Statistical-based System for Morphological Annotation of Arabic Texts

Nabil Khoufi
ANLP Research Group
MIRACL Laboratory
University of Sfax
nabil.khoufi@fsegs.rnu.tn

Manel Boudokhane
ANLP Research Group
MIRACL Laboratory
University of Sfax
manel.boudokhane@gmail.com

Abstract

In this paper, we propose a corpus-based method for the annotation of Arabic texts with morphological information. The proposed method proceeds in two stages: the segmentation stage and the morphological analysis stage. The morphological analysis stage is based on a statistical method using an annotated corpus. In order to evaluate our method, we conducted a comparative analysis between the results generated by our system AMAS (Arabic Morphological Annotation System) and those carried out by a human expert. As input, the system accepts an Arabic text and generates as a result an annotated text with morphological information in XML format.

1 Introduction

In the linguistic field, morphology is the study of the word’s internal structure. It consists in identifying, analysing and describing the structure of morphemes and other units of meaning in a language.

Morphological annotation in Natural Language Processing (NLP) is considered as a preliminary step during any automatic language processing approach. It consists in attributing labels for each word in a text such as the part of speech (POS) (name, verb, adjective, etc.) the gender (feminine, masculine), the number (singular, dual, plural), etc. Such data are useful in the most of applications of NLP such as text analysis, error correction, parsing, machine translation, automatic summarization, etc. Therefore, developing a robust morphological annotation system is needed.

In this paper, we present a brief description of related Arabic morphological ambiguity. Then, we give an overview of the state-of-the-art. The description of the proposed method for annotation of the Arabic text is thereafter introduced. The following section describes our morphological analysis. An example of analysis is then presented with a brief description of AMAS interface. Finally, we provide the evaluation of our system and a discussion of the obtained results.

2 Arabic Morphological Ambiguity

Like all Semitic languages, Arabic is characterised by a complex morphology and a rich vocabulary. Arabic is a derivational, flexional and agglutinative language. In fact, an Arabic word is the result of a combination of a trilateral or quadrilateral root with a specific schema. Moreover, there are many verbal and nominal lemmas that can be derived from an Arabic root. Furthermore, from a verbal or nominal lemma, many flexions are possibly indicating variations in tense (for verbs), in case (for nouns), in gender (for both), etc. Agglutination in Arabic is another specific phenomenon. In fact, in Arabic, articles, prepositions, pronouns, etc. can be affixed to adjectives, nouns, verbs and particles to which they are related. Derivational, flexional and agglutinative aspects of the Arabic language yield significant challenges in NLP. Thus, many morphological ambiguities have to be solved when dealing with Arabic language. In fact, many Arabic words are homographic: they have the same orthographic form, though the pronunciation is different (Attia, 2006). In most cases, these homographs are due to the non vocalization of words. This means that a full vocalization of words can solve these ambiguities, but most of the Arabic texts like books, web pages, news, etc are not vocalized.

Proceedings of the Student Research Workshop associated with RANLP 2013, pages 100–106,
Hissar, Bulgaria, 9-11 September 2013.

We present, in the following, some of these homographs:

Unvocalized word	فرح		
Vocalized forms	فَرُحْ	فَرْحٌ	فَرِحَ
Meaning	so, go	Marriage	Was happy
International Phonetic Alphabet	farho	farhU	fariħa

Table 1 Homographs due to absence of short vowels

In Arabic some conjugated verbs or inflected nouns can have the same orthographic form. Adding short vowels to those words makes differences between them.

Unvocalized word	يهرب	
Vocalized forms	يُهَرِّبُ	يَهْرِبُ
Meaning	He smuggles	He escapes
International Phonetic Alphabet	i:har ~ibu	i:hribu

Table 2: Homographs due to the absence of the character chadda "ّ"

The presence of chadda inside a particular word changes its meaning.

3 State-of-the-art

The annotation task is an important step in NLP. In fact its accuracy strongly influences the results of the following modules in an NLP process such as parsing. Annotation is also used to create a knowledge base such as annotated corpora, which are helpful for the conception of effective NLP software, especially those based on machine learning techniques. Regarding Arabic text annotation, we identify several methods that can be used. All these methods use the same information to annotate a particular word in a given text: its context and its morphology. What differs is the way to represent these elements and prioritise information. In this section, we focus on morphological analysis which is the main task in a morphological annotation system. The overview of the state of the art of Arabic computational morphology shows that there are two main approaches: the knowledge-based approach and the statistical-based approach (Saoudi et al. 2007).

The knowledge-based approach uses symbolic rules and linguistic information. The designer handles all the labelling rules and the linguistic information (such as Root-base, Lexeme-base, Stem-base…) to perform morphological analysis. Some morphological analysers using knowledge-based methods for Arabic have been developed such as Xerox two-level morphology system (Beesley, 2001) ; Sebawai system (Darwish, 2002) for shallow parsing ; Araparse system (Ouersighni, 2002) ; Buckwalter Arabic morphological analyser (Buckwalter, 2004) ; Attia morphological analyser (Attia, 2006) ; ElixirFM analyser(Smrz, 2007) and Abouenour morphological analyser (Abouenour, 2008).

Statistical-based methods utilize machine learning techniques to extract linguistic knowledge from natural language data directly. In fact, the aim of these methods is to learn how to balance between alternative solutions and how to predict useful information for unknown entities through rigorous statistical analysis of the data. Statistical-based analysers can be grouped in two main families: unsupervised learning analysers and supervised learning analysers. Unsupervised learning analysers learn from a raw corpus without any additional information; they use a distributional analysis to automatically group words into classes or groups of words (grammatical categories). This learning method is not being frequently used (Clark, 2003). On the other hand, supervised learning analysers learn from a prelabelled corpus, which allows the preparation of all the necessary data for annotation. These data are created from dictionaries in order to assign to each word of the corpus a set of information: category, lemma, average frequency of occurrence, etc. To the best of our knowledge, among the systems using supervised learning we can mention the morphological analyser developed by Boudlal (Boudlal et al., 2008) ;TBL Analyser (AlGahtani et al., 2007); MADA morphological analyser (Habash et al., 2009) ;(Mansour et al., 2009) analyser, which is an adaptation of MorphTagger to Arabic language, and Diab analyser (Diab, 2010) which is a part of the AMIRA tool kit for Arabic processing.

As far as we know, statistical-based methods remain largely untapped for Arabic language. Furthermore, the comparison of the results of existing analysers shows that a statistical-based analyser gives better results than a knowledge based analyser (see table 3). These good results depend on the use of large amounts of annotated corpora. Since we have access to the Penn Arabic Treebank (ATB) corpus and assume that the statistical analysers provide better results, we opted for a statistical method to build our annotation system of Arabic texts.

Approach	System	accuracy
Statistical based	Diab	**95.49**
	Habash	**97.5**
	Mansour	**96.12**
	AlGahtani	**96.9**
Knowledge based	Ouersigni	76
	Abouenour	82.14

Table 3: Comparison of evaluation results

4 Proposed Method

Our method for morphological annotation of Arabic texts is based on the machine learning approach. Our method consists of two stages: during the first one, the text is segmented into sentences, which are then segmented into words by using punctuation marks and spaces. Then, the obtained words which are the objects of agglutination are also segmented using a stem database to identify the prefixes, suffixes and clitics of each word. The second stage consists of the morphological analysis of the segmented units by referring to the learning corpus; we apply statistical rules to identify the label having the highest probability and supposed to be the most probable one.

A detailed description of these stages will be given in the following section.

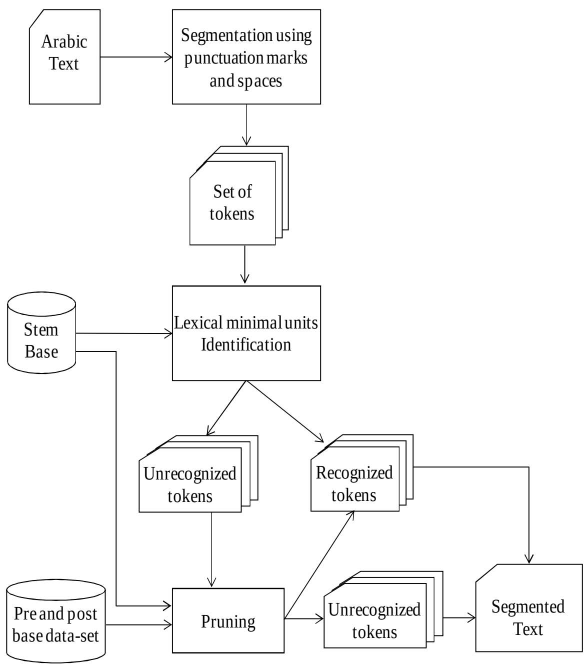
4.1 Principle of The Word's Segmentation

This task is harder for Arabic text than for English or French due to the special features of Arabic as shown in section 2. The main issue of the segmentation is to separate the minimal units such as clitics, prefixes and suffixes from the words.

The principle of our method of segmentation is effective. First of all, we begin by segmenting the text into sentences, then into tokens by using spaces and punctuation marks. We obtain a set of tokens which are compared with the stem database elements to try to identify the lexical minimal units. If the word is recognized in the stem database, it is saved in the segmentation file. Some tokens remain unrecognized in the stem-database. To identify them, we create a pruning process which proceeds as follows:

- Identification of the prefixes, suffixes and clitics in the unrecognized token by referring to the pre- and post-base list.
- Deleting prefixes, suffixes and clitics.
- Comparison of the pruned word with the stem database elements. If the word is found, it is saved as a lexical minimal unit; if not, it is saved as an unknown word.

Figure 1 illustrates the steps of our segmentation stage.



Arabic Text
Segmentation using punctuation marks and spaces
Set of tokens
Stem Base
Lexical minimal units Identification
Unrecognized tokens
Recognized tokens
Pre and post base data-set
Pruning
Unrecognized tokens
Segmented Text

Figure 1: The segmentation stage.

4.2 Principle of The Morphological Analysis

The morphological annotation of a given language consists in assigning the POS (adjective, verb, name ...) and the morphological features (gender, number, person, tense, etc.) to recognized word in the segmented text.

During the morphological analysis, we use an annotated corpus as a knowledge base to predict the morphological category of each word of the input text. This process receives the segmented text (output of the segmentation stage) as an input and generates an annotated text as an output (see figure 2). In this section we begin by presenting the used corpus then we detail the principle of our morphological analysis.

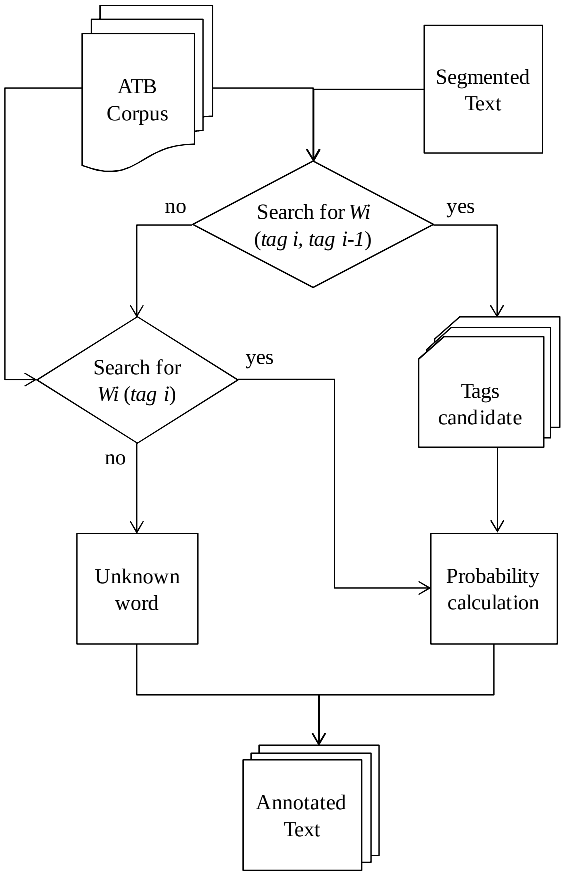


Figure 2: The morphological analysis stage.

Our learning corpus, the Penn Arabic Treebank (ATB), was developed by the LDC at the University of Pennsylvania (Maamouri et al., 2004). It consists of data from linguistic sources written in modern standard Arabic. The corpus consists of 599 unvowelled texts of different stories from a Lebanese news agency publication.

To achieve the morphological annotation of Arabic text, we adopt a statistical method. We use the ATB annotated corpus to extract all possible annotations for a word then we choose the most probable one using the N-gram model, more precisely the first and second order known as unigram and bi-gram. Indeed, (Mustafa and Al-Radaideh, 2004) found that a bi-gram method offers higher overall effectiveness values than the tri-gram method for text processing.

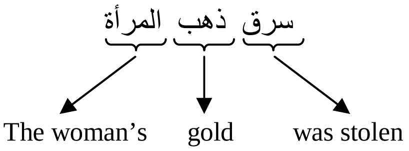
The principle of this model states that the calculation of the probability of occurrence for a given label depends on the label that precedes it. The frequency of the words and the labels will be calculated from the annotated corpus ATB. Probabilities are generated by the conditional probability formula:

$$P\ (t\_i/t\_i\text{-}1)=P(t\_i\text{-}1,t\_i))\ /\ P(t\_i\text{-}1)$$

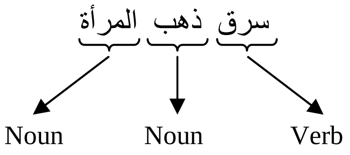
Where *t_i* is the tag of the word *i* and *t_i-1* is the tag of the previous word *i-1*. *P(t_i-1, t_i)* is the sequence frequency of the two words tag *i* and *i-1* . *P (t_i-1)* is the frequency of the tag *t_i-1*.

The analysis process is as follows: We perform a search in the learning corpus for occurrences of the word i (Wi in figure 2). We then extract all the morphological tags of this word from the ATB. Probabilities are then distributed to these tags according to the conditional probability formula. The tag that have the highest probability will be used as the annotation of the word i. There is an exception in the use of the formula for the first word of each sentence and also for each word preceded by an unknown word. If the word is not found in the training corpus, the user has the option to manually annotate the unfound word or to keep it as an unknown word. This process occurs in a sequential mode until the annotation of the whole text. We use the same tag set used in the ATB.

We apply our method to a sentence to show the different results.



In this sentence we have three words; the word سرق is a verb; this annotation is obtained using the frequency calculation (Case of the first word in a sentence). The ambiguity lies in the word ذهب which has two possibilities to be annotated: verb or noun according to our learning corpus. To choose the right annotation, we take into consideration the annotation of the previous word (i.e verb). Probabilities are then calculated and we obtain P (verb/verb) =0.2 and P (noun/verb) =0.8. So ذهب as noun is selected because it has the highest probability. The word المرأة is annotated as a noun because it's the only annotation found in the learning corpus. The sentence will be annotated as follows:



5 The AMAS System

The method that we proposed for the morphological annotation of Arabic texts has been implemented through the AMAS (Arabic Morphological Annotation System) system. In this section, we present the implementation details. The implementation of this method has been done using the Java programming language.

5.1 Used Linguistic Data

The stem database: During the segmentation phase, we used the stem base of the morphological analyser AraMorph[1], which is the Java port of the homonym product developed in Perl by Tim Buckwalter. This stem database contains a large number of stems with a lot of other information such as possible vocalized forms, labels, etc. The entire database is written with the Tim Buckwalter transliteration. We made some changes to its structure, which consists in:

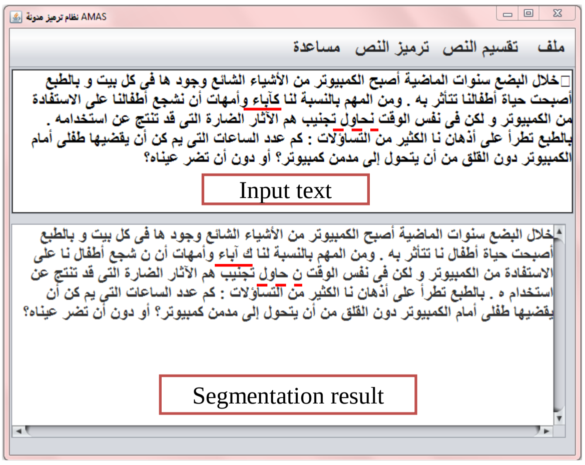
- First, removing unhelpful data such as English translation, syntactic tags and keeping only the stems,
- And second, transliterating the stems from Buckwalter form to the Modern Standard Arabic form.

Pre- and post-base lists: In order to segment agglutinated tokens, a pre- and post-base list is necessary to identify them. The creation of this list was inspired from Abbes's (Abbes, 2004) works. These lists were adapted to our segmentation process. Indeed only pre- and post-bases used in the training corpus were considered in the segmentation process.

5.2 AMAS Interface

As input, the system accepts an Arabic text. Then, the user can segment the text via the segmentation menu. The system then displays the results in a text area as shown in the screen-shot presented in figure 3. The user has the possibility to modify the results if it is necessary to correct some mistakes.

Once the text is segmented and saved in a text file, we proceed to the annotation step using the annotation menu. The user must specify if the analysis should be fully automatic or semi-automatic. If the user chooses the semi automatic option, he must indicate the right annotation to unknown words. The system will be updated with the user annotations. This information can be taken into consideration through the annotation process for the rest of the text. Otherwise, there is no need for the user's contribution and the process will be conducted automatically.
The result is presented in the form of a well-structered XML file. Each Arabic word is presented with its original form, its Buckwalter transliteration, its most probable morphological annotation and its probability (see figure 4).



□خلال البضع سنوات الماضية أصبح الكمبيوتر من الأشياء الشائع وجود ها فى كل بيت و بالطبع
أصبحت حياة أطفالنا تتأثر به . ومن المهم بالنسبة لنا كآباء وأمهات أن نشجع أطفالنا على الاستفادة
من الكمبيوتر و لكن فى نفس الوقت نحاول تجنيب هم الآثار الضارة التى قد تنتج عن استخدامه .
بالطبع تطرأ على أذهان نا الكثير من التساؤلات : كم عدد الساعات التى يم كن أن يقضيها طفلى أمام
الكمبيوتر دون القلق من أن يتحول إلى مدمن كمبيوتر؟ أو دون أن تضر عيناه؟

خلال البضع سنوات الماضية أصبح الكمبيوتر من الأشياء الشائع وجود ها فى كل بيت و بالطبع
أصبحت حياة أطفال نا تتأثر به . ومن المهم بالنسبة لنا ك آباء وأمهات أن ن شجع أطفال نا على
الاستفادة من الكمبيوتر و لكن فى نفس الوقت ن حاول تجنيب هم الآثار الضارة التى قد تنتج عن
استخدام ه . بالطبع تطرأ على أذهان نا الكثير من التساؤلات : كم عدد الساعات التي يم كن أن
يقضيها طفلى أمام الكمبيوتر دون القلق من أن يتحول إلى مدمن كمبيوتر؟ أو دون أن تضر عيناه؟

Figure 3: AMAS's segmentation interface.

6 Obtained Results

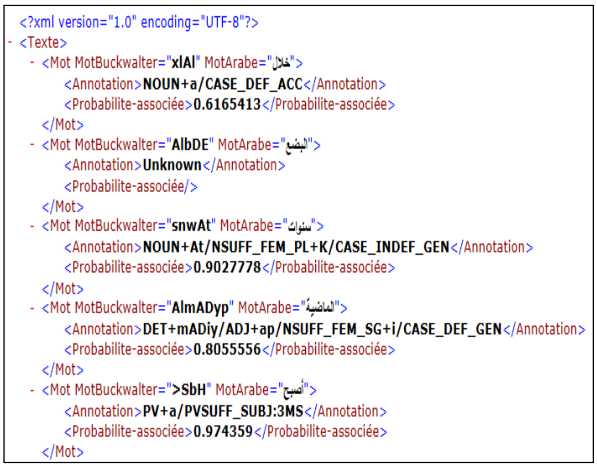
In order to evaluate our system, we used the EASC corpus (The Essex Summaries Arabic Corpus) proposed by (El-Hdj et al., 2010). It contains 153 texts covering different areas such as art, health, politics, tourism, etc. we performed the evaluation on 22 texts containing 10148 segmented words. We then conducted a comparative study between the results generated by our system (automatic annotation process) and those presented by a linguist.
The evaluation operation consists in calculating the recall and precision measures for each domain in the corpus. The average of those measures is then calculated. The average measures for Precision, Recall and F-score are respectively 89.01%, 80.24% and 84.37%.
These results are encouraging and represent a good start for the application of statistical approach for annotation of Arabic texts. Our results are better than Ouersigni and Abouenour systems results which confirm our hypothesis. The difference in performance between our system and state of the art statistical systems is due to the following:

[1] http://www.nongnu.org/aramorph/english/index.html, free software licensed under GPL.

- The propagation of segmentation errors to the morphological analysis involves annotation errors. For example, there is a problem in the segmentation of the agglutination of the article “ال” and the preposition “ل”(e.g التنقل + ل ←للتنقل).
- Unknown words annotation during the morphological analysis, which is likely to influence the annotation of the following word and may decrease the accuracy of the system,
- The way to choose annotation for the first word of a sentence is not precise enough.

Another reason for these results is that the ATB doesn’t contain all words. Some words like “الأشرعة” or “الكمبيوتر” do not exist in the ATB. There is also a difference between the word’s spelling in the ATB and the test corpus. For example, the same word is written “إستخدام” in the test corpus and “ٱستخدام” in the ATB. These two words have the same meaning and same morphological annotation but “ٱستخدام” is annotated as unknown word by our system.

Neverthless, our annotation system produces good results and annotate the majority of the words.



```
  <?xml version="1.0" encoding="UTF-8"?>
- <Texte>
  - <Mot MotBuckwalter="xlAl" MotArabe="خلال">
      <Annotation>NOUN+a/CASE_DEF_ACC</Annotation>
      <Probabilite-associée>0.6165413</Probabilite-associée>
    </Mot>
  - <Mot MotBuckwalter="AlbDE" MotArabe="البضع">
      <Annotation>Unknown</Annotation>
      <Probabilite-associée/>
    </Mot>
  - <Mot MotBuckwalter="snwAt" MotArabe="سنوات">
      <Annotation>NOUN+At/NSUFF_FEM_PL+K/CASE_INDEF_GEN</Annotation>
      <Probabilite-associée>0.9027778</Probabilite-associée>
    </Mot>
  - <Mot MotBuckwalter="AlmADyp" MotArabe="الماضيةِ">
      <Annotation>DET+mADiy/ADJ+ap/NSUFF_FEM_SG+i/CASE_DEF_GEN</Annotation>
      <Probabilite-associée>0.8055556</Probabilite-associée>
    </Mot>
  - <Mot MotBuckwalter=">SbH" MotArabe="أصبح">
      <Annotation>PV+a/PVSUFF_SUBJ:3MS</Annotation>
      <Probabilite-associée>0.974359</Probabilite-associée>
    </Mot>
```

Figure 4: Annotation results.

7 Conclusion and Perspectives

In this paper, we outlined some problems of computational Arabic morphology. Then, we proposed our method for morphological annotation of Arabic texts. We also presented our Arabic morphological annotation system AMAS based on the proposed method. AMAS is implemented using the Java programming language and has been evaluated using EASC corpus. The obtained results are very encouraging (i.e. precision = 89.01% ; recall = 80.24% ; F-measure = 84.37%). As a perspective, we intend to add a stem database to reduce the number of unknown words in the morphological analysis. In addition, we plan to expand n-gram model from 2 to 4. Indeed, It is shown (McNamee and Mayfield, 2004) that the use of n-grams of length 4 is most effective and stable for European languages.

Acknowledgments

The authors would like to thank Mr. Chafik Aloulou and Mrs. Lamia Hadrich Belguith for their valuable comments and participation in the progress of this work within the ANLP-MIRACL laboratory.

References

Ramzi Abbes. 2004. *La conception et la réalisation d'un concordancier électronique pour l'arabe*, Ph.D thesis, ENSSIB/INSA, Lyon, France.

Lahsen Abouenour, Said El Hassani, Tawfiq Yazidy, Karim Bouzouba and Abdelfattah Hamdani. 2008. Building an Arabic Morphological Analyzer as part of an Open Arabic NLP Platform. In *Workshop on HLT and NLP within the Arabic world: Arabic Language and local languages processing Status Updates and Prospects At the 6th Language Resources and Evaluation Conference (LREC’08)*, Marrakech, Morocco.

Shabib AlGahtani, William Black, and John McNaught. 2009. Arabic part-of-speech-tagging using transformation-based learning. In *Proceeedings of the 2nd International Conference on Arabic Language Resources and Tools*, Cairo, Egypt.

Mohammed Attia. 2006. An Ambiguity controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks. *In The challenge of Arabic for NLP/MT conference, the British Computer Society Conference*, pages 48-67, London.

Kenneth Beesley. 2001. Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001. In *Proceedings of the Arabic Language Processing: Status and Prospect-39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France.

Abderrahim Boudlal, Rachid Belahbib, Abdelhak Lakhouaja, Azzeddine Mazroui, Abdelouafi Meziane and Mohamed Ould Abdallahi Ould Bebah. 2008. A Markovian Approach for Arabic Root Extraction. In *The International Arab Conference on Information Technology*, Hammamet, Tunisia.

Tim Buckwalter. 2004. Issues in Arabic Orthography and Morphology Analysis. In *The Workshop on Computational Approaches to Arabic Script-based Languages*, COLING, Geneva.

Alexander Clark. 2003. Combining Distributional and Morphological Information for Part of Speech Induction. In *Proceedings of the tenth Annual Meeting of the European Association for Computational Linguistics EACL 2003*, pages 59-66.

Kareem Darwish. 2002. Building a Shallow Arabic Morphological Analyzer in One Day. In *Proceedings of the ACL-02 workshop on Computational approaches to semitic languages*, Stroudsburg, PA, USA.

Mona Diab. 2009. Second Generation Tools (AMIRA 2.0): Fast and Robust Tokenization, POS tagging, and Base Phrase Chunking, *MEDAR 2nd International Conference on Arabic Language Resources and Tools*, Cairo, Egypt.

Mahmoud El-Haj, Udo Kruschwitz, and Chris Fox. 2010. Using Mechanical Turk to Create a Corpus of Arabic Summaries. In *the Language Resources and Human Language Technologies for Semitic Languages workshop held in conjunction with the 7th International Language Resources and Evaluation Conference*, pages 36-39, Valletta, Malta.

Nizar Habash, Owen Rambow and Ryan Roth. 2009. MADA+TOKAN: A Toolkit for Arabic Tokenization, Diacritization, Morphological Disambiguation, POS Tagging, Stemming and Lemmatization. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR),* Cairo, Egypt.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. In *NEMLAR International Conference on Arabic Language Resources and Tools*.

Saib Mansour, Khalil Sima'an and Yoad Winter. 2007. Smoothing a Lexicon-based POS tagger for Arabic and Hebrew. In *proceedings of ACL 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*. Prague, Czech Republic.

Paul McNamee, James Mayfield. 2004. Character N-Gram Tokenization for European Language Text Retrieval. *In Journal Information Retrieval,* volume 7, issue 1-2, pages 73 – 97.

Suleiman H. Mustafa and Qacem A. Al-Radaideh. 2004. Using N-grams for Arabic text searching. In *Journal of the American Society for Information Science and Technology archive,* Volume 55, Issue 11, Pages 1002-1007 John Wiley & Sons, Inc. New York, USA.

Riadh Ouersighni. 2002. *La conception et la réalisation d'un système d'analyse morpho-syntaxique robuste pour l'arabe : utilisation pour la détection et le diagnostic des fautes d'accord*. Ph.D. thesis, Lumiere-Lyon2 university, France.

Abdelhadi Soudi, Gunter Neumann and Antal Van den Bosch. 2007. Arabic Computational Morphology: Knowledge-based and Empirical Methods. In *Arabic Computational Morphology,* pages 3-14, Springer.

Otakar Smrz. 2007. ElixirFM — Implementation of Functional Arabic Morphology. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources,* pages 1-8, Prague, Czech Republic.