Linguistically Informed and Corpus Informed Morphological Analysis of Arabic

Majdi Sawalha and Eric Atwell
School of Computing
University of Leeds, Leeds, LS2 9JT, UK
sawalha@comp.leeds.ac.uk; eric@comp.leeds.ac.uk

Abstract

Standard English PoS-taggers generally involve tag-assignment (via dictionary-lookup etc) followed by tag-disambiguation (via a context model, e.g. PoS-ngrams or Brill transformations). We want to PoS-tag our Arabic Corpus, but evaluation of existing PoS-taggers has highlighted shortcomings; in particular, about a quarter of all word tokens are not assigned a fully correct morphological analysis. Tag-assignment is significantly more complex for Arabic. An Arabic lemmatiser program can extract the stem or root, but this is not enough for full PoS-tagging; words should be decomposed into five parts: proclitics, prefixes, stem or root, suffixes and postclitics. The morphological analyser should then add the appropriate linguistic information to each of these parts of the word; in effect, instead of a tag for a word, we need a subtag for each part (and possibly multiple subtags if there are multiple proclitics, prefixes, suffixes and postclitics).

Many challenges face the implementation of Arabic morphology, the rich "root-and-pattern" nonconcatenative (or nonlinear) morphology and the highly complex word formation process of root and patterns, especially if one or two long vowels are part of the root letters. Moreover, the orthographic issues of Arabic such as short vowels (ً ٌ ٍ), Hamzah (ئ ؤ أ إ ء), Taa' Marboutah (ة) and Ha' (ه), Ya' (ي) and Alif Maksorah(ى) , Shaddah (ّ) or gemination, and Maddah (آ) or extension which is a compound letter of Hamzah and Alif (أا).

Our morphological analyzer uses linguistic knowledge of the language as well as corpora to verify the linguistic information. To understand the problem, we started by analyzing fifteen established Arabic language dictionaries, to build a broad-coverage lexicon which contains not only roots and single words but also multi-word expressions, idioms, collocations requiring special part-of-speech assignment, and words with special part-of-speech tags. The next stage of research was a detailed analysis and classification of Arabic language roots to address the "tail" of hard cases for existing morphological analyzers, and analysis of the roots, word-root combinations and the coverage of each root category of the Qur'an and the word-root information stored in our lexicon. From authoritative Arabic grammar books, we extracted and generated comprehensive lists of affixes, clitics and patterns. These lists were then cross-checked by analyzing words of three corpora: the Qur'an, the Corpus of Contemporary Arabic and Penn Arabic Treebank (as well as our Lexicon, considered as a fourth cross-check corpus). We also developed a novel algorithm that generates the correct pattern of the words, which deals with the orthographic issues of the Arabic language and other word derivation issues, such as the elimination or substitution of root letters.

1 Introduction[1]

Morphological analysis is the process of assigning the morphological features of a word such as; its root or stem, the morphological pattern of the word, the morphological attributes of the word (part-of-speech of the word whether it is noun, verb or particle). It also involves

specifying the number of the word (singular, dual or plural), the case or mood (nominative, accusative, genitive or jussive). Moreover, it identifies the internal structure of the word such as prefixes, suffixes, clitics and the root or stem.

Generally, there are four main methodologies for developing robust morphological analyzers are: First, the syllable-based Morphology (SBM), which depends on analyzing the syllables of the word. Second, Root-Pattern Methodology depends on the root and the pattern of the word for analysis. Using this method, the root of the word is extracted by matching the word with lists of patterns and affixes. Third, Lexeme-based Morphology where the stem of the word is the crucial information to be extracted from the word. Finally, stem-based Arabic lexicon with grammar and lexis specifications, where stem-grounded lexical databases with entries associated with grammar and lexis specifications, is the most appropriate organization for the storage of Arabic lexical information (Soudi et al, 2007). All these methodologies use pre-stored lists of root, stems, patterns and affixes and grammar and linguistic information encoded with the analyzers. A fifth methodology is using tagged corpora and computer algorithms to build morphological database of the tagged words.

Statistical approaches to stemming have been widely applied to automatic morphological analysis in the field of computational linguistics. Some stemming techniques match the best set of frequently occurring stems and suffixes using information theoretic measures. Some consider the most frequently occurring word-final n-grams to be suffixes. Such systems cannot be expected to perform well on Arabic language in which suffixing is not the only inflectional process. (Larkey et al, 2002)

Some statistical approaches to Arabic language analysis combine word-based and 6-gram based retrieval which performs remarkably well for many languages including Arabic. Another approach is to use clustering on Arabic words to find classes sharing the same root; such clustering is based on morphological similarity using a string similarity metric tailored to Arabic morphology, which is applied after removing “a small number of obvious affixes” (Larkey et al, 2002).

Tim Buckwalter morphological analyzer is one of the most widely used morphological analyzer of Arabic, it uses pre-stored dictionaries of words, stem and affixes constructed manually. It also uses truth tables to determine the correct combinations of prefixes, stem, and suffixes of the word (Thabet, 2004) (Buckwalter, 2004).

An example of root extraction algorithms is Khoja’s Stemmer. This stemmer removes the longest prefix and suffix of the word, then it matches the processed word with lists of noun and verb patterns to extract the correct root of the word. The stemmer has many encoded useful information sources such as: list of diacritics, list of punctuation marks, list of tri-literal and quad-literal roots, list of definite articles and a list of 168 stop words. Khoja’s stemmer has been used in information retrieval applications and it achieved good results which improved results of information retrieval systems, in spite of the mistakes generated (Khoja, 2001) (Larkey & Connell, 2001).

Al-Shalabi et al (2003) have developed a root extraction algorithm for tri-literal roots of Arabic words which does not depend on any pre-stored information. It depends on mathematical calculations of weights assigned to the letters of the word, then multiplying these weights with the position of the letters in the word. Higher weights are assigned to the letters at the beginning and at the end of the word. Then the algorithm selects the letters with lower weights as root letters. They classified the Arabic letters into two groups; the first group is the letters that do not appear in any affix and they assigned the weight (0) to this group, and the second contains letters that appear in affixes, grouped in the word (سألتمونيها), and they assigned different weights to these letters.

2 Arabic Corpora

We used four corpora to study Arabic language roots to address the "tail" of hard cases for existing morphological analyzers, and analysis of roots, word-root combinations and the coverage of each root category in the Qur'an and the word-root information stored in the broad-lexical resource. Moreover, theses corpora are used to cross-check the comprehensive lists of affixes and clitics, by analyzing words of the corpora.

The corpora used are The Qur'an, the Corpus of Contemporary Arabic, the Penn Arabic Treebank and a collection of 15 traditional Arabic dictionary texts. The Qur'an is a special type of corpus of classical Arabic text, which consists of about 78,000 words and about 19,000 vowelized word types and about 15,000 non-vowelized word types. Second, the Corpus of Contemporary Arabic; is a modern Arabic text corpus consisting of 1 million words: the corpus was constructed from magazines and newspaper texts from 14 genres: Autobiography, Short Stories, Children's Stories, Economics, Education, Health and Medicine, Interviews, Politics, Recipes, Religion, Sociology, Science, Sports, Tourist and Travel and Science (Al-Sulaiti & Atwell, 2006). Third, the Penn Arabic Treebank consists of 734 files representing roughly 166,000 words of written Modern Standard Arabic newswire from the Agence France Presse corpus (Maamouri & Bies, 2004). Finally, the text of 15 traditional Arabic language dictionaries can be considered as our fourth corpus. The texts consist of about 11 million words and 2 million word types of both modern and classical Arabic text. The lexicons have been developed over 1,400 years. Figure 1 shows a sample of text taken from the lexicons corpus. Figures 1b, 1c show the google machine translation and the human translation of the sample. Figure 1d is a sample of the Arabic-English lexicon by Edward Lane (Lane, 1968) volume 7, pages 117-119.

Lexicography is the applied part of lexicology. It is concerned with collating, ordering of entries, derivations and their meaning depending on the aim of the lexicon to be constructed and its size. Lexicography is one of the original and deep-rooted arts of Arabic literature. The first lexicon constructed was "*mu'jam al-'ain*" " معجم العين " *al-'ain Lexicon* by *al-farāhydy* (died in 791). Over the past 1200 years, many different kinds of Arabic language lexicons were constructed; these lexicons are different in ordering, size and aim or goal of construction. Many Arabic language linguists and lexicographers studied the construction, development and the different methodologies used to construct these lexicons.

Lexicographers constructing the first Arabic language lexicons are the pioneers in lexicography and lexicon construction. They designed comprehensive lexicography rules. According to these rules and methodologies, Arabic lexicons can be mainly classified into two classes. The first class depends on the meaning of the words or subject to group similar words together; such as, *al-ḡaryb al-muṣnnaf fi al-luḡah* "الغريب المُصنّف في اللغة" The Irregular Classified Language by *abi 'ubayd al-qāsim bin sallām* and "*al-muẖaṣṣaṣ*" " المخصص " The Specified by *ibn sayydah*. The second class depends on the word itself and developed its rules depending on phonology; lexicons were ordered according to the first letter of the words. This class has different ordering methods of lexical entries.

Another classification of Arabic language lexicon distinguishes between four classes of ordering lexical entries in the lexicon. *al-ẖalyl* methodology was developed by *al-ẖalyl bin aḥmad al-farāhydy* (died in 791). His lexicon is called *kitāb al-'ain* "كتاب العين". The *al-'ain* lexicon lists the lexical entries phonologically according to exits of letters sounds from the mouth and throat, from the farthest letter exit to the nearest. The second methodology, *abi 'ubayd* Methodology is developed by *abi 'ubayd al-qāsim bin sallām* " أبي عُبيدٍ القاسم بن سلاَّم" (died in 838). His rules for construction of lexicons depend on the meaning or subjects. He organized his lexicon into chapters and sections for lexical entries that are similar in meaning like a

thesaurus. *abi 'ubayd* wrote many small books, each of which describes one subject or meaning, such as books describing horses, milk, honey, flies, insects, palms, and human creation. Then he collated all these small books into one large lexicon called *al-ḡaryb al-muṣnaf fi al-luḡah* "الغريب المُصنّف في اللغة" The Irregular Classified Language. The third methodology, *al-jawhary* methodology was developed by *'ismā'yl bin ḥammād al-jawhary* (died in 1002) and his lexicon is called *aṣ-ṣiḥāḥ fy al-luḡah* " الصحاح في اللغة" The Correct Language; this uses alphabetical order for ordering the lexical entries. However, he arranged the lexical entries of his lexicon depending on the last letter of the word, and then the first letter. His lexicon was organized into chapters where each chapter corresponds to the last letter of the word. Each chapter includes sections corresponding to the first letter of the word. e.g. the word " بَسَطَ " "*baṣaṭ*" is found in chapter 'ط' 'ṭ' as it represents the last letter of the word, then by looking to section 'ب' 'b' as it represents the first letter. Finally, the *al-barmaky* methodology was developed by *abu al-ma'āly Moḥammed bin tamym al-barmaky* "أبو المعالي محمد بن تميم البرمكي", who lived in the same time period as *al-jawhary*. *al-barmaky* did not construct a new lexicon; but he alphabetically re-arranged a lexicon called *aṣ-ṣiḥāḥ fy al-luḡah* " الصحاح في اللغة" The Correct Language by *al-jawhary*. He added little information to that lexicon. After that, *al-zamaẖšary* "الزمخشري" (died in 1143) followed the same methodology and he constructed his lexicon called "*'asās al-balāḡah*" " أساس البلاغة" Fundamentals of Fluency (al-jawhary, died 1002).

كتب: الكِتابُ: معروف، والجمع كُتُبٌ وكُتْبٌ. كَتَبَ الشيءَ يَكْتُبه كَتْباً وكِتاباً وكِتابةً، وكَتَّبَه: خطَّه؛ قال أَبو النجم: أَقْبَلْتُ من عِنْدِ زِيادٍ كالخَرِفْ، تَخُطُّ رِجْلايَ بخَطٍّ مُخْتَلِفْ، تُكَتِّبانِ في الطَّرِيقِ لامَ أَلِفْ قال: ورأَيت في بعض النسخ تِكَتِّبانِ، بكسر التاء، وهي لغة بَهْرَاءَ، يَكْسِرون التاء، فيقولون: تِعْلَمُونَ، ثم أَتْبَعَ الكافَ كسرةَ التاء.والكِتابُ أَيضاً: الاسمُ، عن اللحياني. الأَزهري: الكِتابُ اسم لما كُتب مَجْمُوعاً؛ والكِتابُ مصدر؛ والكِتابةُ لِـمَنْ تكونُ له صِناعةً، مثل الصِّياغةِ والخِـياطةِ. والكِتْبةُ: اكْتِبائُك كِتاباً تنسخه. ويقال: اكْتَتَبَ فلانٌ فلاناً أَي سأَله أَن يَكْتُبَ له كِتاباً في حاجة. واسْتَكْتَبه الشيءَ أَي سأَله أَن يَكْتُبَه له. ابن سيده: اكْتَتَبَه كَكَتَبَه. وقيل: كَتَبَه خَطَّه؛ واكْتَتَبَه: اسْتَمْلاه، وكذلك اسْتَكْتَبَه. واكْتَتَبه: كَتَبه، واكْتَتَبْته: كَتَبْته. وفي التنزيل العزيز: اكْتَتَبَها فهي تُمْلى عليه بُكْرَةً وأَصِـيلاً؛ أَي اسْتَكْتَبَها. ويقال: اكْتَتَبَ الرجلُ إِذا كَتَبَ نفسَه في دِيوانِ السُّلْطان. وفي الحديث: قال له رجلٌ إِنَّ امرأَتي خَرَجَتْ حاجَّةً، وإِني اكْتُتِبْت في غزوة كذا وكذا؛ أَي كَتَبْتُ اسْمِي في جملة الغُزاة. وتقول: أَكْتِبْنِي هذه القصيدةَ أَي أَمْلِها عليَّ. والكِتابُ: ما كُتِبَ فيه. وفي الحديث: مَن نَظَرَ في كِتابِ أَخيه بغير إِذنه، فكأَنما يَنْظُرُ في النار؛ قال ابن الأَثير: هذا تمثيل، أَي كما يَحْذر النارَ، فَلْـيَحْذَرْ هذا الصنيعَ، قال: وقيل معناه كأَنما يَنْظُر إِلى ما يوجبُ عليه النار؛ قال: ويحتمل أَنه أَرادَ عُقوبةَ البَصرِ لأَن الجناية منه، كما يُعاقَبُ السمعُ إِذا اسْتَمع إِلى قوم، وهم له كارهُونَ؛ قال: وهذا الحديث محمولٌ على الكِتابِ الذي فيه سِرٌّ وأَمانة، يَكْرَه صاحبُه أَن يُطَّلَع عليه؛ وقيل: هو عامٌّ في كل كتاب.

Figure 1a: A sample of text from the traditional Arabic lexicons corpus

Books: Book: well known, the combination of books and books. What books written books and books and writing, and written by: the plan; Abu star: coming from when Ziad Kkherv, adopt various Rgelai handwriting, written in the L. A. said: I saw written in some copies, breaking the sound, the language of Behra, breaking sound, say : you know, and then follow the Kef sound fragment. The book is also: the name of the Alalehyani. Azhari: the name of the book for a total of books; the source of the book; and write to those who have the industry, such as drafting and sewing. And clerks: Akttabk copy book. It is said: subscribed Flana any person asked to write a book in need. Astketbh any thing and asked him to write it. The son of his master: Aktaatbh Kketbh. It was: written by the plan; and Aktaatbh: Astmlah, as well as Astketbh. And Aktaatbh: clerks, Aktaatpth: written. In the download-Aziz: Aktaatbha are dictated by the wheel and integral; any Astketbha. It is said: If men have subscribed the same in the office of the Sultan. In the modern: a man said to him that my wife needed her, and I subscribed to as well as in the conquest, as well as; wrote my name in any other invaders. She says: Oketbni this poem on any hope. The book is: what has been written in it. In the modern: its consideration in the book his brother without his permission, as if seen in the fire; Ibn al-Atheer said: This representation, also warns of any fire, let him beware of doing this, he said: It was meant to be considered if required by the fire; said: It is possible that he wanted the death the sight of it because the crime, and punished if the hearing heard people, who disliked him; he said: This hadeeth portable book in which the secret and the secretariat, to inform the owner hates it; and it was said: It is common in every book.

Figure 1b: (Google) Machine translation of the sample of text from the traditional Arabic lexicons corpus

k t b: [*Alkitab*] the book; is well known. The plural forms are [*kutubun*] and [*kutbun*]. [*kataba Alshay'*] He wrote something. [*yaktubuhu*] the action of writing something. [*katban*], [*kitaban*] and [*kitabatan*] means the art of writing. And [*kattabahu*] writing it means draw it up. Abu Al-Najim said: I returned back from Ziyad's house [after meeting him] and behaved demented, my legs drawn up differently (means walking in a different way). They wrote [*tukattibani*] on the road the letters of *Lam Alif* (describing how he was walking crazily and in a different way). He said: I saw in a different version, the word "they wrote" [*tikittibani*] using the short vowel *kasrah* on the first letter [taa], as it is used by Bahraa' [Arab tribe] dialect. They say: [ti'lamuwn] (you know). Then the short vowel kasrah is propagated to the following letter (kaf). Moreover, [Alkitab] the book is a noun. Al-lihyani Al-Azhari definition is: [Alkitab] The book is the name of a collection of what has been written (a collection of written materials or texts). And the book has gerund [Alkitabatu] writing (art of writing) for whoever has a profession, similar to drafting and sewing. And [Alkitabatu]: is copying a book [copying a book in several copies]. It is said: [iktataba] someone subscribed another means; he asked to write him a letter in something. [istaktabahu] He dictated someone something means to write him something. Ibn Sayyedah: [Iktatabahu] is similar to [katabahu]. It is said: [katabahu] write something down means draw up. And [Iktatabahu] writing something down means dictate someone something, which is the same meaning of [Istaktabahu]. [Iktatabahu] registering (masculine), and [Iktatabathu] registing (feminine). In the Qur'an: [Iktatabaha] He registered it, he has dictated it every sunrise and sunset, which means dictating it. It is said: [Iktataba Al-rajul] The man registered, if he registered himself in the Sultan's office. In Hadith: a man said to him (the prophet): my wife is pilgrimaging (to Mecca), and I have registered [Oktutibtu] in a conquest, which means that I have written my name among the conquerors. And you say: [Aktibny] let me copy this poem, means dictate me the poem. Also, [Alkitab] the book is something which has been written on. And in Hadith: who looks at his brother's book without permission is as looking to hell. Ibn Al-Atheer said: it is a similarity; which means as he avoids hell, he should avoid doing this. He said: the meaning (of the Hadith) is the punishment by hell will be applied if someone looks at a book without permission. He said: it might be the punishment of visual explorers as the crime is done by sight. Hearing explorer is punished if someone intentionally listened to other people who do not like anyone to listen to them. He said: this Hadith is specific for books of secrets and secure books, whose owners hate anybody to look at these books. It is also said: the Hadith is general; applied to any type of books.

Figure 1c: A Human translation of the sample of text from the traditional Arabic lexicons corpus

كتب

1. كَتَبَهُ, aor. ـُ, inf. n. كَتْبٌ and كِتَابٌ and كِتَابَةٌ (S, K) and كُتْبَةٌ; (Msb;) the first of these inf. ns. agreeable with analogy; the second, anomalous; (TA;) or the latter of these two is a subst., like لِبَاسٌ; (Lh;) or originally an inf. n., and afterwards used in the senses given below; (MF;) as also كِتَابَةٌ, and كُتْبَةٌ: (TA:) and ↑ كتّبهُ (K) and ↑ اكتتبهُ; (S, K;) *He wrote it:* (S, K:) or كَتَبَهُ has this signification; and ↑ اكتتبهُ, as also ↑ استكتبهُ, signifies *he asked* [one] *to dictate it* (إِسْتَمْلَاهُ): (K:) ↑ إِكْتَتَبَهَا in the Kur, xxv. 6, signifies *he hath written them* (S) *for himself:* (Bḍ:) or *he hath asked* [one] *to write them for him, or to dictate them* to him. (TA, Bḍ.) — كَتَبَ عَنْهُ [*He wrote what he had heard, or learned from him.*] A phrase of common occurrence in biographies. — كَتَبَ

4. اكتب *He dictated.* (S, K.) Ex. أَكْتِبْنِي هٰذِهِ القَصِيدَةَ *Dictate to me this ode.* (S.) — اكتب and ↑ كتّب *He taught the art of writing.* (K.) = See also 1, in three places.

5. تكتّب ‡ *He girded himself, and drew together his garments upon him.* (TA.) = تكتّب ‡ *It* (an army, S) *collected itself together.* (S, K.)

8. See 1. — كِتْبَةٌ [is a quasi-inf. n. of 8; *syn. with* إِكْتِتَابٌ; and is explained as signifying] The *writing a book, transcribing it* [*from another book*]: (إِكْتِتَابُكَ كِتَابًا تَنْسَخُهُ). (K.) — It also signifies, [as a quasi-inf. n. of 8,] The *writing one's name in* [*the list of those who receive*] *stipend and maintenance* (الاكتتاب فى الفرض والرزق [اصحاب]). (TA.) — اكتتب *He registered himself in the book of the Sultán's army-list, or stipendiaries.* (S, K.) إِكْتَتَبْتُ فِى غَزْوَةِ كَذَا *I wrote down my name in the list of the soldiers of such an expedition.* (TA, from a trad.) — اكتتب كِتَابًا *He asked for a book* (or *the like*) *to be written for him.* (TA.) See also 10. = اكتتب ‡ *His urine was suppressed.* (TA.)

10. استكتبهُ شَيْئًا *He asked him to write a thing for him.* (S.) See also 1 and 8. = With reference to a سِقَاء (or skin), see 1.

كِتَابٌ [inf. n. of 1, q. v. — as a subst.,] *A thing in which, or on which, one writes:* [*a book:*] *a written piece of paper* or [*a record, or register;* and *a written mandate;*] *of skin:* (K:) *a writing,* or *writ,* or *thing written;* as also ↑ كَتِيبَةٌ: and both are applied also to the *revelation from above:* and to *a letter,* or *epistle, which a person writes and sends:* sometimes made fem., as meaning صَحِيفَةٌ

كِتَابَةٌ subst. from 1; signifying The *art of writing.* (IAạr, Msb.) — See also 3.

كَتِيبَةٌ see كِتَابٌ. = *An army; a military force:* (S, K:) or *a collected portion thereof;* (Msb;) [*a body of troops; a corps:*] or *a troop:* or *a troop of horse making a hostile attack* or *incursion, in number from a hundred to a thousand:* (K:) pl. كَتَائِبُ. (S.)

كُتَّابٌ, see مَكْتَبٌ. = The same, (S, K,) as also كُتَّابٌ, q. v., but the former is the more approved: (S: the reverse, however, is said in the TA; and MF says that some authors altogether reject كتّاب, with ت, in the sense here following:) *A kind of small, round-headed, arrow, with which boys learn to shoot.* (S, K.)

كَاتِبٌ [*A writer; a scribe; a secretary*]: pl. كَاتِبُونَ and كُتَّابٌ and كَتَبَةٌ. (S, K.) — *A learned man* (S, K) was so called by the Arabs, (IAạr,) because, in general, he who knew the art of writing was possessed of science and knowledge; and writers among them were few. (TA.)

مَكْتَبٌ (S, K) and ↑ كُتَّابٌ (Lth, S, &c.) *A school; a place where the art of writing is taught:* (S, K, &c.:) accord. to Mbr and F, the assigning this signification to the latter word is an error; it being a pl. of كَاتِبٌ, and signifying, accord. to Mbr, the *boys of a school:*

Figure 1d: A Sample of the definition of the root *ktb* from an Arabic-English Lexicon by Edward Lane, http://www.tyndalearchive.com/TABS/Lane/

3 Arabic Morphological Analyzer

Our main aim of developing a morphological analyzer is to build a tagged Arabic corpus. We stared our research by comparing existing morphological analysers, stemmers and root extraction algorithms, which are freely available for researchers and users. Our study was limited to three of them. These analyzers are: Tim Buckwalter morphological analyzer, Khoja's stemmer, and Tri-literal root extraction algorithm developed by Al-Shalabi and others. A gold standard for evaluation has been developed to compare the results of the different systems and report their accuracy. The gold standard contains two 1000-word documents: the first is taken from chapter 29 of the Qur'an (سورة العنكبوت) (The Spider). The second is a newspaper text document taken from the Corpus of Contemporary Arabic (Al-Sulaiti & Atwell, 2006). We manually extracted the roots of the words in these documents, and had these checked by Arabic language scholars. The results of the three algorithms were compared to their equivalents in the gold standard. The accuracy of these algorithms was computed using four different accuracy measurements. The study showed that the best algorithm failed to achieve an accuracy rate of more than 75%. This proves that more research is required. We can not rely on existing stemming algorithms for further research such as Part-of-Speech tagging and then Parsing because errors from the stemming algorithms will propagate to such systems, accuracy is vital for them. (Sawalha & Atwell, 2008).

3.1 Analytical study of tri-literal roots of Arabic

To understand the nature of Arabic roots, and the derivation process of words from their roots, we classified the tri-literal roots into 22 groups depending on the internal structure of the root itself; whether it contains only consonant letters, Hamza, or defective letters. We studied words and roots of the Qur'an, which contains 45,534 tri-literal root words, and a broad-lexical resource constructed by collecting 15 Arabic language lexicons, which gave us 376,167 word types which are derived from tri-literal roots. Tables 1 & 3 show the results of all root categories. The results show that 68% of the tri-literal roots of Qur'an are intact roots (intact, doubled and contains Hamza), and 61% of the words which are derived from tri-literal roots, belongs to this category. 29% of the tri-literal roots of Qur'an are defective roots (contains one or two vowels in its root) and the percentage of the words belong to this category is 32% of the words of the Qur'an. The third category contains one or two vowels and Hamza in its root. The percentage of tri-literal roots of the Qur'an is 3%, and 7% of the words of the Qur'an belong to this category. Table 2 and figure 2 show these results.

					Roots		Tokens	
	Category				**count**	**Percentage**	**count**	**Percentage**
1	Intact	C1	C2	C3	870	54.04%	20,007	43.94%
2	Doubled	C1	C2	C2	136	8.45%	3,814	8.38%
3	First Letter Hamza	H	C2	C3	44	2.73%	3,243	7.12%
4	Second letter Hamza	C1	H	C3	15	0.93%	281	0.62%
5	Third Letter Hamza	C1	C2	H	32	1.99%	459	1.01%
6	First letter Defective	V	C2	C3	70	4.35%	1,252	2.75%
7	Second Letter Defective	C1	V	C3	198	12.30%	8,162	17.93%
8	Third Letter Defective	C1	C2	V	167	10.37%	3,584	7.87%
9	Separated Mixed Defective	V	C2	V	12	0.12%	710	1.56%
10	Adjacent Mixed defective 1	C1	V1	V2	19	1.18%	473	1.04%
11	Adjacent Mixed defective 2	V1	V2	C3	2	0.12%	445	0.98%
12	First Letter Hamza and Doubled	H	C2	C2	7	0.43%	175	0.38%
13	First letter Defective and Doubled	V	C2	C2	2	0.12%	40	0.09%
14	First letter Hamza and third letter Defective	H	C2	V	13	0.81%	958	2.10%
15	First letter Hamza and second letter Defective	H	V	C3	6	0.37%	153	0.34%
16	Adjacent Mixed defective with Hamza	H	V1	V2	2	0.12%	418	0.92%
17	Second letter Hamza and Third letter Defective	C1	H	V	2	0.12%	330	0.72%
18	Separated Mixed Defective with Hamza	V1	H	V2	0	0.00%	0	0.00%
19	First letter Defective and Second letter Hamza	V	H	C3	3	0.19%	15	0.03%
20	Second Letter Defective and third letter Defective	C1	V	H	8	0.50%	998	2.19%
21	First letter Defective and third Letter Hamza	V	C2	H	2	0.12%	17	0.04%
22	Adjacent Mixed Defective with Hamza	V1	V2	H	0	0.00%	0	0.00%
		Totals			1610	100.00%	45,534	100.00%

Table 1: Category distribution of Root and Tokens extracted from the Qur'an

	Root		Tokens	
Category	**Total**	**Percentage**	**Total**	**Percentage**
Intact	1097	68.14%	27,804	61.06%
Defective	468	29.07%	14,626	32.12%
Compound	45	2.80%	3,104	6.82%
Totals	1610	100.00%	45,534	100.00%

Table 2: summary of category distribution of root and tokens of the Qur'an

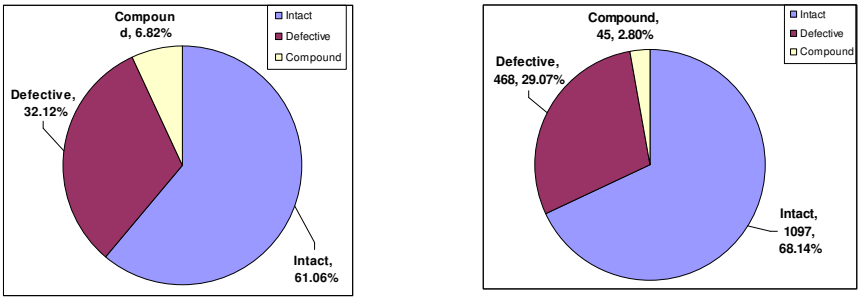

Figure 2: Root distribution (left) and word distribution (right) of the Qur'an

Similar root and word distributions are obtained from the roots and the word types stored in the broad-lexical resource. About 63% of the roots stored in the broad-lexical resource are intact words, and slightly more than 68% of the word types belong to this category. Defective roots forms about 33% of the roots of the broad-lexical resource and 29% of the word types belong to this category. Finally, the compound roots of the broad-lexical resource are approximately 4%, and about 2% of the word types belong to this category. Figure 3 and table 4 shows the root and word types distribution after analyzing the broad-lexical resource. Figure 2 and 3 show similar category distribution in the Qur'an and the broad lexical resource.

					Root		Word Type	
	Category				**Count**	**Percentage**	**Types**	**Percentage**
1	Intact	C1	C2	C3	4147	48.78%	201,385	53.54%
2	Doubled	C1	C2	C2	446	5.25%	32,007	8.51%
3	First Letter Hamza	H	C2	C3	289	3.40%	10,449	2.78%
4	Second letter Hamza	C1	H	C3	216	2.54%	3,909	1.04%
5	Third Letter Hamza	C1	C2	H	270	3.18%	8,985	2.39%
6	First letter Defective	V	C2	C3	386	4.54%	19,219	5.11%
7	Second Letter Defective	C1	V	C3	1115	13.11%	43,512	11.57%
8	Third Letter Defective	C1	C2	V	1151	13.54%	41,295	10.98%
9	Separated Mixed Defective	V	C2	V	45	0.08%	2,372	0.63%
10	Adjacent Mixed defective 1	C1	V1	V2	106	1.25%	4,057	1.08%
11	Adjacent Mixed defective 2	V1	V2	C3	22	0.26%	211	0.06%
12	First Letter Hamza and Doubled	H	C2	C2	30	0.35%	888	0.24%
13	First letter Defective and Doubled	V	C2	C2	29	0.34%	463	0.12%
14	First letter Hamza and third letter Defective	H	C2	V	74	0.87%	2,111	0.56%
15	First letter Hamza and second letter Defective	H	V	C3	47	0.55%	892	0.24%
16	Adjacent Mixed defective with Hamza	H	V1	V2	7	0.08%	135	0.04%
17	Second letter Hamza and Third letter Defective	C1	H	V	42	0.49%	1,041	0.28%
18	Separated Mixed Defective with Hamza	V1	H	V2	2	0.02%	52	0.01%
19	First letter Defective and Second letter Hamza	V	H	C3	15	0.18%	292	0.08%
20	Second Letter Defective and third letter Defective	C1	V	H	42	0.49%	1,590	0.42%
21	First letter Defective and third Letter Hamza	V	C2	H	21	0.25%	1,302	0.35%
22	Adjacent Mixed Defective with Hamza	V1	V2	H	0	0.00%	0	0.00%
		Totals			8502	100.00%	376,167	100.00%

Table 3: Category distribution of Root and Word type extracted from the lexicon

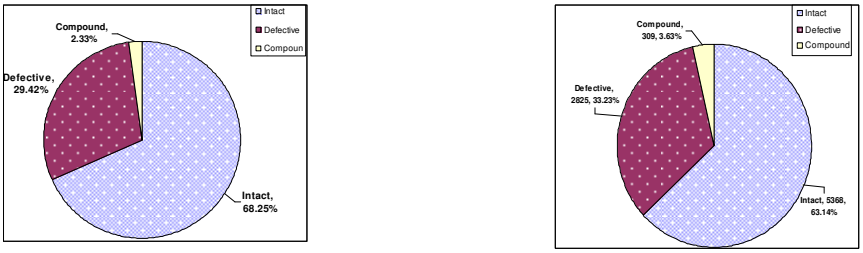

Figure 3: Root distribution (left) and Word type distribution (right) of the broad-lexical resource

	Root		Tokens	
Category	**Total**	**Percentage**	**Total**	**Percentage**
Intact	1097	68.14%	27,804	61.06%
Defective	468	29.07%	14,626	32.12%
Compound	45	2.80%	3,104	6.82%
Totals	1610	100.00%	45,534	100.00%

Table 4: summary of category distribution of root and tokens of Qur'an

3.2 Specifications of the Morphological Analyzer

3.2.1 Inputs

(In the following examples we used Buckwalter transliteration system)

Our morphological analyzer accepts single Arabic word or Arabic text, whether they are vowelized, partially vowelized, or non-vowelized, as inputs to the system. The analyzer deals with both kinds of vowelized and non-vowelized text using one data structure. First, the tokenizer tokenizes and classifies the input text into Arabic word (vowelized, partially vowelized or non-vowelized), number, currency, or punctuation mark. Then the analyzer processes the extracted Arabic words, by resolving the doubled letters (الحروف المضعَّفة) and the extensions (المدّ). The doubled letter marked by *shaddah* (الشَّدَّة) is replaced by two similar letters as the original letter, the first is silent marked by *sukwn*, and the second is vowelized by the same short vowel appears on the original letter. For example the word (وَصَّى) *waS~aY* has the doubled letter (ص) *S* and after processing it will be in this form (وَصْصَى) *waSoSaY*. The extension (المدّ) (آ) is replaced by (Hamza) and (Alif), as in the word (آمَنُوا) *lmanuwA* which will be in this form (ءامَنُوا) *'AmanuwA*.

Only one short vowel can be associated with any letter of the word. Based on this fact we have designed a data structure to process Arabic words. This data structure consists of a one-dimensional array where letters and short vowels are stored. The first letter of the word is stored in the first position of the array followed by its short vowel (if it is present) on the second position, and so on for all letters and short vowels of the word. Figure 4 shows the data structure storing the words (وَصْصَى) *waSoSaY* and (ءامَنُوا) *'AmanuwA*. This data structure is also used to match between the word and the patterns.

12	**11**	**10**	**9**	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**position / word**
				–	ى	ـَ	ص	ـْ	ص	ـَ	و	وَصْصَى
				-	**Y**	**a**	**S**	**o**	**S**	**a**	**w**	**waSoSaY**
–	ا	–	و	ـُ	ن	ـَ	م	–	ا	–	ء	ءامَنُوا
-	**A**	**-**	**w**	**u**	**n**	**a**	**m**	**-**	**A**	**-**	**'**	**'AmanuwA**

Figure 4: The word data structure.

3.2.2 Stop Words (Unambiguous Words)

The system contains a list of 254 unambiguous words (stop words). An unambiguous word has only one morphological analysis wherever it appears on the text. The percentage of unambiguous words in any typical Arabic text is around 40%. The morphological analyzer searches for the word in the unambiguous word list, and if it is found, the analyzer assigns the morphological analysis associated with it. Then the analyzer processes the next word. Figure 5 shows a sample of the unambiguous words.

أنا	*>nA*	me	الذي	*Al*y*	who	حول	*Hwl*	about	عن	*En*	about
نحن	*nHn*	we	على	*ElY*	on	في	*fy*	in	بضع	*bDE*	few
هي	*hy*	she	عند	*End*	next to	مع	*mE*	with	بلى	*blY*	yes
هؤلاء	*h&lA'*	they	ذلك	**lk*	that	بين	*byn*	between	بما	*bmA*	although

Figure 5: Sample of the stop words (unambiguous words).

3.2.3 Prefixes and Suffixes

Using traditional Arabic language grammar books, we have extracted lists of proclitics (conjunctions, prepositions, letters of call, interrogative letters, introduction letters …), prefixes, suffixes, and enclitics (relative pronouns, definite article, prepositions …). These lists were provided to a generating program which generates all the possible combinations of proclitics and prefixes together, and suffixes with enclitics. The generated lists of these combinations were too large. These generated lists were checked by analyzing words in four corpora; the Qur'an text corpus, the Corpus of Contemporary Arabic, the Penn Arabic Treebank, and the text of 15 traditional Arabic lexicons used to construct the broad-lexical resource. Then, we built two lists of prefixes and suffixes, the prefixes list contains 220 prefixes and the suffixes list contains 341 suffixes. Tables 5 & 6 shows samples of these lists with the morphological feature tag assigned to each prefix and suffix in the list. See section 4 for the description of the tags.

Prefix	**Example**	**P1**	**Tag**	**P2**	**Tag**	**P3**	**Tag**
ف ***f***	فقام ***fqAm***	ف ***f***	p--c----------				
فبال ***fbAl***	فبالصدق ***fbAlSdq***	ف ***f***	p--c----------	ب ***b***	p--p----------	ال ***Al***	r---d----------
فست ***fst***	فستذكرون ***fst*krwn***	ف ***f***	p--c----------	س ***s***	p--f----------	ت ***t***	r---a----------
وال ***wAl***	والسماء ***wAlsmA'***	و ***w***	p--c----------	ال ***Al***	r---d----------		
ولت ***wlt***	ولتجدنهم ***wltjdnhm***	و ***w***	p--c----------	ل ***l***	r---a----------	ت ***t***	r---a----------

Table 5: Sample of the prefixes with their morphological tags

Suffix	Example	P1	Tag	P2	Tag	P3	Tag
اتية *Atyp*	معلوماتية *mElwmAtyp*	ات *At*	r---l-fp-v??----------	ي *y*	r---y----------	ة *p*	r---t-fs----------
تموهما *tmwhA*	أورثتموها *>wrvtmwhA*	تم *tm*	r---r-mpssn?----------	و *w*	r---r-mptsnw----------	هما *hmA*	r---r-fstsa?----------
هما *humA*	فأخرجهما *f>xrjhmA*	هما *hmA*	r---r-xdts??----------				
يون *ywn*	الحواريون *AlHwArywm*	ي *y*	r---y----------	ون *wn*	r---m-mp-vnw----------		
هم *hm*	كتابهم *ktAbhm*	هم *hm*	r---r-mpts??----------				

Table 6: Sample of the suffixes and their morphological tags

Moreover, the analyzer divides the word into three parts of different sizes. Then it searches the prefix list for the first part, and the suffix list for the third part. If the first part or the third part are found in the prefixes or suffixes lists, the morphological feature tag associated to the prefix or suffix is assigned to these parts. Then the analyzer selects the analyses of the word where the first part matches one of the prefixes from the list, and the third part matches one of the suffixes from the list. Figure 8 shows the process of matching prefixes and suffixes and the process of selecting the candidate analyses.

3.2.4 Root or Stem

The system uses a list of tri-literal, quad-literal and quint-literal roots, consisting of more than 12,000 roots. These roots were extracted from the 15 traditional Arabic language lexicons. After selecting the candidate analyses that match the first part of the word with the prefixes list, and the third part of the word with the suffixes list, the analyzer matches the second part with the root list. Table 7 shows the matching process between the second part and the root list.

3.2.5 Word Pattern

The process of derivation of words from their roots, whether the root is tri-literal root, quad-literal root or quint-literal root, is done by following specific templates called patterns. These patterns carry linguistic information which is propagated to the derived words. Building on this fact, we provided the analyzer with a list of patterns, containing 2730 verb patterns and 985 noun patterns. Morphological feature tags are assigned to each pattern in the list. Table 9 shows a sample of the pattern list.

An important characteristic of this list is that patterns are fully vowelized. The vowelized patterns will allow the analyzer to add the correct short vowels to the partially vowelized or non-vowelized word. The analyzer uses two algorithms to match between the words and their correct patterns.

Word		First Part		Second Part		Third Part		Prefixes & Suffixes analyses
يَعْمَلُونَ	*yaEomaluwna*			**يعملون**	***yEmlwn***			**Candidate analysis**
يَعْمَلُونَ	*yaEomaluwna*			يعملو	*yEmlw*	ن	*n*	Not accepted
يَعْمَلُونَ	*yaEomaluwna*			**يعمل**	*yEmlw*	**ون**	***wn***	**Candidate analysis**
يَعْمَلُونَ	*yaEomaluwna*			يعم	*yEml*	لون	*lwn*	Not accepted
يَعْمَلُونَ	*yaEomaluwna*			يع	*yE*	ملون	*mlwn*	Not accepted
يَعْمَلُونَ	*yaEomaluwna*			ي	*y*	عملون	*Emlwn*	Not accepted
يَعْمَلُونَ	*yaEomaluwna*	**ي**	***y***	**عملون**	***Emlwn***			**Candidate analysis**
يَعْمَلُونَ	*yaEomaluwna*	ي	*y*	عملو	*Emlw*	ن	*n*	Not accepted
يَعْمَلُونَ	*yaEomaluwna*	**ي**	***y***	**عمل**	***Eml***	**ون**	***wn***	**Candidate analysis**
يَعْمَلُونَ	*yaEomaluwna*	ي	*y*	عم	*Em*	لون	*lwn*	Not accepted
يَعْمَلُونَ	*yaEomaluwna*	ي	*y*	ع	*E*	ملون	*mlwn*	Not accepted
يَعْمَلُونَ	*yaEomaluwna*	يع	*yE*	ملون	*mlwn*			Not accepted
يَعْمَلُونَ	*yaEomaluwna*	يع	*yE*	ملو	*mlw*	ن	*n*	Not accepted
يَعْمَلُونَ	*yaEomaluwna*	يع	*yE*	مل	*ml*	ون	*wn*	Not accepted
يَعْمَلُونَ	*yaEomaluwna*	يع	*yE*	م	*m*	لون	*lwn*	Not accepted
يَعْمَلُونَ	*yaEomaluwna*	يعم	*yEm*	لون	*lwn*			Not accepted
يَعْمَلُونَ	*yaEomaluwna*	يعم	*yEm*	لو	*lw*	ن	*n*	Not accepted
يَعْمَلُونَ	*yaEomaluwna*	يعم	*yEm*	ل	*l*	ون	*wn*	Not accepted
يَعْمَلُونَ	*yaEomaluwna*	يعمل	*yEml*	ون	*wn*			Not accepted

Table 7: Example of the process of selecting the matched prefixes and suffixes

Word		First part		Second part		Third Part		Affixes analyses	Affixes and Root analyses
يَعْمَلُونَ	*yaEomaluwna*			يعملون	*yEmlwn*			Candidate analysis	Not accepted analysis
يَعْمَلُونَ	*yaEomaluwna*			يعمل	*yEml*	ون	*wn*	Candidate analysis	Not accepted analysis
يَعْمَلُونَ	*yaEomaluwna*	ي	*y*	عملون	*Emlwn*			Candidate analysis	Not accepted analysis
يَعْمَلُونَ	*yaEomaluwna*	**ي**	***y***	**عمل**	***Eml***	**ون**	***wn***	**Candidate analysis**	**Accepted Analysis**

Table 8: Example of Affixes and root matching process

Verb Patterns		POS Tag
فَعَلْتُ	*faEalotu*	v-p---nsfs-s-an??dst?-
فَعَلْنَا	*faEalonaA*	v-p---npfs-s-an??dst?-
فَعَلْتَ	*faEalota*	v-p---msss-s-an??dst?-
فَعَلْتِ	*faEaloti*	v-p---fsss-s-an??dst?-
فَعَلْتُمَا	*faEalotumaA*	v-p---xdss-s-an??dst?-
Noun Patterns		**POS Tag**
أُفْعُلاوَى	*>ufoEulAwaY*	n?----??-v???---?dqt-?
اِفْعِيلال	*AifoEiylAl*	ng----??-v???---?dtt-?
فاعُولاء	*fAEuwlA'*	n?----??-v???---?dqt-?
فُعُلْعُلان	*fuEuloEulAn*	n?----??-v???---?dqt-?
فُعَّيْلاء	*fuE~ayolA'*	n?----??-v???---?dqt-?

Table 9: Sample of the pattern list

3.2.5.1 The first algorithm (Word and its root)

The first algorithm to extract the pattern of the word depends on the word itself and its root as inputs. After selecting the analyses from the previous step which match the first part with the prefixes list, the second part with the roots list, and the third parts with the suffixes list, the algorithm replaces the root letters in the word with the pattern letters (fa', Ain, and lam) (ف، ع، ل).

This process is not easy; as some root letters might be changed. The changes include incorporation, turnover, defection and replacement. The algorithm must deal with these changes and extract the correct pattern of the word. Finally, the pattern list is searched for the candidate pattern. If the pattern is found in the list, the morphological feature tag associated with the pattern in the list is assigned to the analyzed word. Figure 7 shows examples of extracting the pattern using this method.

Word	أَحَسِبَ	Letters	أ >	ح H	س s	ب b		Root	حسب	ح H	س s	ب b
	>aHasiba	Index	0	1	2	3			Hsb	1	2	3
Root letters indices												
First letter (ح H) = [1]			Second letter (س s) = [2]				Third letter (ب b) = [3]					
Candidate indices list = [1,2,3]												
Pattern			Prefix				Stem			Suffix		
أفعل >fEl			أ >				حسب Hsb					

Word	آمَنُوا	Letters	أ >	ا A	م m	ن m	و w	ا A	Root	أمن	أ >	م m	ن n
	|manuwA	Index	0	1	2	3	4	5		>mn	1	2	3
Root letters indices													
First letter (أ >) = [-1, 0]			Second letter (م m) = [2]				Third letter (ن n) = [3]						
Indices [-1, 2, 3] , [0, 2, 3]													
Candidate indices list = [-1 , 2 , 3]													
Pattern			Prefix				Stem			Suffix			
أاعلوا >AElwA			أا >A				من mn			وا wA			
Candidate indices list = [0, 2, 3]													
Pattern			Prefix				Stem			Suffix			
فاعلوا fAElwA							أامن >Amn			وا wA			

Word	الْعَلِيم	Letters	ا A	ل l	ع E	ل l	ي y	م m	Root	علم	ع E	ل l	م m
	AloEaliym	Index	0	1	2	3	4	5		Elm	1	2	3
Root letters indices													
First letter (ع E) = [2]			Second letter (ل l) = [1,3]						Third letter (م m) = [5]				
Candidate indices list = [2 ,1 , 3] False [2,3,5] True													
Pattern			Prefix				Stem			Suffix			
فعيل fEyl			ال Al				عليم Elym						

Figure 7: Examples of extracting the pattern of the words using the first method (the word and its root)

3.2.5.2 The second algorithm

The second method of extracting the pattern of the word mainly depends on the Pattern Matching Algorithm (PMA) (Alqrainy, 2008). This algorithm matches a partially vowelized word, with the last diacritic mark only, with a pattern lexicon without doing any analyses for the prefixes or suffixes of the word.

However, our pattern matching algorithm searches the patterns list for patterns of similar size to the analyzed word after removing the prefixes and suffixes of the word. For example, the word (كتب) (*ktb*) has a size of 6 according to the data structure we used, whether the word is fully-vowelized, partially-vowelized or non-vowelized. And it matches the following patterns (فَعْل *FaEol*, فَعَل *faEal*, فَعُل *faEul*, فَعِل *faEil*, فُعْل *fuEol*, فُعَل *fuEal*, فُعُل *fuEul*, فُعِل *fuEil*, فِعْل *fiEol*). In the second step, the algorithm replaces the letters of the word

corresponding to the letters (Fa', Ain, Lam) (ف، ع، ل) of the pattern. Then these generated patterns are searched for in the pattern list. If the pattern is found in the pattern list, then it is a candidate pattern of the word, and the morphological tag associated with the pattern in the list is assigned to the analyzed word. Figure 8 shows example of extracting the pattern of the word using this method.

Word		Pattern		Tag
يَعْمَلُونَ	*yaEomaluwna*	يَفْعُلُونَ	*yafoEuluwna*	v-c---mptdnn-an??dst?-
يَعْمَلُونَ	*yaEomaluwna*	يَفْعِلُونَ	*yafoEiluwna*	v-c---mptdnn-an??dst?-
يَعْمَلُونَ	*yaEomaluwna*	يَفْعَلُونَ	*yafoEaluwna*	v-c---mptdnn-an??dst?-
يَعْمَلُونَ	*yaEomaluwna*	يُفْعِلُونَ	*yufoEiluwna*	v-c---mptdnn-an??dat?-
يَعْمَلُونَ	*yaEomaluwna*	يُفْعَلُونَ	*yufoEaluwna*	v-c---mptdnn-pn??dtt?-
كتب	*ktb*	فَعَلَ	*faEala*	v-p---msts-a-an??dst?-
كتب	*ktb*	فَعِلَ	*faEila*	v-p---msts-f-an??dst--
كتب	*ktb*	فَعُلَ	*faEula*	v-p---msts-f-an??dst--
كتب	*ktb*	فَعِلَ	*faEila*	v-p---msts-f-an??dst--
كتب	*ktb*	فُعِلَ	*fuEila*	v-p---msts-f-pn??dtt--
كتب	*ktb*	فَعْل	*faEol*	n?----??-v???---?dst-?
كتب	*ktb*	فَعَل	*FaEal*	ng----f?-v???---?dst-?
كتب	*ktb*	فَعُل	*faEul*	n?----??-v???---?dst-?
كتب	*ktb*	فَعِل	*faEl*	n?----??-v???---?dst-?
كتب	*ktb*	فُعْل	*fuEol*	n?----??-v???---?dst-?
كتب	*ktb*	فُعَل	*fuEal*	n?----??-v???---?dst-?
كتب	*ktb*	فُعُل	*fuEul*	n?----??-v???---?dst-?
كتب	*ktb*	فُعِل	*fuEil*	n?----??-v???---?dst-?
كتب	*ktb*	فِعْل	*fiEol*	n?----??-v???---?dst-?
كتب	*ktb*	فِعِل	*fiEil*	n?----??-v???---?dst-?
كتب	*ktb*	فَعِل	*faEil*	nx----??-v???---?dst-?

Figure 8: example of using the second method for extracting the patterns of the word

3.2.6 Vowelization

Vowelization is an important characteristic of the Arabic word. Vowelization helps in determining some morphological features of the words. The presence of the short vowel on the last letter helps in determining the case or mood of the word. And the presence of the vowels on the first letter determines whether the verb is active or passive. The presence of other diacritics such as Shaddah and maddah (extention) solve some ambiguities of words.

After matching the patterns and the analyzed word, in the previous step, taking into account that the patterns are fully vowelized, the analyzer adds the short vowels which appear on the patterns to the analyzed word, whether it is partially-vowelized or non-vowelized. The result is a correctly vowelized list of the possible analyses. Figure 9 shows the process of adding vowels to the non-vowelized words.

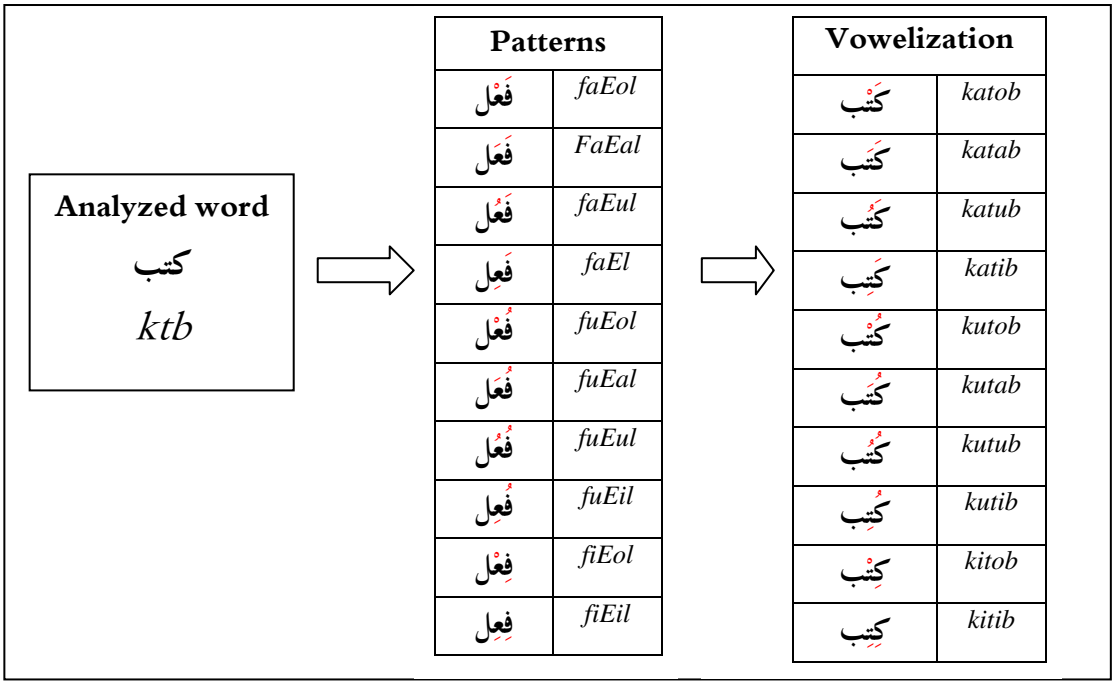

Figure 9: Vowelization process example

4. Morphological features of Arabic words and morphological features Tag Set

Scholars of the Arabic language classify Arabic words into three main parts of speech; nouns, verbs and particles, see tags example (Atwell, 2008). Each part of speech has been described in detail. Morphological features of each part of speech have been comprehensively determined. Nouns include many sub classifications such as: original nouns, pronouns, adjectives, demonstrative nouns, relative nouns, proper nouns, nouns of places and time, and others. Verbs include past verb, progress verb and imperative verb. Particles include prepositions, conjunctions, call letters and others. Morphological features of the words include gender (masculine and feminine), number (singular, dual and plural), person (first person, second person and third person), case, mood, definiteness, active or passive verbs, emphasizing, and transitivity. Other features are: stripped of augmented words, number of root letters and the internal structure of the verb.

Building on these traditional part of speech and features, we have designed a Morphological Features Part-of-Speech Tag Set, to be used in a part-of-speech tagging system, to annotate Arabic corpora. The annotation scheme is a detailed annotation in a way which includes all the morphological features of the words in the corpora. This tag set can be used to study, develop and evaluate Arabic morphological analyzers in a simple and direct way. The morphological features tag set is designed to contain 22 morphological features of the Arabic word in a single tag. Table 10 shows the 22 morphological features which have been used in the design of the morphological features tag set. The detailed Arabic morphological features Tag Set is found on the website http://www.comp.leeds.ac.uk/sawalha/tagset.html and in the paper (Sawalha & Atwell, 2009).

The tag string consists of 22 characters. Each character represents a value or attribute which belongs to a morphological feature category. The position of the character in the tag string is important to identify the morphological feature category. Morphological feature category attribute is represented by one lowercase letter, which is still readable, such as: **v** in the first position to indicate ***verb***, **n** in the second position to indicate ***name***, gender category values in the seventh position such as: ***masculine*** is represented by **m**, ***feminine*** is represented by **f** and ***neuter*** is represented by **x**. If the value of a certain feature is not applicable for the tagged word

then dash '**-**' is used to indicate that. Question mark '**?**' is interpreted as a certain feature belongs to the word but at the moment is not available or the automatic tagger could not guess it.

The interpretation of the tag is handled by referring to the value and its position in the tag string, to identify the morphological feature category that the value belongs to. Then, all these single interpretations of attributes are grouped together to represent the full tag of the word. This will make the tag more readable when it includes the other morphological features. Figure 10 shows samples of tagged text using the morphological feature tag set, taken from the Qur'an and the Penn Arabic Treebank.

Position	**Morphological Features Categories**	
1	Main Part-of-Speech	أَقسام الكلام الرئيسيَّة
2	Part-of-Speech of Noun	أقسام الكلام الفرعيَّة (الاسم)
3	Part-of-Speech of Verb	أقسام الكلام الفرعيَّة (الفعل)
4	Part-of-Speech of Particle	أقسام الكلام الفرعيَّة (الحرف)
5	Residuals	أقسام الكلام الفرعيَّة (أخرى)
6	Punctuation marks	أقسام الكلام الفرعيَّة (علامات الترقيم)
7	Gender	الجنس
8	Number	العدد
9	Person	الشخص
10	Morphology	الصَّرف
11	Case and Mood	الحالة الإعرابية للاسم أو الفعل
12	Case and Mood marks	علامة الإعراب أو البناء
13	Definiteness	المَعْرِفَة والنَّكِرَة
14	Voice	المَبْني لِلمَعْلُوم و المَبْني لِلمَجْهُول
15	Emphasize	المُؤكَّد وغيرُ المُؤكَّد
16	Transitivity	اللازم والمتعدي
17	Humanness	العاقل وغير العاقل
18	Variability & Conjungation	التَّصريف
19	Augmented and Unaugmented	المجرَّد والمزيد
20	Root letters	عَدَد أحْرُف الجَذْر
21	Verb Internal Structure	بُنية الفعل
22	Noun finals	أقسام الأسم تبعاً للفظ آخره

Table 10: Morphological feature categories of the Tag Set

Word			Tag
وَ	*wa*	*And*	p--c----------
وَصَّيْ	*waS~ayo*	*Recommended*	v-p------s-s-amohdst&-
نَا	*naA*	*We*	r---r-xpfs-f----hn----
الْإِنسَانَ	*Alo<insaAna*	*human*	nq----mb-pafd---hcbt-s
بِ	*bi*	*to*	p--p----------
وَالِدَيْ	*waAlidayo*	*parents*	nu----md-dgyd---hdat-s
هِ	*hi*	*his*	r---r-msts-k----hn----
حُسْنًا	*HusonAF*	*well*	ng----xs-vafi----ast-s

Word			Tag
تم	*tm*	*Accomplished*	v-p------s-f-amihdstb-
اعداد	*AEdAd*	*Preparing*	ng----??-vndi---?db3-s
الوثائق	*AlwvA'q*	*Documents*	nq----fb-vafd---ndbt-s
المتوفرة	*Almtwfrp*	*Available*	nj----f?-vafd---ndtt-s
ب	*b*	*In*	p--p----------
كثرة	*kvrp*	*Many*	nj----fb-vgki----dat-s
حول	*Hwl*	*About*	nv-------s-fi----nst-s
أول	*>wl*	*First*	n+----ms-vgki----dst-s
رحلة	*rHlp*	*Trip*	no----fs-vgki----dat-s
طيران	*TyrAn*	*Flight*	ng----??-vgki----dbt-s
عثمانية	*EvmAnyp*	*Ottomani*	n*----fs-pgki----daq-s
فوق	*fwq*	*Over*	nv-------s-fi----nst-s
البلاد	*AlblAd*	*Countries*	nl----mb-vgkd---ndat-s
العربية	*AlErbyp*	*Arabian*	n*----fb-vgkd---hdst-s

Figure 10: Samples of Tagged text from the Qur'an and the Penn Arabic Treebank using the Morphological feature tag set

5. Evaluation and Results

5.1 Gold Standard for Evaluation

Gold standards are used to evaluate and measure the actual accuracy of automatic systems. The evaluation can be used to compare different systems or algorithms of the same problem domain. It precisely shows the successes and failures of an algorithm. Gold standards can be used to compute similarity between systems by highlighting the cases of agreed analyses and the cases when a tie resulted.

To construct a gold standard for evaluation, we need to determine the problem domain of the algorithms to be evaluated, the texts to be used as gold standard, the format of the gold standard, its size, the script used and transliteration scheme, and the phases of constructing the gold standard.

5.1.1 Problem domain

Our gold standard will be used to evaluate morphological analyzers and part-of-speech taggers. The gold standard should have morphological information and part-of-speech tags for each word of the selected corpora.

5.1.2 The Corpora

Corpora are used to build gold standards. Many Arabic language corpora have been developed. But to build a widely used general purpose gold standard, we have to select corpora of different text domains, formats and genres of both vowelized and non-vowelized Arabic text. First, we selected the Qur'an corpus to be used in the construction of the gold standard. We have two versions of the Qur'an text, vowelized Qur'an text, where diacritics appear above or below each letter of the Qur'an text, and a non-vowelized one, where diacritics are omitted from the vowelized text of Qur'an. Second, we want to use the Corpus of Contemporary Arabic (Al-Sulaiti & Atwell, 2006). This corpus contains 1 million words taken from different genres collected from newspapers and magazines. It contains the following domains; Autobiography, Short Stories, Children's Stories, Economics, Education, Health and Medicine, Interviews, Politics, Recipes, Religion, Sociology, Science, Sports, Tourist and Travel and Science.

5.1.3 Gold Standard Format

The gold standard will include morphological and part-of-speech information for each word of the gold standard. The analysis divides the words into their morphemes; conjunctions, prepositions, prefixes, stem or root, suffixes and relative pronouns. For each morpheme, part-of-speech tagging information will be provided. A compound part-of-speech tag of the whole word or lexical entry can be generated by combining the part-of-speech tag information of every morpheme of the word. Moreover, the gold standard will contain the root and the pattern information of the words. The gold standard will be stored using flat text files, using Unicode utf-8 encoding, each word and its morphological and part-of-speech information in a line separated by tabs.

5.1.4 Gold Standard Size

The gold standard must be relatively large, so, it can cover most cases that morphological analyzer have to handle. The gold standard size is measured by the number of words it contains.

5.2 Qur'an gold standard of MorphoChallenge 2009

We developed a gold standard of the Qur'an to be used to evaluate morphological analyzers in the Morphochallenge 2009 competition, which aims to develop an unsupervised morphological analyzer to be used for different languages including Arabic http://www.cis.hut.fi/morphochallenge2009/datasets.shtml. The gold standard size is 78,004 words. The gold standard of Qur'an contains the full morphological analysis for each word, according to the Tagged database of the Qur'an developed at the University of Haifa (Dror et al, 2004) but reformatted to match other Morphochallenge test sets in other languages. Figure 11 shows a sample of the Qur'an gold standard.

Moreover, gold standard can be used to determine the specifications of the morphological analyzers by specifying which morphological features or which it can not handle. And this is another way to evaluate morphological analyzers by describing their specifications.

بِسْمِ	سم	None	ب+Prep , سم+Noun+Triptotic+Sg+Masc+Gen ,
اللّهِ	None	None	للاه+Noun+ProperName+Gen+Def ,
الرَّحْمـــنِ	رحم	فَعلَان	رَحمَان+Noun+Triptotic+Adjective+Sg+Masc+Gen+Def ,
الرَّحِيمِ	رحم	فَعِيل	رَحِيم+Noun+Triptotic+Adjective+Sg+Masc+Gen+Def ,
الْحَمْدُ	حمد	فَعل	حَمد+Noun+Triptotic+Sg+Masc+Nom+Def ,
للّهِ	None	None	ل+Prep , للاه+Noun+ProperName+Gen+Def ,
رَبِّ	ربب	فَعل	رَبب+Noun+Triptotic+Sg+Masc , Pron+Dependent+1P+Sg , رَبب +Noun+Triptotic+Sg+Masc+Gen ,
الْعَالَمِينَ	علم	فَاعَل	عَالَم+Noun+Triptotic+Pl+Masc+Obliquus+Def ,

(vowelized Arabic script)

بسم	سم	None	ب+Prep , سم+Noun+Triptotic+Sg+Masc+Gen ,
الله	None	None	للاه+Noun+ProperName+Gen+Def ,
الرحمـــن	رحم	فعلان	رحمان+Noun+Triptotic+Adjective+Sg+Masc+Gen+Def ,
الرحيم	رحم	فعيل	رحيم+Noun+Triptotic+Adjective+Sg+Masc+Gen+Def ,
الحمد	حمد	فعل	حمد+Noun+Triptotic+Sg+Masc+Nom+Def ,
لله	None	None	ل+Prep , للاه+Noun+ProperName+Gen+Def ,
رب	ربب	فعل	ربب+Noun+Triptotic+Sg+Masc , + Pron + Dependent+1P+Sg , ربب+Noun+Triptotic+Sg+Masc+Gen ,
العالمين	علم	فاعل	عالم+Noun+Triptotic+Pl+Masc+Obliquus+Def ,

(Non-Vowelized Arabic script)

bisomi	sm	None	b+Prep , sm+Noun+Triptotic+Sg+Masc+Gen ,
All~hi	None	None	llaah+Noun+ProperName+Gen+Def ,
Alr~aHom_ani	rHm	faElaAn	raHmaan+Noun+Triptotic+Adjective+Sg+Masc+Gen+Def ,
Alr~aHiymi	rHm	faEiyl	raHiim+Noun+Triptotic+Adjective+Sg+Masc+Gen+Def ,
AloHamodu	Hmd	faEl	Hamd+Noun+Triptotic+Sg+Masc+Nom+Def ,
ll~hi	None	None	l+Prep , llaah+Noun+ProperName+Gen+Def ,
rab~i	rbb	faEl	rabb+Noun+Triptotic+Sg+Masc , + Pron + Dependent+1P+Sg, rabb+Noun+Triptotic+Sg+Masc+Gen ,
AloEaAlamiyna	Elm	faAEal	&aalam+Noun+Triptotic+Pl+Masc+Obliquus+Def ,

(Vowelized Romanized script using Buckwalter transliteration scheme)

bsm	sm	None	b+Prep , sm+Noun+Triptotic+Sg+Masc+Gen ,
Allh	None	None	llAh+Noun+ProperName+Gen+Def ,
AlrHm_n	rHm	fElAn	rHmAn+Noun+Triptotic+Adjective+Sg+Masc+Gen+Def ,
AlrHym	rHm	fEyl	rHym+Noun+Triptotic+Adjective+Sg+Masc+Gen+Def ,
AlHmd	Hmd	fEl	Hmd+Noun+Triptotic+Sg+Masc+Nom+Def ,
llh	None	None	l+Prep , llAh+Noun+ProperName+Gen+Def ,
rb	rbb	fEl	rbb+Noun+Triptotic+Sg+Masc , + Pron + Dependent+1P+Sg , rbb+Noun+Triptotic+Sg+Masc+Gen ,
AlEAlmyn	Elm	fAEl	EAlm+Noun+Triptotic+Pl+Masc+Obliquus+Def ,

(Von-vowelized Romanized script using Buckwalter transliteration scheme)

Figure 11: a sample of the Qur'an Gold Standard for evaluating morphological analyzers in the Morphochallenge2009 competition.

6. Conclusions

In this paper, we reviewed the morphological analyzers required to build a tagged corpus tagged with the morphological features analyses for each word. This paper showed the results of comparing three different freely available morphological analyzers and stemmers. The comparison depended on a gold standard for evaluation which contains two 1000-word documents from the Qur'an and the Corpus of Contemporary Arabic. The results showed that

morphological analyzers and stemmers have failed to analyze about quarter of the words of the test documents. So, we started to search for other methods that improve the accuracy of the morphological analyzers. To understand the morphology problem well, we analyzed the tri-literal roots of the Qur'an and the word types stored in the broad-lexical resource. The results of this analysis showed that about 40% of these tri-literal roots are defective roots which add more challenge on developing a robust morphological analyzer.

We have developed a morphological analyzer for Arabic text which depends on pre-stored lists of prefixes, suffixes, roots and patterns. These lists were extracted by referring to traditional grammar books. The affixes lists have been verified by analyzing the Qur'an, the Corpus of Contemporary Arabic, the Penn Arabic Tree bank and the text of 15 traditional Arabic language lexicons as our fourth corpus. The prefixes list contains 215 prefixes. The suffixes list contains 127 suffixes and the patterns list contains 2730 verb patterns and 985 nouns patterns.

The morphological analyzer was developed to analyze the word and specify its morphological features. We have distinguished between many morphological features, which we hope that a morphological analyzer for Arabic text can handle. For this purpose, we have developed a Morphological Features Part-of-Speech Tag Set, which can be used in developing morphological analyzers. Also, it can be used to morphologically annotate corpora. The morphological features tag consists of string of 22 characters, where each character in a specific position in the tag represents a morphological feature for the analyzed word.

To evaluate the results of different morphological analyzers, we propose developing a gold standard for evaluation. The text of the gold standard is selected from different types, domains and genres of vowelized and non vowelized text.

---

[1] This paper is based on the Arabic version of the paper presented in the workshop of morphological analyzer experts for Arabic language. Organized by Arab League Educational, Cultural and Scientific Organization (ALECSO), King Abdul-Aziz City of Technology (KACT) and Arabic Language Academy. Damascus, Syria. 26-28 April 2009.

References

Al-Ghalayyeni, A.-S. M. الغلاييني, م. ا (2005) *Jami' Al-Duroos Al-Arabia "جامع الدروس العربية"*, Saida - Lebanon, Al-Maktaba Al-Asriyiah "المكتبة العصرية".

Al-Jawhari "ابو النصر اسماعيل بن حماد الجوهري الفارابي" , Al-Sihah fi Al-lughah " الصحاح في اللغة" The Correct Language , died in 1002 A.D, Al-Meshkat Islamic Library (online-library) http://www.almeshkat.net/books/archive/books/alsehah%20g.zip

Alqrainy, S. (2008) *A Morphological-Syntactical Analysis Approach For Arabic Textual Tagging. 2008.* PhD Thesis, De Montfort University, Leicester, UK.

Al-Shalabi, R., Kanaan, G., & Al-Serhan, H. (2003). *New approach for extracting Arabic roots*. Paper presented at the International Arab Conference on Information Technology (ACIT'2003), Alexandria, Egypt.

Al-Sulaiti, Latifa & Atwell, Eric (2006). *The design of a corpus of contemporary Arabic*. International Journal of Corpus Linguistics, vol. 11, pp. 135-171.

Atwell, E. (2008) *Development of tag sets for part-of-speech tagging*. In Ludeling, A. & Kyto, M. (Eds.) Corpus Linguistics: An International Handbook Volume 1. Mouton de Gruyter.

Buckwalter, T. (2004) *Buckwalter Arabic Morphological Analyzer Version 2.0*. Linguistic Data Consortium, catalog number LDC2004L02 and ISBN 1-58563-324-0.

Dahdah, A. (1987) *A dictionary of Arabic Grammar in Charts and Tables " معجم قواعد اللغة العربيه – في جداول ولوحات ",* Beirut, Lebanon, Librairie du Liban.

Dahdah, A. (1993) *A dictionary of Arabic Grammatical nomenclature Arabic – English " معجم لغة النحو العربي عربي-انكليزي",* Beirut, Lebanon, Librairie du Liban.

Dror Judith, Shaharabani Dudu, Talmon Rafi & Wintner Shuly. (2004) *Morphological Analysis of the Qur'an.* Literary and Linguistic Computing, 19(4):431-452.

Lane, E. W. (1968). *An Arabic-English Lexicon*. Beirut, Librarie Du Liban.

Larkey Leah. S. & Connell Margrate. E. (2001). *Arabic information retrieval at UMass*. In Proceedings of TREC 2001, Gaithersburg: NIST.

Larkey Leah S. Ballesteros Lisa & E.Connell Margrate. (2002). *Improving stemming for Arabic information retrieval: Light Stemming and co-occurrence analysis*. In *SIGIR 2002*, Tampere, Finland: ACM.

Maamouri, M. & Bies, A. (2004) *Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools*. Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004).

Sawalha, Majdi. & Awell, Eric. (2009) *توظيف قواعد النحو والصرف في بناء محلل صرفي للغة العربية (Adapting Language Grammar Rules for Building a Morphological Analyzer for Arabic Language).* Proceedings of the workshop of morphological analyzer experts for Arabic language, organized by Arab League Educational, Cultural and Scientific Organization (ALECSO), King Abdul-Aziz City of Technology (KACT) and Arabic Language Academy. Damascus, Syria. 26-28 April 2009.

Sawalha, Majdi. & Atwell, Eric. (2008) Comparative evaluation of Arabic language morphological analysers and stemmers. *Proceedings of COLING 2008 22nd International Conference on Comptational Linguistics.*

Soudi, A., Bosch, A. V. D. & Neumann, G. (Eds.) (2007) *Arabic Computational Morphology: Knowledge-Based and Empirical Methods*, Springer Netherlands.

Thabet, N. (2004) *Stemming the Qur'an*. COLING 2004, Workshop on computational approaches to Arabic script-based languages.