

# A Linguistic Method into Stemming of Arabic for Data Compression

Hussein Soori, Jan Platoš, and Václav Snášel

Faculty of Electrical Engineering and Computer Science  
VSB-Technical University of Ostrava, Czech Republic  
{sen.soori, jan.platos,vaclav.snasel}@vsb.cz

**Abstract.** Creating good stemming rules for the Arabic language comes from the importance of Arabic language as the sixth most used language in the world. Stemming is very important in information retrieval, data mining and language processing. With Arabic having complex morphology and grammatical properties, this poses a challenge for researchers in this field. In this paper, we try to use an online morphological parser to distinguish parts of speech (POS), and then set some extracting rules to produce stems, and finally, mismatch these stems with an electronic dictionary. As a pilot study for this method, in this paper we deal with three POS: nouns, verbs and adjectives.

**Keywords:** Stanford Online Parser, data compression for Arabic, Arabic natural language processing, Arabic data mining, Arabic morphology, stemming of Arabic.

## 1. Introduction

The rapidly growing number of computer and Internet users in the Arab world and the fact that the Arabic language is the sixth most used language in the world today creates a demand for more research in the area of data mining and natural language processing in Arabic language. Another two factors maybe that Arabic alphabet is the second-most widely used alphabet around the world - Arabic script has been used and adapted to such diverse languages as Amazigh (Berber), Hausa, and Mandinka (in West Africa), Hebrew, Malay (Jawi in Malaysi and Indonesia), Persian, the Slavic tongues (also known as Slavic languages), Spanish, Sudanese, and some other languages, Swahili (in East Africa), Turkish, Urdu [10], and that Arabic is one of the six languages used in the United Nations [11] after the Latin alphabet.

### 1.1 Arabic Complex Morphological and Grammatical Properties

A few challenges may face researchers as for as the special nature of Arabic script is concerned. Arabic is considered as one of the highly inflectional languages with complex morphology. Unlike most other languages, it is written horizontally from right to left. It consists of 28 main letters. The shape of each letter depends on its position in a

word—initial, medial, and final. There is a fourth form of the letter when written alone. One example of this can be given for the letter (ع) as follow:

Initial	Medial	Final	Separate
ع	ع	ع	ع

Fig. 1. Arabic Alphabets

Moreover, the letters alif, waw, and ya (standing for glottal stop, w, and y, respectively) are used to represent the long vowels a, u, and i. This is very much different from Roman alphabet which is naturally not linked. Other orthographic challenges can be the the persistent and widespread variation in the spelling of letters such as hamza (ء) and ta' marbuTa (ة), as well as, the increasing lack of differentiation between word-final ya (ي) and alif maqSura (ى). Typists often neglect to insert a space after words that end with a non-connector letter such as و, ز, ر [3]. In addition to that, Arabic has eight short vowels and diacritics (َ, ِ, ُ, ً, ٌ, ٍ, ً, ٍ). Typists normally ignore putting them in a text, but in case of texts where typists do put them, they are pre-normalized –in value- to avoid any mismatching with the dictionary or corpus in light stemming. As a result, the letters in the decompressed text, appear without these special diacritics.

Diacritization has always been a problem for researches. According to Habash [12], since diacritical problems in Arabic occur so infrequently, they are removed from the text by most researchers. Other text recognition studies in Arabic include, Andrew Gillies *et al.* [11], John Trenkle *et al.* [30] and Maamouri *et al.* [20].

Other than letters, another factor determain the word identity and in many instances can change the meaning and part of speech. This factor is the eight short vowels and diacritics (َ, ِ, ُ, ً, ٌ, ٍ, ً, ٍ). An example for (رجل) is given in the following table where we can see the total change in word category and meaning as a result of adding the diacriticals which resulted in producing three different words in meaning and three different parts of speech for the same three letter رجل :

Word	Meaning	Part of Speech
رَجُلٌ	man	noun (subject)
رَجُلًا	man	noun (object)
رِجْلٌ	foot	noun
رَجَلَ	to go on foot (rather than, e. g., ride a bike)	verb

Never the less, it is always advised that these vowels and diacritics are often normalized before processing in most light stemming or morphological approaches [4]. Mainly the reasons for not including them in the word processing is the claim that they do occur so infrequently, and that in Modern Standard Arabic (MSA), people

tend not to use them and, as a result of that, the meaning is left for the native speaker's intuition, or, in some cases, can be determined from the context. This problem is still waiting for a challenging attempt where the processor is ready to process words with or without diacritics, without needing to normalize words.

Another morphological feature in Arabic is that, unlike Roman letters which are separated naturally, Arabic has an agglutinated nature (as mentioned above) where letters are linked to each other in some cases, while unlinked in some other case, depending on position of the letter in the root, stem and word level. For example, in English the pronoun (he) in (he plays) is separated from the following noun (plays), while in Arabic the pronoun is represented by the letter (ي) which is linked to the root verb لعب to form يلعب (he plays). The same is true when it comes to different kinds of Affixes.

Arabic has four types of affixes. Prefixes: these are letters (normally one) that change the tense of the verb from past to present, such as the letter (ي) in case of the verb لعب and يلعب above. Suffixes: these represent the inflectional terminations (endings) of verbs, as well as, the female and dual/plural markers for the nouns. Postfixes: these are the pronouns attached at the end of the word. Antefixes: these are prepositions agglutinated to the beginning of words.

## 1.2 The Problem at Hand:

This paper is trying to improve the rules for stemming of Arabic texts for data compression. A few different linguistic methods were used by us in the past, for example: the vowel letter method [2]. This method was mainly dependent on syllabification of words and focused on splitting words according to vowel letters. The second approach [8] was a simple approach into stemming rules, where 4 category of words were selected (nouns, verbs, adjectives and adverbs) from short news item texts. These two approaches produced some good results. However, two major problems showed up.

The first problem had to do with parts of speech (POS) recognition problem. For example, the verb يلعب (plays) starts with the letter (ي). In Arabic, adding the suffix (ي) is a very common way to change the word from its past form into its present form. When some rules are set to remove the letter (ي) so to produce the root form of لعب, these rules always removed the letter (ي) from other POS as well, such as the word يمن (Yemen) where the letter (ي) is part of the root word.

The second problem occurs within the sub-POSs when, for example, trying to remove the determiner ال (the definite article 'the') from common nouns as in الطالب (the student). The rules set remove the ال from all nouns including proper nouns such as, ألمانيا (Germany) where the ال is part of the original noun and not a determiner.

For these reasons, in this paper we try to use Stanford online [9] to better categorize the different POS and later to be mismatch the output words -after stemming- with an electronic dictionary.

## 1.3 The Stanford Online Parser

The Stanford parser is a powerful online parser that parses texts in three languages: Arabic, Chinese and English. This parser is using dependency grammar. The Arabic parts of the parser [9] is depending on the Penn Treebank project that was launched in

2001 in the University of Pennsylvania and headed by Prof. Mohamed Maamouri. According to this corpus documentation [10], this corpus is designed for those who study or use languages professionally or academically, as well as, for those who need text corpora in their work. The Penn Arabic Treebank is particularly suitable for language developers, computational linguists and computer scientists who are interested in various aspects of natural language processing.

**Table 1:** English transliteration of Arabic alphabets

Arabic Alphabet	Transliteration	Arabic Alphabet	Transliteration
ا	alif	ع	Ayn
ب	baa	غ	ghayn
ت	ta	ف	faa
ث	tha	ق	qaaf
ج	jiim	ك	kaaf
ح	haa	ل	laam
خ	kha	م	miim
د	daal	ن	nuun
ذ	thal	هـ	haa
ر	raa	ة	taMarboota
ز	zay	و	waaw
س	siin	لا	laamAlif
ش	shiin	ء	hamza
ص	Saad	ئا	hamzaONyaa
ض	Daad	ؤ	hamzaONwaaw
ط	Taa	ي	yaa
ظ	Dhaa	ى	alifMaqsoora

## 1.4 The Arabic Alphabets Transliteration System

In this study, we use a transliteration system for Arabic Alphabets so to enable non-Arabic speakers identify Arabic alphabets and to understand the rules proposed. A legend of Arabic Alphabets and their English transliterations is provided in Table 1.

## 2. Stemming Rules

According to Stanford Online Parser for Arabic language, there are 27 different POSs. In this paper, a number of rules are set for 3 main POSs: nouns, verbs and adjectives as follows:

The rule for every POS or sub-POS is divided into steps as shown below. Every step is to be implemented in the order of numbering:

### Specifications

W – any word or its part (word refers to any POS in the rule: noun, verb, adjective, etc.)

[] – arabic letter

Ins(x, y) – return true when x is anywhere in y

|x| - length of word x

[x]W – letter x is at the beginning of the word

### Nouns Rules:

#### a) DTNN: determiner + singular common noun

**Step 1:** [alif laamAlif laamAlif]W -> [alif laam]W

**Step 2:** [alif laamAlif]Wxy -> [alif laam]Wy

#### b) DTNNP: determiner + singular proper noun

**Step 1:** [alif laam]W -> W

#### c) DTNNS: determiner + plural common noun

**Step 1:** [alif laam]W -> W

#### d) NNPS: common noun, plural or dual

**Step 1:** W[ta] -> W

W[yaa nuun] -> W

**Step 2:** |W| < 5 -> W[taMarboota]

**Step 3:** W[waaw][taMarboota] -> W[taMarboota]

**Verbs Rules:**

**a) VBD: perfect verb (\*\*nb: perfect rather than past tense)**

- Step 1:** |[waaw]W|>2 -> W
- Step 2:** W[alif] -> W  
W[ta] -> W  
W[waaw nuun] -> W
- Step 3:** W[alif haa] -> W[alifMaqsoora]  
W[ta haa] -> W[alifMaqsoora]

**b) VBN: passive verb (\*\*nb: passive rather than past participle)**

- Step 1:** [yaa]W -> W
- Step 2:** |W| = 4 & [ta]W -> [alif]W

**c) VBP: imperfect verb (\*\*nb: imperfect rather than present tense)**

- Step 1:** [ta]W -> W  
[ta ta]W -> W  
[yaa]W -> W
- Step 2:** W[waaw] -> W
- Step 3:** [nuun]W -> W  
[waaw nuun]W -> W  
[haa]W -> W  
[haa alif]W -> W
- Step 4:** |W| = 2 -> W[alifMaqsoora]
- Step 5:** W[yaa] -> [alif]W[alifMaqsoora]
- Step 6:** [siin]W & ins(W, [ta]) -> [alif][siin]W
- Step 7:** W[waaw laam] -> W[alif laam]  
W[waaw laam waaw nuun] -> W[alif laam]  
W[waaw nuun] -> W[alif laam]
- Step 8:** [nuun][ta]W & |[nuun][ta]W| > 3 -> [nuun]W

**Adjectives Rules:**

**a) DTJJ: determiner + adjective**

- Step 1:** [alif laam]W -> W
- Step 2:** W[taMarboota] -> W

**3. Experiments**

The suggested rules must be tested against real data. For this purpose, we use some news articles, from the BBC Arabic and Al Jazeera Arabic news portals. These articles are parsed by Stanford Online Parser and the results are shown in table 2. In the

following table, repeated words are deleted and sample words of every POS or sub-POS are shown in the table.

**Table 2.** List of words used in our experiments

<b>Nouns</b>
<b>a) DTNN: determiner + singular common noun</b>
البحر البنزين البيئة الجرف الحد الدفاع السيناتورين الشرق الشؤون البر المصادر السابق الحكوم المحيط النفط الوصول العالم الاحتياطي التنقيب العام الدولة الابار الاحتياطي الامر الامور الاميال الانسان بالكلاب الطاقة الطلب العواصف المثبت المحيط المشاغل الموارد المياه النفط الوحل الانواع البيئة العاصمة الاخطار الفصائل البيئة الدردشة السهل
<b>b) DTNNP: determiner + singular proper noun</b>
الانترنت البرازيل الدوحة العاج المكسيك
<b>c) DTNNS: determiner + plural common noun</b>
الامريكيبين الاولويات الجمهوريون الديموقراطيون الشركات العشرات لتحقيق للحماية المحافظون المشتريين المعاملات الملوثات المنتجات المندوبون المنصات الولايات
<b>d) NNPS: common noun, plural or dual</b>
تغييرات تقديرات جماعات سنوات طبقات طنين عشرات عمليات كميات مجتمعات مقترحات منصات
<b>Verbs</b>
<b>a) VBD: perfect verb (**nb: perfect rather than past tense)</b>
اجراها اجرته ادى استخرجت اصبح اصيحت اقترع ايد بامكاننا تسبب تسربا جعلت فقد كان مضى وتزيد وقال وقد وقدمت ويضيفون ويقول
<b>b) VBN: passive verb (**nb: passive rather than past participle)</b>
يوجد تسحب يرجح تستخدم يدرج
<b>c) VBP: imperfect verb (**nb: imperfect rather than present tense)</b>
تبدو تتنافس تتوزع تتوقف تتيحها تحتوي تربض تستخرج تسجل تشارك تشير تصبح تطا تطفو تعد تلاحظه تنتج تهدد يبدو يبلغ يتعرض يتفوق يتم يتناولها يحاولون يحتوي يخبئني يخرجها يزال يساهم يستحق يستفيد يسمون يشكل يصبح يقترع يقع يقول يقولون يكاد يكون يمارس يمثل يمكن ينتشر ينتهي ينجح يهدد
<b>Adjectives</b>
<b>a) DTJJ: determiner + adjective</b>

الاستقلالية	الاشعاعية	الامريكي	الامريكية	الاولى	البرية	البيئية
التجارية	الجاري	الجديد	الحمراء	الحيوانية	الخارجي	الداخلية
الدولي	الدولية	الطبيعية	العالمي	العالمية	القاري	القانونية
القطبية	القطرية	المتبقية	المتحدة	المحلية	المحمية	المرجانية
المهددة	المهيمن					
النادرة	النفطية	الواسعة				

Before any rule is applied, all words must be normalized and preprocessed. We store all words in plain text files using codepage 1256 – Arabic. Because all our software is written in C+, we read these text files into Unicode representation.

Our results for the nouns list are depicted in Tables 3, 4, 5 and 6. The results for the noun rules produced very good results in case of DTNNP and DTNN. Very few undesirable results were produced because some words were wrongly parsed by the parser such as (بالكلاب). As for DTNNS, some more rules needed to deal with the plural and dual suffixes. NNPS produced very good results.

**Table 3.** Processed Nouns - DTNN: determiner + singular common noun

حكوم	سابق	مصادر	بر	شؤون	شرق	سيناتور	دفاع	حد	جرف	بيئة	بنزين	بحر
طاقة	بالكلاب	مر	ميل	البر	دولة	عام	تنقيب	عالم	وصول	نفط	محيط	
عاصمة	بيئة	النوع	وحل	نفط	مياه	موارد	مشاغل	محيط	مثبت	عواصف	طلب	
دردشة	سهل	بيئة	فصائل	الخطر								

**Table 4.** Processed nouns - DTNNP: determiner + singular proper noun

عاج	مكسيك	دوحة	برازيل	انترنت
-----	-------	------	--------	--------

**Table 5.** Processed nouns - DTNNS determiner + plural common noun

لتحقيق	عشرات	شركات	ديموقراطيون	جمهوريون	اولويات	امريكبين
منصات	ولايات	مندوبون	منتجات	ملوثات	معاملات	مشتريين
						محافظون
						للحماية

**Table 5.** Processed Nouns -NNPS: common noun, plural or dual

مقترح	منصة	مجتمع	كمية	عملية	عشرة	طنة	طبقة	سنة	جماعة	تقدير	تغيير
-------	------	-------	------	-------	------	-----	------	-----	-------	-------	-------

The verbs' rules results are depicted in Tables 7, 8 and 9. The verbs' rules produced good results in case of VBD and VBN. However, in case of VNP, a few bad results show up and the rules have to be enhanced in the future.

**Table 7.** Processed Verb - VBD: perfect verb

جعلت	تسرب	تسبب	بامكانن	ايد	اقترع	اصبح	اصبح	استخرج	ادى	اجرى	اجرى
يضيف	يقول	قدمت	قد	قال	تزيد	مضى	كان	فقد			

**Table 8.** Processed verbs – VBN: passive verb

درج استخدم رجح سحب وجد
------------------------

**Table 9.** Processed Verb – VNP: imperfect verb

طای اصبح شارک سجل استخراج ریض احتوی تیجها توقف توزع تنافس بدی ختبئی احتوی حال تناولها تمی تفوق تعرض بلغ بدی تجی لاحظ عدی طفی کال کاد قال قال قعی اقترع اصبح شکل استفيد استحق ساهم زال خرجها مکن مثل مارس
---

The results for the adjectives' rules are depicted In Table 10. Almost all rules made for adjectives produced successful results.

**Table 10.** Processed adjectives - DTJJ

جدید جاری تجاری بیئی بری اولی امریکی امریکی اشعاعی استقلالی قاری عالمی عالمی طبیعی دولی دولی داخلی خارجی حیوانی حمراء نادر مهیمن مهدد مرجانی محمی محلی متحد متبقي قطري قطبي قانونی نفطی واسع
---

#### 4. Conclusion

In this paper we set rules for POS and to parse our training data, we used Stanford Online Parser for Arabic language, which identifies 27 different POSs. In this paper, the rules set are for 3 main POSs: nouns, verbs and adjectives. Every rule for every POS or sub-POS is divided into one or more steps.

The results for the noun rules produced very good results in case of DTNNP and DTNN. Very few undesirable results occur because some words were wrongly parsed by the parser such as (بالکلاب). As for DTNNS, some more rules needed to deal with the plural and dual suffixes. NNPS produced very good results. The verbs' rules results are depicted in Tables 7, 8 and 9. The verbs' rules produced very good results in case of VBD and VBN. However, in case of VNP, a few bad results show up and the rules have to be enhanced in the future. The results for the adjectives's rules are depicted In Table 10. Almost all rules made for adjectives produced very good results. Most errors occurred in case of VBP. However, the overall evaluation of these rules proved that the rules produced very good results. In the future, these rules must be improved and enhanced to include more POSs and should be tested against wider variety of vocabulary and bigger corpora.

**Acknowledgments:** This work was partially supported by the Grant Agency of the Czech Republic under grant no. P202/11/P142, SGS in VSB – Technical University of Ostrava, Czech Republic, under the grant No. SP2013/70, and has been elaborated in the framework of the IT4Innovations Centre of Excellence project, reg. no. CZ.1.05/1.1.00/02.0070 supported by Operational Programme 'Research and Development for Innovations' funded by Structural Funds of the European Union and state

budget of the Czech Republic and by the Bio-Inspired Methods: research, development and knowledge transfer project, reg. no. CZ.1.07/2.3.00/20.0073 funded by Operational Programme Education for Competitiveness, co-financed by ESF and state budget of the Czech Republic.

## References

1. Encyclopedia Britannica Online. Alphabet. Online (2011). URL: <http://www.britannica.com/EBchecked/topic/17212/alphabet>
2. H. Soori, J. Platos, V. Snasel, H. Abdulla, in *Digital Information Processing and Communications, Communications in Computer and Information Science*, vol. 188, ed. By V. Snasel, J. Platos, E. El-Qawasmeh (Springer Berlin Heidelberg, 2011), pp. 97{105. URL [http://dx.doi.org/10.1007/978-3-642-22389-1\\_9](http://dx.doi.org/10.1007/978-3-642-22389-1_9) 10.1007/978-3-642-22389-1\_9
3. T. Buckwalter, in *Arabic Computational Morphology, Text, Speech and Language Technology*, vol. 38, ed. by N. Ide, J. Veronis, A. Soudi, A.v.d. Bosch, G. Neumann (Springer Netherlands, 2007), pp. 23{41. URL [http://dx.doi.org/10.1007/978-1-4020-6046-5\\_3](http://dx.doi.org/10.1007/978-1-4020-6046-5_3) 10.1007/978-1-4020-6046-5\_3
4. N.Y. Habash, *Synthesis Lectures on Human Language Technologies* 3(1), 1 (2010). DOI 10.2200/S00277ED1V01Y201008HLT010. URL <http://www.morganclaypool.com/doi/abs/10.2200/S00277ED1V01Y201008HLT010> (last accessed 10/12/2012)
5. A. Gillies, E. Erl, J. Trenkle, S. Schlosser, in *Proceedings of the Symposium on Document Image Understanding Technology* (1999)
6. J. Trenkle, A. Gilles, E. Eriandson, S. Schlosser, S. Cavin, in *Symposium on Document Image Understanding Technology* (2001), pp. 159{168
7. M. Maamouri, A. Bies, S. Kulick, in *Proceedings of the British Computer Society Arabic NLP/MT Conference* (2006).
8. Soori, H. , Platoš, J. , Snášel, V.: Simple stemming rules for Arabic language, *Advances in Intelligent Systems and Computing*, Volume 179 AISC, 2012, Pages 99-108, ISBN: 978-364231602-9
9. Spence Green and Christopher D. Manning. 2010. Better Arabic Parsing: Baselines, valuations, and analysis. In *23rd Conference on Computational Linguistics*, pages 394–402, Beijing, China.
10. <http://www.ircs.upenn.edu/arabic/Jan03release/README.txt> (last accessed 10/03/2013)
11. <http://www.un.org/> (last accessed 10/03/2013)