Figure 1-** Khoja Stemmer Algorithm

## ISRI Arabic Stemmer:

The Information Science Research Institute's (ISRI) Arabic stemmer shares many features with the Khoja stemmer. However, the main difference is that ISRI stemmer does not use root dictionary. Also, if a root is not found, ISRI stemmer returned normalized form, rather than returning the original unmodified word.



- Additional adjustments were made to improve the algorithm:

- 1- Adding 60 stop words.

- 2- Adding the pattern (تفاعيل) to ISRI pattern set.

- 3- The step 2 in the original algorithm was normalizing all hamza. This step is discarded because it increases the word ambiguities and changes the original root.

- The ISRI Stemmer requires that all tokens have Unicode string types.



# ISRI Arabic Stemmer:

```
>>> st=nltk.ISRIStemmer()  
>>> resultstem=[st.stem(w) for w in tokens]  
>>> for w in resultstem:  
    print w + '-'
```

حكى- بعض- قال- :- كنت- في- سفر- ضل- عن- طرق- ، - فرأ- بيت- في- فلة- ، - فأت- فإ-  
- ، - ذ- به- عرب- ، - فلم- رأت- قلت- من- تكون- ؟- قلت- ضيف- . - قلت- اهل- رحب- ضيف  
-نزل- على- رحب- سعة- . - قال- نزل- قدم- لي- طعا- أكل- ، - وم- شرب- ، - فبن- انا-  
-على- ذلك- اذ- قبل- صحب- بيت- . - فقل- من- هذا- ؟- فقل- ضيف- . - فقل- لا- اهل- ولا  
رحب- ، - ما- لنا- ضيف- ، - فلم- سمع- كلم- ركب- من- سعت- وسر- ، - فلم- كان- من- لغ  
د- رأت- بيت- في- فلة- قصد- فإذ- فيه- عرب- فلم- رأت- قلت- من- تكون- ؟- قلت- ضيف  
قلت- لا- اهل- ولا- رحب- ضيف- ، - ما- لنا- ضيف- ، - فبن- هي- كلم- اذ- قبل- صحب- . -  
- . - بيت- فلم- رآن- قال- من- هذا- ؟- قلت- ضيف-

```
>>> st= nltk.ISRIStemmer()
```

```
>>> st.stem('معلومات')
```

```
u'\u0639\u0644\u0645'
```

```
>>> print u'\u0639\u0644\u0645'
```

علم

```
>>> st= nltk.ISRIStemmer()
```

```
>>> st.stem('معلومات')
```

```
u'\u0639\u0644\u0645'
```

```
>>> print u'\u0639\u0644\u0645'
```

علم

# Agenda

- Text Normalization Tasks:
  - Stemming
  - Lemmatization
- Arabic Stemming and Stemmers.
- How to build Arabic stemmer.

# Arabic Root Word Resources:

www.tyndalearchive.com/TABS/Lane/index.htm

## Arabic-English Lexicon by Edward William Lane (London: Willams & Norgate 1863)

Contents

رجل

رجل

رجل

رجل

رجل

رجل

رجل

رجل

رجل

رجل

رجل

رجل

رجل

رجل

رجل

رجل

رجل

رجل

رجل

رجل

رجل

Prev. Page

Magnify Box Smallest

[Book I.]

The tenth letter of the alphabet: called راء and راء: pl. [of the former] راء and [of the latter] راء. (TA in الالف اللينة.) It is one of the letters termed مَجْهُورَة [or vocal, i. e. pronounced with the voice, not with the breath only]; and of the letters termed ذَلَقِي, which are ر and ل and ن, [also termed ذَوَقِيَة, or pronounced with the extremity of the tongue, and ب and ف and م, which are also termed شَفِيحَة, or pronounced with the lips:] these letters which are pronounced with the tip of the tongue and with the lips abound in the composition of Arabic words: (L:) and hence ر is termed, in a vulgar prov., جَبَّارُ الشُّعْرَاءِ ["the ass of the poets"]. (TA in الالف اللينة.) ر is substituted for ل, in نَشْرَة for نَشْرَة, and in نَعْلَ for أَوْجَلْ and وَجَلْ and أَوْجَرْ and وَجَرْ, and نَعْلَ; and this substitution is a peculiarity of the dial. of Kays; wherefore some assert that the ر in these cases is an original radical letter. (MF.)=[As a numeral, it denotes Two hundred.]

ر is an imperative of رَأَى [q. v.]. (AZ, T and S and M in art. رَأَى.)

راء

راء and راء: see the preceding paragraph, and art. رَأَى and رَأَى. راء is also said by some for رَأَى [q. v.]. (M in art. رَأَى.)

راء

R. Q. I. راء السراب, (Sgh, and so in a copy of the S,) or السحاب, (M, and so in a copy of the S,) or both, (K,) The mirage, or the clouds, or both, shone, or glistened. (S, M, Sgh, K.) — [Hence, probably,] راء عَيْنَاهُ [app. meaning His eyes glanced] is said when one turns his

or an adulteress, she moved about the blacks of her eyes [as a sign] to the man seeking her: (T:) or راءت بعينها, said of a woman, (S, M,) she glistened with her eye, by reason of looking hard, or intently: (S:) or she opened her eye wide, and looked sharply, or intently. (M.) Also, said of a woman, She looked at her face in a mirror. (K, TA.) — راءات الظبابة The gazelles wagged their tails: (K:) or so راءات بأذنابها; like راءت. (T.) — راءا بالغدير (K, TA.) inf. n. راءا. (T.) He called the sheep, or goats, to water: (T:) or he called them

the cry راءا, or [rather] راءا. (M,) or by the cry راءا. (K:) accord. to analogy, the verb [derived from the cry] should be راءا: (M:) طرطط به, inf. n. طرطط, signifies "he called them [to be milked by making a sound] with his lips." (T.)

راءا الغنم (S, M) and راءا راءا (T, S, M, A, K,) and راءا راءا (Kr, M,) A man who turns about the black of the eye much. (T, S, M, K.) And راءا راءا (T, M, K,) with medd. and without s, (T,) and راءا راءا (M, K,) A woman who opens her eye wide, (M,) or who glistens with her eyes, (K,) looking sharply, or intently. (M, K.)

راءا: see the next preceding paragraph, in three places.

راء

1. راء, (T, S, M, A, K,) aor. ٴ, (M, A, K,) inf. n. راء, (M, TA,) He repaired, or mended, (T, S, M, A, K,) a [cracked, or broken,] vessel, (S,) or a crack, or fissure; (M, A, K,) as also راء, راء, راء (S, A) + O God, effect a reconciliation, or make peace, between them: (S:) or rectify the matter, or affair, between them. (A.) And راء راءا: [O God, rectify, or amend, our state, or condition]. (TA.) — Also, inf. n. as above, + He collected a thing together, and bound it gently. (TA.) — راءت الأرض + The land produced its [trefoil called] راءة, or راءة, [so accord. to different copies of the K,] after the cutting [of a crop thereof]. (K.)

2 and 4 and 8: see above, first sentence.

in the sense of [the act. the saying, راءا]

راءا راءا (S, A) + O God, effect a reconciliation, or make peace, between them: (S:) or rectify the matter, or affair, between them. (A.) And راء راءا: [O God, rectify, or amend, our state, or condition]. (TA.) — Also, inf. n. as above, + He collected a thing together, and bound it gently. (TA.) — راءت الأرض + The land produced its [trefoil called] راءة, or راءة, [so accord. to different copies of the K,] after the cutting [of a crop thereof]. (K.)

راءا راءا (S, A) + O God, effect a reconciliation, or make peace, between them: (S:) or rectify the matter, or affair, between them. (A.) And راء راءا: [O God, rectify, or amend, our state, or condition]. (TA.) — Also, inf. n. as above, + He collected a thing together, and bound it gently. (TA.) — راءت الأرض + The land produced its [trefoil called] راءة, or راءة, [so accord. to different copies of the K,] after the cutting [of a crop thereof]. (K.)



# Arabic Root Word Resources:

← → ↻ 🏠 [https://wahiduddin.net/words/arabic\\_glossary.htm](https://wahiduddin.net/words/arabic_glossary.htm)

[Spirituality of Buddhism ?](#)  
[One Nation Under God ?](#)  
[Daily Resolutions](#)  
[Differences of opinion](#)  
[Celebration of Christmas](#)

## MUSIC...

by wahiduddin

[The Music Page](#)

## POETRY...

by wahiduddin

[Journey of the Heart](#)  
[Resurrection](#)  
[In the Garden of Lovers](#)  
[Rendezvous with the Beloved](#)

## ROOTS OF WORDS...

by wahiduddin

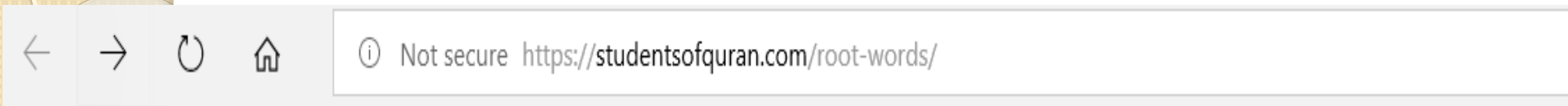
[Vanity of vanities, all is vanity](#)  
[Bismillah ir rahman ir rahim](#)  
[Fear the Lord ?](#)  
[Jesus...Iesous...Yeshua](#)  
[Why hast thou forsaken me ?](#)  
[The Name "God"](#)  
[Shalom](#)  
[Arabic Devotional Terms](#)  
[Arabic Roots](#)  
[99 Beautiful Names](#)  
[la ilaha illa allah](#)  
[la hawla wa la quwwata](#)  
[the name wahiduddin](#)

## QUR'AN

Examples of typical usage of the root are shown in parenthesis.

a-b	to be a father, ancestor, forefather (ab, abū)
'a-b-d	to serve, worship, be devoted to, show veneration (‘abd, ‘ibāda, ma‘būd)
'a-d-l	to act justly, equitably or to make straight, set in order (‘adl, a‘dāla, ta‘dīl)
'a-d-m	to be non-existent, disappeared, destroyed, devoid of (‘adam, ‘adīm)
'a-f-w	to be obliterated, effaced, eliminated (al-‘afūw, ‘afwīya, ‘afā’, isit‘fā’, ‘āfin, mu‘fan)
a-ḥ-d	to unify, be one (al-aḥad, aḥadīya, uḥādī)
a-kh-r	to postpone, defer, be last, final, ultimate (al-ākhir, ākhar, ukhrā, ta‘khīr, mu‘akhkhara)
a-l-h	to adore, deify, turn to another with intense feeling (ilāh, ilāhī)
'a-l-m	to know, have knowledge, be informed, teach, notify (al-‘alīm, ‘ilm, ‘ilmiya, ‘allam, u‘lūma)
a-m-l	to hope, to look attentively, meditate, consider (amal, āmāl, āmil, muta‘ammil)
a-n-s	to be familiar, friendly, sociable (uns, insī, ins, anīs)
'a-q-d	to tie, knit, make a knot, put together, join (‘aqd, ‘aqīda)
'a-q-l	to have the faculty of reasoning, comprehension (‘aql, ‘aqlī)

# Arabic Root Word Resources:



A Directory of arabic root words and letters, singular and plurals from Juz 1 – Juz 30

Root Words

Singular/Plural

Blue Juz 1-30



**Juz 1**



**Juz 2**



**Juz 3**



**Juz 4**



**Juz 5**



# Arabic Root Word Resources:

- 1- <http://www.tyndalearchive.com/TABS/Lane/index.htm>
- 2- [https://wahiduddin.net/words/arabic\\_glossary.htm](https://wahiduddin.net/words/arabic_glossary.htm)
- 3- <https://studentsofquran.com/root-words/>