ARABIC PART-OF-SPEECH TAGGING USING THE SENTENCE STRUCTURE

Y.O. Mohamed El Hadj[1], I.A. Al-Sughayeir[1], A.M. Al-Ansari[2]

[1]Center of Research at the College of Computer & Information Sciences
[2]College of Arabic Language
Imam University
P.O.Box. 8488, Riyadh 11681, KSA
m_e_hadj@hotmail.com, imadas@gmail.com, ansary_22@hotmail.com

Abstract

This paper presents a system for Arabic Part-Of-Speech Tagging, which combines morphological analysis with Hidden Markov Model (HMM) and relies on the Arabic sentence structure. On the one hand, the morphological analysis is used to reduce the size of the tags lexicon by segmenting Arabic words in their prefixes, stems, and suffixes due to the fact that Arabic is a derivational language. On the other hand, HMM is used to represent the Arabic sentence structure in order to take into account the logical linguistic sequencing. For these purposes, an appropriate tagging system has been proposed to represent the main Arabic part of speech in a hierarchical manner allowing an easy expansion whenever it is needed. Each tag in this system is used to represent a possible state of the HMM and the transitions between tags (states) are governed by the syntax of the sentence.

A corpus of some old texts, extracted from Books of third century (Hijri), is manually tagged using our developed tagset. and then used for training and testing this system. First experiments conducted on the dataset give a recognition rate of 96% and thus are very promising compared to the data size tagged till now and used in the training.

INTRODUCTION

The computational Processing of the Arabic has gained a more interest in the last few years due to a massive need of computer tools necessary to deal with the huge amount of Arabic data electronically available and, which is dramatically increasing daily (Abdelali et al, 2005). A report published by Madar Research Journal in the year 2005, which includes statistics and forecasts on Internet users in 17 Arab countries, estimated the size of the Internet community in the Arab world in excess of 25 millions (Madar). An update of this study published in march 2008 brings significant news such as a 20- fold increase in the total number of Arabic Web pages produced collectively by 12 countries in the two- year period (2006 and 2007), with growth ranging from as little as 11 fold in Saudi Arabia to an outstanding 163 fold in Syria. Moreover, a study from the Research Unit of Internet Arab World magazine states that there are 1.9 million online websites in Arabic and that number is expected to double every year (IAWRU). In addition of Arabic content on the web, there are many initiatives for developing electronic libraries and corpora of various types for wide range of research purposes (Alansary et al, 2007; Sulaiti & Atwell, 2006).

Providing users with a high quality tools for linguistic processing is essential to keep up with the growth, and still need contribution from all the scientific community.

One of the basic tools and components necessary for any robust Natural Language Processing infrastructure of a given language, is Part-Of-Speech tagging (POST) also known PoS-tagging or just Tagging (Atwel et al, 2004; Alansary et al, 2008). It is considered as one of the basic tools needed in speech recognition, natural language parsing, information retrieval and information extraction. Moreover, POST is also considered as first stage for analyzing and annotating corpora.

Our contribution in this paper concerns the development of an Arabic Part-Of-Speech Tagging system, which combines morphological analysis with a statistical approach that relies on the Arabic sentence structure.

POS-TAGGING TECHNIQUES

POST is the process by which a specific tag is assigned to each word of a sentence to indicate the function of that word in the specific context (Jurafsky & Martin, 2008). Arabic POST (APOST) is not an easy task due to the high ambiguity results from the absence of diacritics and also from the complexity of the Arabic morphology. Consider the following example: "عالم علم رجالا". Each word in the above example has more than one morphological analysis. The APOST is responsible for assigning to each word the most appropriate morphological tag.

There are three general approaches to deal with the tagging problem:

1. **Rule-based approach:** consists of developing a knowledge base of rules written by linguists to define precisely how and where to assign the various POS tags.
2. **Statistical approach:** consists of building a trainable model and to use previously-tagged corpus to estimate its parameters. Once this is done, the model can be used to automatically tagging other texts. Successful statistical taggers were built during the last years and are mainly based on Hidden Markov Models (HMMs).
3. **Hybrid approach:** Consists in combining rule-based approach with a statistical one. Most of the recent works use this approach as it gives better results.

Different Arabic taggers have recently emerged, some of them are developed by companies (Xerox, Sakhr, RDI) as commercial products, while others are a result of research efforts in the scientific community (Khoja, 2001; Freeman, 2001; Maamouri & Cieri, 2002; Diab et al, 2004; Banko & Moore, 2004; Tlili-Guiassa, 2006). Among these works, Khoja (2001) combines statistical and rule-based techniques and uses a tagset of 131 basically derived from the BNC English tagset. (Freeman, 2001) is based on the Brill tagger and uses a machine learning approach. A tagset of 146 tags, based on that of Brown corpus for English is used. (Maamouri & Cieri, 2002) is based on the automatic annotation output produced by the morphological analyzer of Tim Buckwalter (Buckwalter, 2004); it achieved an accuracy of 96%. Diab et al (2004) use Support Vector Machine (SVM) method and the LDC's POS tagset, which consists of 24 tags. Banko and Moore (2004) presents a HMM tagger that exploits context on both sides of a word to be tagged. It is evaluated in both the unsupervised and supervised cases and achieves an accuracy of about 96%. Tlili-Guiassa (2006) uses a hybrid method of based-rules and a memory-based learning method. A tagset composed of symbols from Khoja's tagger and new ones is used and a performance of 85% was reported.

Almost all of these taggers, either use tagsets derived from English which is not appropriate for Arabic, either they rely on a transliteration of the Arabic input text. An other important point is that the structure of the Arabic sentence does not generally taken into account during the tagging process and, in our knowledge, few works are interested to that (Shamsi & Guessoum, 2006).
In this paper, we present a system for Arabic Part-Of-Speech Tagging that relies on the Arabic sentence structure and combines morphological analysis with Hidden Markov Models (HMMs) as we will explain in the following section.

OUR APPROACH FOR ARABIC POST

In this work, a form of combination between statistical and linguistic approaches will be employed, so that the processing will be performed in two levels. In the first level, text is firstly normalized and tokenized into words, and then morphologically analyzed. The morphological analysis is used as input module to reduce the size of the needed tags' lexicon by segmenting Arabic words in their prefixes, stems, and suffixes. This is very important due to the fact that Arabic is a derivational language. For this purpose, an appropriate tagging system has been proposed to represent the main Arabic part of speech in a hierarchical manner allowing an easy expansion whenever it is needed.

In the second level, an appropriate statistical model based on the internal structure of the Arabic sentence is used to recognize the morphological characteristics of the words for the entered text. The use of the linguistic internal structure of the Arabic sentence will allow us to identify logical sequences of words, and consequently their corresponding tags. Since the probability of a certain word (or its tag) occurrence depends on the words preceding it in a given context, the HMM will be the best suitable statistical model to keep track of this history. A linguistic study is conducted to determine the Arabic sentence structure by identifying the different main forms of both nominal and verbal sentences. Based on that, a HMM model is then used to represent this structure. Each state of the HMM is represented by a possible tag in the lexicon and the transitions between states (tags) are governed by the syntax of the sentence. Transition' probabilities are calculated using a smoothed tri-gram and a special processing is used to handle unknown words to determine their lexical probabilities.

Before giving the details of our Arabic POS tagger, a linguistic study of Arabic words and grammatical structures will be required for the purpose of coding morphological characteristics and for extracting the most appropriate structure for common Arabic sentence' forms.

DESCRIPTION OF THE TAGGING SYSTEM

We investigated the principle aspects of Arabic morphology and grammar. The following is a brief review of those aspects. The Arabic verbal structures are composed of three classes: noun (اسم), verb (فِعل) and that we will call particle (حَرف).

NOUN

It is either a name or a word that describes a person, thing or idea. It could be definite or indefinite and can be subcategorized by the person (narrator, interlocutor and absent), number (Singular, Dual, Plural), gender (Masculine, Feminine), and grammatical cases ("الرفع"، "النصب"، "الجر"). Fig1 gives a main classification of the noun and its prominent ramifications.

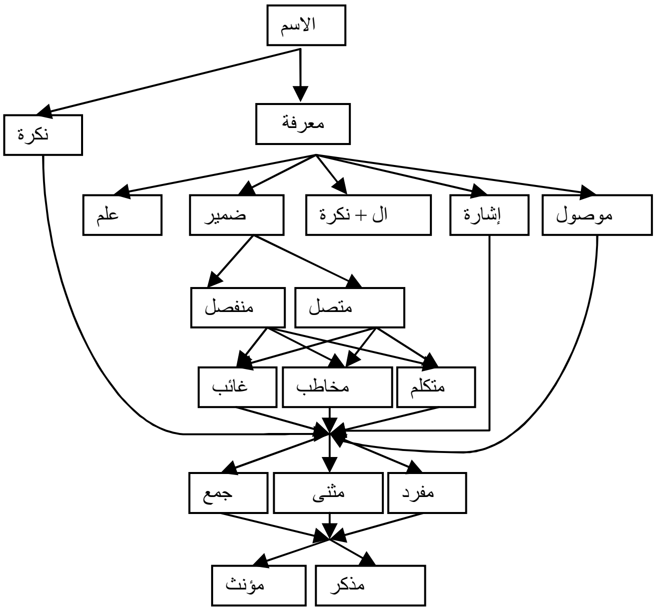

Fig. 1: Noun and its sub-categories

VERB

It is a word that denotes an action and could be combined with some particles. In term of tense (see Fig. 2), the verb could be past (imperative), present (imperfect) or imperative. A future verb tense exists, but it's a derivative of the present tense that you achieve by attaching a prefix to the present tense of the verb. Particles can be added as prefixes and/or suffixes indicating the number, gender, and person of the subject, like for example: قالت, يقول, يقولان, يقولون.
Three moods are possible for verbs: indicative "الرفع", subjunctive "النصب", and jussive "الجزم"

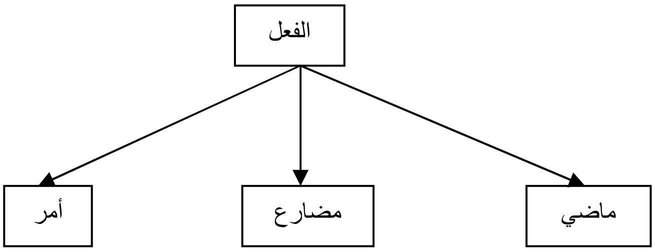

Fig. 2: verb and its temporal-forms

PARTICULE

This class includes everything that is neither a verb nor a noun. It contains the “jarr” prepositions, the coordination prepositions and the functional words like “inna wa akhawatuha إن وأخواتها” which influences the upcoming words analysis. There are many prepositions, but we do not really need, at least in this phase of work, to give an exhaustive list of them. In fact, our objective is not to know them in detail. Fig. 3 gives an example of the classification of particles, according to their functions.

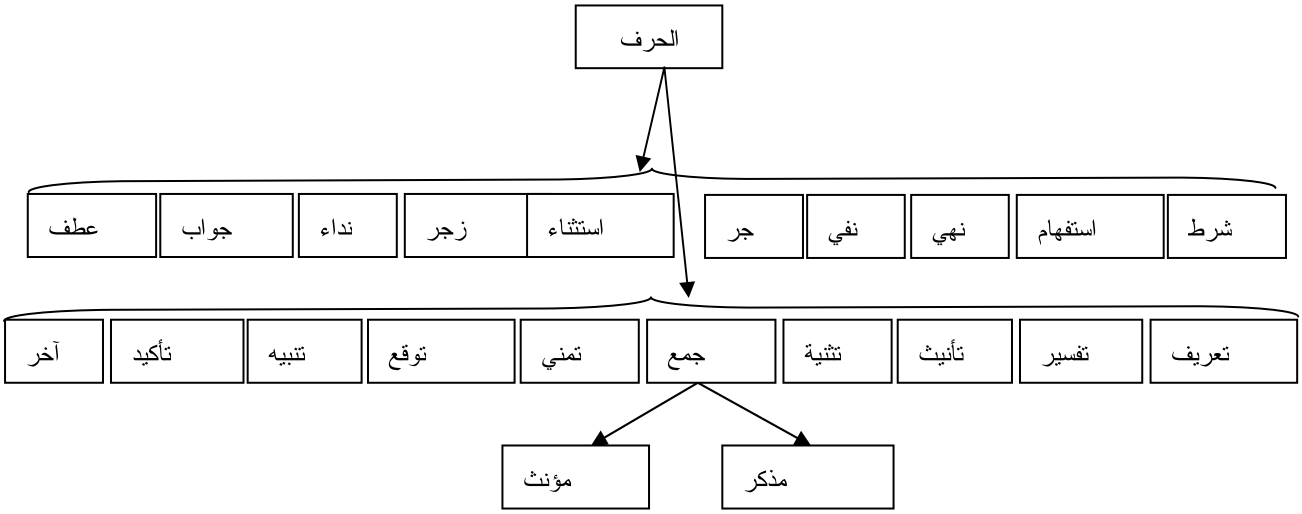

Fig. 3: main groups of particles

PROPOSED TAGSET

The previous classification is used to develop an appropriate tagging scheme considering the parts of speech hierarchy in order to make it meaningful and easily expandable to include more details and precision about the Arabic units whenever it is needed.

As we have seen before, the noun could be defined or undefined. We will give the noun in its global format the symbol “NoIf”. In its defined format, it will get the symbol “NoPr” if it is a proper name, the symbol “NoPn” if it's a pronoun, “NoDe” if it's a demonstrative pronoun, “NoCn” if it's "اسم موصول". Because the pronoun could be attached "متصل" or not attached "منفصل" to another word, So we will use “NoPnAt” to tag the first one and “NoPnSe” to tag the second one. To indicate the gender, number and person, we will add respectively the letters M, F, S, D, P, 1, 2 or 3.

As far as the verb is concerned, it will be given the symbol “Ve” globally, “VePe” for the Perfect, “VeIf” for the Imperfect and “VeIa” for the imperative.

Regarding the class of particles, tags are specified only for some ones that are of subject matter for our work in its initial phase. Among those, “PaDe” is used to tag the identifier (أل), “PaDu” and “PaPl” are respectively used for tagging particles indicating the number (dual and plural). For indicating the gender, the letters M or F can be used. The remaining particles are assigned the tag “PaOt”, but they can be tagged separately following the same logic. “Pa” is the global tag given to the particle if we do not need to distinguish a particular one.

Finally, we will assign to the punctuation signs (., ?, !, etc) the symbol “Pu”. The digits and dates are denoted by the symbol “Nu”.

SPECIFICATION OF THE SENTENCE STRUCTURE: MODEL ARCHITECTURE

A linguistic study has been conducted to extract common types of formulations of the Arabic sentence, so that it can serve as architecture of the statistical model. The references of this study were the old morphology books and modern studies concerned with sentence structures in the Arabic language such as the following references [Harkat, Mutawakkil, Al-Rahhali, Al-Shukri, Yaqut].

The sentence in the Arabic language is either nominal like in “الشمس ساطعة” or verbal like in “يلعب الأطفال الكرة”. Each of them may have different forms and styles. A list of more than 100 ways of common grammatical structures in the Arabic language has been surveyed. It covers the general syntactical analysis and detailed morphological analysis of the nouns and verbs.

STRUCTURE OF NOMINAL SENTENCES

Different forms of formulation have been identified for nominal sentences. They can be represented by the following figure (fig. 4) in terms of sequences, where V, N, and P respectively denote NOUN, VERB, and PARTICLE. S and E are special states, used to represent the start and the end of the nominal phrase. Notice that a loop on a state indicates a certain number of repetitions of this symbol, and an arrow between two sates, means that first one may be followed by the second one depending on its direction.

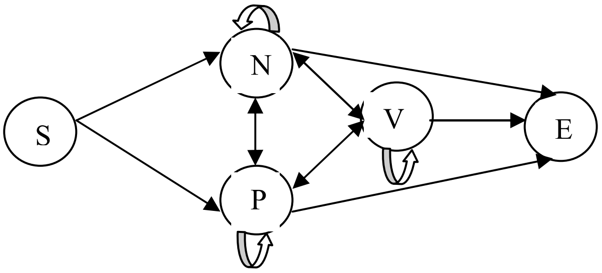

Fig. 4: Structure of Nominal sentences

STRUCTURE OF VERBAL SENTENCES

Verbal sentence structure can be represented by a graph as in the following figure (Fig. 5). This means that a verbal sentence starts either by a verb or a particle and is fallowed by any combination of the main parts of speech.

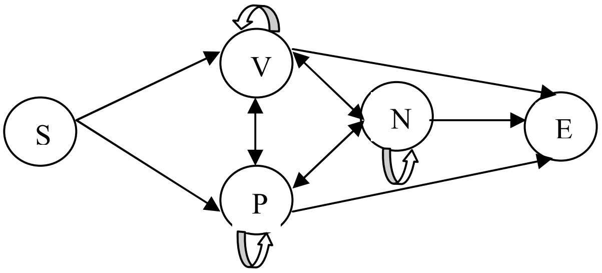

Fig. 5: Structure of Verbal sentences

ARCHITECTURE OF THE STATISTICAL MODEL

Although the previous representation of both nominal and verbal sentences' structures can be seen as trivial and straightforward, they are very interesting for specifying the architecture of our HMM model. It suffuses to combine them in a one graph and to replace each state by the underlying part of speech, and then expand it to include its subcategories as we have specified in the description of the tagset. Each state in the new model (HMM) is representing a valid tag from our lexicon. Determination of the model parameters will be discussed in the following section.

THE HMM-BASED POS TAGGER

The use of a Hidden Markov Model to do part-of-speech-tagging can be seen as a special case of Bayesian inference. It can be formalized as follows: for a given sequence of words, what is the best sequence of tags which corresponds to this sequence of words? If we represent an entered text (sequence of morphological units in our case) by $W = (w_i)_{1 \le i \le n}$ and a sequence of tags from the lexicon by $T = (t_i)_{1 \le i \le n}$, we have to compute:

$$\max_T \left[P(T \mid W)\right].$$

By using the Bayesian rule and then eliminating the constant part $P(W)$, the equation can be transformed to this new one:

$$\max_T \left[P(W \mid T) * P(T)\right].$$

$P(T)$ represents the probability of the tag sequence (tag transition probabilities), and can be computed using an N-gram model (trigram in our case), as follows:

$$P(T = t_1 t_2 \cdots t_n) = \prod_{i=1}^{n} P(t_i \mid t_{i-2} t_{i-1}).$$

A tagged training corpus is used to compute $P(t_i \mid t_{i-2} t_{i-1})$, by calculating frequencies of trigrams and bigrams (respectively $f(t_{i-2} t_{i-1} t_i)$ and $f(t_{i-2} t_{i-1})$) as follows:

$$P(t_i \mid t_{i-2} t_{i-1}) = f(t_{i-2} t_{i-1} t_i) / f(t_{i-2} t_{i-1}).$$

However, it can happens that some trigrams (bigrams) will never appear in the training set; so, to avoid assigning null probabilities to unseen trigrams (bigrams), we used a **deleted interpolation** developed by (Brants, 2000):

$$\lambda_1 * P(t_i \mid t_{i-2} t_{i-1}) + \lambda_2 * P(t_i \mid t_{i-1}) + \lambda_3 * P(t_i),$$

Where $\lambda_1 + \lambda_2 + \lambda_3 = 1$.

Now, for calculating the likelihood of the word sequence given tags $P(W \mid T)$, the probability of a word appearing is generally supposed to be dependent only on its own part-of-speech tag. So, it can be written as follows:

$$P(W \mid T) = \prod_{i=1}^{n} P(w_i \mid t_i).$$

Here also, a tagged training set has to be used for computing these probabilities, as follows:

$$P(w_i \mid t_i) = f(w_i, t_i) / f(t_i),$$

Where $f(w_i, t_i)$ and $f(t_i)$ represent respectively how many times w_i is tagged as t_i and the frequency of the tag t_i itself.

Tag sequence probabilities and word likelihoods represent the HMM model' parameters: transition probabilities and emission (observation) probabilities. Once these parameters are set, the HMM model can be used to find the best sequence of tags given a sequence of input words. The Viterbi algorithm is used to perform this task.

PERFORMANCE EVALUATION

CORPUS PREPARATION

We remember that our ultimate goal is to build an Arabic POS tagger that can be used for relatively old books (from the third century Hijri). Although these texts may be classified as MSA, their styles can vary greatly from those of nowdays. So, we have created a corpus composed of some texts extracted from ALJAHEZ's book "Albayan-wa-tabyin" (255 Hijri). It is obtained from "Ashamila" library, which is downloadable from this link: http://www.shamela.ws. A manual tagging of this corpus using our own tagset is currently running. Due to the complexity of the manual tagging, only a subset of the corpus has been finished till now. It counts a total words of 21882 with a 3565 unique words ranged in more than 1600 sentences. Among these counts, there are 10258 nouns, 2587 verbs, and 9037 particles.

DATA-SETS AND EVALUATION

Our model is trained on 95% of the tagged corpus previously described, using 13 tags: 3 subcategories of verbs, 6 subcategories of nouns, and 4 subcategories of particles. It is tested on the remaining 5%, which represents about 1000 words. To evaluate its performance, we have used the F-measure defined as follows: $(2 * P * R)/(P + R)$, where P and R denotes precision and Recall respectively. They are calculated , using the total number of correct assigned tags (Nc), total number of assigned tags (Na), and the total number of the assigned tags in the test-set (Nt): $P = Nc / Na$ and $R = Nc / Nt$.

We have obtained an accuracy of 96%, which is very encouraging compared to the size of the tagset used till now.

CONCLUSION

In this paper we have presented an Arabic Part-Of-Speech tagger that uses a HMM model to represent the internal linguistic structure of the Arabic sentence. We have conducted a linguistic study to determine the main Arabic POS and to specify different common forms of Arabic sentence. After that, an appropriate tagging system has been proposed to represent these main Arabic parts of speech in a hierarchical manner allowing an easy expansion whenever it is needed. Next, a suitable architecture of the HMM model is specified based-on the structure of both nominal and verbal sentence. Having done this, a corpus composed of old texts extracted from books of third century Hijri is created. A part of it is manually tagged and used to train and to test the tagger. Performance evaluation has shown an accuracy of 96%. However, although this is represents a very good result compared to the size of the training corpus, we have to increase our tagged corpus and to conduct further tests on more interesting dataset to evaluate the real performance of this approach.

We plan to use the developed tagger for our research activities in a variety of ways, especially for applications dealing with old texts "النصوص التراثية".

REFERENCES

Abdelali A., Cowie J., Soliman H.S. (2005). *Building A Modern Standard Arabic Corpus*. Workshop on Computational Modeling of Lexical Acquisation, the split meeting, Croatia.

Alansary S, Nagi M, Adly N. (2008). *Towards Analyzing the International Corpus of Arabic (ICA).* 8th International Conference on Language Engineering, Egypt.

Alansary S, Nagi M, Adly N. (2007). *Building an International Corpus of Arabic (ICA)*. 7th International Conference on Language Engineering, Egypt.

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

Atwell E, Al-Sulaiti L, Al-Osaimi S, Abu-Shawar B. (2004). *A Review of Arabic Corpus Analysis Tools*. Proceedings of JEP-TALN'04 Arabic Language Processing.

Banko M, Moore R. C. (2004). *Part of Speech Tagging in Context*. Proc of the 20th international conference on Computational Linguistics, Switzerland.

Brants T. TnT. *A statistical part of speech tagger*. In *proc. of ANLP'2000, the 6th Conference on Applied Natural Language Processing:* 224-231, Seattle, Washington, Morgan Kaufmann Publishers Inc. 2000.

Diab M., Hacioglu K. and Jurafsky D. (2004). *Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks*. *proc. of HLTNAACL'04*: 149–152.

Freeman A (2001). *Brill's POS tagger and a morphology parser for Arabic*. In ACL'01 Workshop on Arabic language processing.

Internet Arab World research Unit (IAWRU): http://www.teckies.com/lebanon/

Jurafsky D., Martin J.H. (2008). *Speech and Language Processing*: *An introduction to speech recognition, computational linguistics and natural language processing*. 2nd Edition.

[Madar] Madra Research: http://www.madarresearch.com/archive/archive_toc.aspx?id=50.

Maamouri M, Cieri C. (2002). *Resources for Arabic Natural Language Processing at the LDC*. Proceedings of the International Symposium on the Processing of Arabic ,Tunisia, pp.125-146.

Shamsi F, Guessoum A. (2006). *A Hidden Markov Model –Based POS Tagger for Arabic*, JADT'06.

Tlili-Guiassa Y. (2006). *Hybrid Method for Tagging Arabic Text*. Journal of Computer Science 2 (3): 245-248.

Tim Buckwalter. (2004). *Buckwalter Arabic Morphological Analyzer, Version 2.0.* LDC Catalog No. LDC2004L02, Linguistic Data Consortium, www.ldc.upenn.edu/Catalog.