SALMA: Standard Arabic Language Morphological Analysis

Majdi Sawalha
Computer Information Systems Department,
King Abdullah II School of Information Technology,
University of Jordan, Amman, Jordan
sawalha.majdi@gmail.com

Eric Atwell
I-AIBS Institute for Artificial intelligence and Biological Systems, School of Computing,
University of Leeds, Leeds LS2 9JT, UK.
e.s.atwell@leeds.ac.uk

Mohammad A. M. Abushariah
Computer Information Systems Department,
King Abdullah II School of Information Technology,
University of Jordan, Amman, Jordan
m.abushariah@ju.edu.jo

Abstract— **Morphological analyzers are preprocessors for text analysis. Many Text Analytics applications need them to perform their tasks. This paper reviews the SALMA-Tools (Standard Arabic Language Morphological Analysis) [1]. The SALMA-Tools is a collection of open-source standards, tools and resources that widen the scope of Arabic word structure analysis - particularly morphological analysis, to process Arabic text corpora of different domains, formats and genres, of both vowelized and non-vowelized text. Tag-assignment is significantly more complex for Arabic than for many languages. The morphological analyzer should add the appropriate linguistic information to each part or morpheme of the word (proclitic, prefix, stem, suffix and enclitic); in effect, instead of a tag for a word, we need a subtag for each part. Very fine-grained distinctions may cause problems for automatic morphosyntactic analysis – particularly probabilistic taggers which require training data, if some words can change grammatical tag depending on function and context; on the other hand, fine-grained distinctions may actually help to disambiguate other words in the local context.**

The SALMA – Tagger is a fine grained morphological analyzer which is mainly depends on linguistic information extracted from traditional Arabic grammar books and prior-knowledge broad-coverage lexical resources; the SALMA – ABCLexicon. More fine-grained tag sets may be more appropriate for some tasks. The SALMA – Tag Set is a standard tag set for encoding, which captures long-established traditional fine-grained morphological features of Arabic, in a notation format intended to be compact yet transparent.

Keywords—Morphological analysis; Tag Set; Fine-grain; Traditional Arabic Grammar; Traditional Arabic Lexicons

I. INTRODUCTION

Morphological analysis for text corpora is a prerequisite for many text analytics applications, which has attracted many researchers from different disciplines such as linguistics (computational and corpus linguistics), artificial intelligence, and natural language processing, to morphosyntactically analyze text of different languages including Arabic. Recently, several researchers have investigated different approaches to morphological and syntactic analysis for Arabic text. Many systems have been developed which vary in complexity from light stemmers, root extraction systems, lemmatizers, complex morphological analyzers, part-of-speech taggers and parsers.

Morphology is the study, identification, analysis and description of the minimal meaning bearing units that constitute a word. The minimal meaning bearing unit of a word is called a morpheme. Categorizing and building a representative structure of the component morphemes is called morphological analysis. Both orthographic rules and morphological rules are important for categorizing a word's morphemes. For instance, orthographic rules for pluralizing English words ending with *–y* such as *party* indicates changing the *–y* to *-i-* and adding *–es*. And morphological rules tell us that *fish* has null plural and the plural of *goose* is formed by a vowel change. Morphological analysis of the surface or input form *going* is the verbal stem *go* plus the *–ing* morpheme VERB-go + GERUND-ing [2].

Computational morphology is a branch of computational linguistics (*i.e.* natural language processing or language engineering). The main concern of computational morphology is to develop computer applications (*i.e.* toolkits) that analyze words of a given text and deal with the internal structure of words such as determining their part-of-speech and morphological features (*e.g.* gender, number, person, case, mood, voice, etc) [3].

Morphological analysis has many applications throughout speech and language processing. In web searching for morphologically complex languages, morphological analysis enables searching for the inflected form of the word even if the search query contains only the base form. Morphological analysis gives the most important information for a part-of-speech tagger to select the most suitable analysis for a given context. Dictionary construction and spell-checking applications rely on a robust morphological analysis. Machine translation systems rely on highly accurate morphological analysis to specify the correct translation of an input sentence[2]. Lemmatization is an aspect of morphological analysis. Google's search facilities use lemmatization to produce hits of all inflectional forms of the input word. Statistical models of language in machine translation and speech recognition also use lemmatization. Lexicographic applications use lemmatizers as an essential tool for corpus-based compilation [4]. Morphological analysis techniques form the basis of most natural language processing systems. Such techniques are very useful for many applications, such as information retrieval, text categorization, dictionary automation, text compression, data encryption, vowelization and spelling aids, automatic translation, and computer-aided instruction [5].

II. II. ARABIC COMPUTATIONAL MORPHOLOGY

Arabic is a living language that belongs to the Semitic group of languages. The main characteristic feature of Semitic languages is their nonconcatenative morphology where words are derived from their basis of mostly triliteral consonantal roots. Roots of Semitic languages carry the basic conceptual meanings, while varying the vowelling of the simple root and adding prefixes, suffixes and infixes to produce the different variations in shade of meaning [6]. For example, from the Arabic root كتب *k-t-b* 'wrote' we can derive the following words by filling in the vowels: كِتَاب *kitāb* 'book', كُتُب *kutub* 'books', كَاتِب *kātib* 'writer', كُتَّاب *kuttāb* 'writers', كَتَبَ *kataba* 'he wrote', يَكْتُبُ *yaktubu* 'he writes', etc.

Modern linguistic theories of Arabic morphology have studied the derivation process of Arabic words from two points of view: root-based and stem-based (or word-based). The theory of Prosodic Morphology [7, 8] defines the basic character of phonological structure and its consequences for morphology. The true templatic morphology is represented by the derivational categories of the Arabic verbs. Using multiple levels of representation, Arabic verbs have three auto-segmental tiers: consonantal tier (*i.e.* the root), CV skeleton (*i.e.* patterns) and vocalic melody (*i.e.* short vowels).

Benmamoun (1999) studied the nature and role of the imperfective verb in Arabic. The imperfective verb is not specified for tense. Hence, it is the default form of the verb that does not carry temporal features. This feature of unmarked status for imperfective verbs is consistent with its central role in word formation which allows for a unified analysis of nominal and verbal morphology. In conclusion, a word-based approach for Arabic word formation is more important than root-based.

Morphological analysis for Arabic entails computer applications that analyze Arabic words of a given text and deal with the internal structure. It involves a series of processes that identify all possible analyses of the orthographic word. These processes are both form-based and function-based [9-12]. Morphological analyzers for Arabic text are required to develop processes that deal with both the *form* and the *function* of the word. These processes include tokenization, spell-checking, stemming and lemmatization, pattern matching, diacritization, predicting the morphological features of the word's morphemes, part-of-speech tagging and parsing. Many morphological analyzers for Arabic text were developed using a range of methodologies. These methodologies are: Syllable-Based Morphology (SBM), which depends on analyzing the syllables of the word; Root-Pattern Methodology, which depends on the root and the pattern of the word for analysis; Lexeme-based Morphology, where the stem of the word is the crucial information that needs to be extracted from the word; and Stem-based Arabic lexicons with grammar and lexis specifications [13, 14].

Morphological analyzers are different in their methodologies and their tasks. **Stemmers** are responsible for extracting the stem/root of words [15-21]. **Lemmatizers** identify the canonical form, dictionary form, or citation form, which is also called the lemma for words [22, 23]. **Pattern matching algorithms** generate the templatic form (*i.e.* patterns) and vocalism of the analysed words. However, the representation of the templatic forms and vocalism might vary from one algorithm to another [17, 24-26]. General purpose **morphological analyzers** generate all possible analyses of the words out of their contexts. Key morphological analyzers for Arabic text are: Xerox system [27, 28], Buckwalter's Morphological Analyzer (BAMA) [29, 30], ElixirMF [31], AlKhalil [32], MORPH2 [33, 34] and MIDAD [35].

III. THE PRACTICAL CHALLENGE OF MORPHOLOGICAL ANALYSIS FOR ARABIC TEXT

Several stemming algorithms for Arabic already exist, but each researcher proposes an evaluation methodology based on different text corpora. Therefore, direct comparisons between these evaluations cannot be made. At the time of the experiment, only three stemming algorithms and morphological analyzers for Arabic text were readily accessible to assess their implementation and/or performance results. The three selected algorithms are Khoja's stemmer [16], Buckwalter's morphological Analyzer (BAMA) [29] and the triliteral root extraction algorithm [18].

A range of four fair and precise evaluation experiments was conducted using a gold standard for evaluation consisting of two 1000-word text documents from the Holy Qur'an and the Corpus of Contemporary Arabic. The four experiments on both text samples show the same accuracy rank for the stemming algorithms: Khoja's stemmer achieved the highest accuracy, then the triliteral root extraction algorithm, and finally BAMA. The results show that :

- The stemming algorithms used in the experiments work better on MSA text (*i.e.* newspaper text) than Classical Arabic (*i.e.* Qur'an text), not unexpectedly as they were originally designed for stemming MSA text (*i.e.* newspaper text). The SALMA – Tagger is designed for wide coverage and so can deal with both genres.
- All stemming algorithms involved in the experiments agree and generate correct analysis for simple roots that do not require detailed analysis.

- Most stemming algorithms are designed for information retrieval systems where accuracy of the stemmers is not such an important issue.
- Accuracy rates surveyed show that even the best algorithm failed to achieve an accuracy rate of more than 75%.

IV. Resources for Improving Arabic Morphological Analysis

The previous section raises the following question: How can we improve stemming and morphological analysis for Arabic so the algorithm can deal successfully with the hard cases? We chose to construct a broad-coverage lexical resource, the SALMA - ABCLexicon to improve the accuracy of Arabic morphological analysis.

The SALMA-ABCLexicon was constructed by analysing the text of 23 traditional Arabic lexicons, all of which are freely available open-source documents, and by following an agreed standard for constructing a morphological lexicon from raw text. However, three factors directed the selection of traditional Arabic lexicons as our raw text corpus: (i) the absence of an open-source, large, representative Arabic corpus; (ii) the absence of an open-source generation program; and (iii) the generation programme problems of over-generation and under-generation. The major advantages of using the traditional Arabic lexicons text as a corpus are: the corpus contains a large number of words (14,369,570) and word types (2,184,315), and the possibility of finding the different forms of the derived words of a given root.

The SALMA-ABCLexicon contains 2,781,796 vowelized word-root pairs which represent 509,506 different non-vowelized words. The lexicon is stored in three different formats: tab-separated column files, XML files, and a relational database. It is also provided with access and searching facilities and a web interface that provides a facility for searching a certain root and retrieving the original root definitions of the analyzed traditional Arabic lexicons http://www.comp.leeds.ac.uk/cgi-bin/scmss/arabic_roots.py.

V. Standards for Arabic Morphosyntactic Analysis

The initial evaluation of morphological analyzers and stemmers for Arabic text pointed out the lack of standardization and guidelines for morphosyntactic annotation for Arabic text. These standards and guidelines are the prerequisites for morphosyntactic annotation of corpora. Therefore, eight existing Arabic tag sets were surveyed and compared in terms of purpose of design, characteristics, tag-set size, and their applications.

For a morphologically rich language like Arabic, the Part-of-Speech tag set should be defined in terms of morphological features characterizing word structure. The SALMA – Tag Set has the following characteristics:

- The SALMA – Tag Set captures long-established traditional morphological features of Arabic, in a notation format intended to be compact yet transparent.
- A detailed description of the SALMA – Tag Set explains and illustrates each feature and its possible values.
- A tag consists of 22 characters; each position represents a feature and the letter at that location represents a value or attribute of the morphological feature; the dash “-” represents a feature not relevant to a given word.
- The SALMA – Tag Set is not tied to a specific tagging algorithm or theory, and other tag sets could be mapped onto this standard, to simplify and promote comparisons between and reuse of Arabic taggers and tagged corpora.

The SALMA – Tag Set has been validated in two ways. First, it was validated by proposing it as a standard for the Arabic language computing community, and it has been adopted in Arabic language processing systems. (i) It has been used in the SALMA – Tagger to encode the morphological features of each morpheme [36, 37]. (ii) Parts of The SALMA Tag Set were also used in the Arabic morphological analyzer and part-of-speech tagger Qutuf [38]. (iii) It has been reported as a standard for evaluating morphological analyzers for Arabic text and for building a gold standard for evaluating morphological analyzers and part-of-speech taggers for Arabic text [11].

Second, an empirical approach to evaluating the SALMA Tag Set of Arabic showed that it can be applied to an Arabic text corpus, by mapping from an existing tag set to the more detailed SALMA Tag Set. The morphological tags of a 1000-word test text, chapter 29 of the Quranic Arabic Corpus, were automatically mapped to SALMA tags. Then, the mapped tags were proofread and corrected.

VI. Applications and Implementations

Morphosyntactic analysis is a very important and basic application of Natural Language Processing which can be integrated into a wide range of NLP applications. Arabic has many morphological and grammatical features, including sub-categories, person, number, gender, case, mood, etc. More fine-grained tag sets are often considered more appropriate. The additional information may also help to disambiguate the (base) part of speech.

The SALMA – Tagger is an open-source fine-grain morphological analyzer for Arabic text which puts together the developed resources (*i.e.* mainly the SALMA – ABCLexicon) and standards (the SALMA – Tag Set). It also depends on pre-stored lists (*i.e.* prefixes, suffixes, roots, patterns, function words, broken plurals, named entities, etc.) which were extracted from traditional grammar books. The morphological analyzer was developed to analyze the word and specify its morphological features. It uses a tokenization scheme for Arabic words that distinguishes between five parts of a word’s morphemes as defined by the SALMA – Tag Set. Each part is given a fine-grained SALMA Tag that encodes 22 morphosyntactic categories of the morpheme (or possibly multiple tags if the part has multiple clitics or affixes). The SALMA – Tagger consists of several modules which can be used independently to perform a specific task such as root extraction, lemmatizing and pattern extraction. Or, they can be used together to produce full detailed analyses of the words. Figure 1 shows the SALMA – Tagger modules.

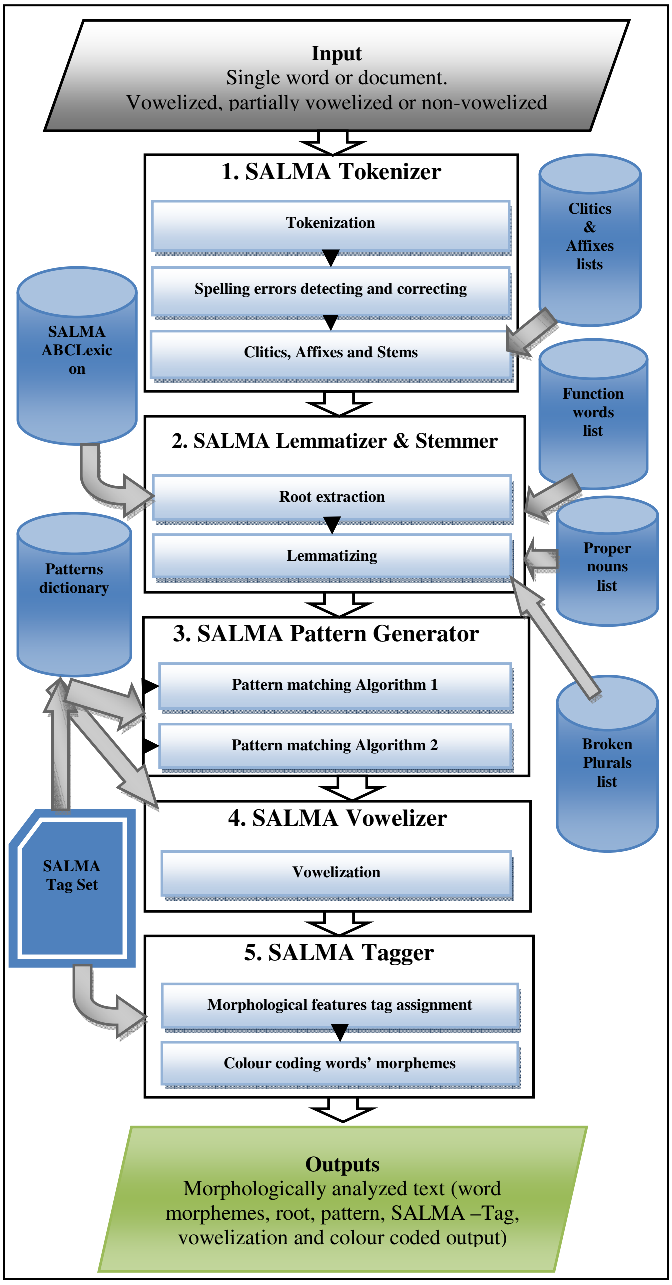

Figure 1: The SALMA Tagger algorithm

VII. Evaluation

The evaluation for the SALMA – Tagger showed that evaluation methodologies for morphological analyzers are not standardized yet. Therefore, we developed agreed standards for evaluating morphological analyzers for Arabic text, based on our experiences and participation in two community-based evaluation contests: The Arab League Educational, Cultural and Scientific Organization (ALECSO) and King Abdul-Aziz City of Science and Technology (KACST) initiative on morphological analysers of Arabic text; and the MorphoChallenge 2009 competition.

The developed evaluation standards depend on using gold standards for evaluating morphological analyzers for Arabic text. A reusable general purpose gold standard (the SALMA – Gold Standard) was constructed to evaluate various morphological analyzers for Arabic text and to allow comparisons between the different analyzers. The SALMA – Gold Standard is adherent to standards, and enriched with fine-grained morphological information for each morpheme of the gold standard text samples. The detailed information is: the input word, its root, lemma, pattern, word type and the word's morphemes. For each of the word's morphemes, the morpheme type is classified into proclitic, prefix, stem, suffix and enclitic, and a fine-grain SALMA Tag which encodes 22 morphological feature categories of each morpheme, is also included.

The SALMA – Gold Standard contains two text samples of about 1000-words each representing two different text domains and genres of both vowelized and non-vowelized text taken from the Qur'an – chapter 29 representing Classical Arabic, and from the CCA representing Modern Standard Arabic. Figure 2 shows a sample of the Qur'an gold standard in tab separated column file.

وَوَصَّيْنَا	وصي	فَعَّلْنَا	وَ p--c------------------
			وَصَّيْ v-p---mpfs-s-amohvtt&-
			نَا r---r-xpfs-s----------
الْإِنْسَانَ	أنس	فِعْلَان	الْ r---d----------------
			إِنْسَانَ nq----ms-pafd---hdbt-s
بِوَالِدَيْهِ	ولد	فَاعِل	بِ p--p-----------------
			وَالِدَ nq----ms-pafd---hdbt-s
			يْ r---r-xdts-s----------
			هِ r---r-msts-k----------
حُسْنًا	حسن	فُعْل	حُسْنً ng----ms-vafi---ndst-s
			ا r---k------f----------

Figure 2: A sample of the SALMA – Gold Standard, Qur'an part, stored using text file

The evaluation using the SALMA – Gold Standard focused on measuring the prediction accuracy of the 22 morphological features encoded in the SALMA – Tags for each of the gold standard's text samples morphemes. The results show that 53.50% of the Qur'an text sample morphemes and 71.21% of the CCA text sample were correctly tagged using "exact match" of the gold standard's morpheme tags, but some of the errors were very minor such as replacing '?' by '-'. These results of applying the SALMA – Tagger show that fine-grained morphological analysis for Arabic text is practical. The results show the applicability of the SALMA – Tagger to process different types of text types, domains and genres of both vowelized and non-vowelized Arabic text. The SALMA – Tagger can be used to POS-tag Arabic text corpora and to provide detailed fine-grained analysis for each morpheme of the corpus words.

Moreover, these general results and the individual accuracy rates reported for each morphological feature show that the linguistically-informed knowledge-based system for predicting the values of the morphological feature categories is applicable to Arabic morphological analysis. The traditional Arabic grammar rules are leveraged to inform and construct the knowledge-based system for predicting the attribute values of the morphological feature categories.

The individual category accuracy results are useful for users who will use/reuse the SALMA – Tagger or parts of it, to know in advance the prediction accuracy of the attributes of each morphological feature category. Prediction accuracy was high for 15 morphological feature categories: namely, 98.53%-100%for the CCA test sample and 90.11%-100% for the Qur'an test sample. These categories are: main part-of-speech; subcategory of verb; subcategory of particle; subcategory of other (residual); punctuation; definiteness; voice; emphasized and non-emphasized; transitivity; declension and conjugation; unaugmented and augmented; number of root letters; verb roots; and noun finals.

The remaining 7 morphological feature categories, namely: the subcategory of noun; gender; number; person; inflectional morphology; case or mood; case and mood marks; and the morphological feature of rational, achieved slightly lower prediction accuracy: 81.35%-97.51%for the CCA test sample and 74.25%-89.03% for the Qur'an test sample.

VIII. CONCLUSIONS

Several computational linguists have designed and developed algorithms to address problems in automatic morphological annotation of Arabic text. This paper has surveyed current Arabic morphological analyzers, and conducted experiments to discover the theoretical and practical challenges of morphological analysis for Arabic. Practical work includes the development of resources to enhance the accuracy of such systems, where these resources can also be reused in diverse Arabic text analytics applications. It also includes the development of linguistically-informed standards for Arabic morphological analysis which draw on long-established traditions of Arabic grammar. Finally, resources and standards are brought together in the development of the SALMA – Tagger: a fine-grained morphological analyzer for Arabic text of different domains, formats and genres.

REFERENCES

1. Sawalha, M., *Open-source Resources and Standards for Arabic Word Structure Analysis*, in *School of Computing*. 2011, University of Leeds: Leeds.
2. Jurafsky, D. and J.H. Martin, *Speech and Language Processing*. Second Edition ed. Prentice Hall Series in Artificial Intelligence, ed. S. Russell and P. Norvig. 2008, New Jersey: Prentice Hall. 1024.
3. Kiraz, G.A., *Computational Nonlinear Morphology with Emphasis on Sematic Languages*. Studies in Natural Language Processing, ed. B. Boguraev and S. Bird. 2001, Cambridge: Cambridge Univeristy Press.
4. Pauw, G.D. and G.-M.D. Schryver, *Improving the Computational Morphological Analysis of a Swahili Corpus for Lexicographic Purposes.* Lexikos 18 (AFRILEX-reeks/series 18: 2008), 2008: p. 303-318.
5. Al-Sughaiyer, I.A. and I.A. Al-Kharashi, *Arabic morphological analysis techniques: A comprehensive survey.* Journal of the American Society for Information Science and Technology, 2004. **55**(3): p. 189-213.
6. Haywood, J.A. and H.M. Nahmad, *A New Arabic Grammar of the Written Language*. Second edition ed. 1965, London: Lund Humphries. 687.
7. McCarthy, J. and A. Prince. *Prosodic morphology and templatic morphology*. in *Perspectives on Arabic Linguistics: Papers from the Second Symposium*. 1990. Amsterdam: Benjamins, Amsterdam.
8. McCarthy, J. and A. Prince, *Foot and word in prosodic morphology: The Arabic broken plurals.* Natural Language & Linguistic Theory, 1990. **8**: p. 209–282.
9. Habash, N.Y., *Introduction to Arabic Natural Language Processing*. Synthesis lectures on Human Language Technologies #10. 2010: Morgan & Claypool Publishers.
10. Hamada, S. المحللات الصرفية للغة العربية *"Morphological Analyzers for Arabic"*. in *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.* 2009. Damascus, Syria.
11. Hamada, S., مقترح لمعايير وضوابط تقييم المحلِّلات الصرفية *Evaluation of the Arabic Morphological Analyzers* in *Proceedings of The Sixth International Computing science Conference ICCA*. 2010: Hammamet, Tunisia.
12. Thabet, N. *Stemming the Qur'an*. in *COLING 2004, Workshop on computational approaches to Arabic script-based languages.August 28,2004*. 2004.
13. Soudi, A., A.v.d. Bosch, and G. Neumann, eds. *Arabic Computational Morphology. Knowledge-based and Empirical Methods*. Text, Speech and Language Technology 38, ed. N. Ide and J. Véronis. Vol. 38. 2007, Springer: Dordrecht, The Netherlands.
14. Soudi, A., V. Cavalli-Sforza, and A. Jamari, *A Computational Lexeme-Based Treatment of Arabic Morphology*, in *ACL/EACL 2001 Workshop on Arabic NLP*. 2001: Toulouse, France, Friday 6 July 2001.
15. Khoja, S. *APT: Arabic Part-of-Speech Tagger*. in *Student Workshop at the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL2001)*. 2001. Carnegie Mellon University, Pittsburgh, Pennsylvania.
16. Khoja, S., *APT: An Automatic Arabic Part-of-Speech Tagger*, in *Computing Department*. 2003, Lancaster University: Lancaster, UK. p. 157.

17. Al-Shalabi, R., *Pattern-based Stemmer for Finding Arabic Roots.* Information Technology Journal, 2005. **4**(1): p. 38-43.
18. Al-Shalabi, R., G. Kanaan, and H. Al-Serhan, *New approach for extracting Arabic roots*, in *in ACIT '2003: Proceedings of The 2003 Arab conference on Information Technology*. 2003: Alexandria, Egypt.
19. Al-Sughaiyer, I.A. and I.A. Al-Kharashi, *Rule Parser for Arabic Stemmer* in *Text, Speech and Dialogue*. 2002, Springer Berlin / Heidelberg. p. 11-18.
20. AlSerhan, H. and A. Ayesh, *A Triliteral Word Roots Extraction Using Neural Network For Arabic*, in *IEEE International Conference on Computer Engineering and Systems (ICCES06)*. 2006: Cairo, Egypt. p. 436-440.
21. Boudlal, A., et al., *A Markovian Approach for Arabic Root Extraction.* The International Arab Journal of Information Technology, 2011. **8**(1): p. 91-98.
22. Al-Shammari, E. and J. Lin. *A novel Arabic lemmatization algorithm*. in *AND '08: Proceedings of the second workshop on Analytics for noisy unstructured text data*. 2008. Singapore: ACM.
23. Dichy, J., *On lemmatization in Arabic, A formal definition of the Arabic entries of multilingual lexical databases*, in *ACL/EACL 2001 Workshop on Arabic NLP*. 2001: Toulouse, France, Friday 6 July 2001.
24. Dichy, J. and A. Farghaly, *Roots & patterns vs. stems plus grammar-lexis specifications: on what basis should a multilingual database centred on Arabic be built?*, in *MT Summit IX -- workshop: Machine translation for semitic languages*. 2003: New Orleans, USA.
25. Yousfi, A., *The morphological analysis of Arabic verbs by using the surface patterns.* IJCSI International Journal of Computer Science Issues, 2010. **7**(3(11)): p. 33-36.
26. Alqrainy, S., *A Morphological-Syntactical Analysis Approach For Arabic Textual Tagging*, in *2008*. 2008, De Montfort University: Leicester, UK. p. 197.
27. Beesley, K.R., *Arabic finite-state morphological analysis and generation*, in *Proceedings of the 16th conference on Computational linguistics - Volume 1*. 1996, Association for Computational Linguistics: Copenhagen, Denmark.
28. Beesley, K.R., *Arabic morphology using only finite-state operations*, in *Proceedings of the Workshop on Computational Approaches to Semitic Languages*. 1998, Association for Computational Linguistics: Montreal, Quebec, Canada.
29. Buckwalter, T., *Buckwalter Arabic Morphological Analyzer Version 1.0*. 2002, Linguistic Data Consortium, catalog number LDC2002L49 and ISBN 1-58563-257-0.
30. Buckwalter, T., *Buckwalter Arabic Morphological Analyzer Version 2.0*. 2004, Linguistic Data Consortium, catalog number LDC2004L02 and ISBN 1-58563-324-0.
31. Smrz, O., *Functional Arabic Morphology: Formal System and Implementation*, in *INSTITUTE OF FORMAL AND APPLIED LINGUISTICS, FACULTY OF MATHEMATICS AND PHYSICS*. 2007, CHARLES UNIVERSITY IN PRAGUE: Prague. p. 104.
32. Boudlal, A., et al., *Alkhalil Morpho Sys: A Morphosyntactic analysis system for Arabic texts.* IJCSI International Journal of Computer Science Issues, 2010.
33. Hamado, A.-M.B., L. Belghayth, and N. Sha'baan, *MORPH, "الصرفي للغة العربية لمخبر "ميراكل morphological analyzer for Arabic text developed at MIRACL Labs*, in *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.* 2009: Damascus, Syria.
34. Kammoun, N.C., L.H. Belguith, and A.B. Hamadou, *The MORPH2 new version: A rubust morphological analyzer for Arabic text*, in *JADT 2010: 10th International Conference on Statistical Analysis of Textual Data*. 2010: SAPIENZA, Italy.
35. Sabir, M. and A.-M.i. Abdul-Mun'im, برنامج (مداد) للتحليل الصرفي للكلمات العربية *MIDAD morphological analyzer for Arabic text*, in *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.* 2009: Damascus, Syria.
36. Sawalha, M. and E. Atwell. *Linguistically Informed and Corpus Informed Morphological Analysis of Arabic*. in *Proceedings of the 5th International Corpus Linguistics Conference CL2009*. 2009. Liverpool, UK.
37. Sawalha, M. and E. Atwell. *Fine-Grain Morphological Analyzer and Part-of-Speech Tagger for Arabic Text.* in *Language Resource and Evaluation Conference LREC 2010* 2010. Valleta, Malta: European Language Resources Association (ELRA).
38. Altabbaa, M., A. Al-Zaraee, and M.A. Shukairy, *An Arabic Morphological Analyzer and Part-Of-Speech Tagger Qutuf 'قُطُوف'*, in *Faculty of Informatics Engineering*. 2010, Arab International University: Damascus. p. 100.