Arabic Natural Language Processing: Challenges and Solutions

Article · January 2009

CITATIONS
379

READS
10,921

4 authors, including:

Ali Farghaly
Monterey Peninsula College
14 PUBLICATIONS **515** CITATIONS
SEE PROFILE

Ali Farghaly
Appen
15 PUBLICATIONS **665** CITATIONS
SEE PROFILE

Khaled Shaalan
British University in Dubai
333 PUBLICATIONS **8,545** CITATIONS
SEE PROFILE

Arabic Natural Language Processing: Challenges and Solutions

ALI FARGHALY
Monterey Institute of International Studies
and
KHALED SHAALAN
The British University in Dubai

The Arabic language presents researchers and developers of natural language processing (NLP) applications for Arabic text and speech with serious challenges. The purpose of this article is to describe some of these challenges and to present some solutions that would guide current and future practitioners in the field of Arabic natural language processing (ANLP). We begin with general features of the Arabic language in Sections 1, 2, and 3 and then we move to more specific properties of the language in the rest of the article. In Section 1 of this article we highlight the significance of the Arabic language today and describe its general properties. Section 2 presents the feature of Arabic Diglossia showing how the sociolinguistic aspects of the Arabic language differ from other languages. The stability of Arabic Diglossia and its implications for ANLP applications are discussed and ways to deal with this problematic property are proposed. Section 3 deals with the properties of the Arabic script and the explosion of ambiguity that results from the absence of short vowel representations and overt case markers in contemporary Arabic texts. We present in Section 4 specific features of the Arabic language such as the nonconcatenative property of Arabic morphology, Arabic as an agglutinative language, Arabic as a pro-drop language, and the challenge these properties pose to ANLP. We also present solutions that have already been adopted by some pioneering researchers in the field. In Section 5 we point out to the lack of formal and explicit grammars of Modern Standard Arabic which impedes the progress of more advanced ANLP systems. In Section 6 we draw our conclusion.

Categories and Subject Descriptors: I 2.7 [**Artificial Intelligence**]: Natural language processing

General Terms: Languages

Additional Key Words and Phrases: Arabic script, Modern Standard Arabic, Arabic dialects

Authors' addresses: A. Farghaly, Monterey Institute of International Studies, Monterey, CA 93940; email: afarghal@miis.edu; K. Shaalan, The British University in Dubai, AlSufuh St., Knowledge Village, Block 17, Dubai United Arab Emirates (UAE), P.O. Box 502216, Dubai; email: khaled.shaalan@buid.ac.ae.

ACM Reference Format:
Farghaly, A. and Shaalan, K. 2009. Arabic natural language processing: Challenges and solutions. ACM Trans. Asian Lang. Inform. Process. 8, 4, Article 14 (December 2009), 22 pages.
DOI = 10.1145/1644879.1644881. http://doi.acm.org/10.1145/1644879.1644881.

1. THE ARABIC LANGUAGE

The Arabic language is both challenging and interesting. It is interesting due to its history [Versteegh 1997], the strategic importance of its people and the region they occupy, and its cultural and literary heritage [Bakalla 2002]. It is also challenging because of its complex linguistic structure [Attia 2008].

At the historical level, Classical Arabic has remained unchanged, intelligible and functional for more than fifteen centuries. Culturally, the Arabic language is closely associated with Islam and with a highly esteemed body of literature. Strategically, it is the native language of more than 330 million speakers [CIA 2008] living in an important region with huge oil reserves crucial to the world economy, and home as well to the sacred sites of the world three monotheistic religions. It is also the language in which 1.4 billion Muslims perform their prayers five times daily. Linguistically, it is characterized by a complex Diglossia situation [Diab and Habash 2007; Farghaly 1999; Ferguson 1959, 1996]. Chronologically Classical Arabic represents the language spoken by the Arabs more than fourteen centuries ago, while Modern Standard Arabic is an evolving variety of Arabic with constant borrowings and innovations proving that Arabic reinvents itself to meet the changing needs of its speakers. At the regional level there are as many Arab dialects as there are members of the Arab league. The diglossic nature of the Arabic language will be discussed in detail in Section 2.

Arabic is a Semitic language spoken by more than 330 million people as a native language, in an area extending from the Arabian/Persian Gulf in the East to the Atlantic Ocean in the West. Arabic is a highly structured and derivational language where morphology plays a very important role [Attia 1999; Beesley 2001; Buckwalter 2004; Farghaly 1987; McCarthy 1981; Soudi et al. 2007]. Arabic NLP applications must deal with several complex problems pertinent to the nature and structure of the Arabic language. For example, Arabic is written from right to left. Like Chinese, Japanese, and Korean there is no capitalization in Arabic. In addition, Arabic letters change shape according to their position in the word. Modern Standard Arabic does not have orthographic representation of short letters which requires a high degree of homograph resolution and word sense disambiguation. Like Italian, Spanish, Chinese, and Japanese, Arabic is a pro-drop language, that is, it allows subject pronouns to drop [Farghaly 1982] subject to recoverability of deletion [Chomsky 1965].

As a natural language, Arabic has much in common with other languages such as English. However, it also is unique in terms of its history, diglossic nature, internal structure, inseparable link with Islam, and the Arabic culture and identity. Any Arabic NLP system that does not take the specific features of the Arabic language into account is certain to be inadequate [Shaalan 2005a;

2005b]. The challenge the Arabic language poses to researchers is not limited to the social aspects of the language, but also extends to its inherent linguistic structure which will be elaborated below.

Over the last few years, Arabic natural language processing (ANLP) has gained increasing importance, and several state-of-the-art systems have been developed for a wide range of applications, including machine translation, information retrieval and extraction, speech synthesis and recognition, localization and multilingual information retrieval systems, text to speech, and tutoring systems. These applications had to deal with several complex problems pertinent to the nature and structure of the Arabic language. Most ANLP systems developed in the Western world focus on tools to enable non-Arabic speakers make sense of Arabic texts. Arabic Tools such as Arabic named entity recognition, machine translation and sentiment analysis are very useful to intelligence and security agencies. Because the need for such tools was urgent, they were developed using machine learning approaches. Machine learning does not usually require deep linguistic knowledge and fast and inexpensive. Developers of such tools had to deal with difficult issues. One problem is when Arabic texts include many translated and transliterated named entities whose spelling in general tends to be inconsistent in Arabic texts [Shaalan and Raza 2008]. For example a named entity such as the city of Washington could be spelled 'واشنجطن' ، 'واشنغطن'، "واشنطن' ، 'وشنطن'. Another problem is the lack of a sizable corpus of Arabic-named entities which would have helped both in rule-based and statistical named entity recognition systems. Efforts are being made to remedy this. For example, the LDC released in May 2009 an entity translation training/dev test for Arabic, English, and Mandarin Chinese. A third limitation is that NLP tools developed for Western languages are not easily adaptable to Arabic due to the specific features of the Arabic language. Recognizing that developing tools for Arabic is vital for the progress in ANLP, the MEDAR consortium has started an initiative for cooperation between Arabic and European Union countries for developing Arabic language Resources [Choukri 2009].

1.1 Significance of ANLP for the Arabic-Speaking Population

Funding for the development of ANLP applications has surged in the U.S. since September 11, 2001. The U.S. Department of Homeland Security was confronted with very difficult tasks ranging from identifying Arabic names correctly at airport security and in Arabic documents seized by the American authorities in the U.S. and abroad. They also had accumulated an enormous volume of Arabic texts that they had no clue as to whether they were relevant or not. They had neither the human expertise needed to perform the task nor the time to wait for human translators to complete the task. ANLP tools that could scan such documents to recognize names, places, dates, etc., of interest soon became essential. As a result, funding became available for companies and research centers to develop tools such as named entity recognition, machine translation, especially spoken machine translation, document categorization, etc.

Because time was of the essence, and in the absence of complete computationally viable grammars of Arabic, statistical approaches that rely primarily on training data and parallel texts gained momentum. Machine learning systems usually give good results when the training set and the testing data are similar. There is also a point at which more training data does not make significant improvement. Moreover, there may be some structures or entities that are sparse. In this case the machine learning component does not have enough data to make the right generalization.

On the other hand, ANLP applications developed in the Arab World have different objectives and usually employ both rule-based and machine-learning approaches. The following are some of the objectives of ANLP for the Arab World:

1. *Transfer of knowledge and technology to the Arab World*. Most recent publications in science and technology are published in the English language and are not accessible to Arab readers with little or no competence in English. To use human translators to translate such an enormous amount of data to Arabic is very costly and time consuming. So Arabic NLP could help reduce the time and cost of translating, summarizing, and retrieving information in Arabic for Arab speakers.
2. *Modernize and fertilize the Arabic language*. This follows from (1) above. Translating new concepts and terminology into Arabic involves coinage, arabization, and making use of lexical gaps in the Arabic language. This will positively effect the revitalization of the Arabic language and enable it to fulfill the essential needs for its speakers.
3. *Improve and modernize Arabic linguistics*. Arabic NLP needs a more formal and precise grammar of Arabic than the traditional grammar so widely employed today. Innovation is needed as well to preserve the valuable heritage of traditional Arab grammarians.
4. *Make information retrieval, extraction, summarization, and translation available to the Arab user*. The hope is to bridge the gap between peoples of the Arab world and their peers in more technically advanced countries. By making information available to Arabic speakers in their native language, Arabic NLP tools empower the present generation of educated Arabs. Thus Arabic NLP tools are indispensable in the struggle of Arabic speakers to attain parity with the rest of the world which is, in turn a matter of national security to the Arab World [Farghaly 2008].

2. ARABIC DIGLOSSIA

Diglossia is a phenomenon [Ferguson 1959, 1996] whereby two or more varieties of the same language exist side-by-side in the same speech community. Each is used for a specific purpose and in a distinct situation. Using the wrong variety in a situation is usually ridiculed. Thus, it differs from the more familiar examples of regional dialects in Italian or Persian where many speakers use their local dialect at home and within the community, but use the standard

language when communicating more formally or with speakers from other regions.

Arabic, however, exhibits a true diglossic situation where at least three varieties of the same language are used within a speech community [Farghaly 2005] and in circumscribed situations. Classical Arabic is the language of religion and is used by Arabic speakers in their daily prayers while Modern Standard Arabic (MSA), a more recent variety of Classical Arabic, is used by educated people in more formal settings such as in the media, classroom, and business. With family, friends, and in the community, people speak their own regional dialect which varies considerably from region to region. These three varieties are available to every Arab on a daily basis. For example, on any given day an Arabic speaker will use Classical Arabic while reciting his daily prayers; MSA when listening to or reading the news, and his particular dialect at home with family or friends.

How then to define diglossia? First, it is not an evolutionary process toward a more standard form of a language. It may develop from different origins and is quite stable. With respect to Arabic, a diglossic situation can be traced to the earliest knowledge of the language and Classical Arabic itself and has remained very stable for over fifteen centuries.

In a diglossic situation a superposed variety is a high variety; in this case Classical Arabic, while the regional dialects are low varieties with MSA occupying an intermediate position. The same can be said for other diglossic situations such as Greek, Creole, and Swiss German [Ferguson 1959]. But what is important is that with each situation, including Arabic, the same features can be found that more clearly define true diglossia.

First, there is a specialization of function for a high or low variety. In one situation it is only appropriate to use the high variety, for example, in a speech, news broadcast, or lecture hall; and the low variety when communicating with family and in more personal settings.

Second, prestige is accorded to the high variety and knowledge of it is assumed to be representative of a speaker's educational level and/or social standing. With Arabic, the prestige accorded Classical Arabic is associated with religion, that is, Islam and the Quran and a highly esteemed literary heritage.

Third, it has been found that in each diglossic situation there is a sizable body of literature written in the high variety that is esteemed by its speech community. For Arabic, there is an extensive legacy of poetry and philosophical and scientific treatises.

Fourth, the way in which each variety is acquired is very important. In Arabic, as well as other diglossia, the low variety or dialect is acquired by children at home and without explicit rules of grammar and is assumed to be acquired "naturally." The high variety, Classical Arabic and MSA are learned at school in the same way any other foreign language would be acquired.

A fifth feature is that in each diglossic situation the high variety has a strong tradition of grammar as is the case with Classical Arabic. There are grammar texts, dictionaries, and works on style and pronunciation. There is an accepted and established norm for grammar, vocabulary, and pronunciation which has minimal variation. In contrast, the same body of standards does not exist

for the low variety and as a result there is considerable variation in grammar, vocabulary and pronunciation. This poses a significant problem with respect to the teaching of Arabic to non-native speakers who are generally taught MSA in the classroom but are frustrated by their inability to communicate more naturally with native speakers. When they attempt to learn a dialect such as Egyptian, Levantine, or Gulf Arabic there are few, if any, resources available to them to convey the proper grammar and method of vocabulary building.

As mentioned earlier, diglossia is not an evolutionary process toward standardization of a language, but in fact the diglossic situation is remarkably stable. With respect to Arabic the two varieties have existed side by side for more than 1,500 years.

The grammar structure demonstrates one of the most explicit differences between varieties as evidenced by certain features present in MSA or Classical Arabic [Ryding 2005] while absent from any dialect. For example, Classical Arabic has three cases marked by case endings where the dialects have none. Further the word order differs as well in that Classical Arabic and MSA have mostly VSO (Verb-Subject-Object) word order while most Arabic dialects are SVO. There are differences in the Wh-construct as well with MSA fronting it and in Egyptian Arabic it is not fronted. And in general, the low variety has a simpler grammatical structure than MSA and less complex morphological structure.

Having described the features of diglossia it can be succinctly defined according to Ferguson [1959]:

> "a relatively stable language situation in which, in addition to the primary dialects of the language, (which may include a standard or regional standards), there is a very divergent, highly codified (often more grammatically complex) superposed variety, the vehicle of a large and respected body of written literature, either of an earlier period or in another speech community, which is learned largely by formal education and is used for most written and formal spoken purposes but is not used by any sector of the community for ordinary conversation."

2.1 Diglossia and ANLP

There are of course, significant implications for developing NLP systems in a diglossic situation like Arabic. First, it is very difficult and almost impossible for any one ANLP single application to process data from all the varieties of Arabic. Each variety has its own grammar, lexicon, and morphology even though they have some properties in common. An ANLP application has to specify beforehand which variety it is aiming to address. Moreover, the application has to have a good "understanding" of the linguistic properties of the particular variety it aims at. An understanding of the complex sociolinguistic situation of Arabic can be very useful for ANLP researchers and developers.

Most of the research and tools have been developed to handle written text that is primarily written in MSA. However, when attempting to apply these tools to transcribed Egyptian or Levantine text, they are far from accurate

since there are significant variations in grammar, syntax, and expressions from one variety to another and between the dialects themselves. Furthermore, there are few resources available in the form of grammars or dictionaries as the basis for developing NLP for the dialects.

One interesting approach to deal with this problem has been that undertaken by researchers at Columbia University [Habash et al. 2005]. In their approach, they have made the assumption that is it simpler to develop NLP systems for the dialects by first extracting and categorizing the systematic grammatical features of a dialect, making it more like MSA and then applying MSA natural language processing tools to process a text. Another approach, built upon the same assumption, is to create Dialect Treebanks that resemble MSA Treebanks by exploiting systematic regularities within a dialect and among dialects. For example the work reported in Shaalan et al. [2007] transfers Egyptian Arabic texts to MSA using a lexical transfer approach in addition to changing the SVO Egyptian order into the MSA VSO order. They also enhanced the tables of Buckwalter's morphological analyzer to transform Egyptian Arabic words into MSA words. Following then same approach one could also reuse MSA tools to process colloquial Arabic.

Farghaly [2005] proposed initiating a line of research that attempts to define what constitutes the Arabic language, and, in turn, assumes that there is a core entity that has well-defined phonological, morphological and syntactic properties. It also proposes that the three main varieties of Arabic, Classical Arabic, the colloquials, and Modern Standard Arabic share a common core or an inter-Arabic grammar. This is likely, given the mutual intelligibility among all speakers of Arabic and the ability of illiterate Muslims to understand the Quran. A core inter-Arabic grammar in addition to Dialect Treebanks and lexical resources for the dialects would greatly facilitate the development of NLP tools and systems for transcribed texts of the Arabic dialects.

2.2 Solutions

The most important solution for Arabic Diglossia is to build resources for the various varieties of Arabic. The LDC has already built corpora for Egyptian, Levantine, and Iraqi Arabic. In addition, there is an important project at Columbia University to build a Treebank of Arabic dialects using resources already available for Modern Standard Arabic. The project exploits systematic mapping of Modern Standard Arabic to some dialects at the phonological and morphological, and lexical levels.

ANLP researchers and developers should be aware of the implications of Arabic Diglossia for their applications since it is hard to build a system that can handle all of the varieties of Arabic simultaneously. Developers must be clear as to which variety of Arabic is appropriate for their specific applications. For example, an application for speech recognition of Arabic telephone conversations will most probably need dialect resources while another for processing Arabic news broadcasts would require Modern Standard Arabic resources whether in the form of linguistic knowledge or in corpora for training purposes.

3. THE ARABIC SCRIPT

One of the key linguistic properties of the Arabic language that poses a challenge to the automatic processing of Arabic is the Arabic script itself. Although Arabic is a phonetic language in the sense that there is one-to-one mapping between the letters in the language and the sounds they are associated with, Arabic is far from being an easy language to read due to the lack of dedicated letters to represent short vowels, changes in the form of the letter depending on its place in the word, and the absence of capitalization and minimal punctuation.

As an example of the regularity of the association of letters to sounds, the letter ('ب') *b* is always pronounced as "baa," unlike letters in English that have more than one pronunciation. For example, the letter "s" in English may be pronounced as /*z*/ as in "cause" or /*s*/ as in "sail" or /*sh*/ as in "sure." Further, while English has silent letters such as the "p" in "pneumatic" the "b" in "doubt," the "k" in "know" and the "gh" in "weight," Arabic has no silent letters. Moreover, Arabic does not combine two letters to produce a new sound. For example, combining the letters "t" and "h" in English sometimes produces a voiceless interdental fricative as in "think" or a voiced interdental fricative as in "though." However, this distinction is highly systematic in English since in most lexical words the "th" is pronounced as it is in "think" while most functional words assume the other pronunciation.

While the Arabic script does not have dedicated letters to represent the short vowels in the language, short vowels have been represented by diacritics which are marks above or below the letters. However, these diacritics have been disappearing in contemporary writings and readers are expected to fill in the missing short vowels through their knowledge of the language. But the absence of short vowels from MSA texts makes it difficult for non-native speakers of Arabic to learn the language and presents challenges to the automatic processing of Arabic.

Arabic letters have different shapes depending on the position of the letter in the word. For example, the letter (ع) "'ain" has an initial shape (عـ), a median shape (ـعـ), a final connecting shape (ـع) , and a final non-connecting shape (ع). The selection of the correct shape relative to its position in the word is rule governed. All Arabic word processors implement these rules so that the user does not have to manually select the correct shape. Hence there is only one key for each letter, and the encoded rules both recognize the context and insert the correct shape automatically. Moreover, there are shapes which the morphological processing tool should handle. For example, the Hamza letter is changed to other forms during the morphological and syntactic generations of the inflected word. For example, the use of the letter "ي" (Yeh to indicate my), with the irregular (broken) plural "زملاء" (colleagues) produces "زملائي" (my-colleagues) instead of "زملاءي."

In English and other Latin script-based languages, most sentences begin with an uppercase letter and end with a period. In NLP applications such as machine translation, information retrieval, clustering, and classification it is necessary to split a running text correctly into sentences and a sentence

splitter capitalizes on these features. But scripts such as Arabic, Chinese, Japanese, and Korean have neither capitalization nor strict rules of punctuation and their absence makes the task of preprocessing a text much more difficult.

Recognizing sentence boundaries in a running text is a more difficult task in languages such as Arabic than it is in languages like English due to the absence of strict punctuation rules. In fact, it is common in Arabic discourse to write an entire paragraph without a single period except at the end of that paragraph. Sentences are often conjoined via the Arabic coordinators (و) *wa* and (ف) *fa* and Arabic discourse is characterized by excessive use of coordination, subordination and logical connectives.

Capitalization and punctuation not only facilitate recognizing sentence boundaries, but also play an important role in the task of named entity recognition (NER) [Benajiba et al. 2008; Farghaly 2007; Shaalan and Raza 2009], which has become an essential component of many NLP applications. Beginning with the 1987 Message Understanding Conference (MUC) and the subsequent series of conferences [Grisham and Sundheim 1996], Information Extraction (IE) has become the focus of research in many NLP applications. While Information Retrieval involves identifying relevant documents in response to queries and ranking them such that the most relevant documents are placed at the beginning, the focus of IE is to extract words and phrases that denote entities, actions, or relations of interest to the user. The task involves extracting from unstructured texts entities such as person names, postal addresses, zip codes, person titles, cities, regions, buildings, etc. Because some of these entities have a strict format, practitioners in the field acknowledge that capitalization and punctuation facilitate recognition of these patterns.

In a language like English, one could easily write a script to recognize a pattern consisting of an uppercase word followed by an initial with an optional period followed by an uppercase word to extract person names such as Hillary R. Clinton and Mary A. Hoffman. There are many patterns that can be recognized and extracted with a high degree of confidence by utilizing capitalization and punctuation rules: street addresses, some company names, some person names, zip codes, phone numbers, Social Security numbers, and e-mail addresses to name but a few. But a computational linguist developing IE applications for languages like Arabic, where the script does not allow for capitalization nor does it follow strict punctuation rules, must have insights into the structure and syntax of the Arabic language to identify patterns in the absence of these rules [Shaalan and Raza 2008, 2009]. This presents a much more challenging task for the development of IE systems for Chinese, Korean, and Arabic.

3.1 Normalization of the Arabic Script

Another challenge facing researchers and developers of Arabic computational linguistics is the dilemma of normalization. The problem arises because of the inconsistency in the use of diacritic marks and certain letters in contemporary

Arabic texts. Some Arabic letters share the same shape and are only differentiated by adding certain marks such as a dot, a hamza or a madda placed above or below the letter. For example, the "alif" in Arabic (ا) may be three different letters depending on whether it has a hamza above as in (أ) or a hamza below as in (إ) or a madda above as in (آ). Recognizing these marks above or below a letter is essential to be able to distinguish between apparently similar letters.

But texts written in MSA often do not incorporate voweling as mentioned earlier nor do they adhere to the "proper" inclusion of marks above or beneath some Arabic letters. To manage this problem, the common practice in Arabic NLP systems is to normalize the input text [Larkey and Connell 2001]. For example, in order to handle the different variations in Arabic script, Larkey and Connell [2001] replace the initial alif with a hamza above or below with simply an alif, or bare alif. They also normalize the alif madda with a bare alif. Further, they normalize the final taa marbuuTa (ة or ـة) with a final haa (ه or ـه) and the alif maqsuura (ى) with the yaa (ي).

Following a similar approach, the Stanford Arabic Statistical Parser designed by The Stanford Natural Language Processing Group implements a similar normalization strategy for Arabic texts.

The SYSTRAN Arabic-to-English machine translation system [Farghaly and Senellart 2003] also incorporated normalization. But it soon became apparent that although normalization improves recognition by solving the variability in input, it increases the probability of ambiguity [Farghaly 2010]. For example, normalizing an initial alif with a hamza above or below it, removes an important distinction between (أن) ann and (إن) inn. The first translates into "that" and must be followed by a nominal sentence. The second could translate into "to" which indicates the English infinitive, but whose translation is meaningless if followed by a noun. In short, although normalization solves recognition problems, it creates the unintended effect of increased ambiguity.

3.2 Ambiguity and NLP Systems

The many levels of ambiguity pose a significant challenge to researchers developing NLP systems for Arabic [Attia 2008]. The reason is that ambiguity exists on many levels as evidenced by Maamouri and Bies [2010] who show 21 different analyses of the Arabic word (ثمن) tmn, produced by BAMA. At SYSTRAN, which has been developing machine translation systems for over 40 years, it was estimated that the average number of ambiguities for a token in most languages was 2.3, whereas in MSA it reaches 19.2. Although ambiguity is caused primarily by the absence of short vowels, at SYSTRAN researchers have found ambiguity in Arabic to be present at every level.

(1) Homographs: A word belonging to more than one part of speech such as قدم *qdm* which could be a verb of Form II meaning "to introduce" or a verb of Form 1 meaning "to arrive from" or a noun meaning "foot." Some

homograph ambiguity can be resolved by contextual rules. For example, an Arabic word that could be either a noun or a verb can be disambiguated by the following rule which says that such a word will be disambiguated to a noun when preceded by a preposition.

Contextual homograph resolution
e.g., [كتب] N | V -> N / Prep ___

(2) Internal word structure ambiguity: That is, when a complex Arabic word could be segmented in different ways. For example, "ولي" wly could be segmented into "و+ ل+ ي" corresponding to coordinate-prep-pronoun meaning "and for me," or may not be segmented at all meaning "a pious person favored by God".
Consider also the Arabic word "بعقوبة" when written in MSA without the short vowels, at least two analyses are valid. The first analysis is that it is a proper noun referring to a town in Iraq */Ba'quuba/* which was the center for the Iraqi resistance to the American occupation. The second analysis is that it is a preposition followed by a noun (MaSdar-infinitive) meaning "with the punishment of". An Arabic machine translation system or an Arabic named entity extraction system should select the correct analysis. The following rule is an example of how to give a preference to the proper noun analysis when it is preceded by lexical triggers such as "in" or "to". Resolving Arabic word segmentation ambiguity can be achieved by contextual rules like the following:

Arabic Word Segmentation
for example, [بعقوبة] PN | PrepP -> PN / في|الى ___

(3) Syntactic ambiguity: As in the case of a prepositional attachment as in "قابلت مدير البنك الجديد"*/qabaltu mudiir al-bank al-jadiid/* which could mean "I met with the new bank manager" or "I met with the manager of the new bank" depending on the internal analysis of the noun phrase.
(4) Semantic ambiguity: Sentences and phrases may be interpreted in different ways. For example, "يحب علي احمد أكثر من ابراهيم"*/yhb 'ly ahmd aktr mn abrahy*m/ "Ali likes Ahmed more than Ibrahim." Does this mean that Ali likes Ahmed more than Ali likes Ibrahim, or do Ali and Ibrahim like Ahmed, but Ali likes Ahmed more than Ibrahim likes Ahmed?
(5) Constituent boundary ambiguity: For example "مدير البنك الجديد" *mdyr albnk algydyd* could mean "the new manager of the bank" or "the manager of the new bank" depending on the boundary of the adjective phrase within this noun construct.
(6) Anaphoric ambiguity: As in قال علي أنه نجح*/qala Ali annahu najah*/Ali said that he succeeded. This sentence is ambiguous both in English and Arabic. Chomsky's Binding principles account for sentences like this.

The question here is does "he" refer to Ali or to someone else? Another interesting example is the following:

قابل الصحفي الوزير الذي انتقده
The journalist$_i$ met the minister$_j$ who$_{i/j}$ criticized him$_{i/j}$.
(Who criticized who?)

There are two lexical NPs: the journalist and the minister. Each has a different index. However, each of "who" and "him" can refer to any of the two lexical NPs. This is represented by the indices each has. Both have the indices $_i$ and $_j$ indicating the ambiguity they have in that each may refer to the journalist or to the minister. Only discourse information could disambiguate such sentences.

In addition to these levels of ambiguity the process of normalization plus features of Arabic such as the pro-drop structure, complex word structure, lack of capitalization, and minimal punctuation contribute to ambiguity, but it is the absence of short vowels that contributes most significantly to ambiguity.

With the absence of short vowels, two types of linguistic information are lost. The first is most of the case markers that define the grammatical function of Arabic nouns and adjectives. For example, a *Damma*, which is a high back rounded vowel at the end of a common noun or adjective marks the nominative case whereas a *fatHa*, which is a low front vowel in the final position of a common noun, marks the accusative case and a *kasra* which is a high front vowel marks the genitive case. The absence of case markers and thus the grammatical function of a word, creates multiple ambiguities due to the relatively free word order in Arabic and because Arabic is a pro-drop language.

The second type of information that is lost due to the nature of the Arabic script is the lexical and part of speech information. Thus, in the absence of internal voweling it is sometimes impossible to determine the part of speech (POS) without contextual clues. For example, without contextual clues a word like (من) *mn* could be a preposition meaning "from," a wh-phrase meaning "who" or a verb meaning "granted." An Arabic token such as (كتب) *ktb* without internal voweling could be a plural noun "books," an active past tense verb "wrote," a passive past tense verb "was written" or a causative past tense verb "he made him write."

While ambiguity is a challenge in any language, what makes Arabic so challenging is that all of these features are present in one.

3.3 Solutions

Tokenization in Arabic presents a problem because of the rich and complex morphology of Arabic. A token is usually defined as a sequence of one or more letters preceded and followed by space. This definition works well for non-agglutinative languages like English. Attia [2007] points out that tokenization of Arabic texts is a non-trivial task. For example a single Arabic word may contain up to four different tokens. Thus, tokenization requires knowledge of the constraints on concatenating affixes and clitics within Arabic words. A

distinction needs to be made between clitics which are syntactic units and thus have their own part of speech but do not stand alone, and affixes that mark grammatical inflections such as tense, number and person agreement. Attia's solution to the Arabic tokenization problem involves combining the morphological analysis and tokenization in one process.

The absence of capitalization in Arabic can be compensated for by looking deeper into the language to detect regularities that could help in information extraction. For example, many Arabic names have a middle token such as "بن" /bin/ meaning "son of." This could be a linguistic trigger that recognizes names such as "Osama bin Laden" even if they are not in the names dictionary. Recognizing patterns of Arabic names, dates, addresses, etc., can improve recall of Arabic entity recognition.

4. THE NONCONCATENATIVE ARABIC MORPHOLOGY

Arabic is characterized by its nonconcatenative morphology [McCarthy 1981] which presents a challenge to the structuralists theory of the morpheme. They defined the morpheme as a minimal linguistic unit that has a meaning. By minimal it is always meant that a morpheme cannot have a morpheme boundary within it. This definition works well for languages with concatenative morphology like English. McCarthy [1981] points out that the building blocks of Arabic words are the consonantal root which represents a semantic field such as "KTB" "writing" and a vocalism that represent a grammatical form. Arabic stems are described in terms of prosodic templates such as CVCVC. The Cs represent the root radicals and Vs represent the vocalism [Cavalli-Sforza et al. 2000]. Thus words such as "كتب"/katab/ is formed by an association of the radicals to the vocalism. While McCarthy proposes that Arabic words are analyzed at tiers (the root and the vocalism), Farghaly [1987] proposed a three tiered Arabic morphology by adding a third level for catenative affixation to Arabic stems.

Much of the work in Arabic linguistics has focused on the field of morphology and morphological analysis and early NLP systems have benefited from this work [Al-Sughaiyer and Al-Kharashi 2004; Soudi et al. 2007]. A pioneering work in Arabic computational morphology has been that of Hlal [1985] which was based on a lexicon of Arabic roots, prefixes, and suffixes and taking Arabic words as input, decomposing each word by identifying all prefixes, suffixes, and infixes and then recovering the root. Many Arabic morphological systems followed a similar approach while improving the performance [Beesley 2001; Rafea and Shaalan 1993]. Almost all computational treatment focused on recovering the roots from Arabic words.

Another approach was adopted in the development of the Buckwalter Arabic Morphological Analyzer (BAMA) [Buckwalter 2002], begun in the 1980s and made commercially available in 2000. The BAMA system is based on three tables: a table each for Arabic stems, Arabic prefixes, and Arabic suffixes. There are constraints placed on the prefixes and suffixes that can combine with a stem to form a legitimate Arabic word. The BAMA's dictionary of stems

is extensive and has a very high coverage. The BAMA system became available at UPenn LDC and soon became the morphological module of choice for most statistically-based ANLP applications in the U.S. It became the preferred module because it was possible for developers with no knowledge of the Arabic language to process unstructured Arabic texts due to its brilliant bidirectional transliteration schema from the Arabic script to the Latin script. The BAMA became the foundation for subsequent ANLP systems and expedited the development of an Arabic NLP system for machine translation and information retrieval during the last few years.

The BAMA approach was novel since, as mentioned before, most prior approaches to Arabic morphology were based on theoretical considerations and aimed to recover the Arabic consonantal root from Arabic words. In contrast, the Buckwalter system was pioneering because it implemented a stem-based approach to Arabic morphology. Buckwalter showed that it is simpler to consider the stem rather than the root as the basic unit of Arabic lexicon, but the users of the BAMA also have access to root information. The system incorporates three separate lexicons: prefixes, stems, and suffixes with tables to assess the compatibility of stems, prefixes, and suffixes. In addition to segmentation and stemming, the BAMA provides English glosses, full case endings, and noun case endings. It does not perform context-sensitive analysis but does give all possible analyses of the words in the input text. The MADA system [Habash and Owen 2005] goes one step further by using a disambiguation module that determines the correct POS tag in a specific context.

4.1 Systran's Stem-Based Morphological Generator

The traditional Arab grammarians' account of Arabic morphology in terms of roots and patterns is very precise and explicit and most work on Arabic morphology aims to identify and separate the prefixes and suffixes from the surface word and recover the root or stem that may have undergone morphemic changes. SYSTRAN's system [Farghaly and Senellart 2003] made a fundamental distinction between two kinds of affixes that can be attached to roots or stems. The first is an affix with only a grammatical meaning such as subject-verb agreement, tense or mood markers. While not part of the SYSTRAN dictionary, they are generated by its Arabic morphological generator that produces all the surface forms each stem could assume. Other researchers [Beesley 1996; Cavalli-Sforza et al. 2000; Habash 2004; Hosny et al. 2008; Shaalan et al. 2006] also developed morphological generators for Arabic.

Arabic is an agglutinative language and affixes that represent different parts of speech can be attached to a stem or root to form a token that has a syntactic structure. For example, the token (بالمدينه) *bi'lmmdynh* (in the city) has a stem (مدينه) *mdynh* (a city) and two prefixes; the first is (بـ) *b* (in) which is a preposition and ('ال') *al* (the) which is the definite article. In this way, external morphology describes the way the rule-governed affixes representing the different parts of speech are attached to Arabic stems.

4.2 Morphological Processing and the Dialects

Although most work in Arabic NLP focuses on developing NLP tools and systems for MSA, there is both a strong need and interest to develop systems to analyze Arabic dialects. But because there are limited parallel MSA-dialect resources and few annotated Arabic dialect texts, it has been difficult to develop these tools.

To address this problem, several novel approaches have been undertaken by researchers. One of the early morphological analyzers and generators for Arabic dialects, MAGEAD, uses an analyzer without a lexicon by exploiting regularities among dialects through sound changes at the radical level and thus, explicit analysis of roots and patterns [Habash and Owen 2005]. Needed to develop this system are representations of phonology and orthography to be able to both analyze and generate morphology for applications to NLP. In the MAGEAD system, an Arabic lexeme is defined as a root, a meaning index, and a morphological behavior class, and it is this definition that allows operation without a lexicon. The hypothesis is that morphological behavior class is variant independent enabling a lexeme-based representation to operate without a lexicon.

4.3 Arabic as an Agglutinative Language

The development of Arabic NLP systems is further challenged by two additional linguistic properties: word agglutination and the pro-drop feature.

Unlike English and most languages, Arabic has a complex word structure. An agglutinative language constructs complex words that often contain affixes and clitics representing various parts of speech. For example, a verb may embed within itself its subject and object as well as other clitics signifying tense, gender, person, number, and voice.

As mentioned earlier, in Arabic, verb stems are formed from a discontinuous consonantal root that represents a semantic field and a discontinuous melody (vocalism) that carries a grammatical meaning. The root and the melody are represented at different tiers and together they form a prosodic template [McCarthy 1981] such as CVCVC as in رَحَلَ "*raHal*" (travel) and عَلِم "*alim*" (learn).

The consonantal root in فهم/*fahim+a*/ "he understood" represents the field of "understanding" while the melody which consists of the two discontinuous short vowels "*a* – *i*" represents the past tense. The "*a*" suffix represents the third person, singular and masculine. The root and the melody together form the verb stem فهم /*fahim+a*/. But a different melody such as "*aa-i*" on the same root creates the present participle فاهم *faahim* meaning "understanding" as in أنا فاهم *ana faahim*, or "I am understanding."

An Arabic word, defined as a string of characters delimited by spaces, may be deconstructed into as many as four different parts of speech or morphemes. For example, the Arabic sentence ورأيتهم/*wra'aytuhum*/ "and I saw them" is written as one word and may be decomposed into the following four morphemes:

1. و /*wa*/ Conjunction "and"

2. رأى /*r'aa*/ Past tense Verb "saw"
3. ت /*tu*/ Subject Pronoun "I"
4. هُم /*hum*/ Object Pronoun "them"

The interaction between the phonological rules and morphological derivations of Arabic words makes the deconstruction of Arabic words even more difficult. For example, the *alif maqsuura* ى /aa/ in the stem only occurs in the final position. A phonological rule changes the alif maqsuura into a yaa' ي */ii/* when it is attached to a suffix or a clitic as is seen in the above example. Thus the complex internal structure of Arabic words makes tokenization, usually one of the early preprocessing tasks of a text, very challenging [Attia 2007].

The order in which affixes are attached to stems is rule governed which makes the decomposition of Arabic words possible. However, it is far from an easy process due to the high degree of ambiguity in Arabic [Attia 2008]. For example, the word وهم/*whm*/ has at least four valid analyses:

1. هم +و /*wa*+ *hum*/ CONJ SUBJPRON/OBJPRON "and they"
2. هم +و /*wa*+*hammun*/ CONJCOMMON NOUN "and worry"
3. وهم /*wahm*/ COMMON NOUN "illusion"
4. هم +و /*wa*+ *hamma*/ CONJ PVERB "and he initiated"

Furthermore, هم *hum* as a functional word is ambiguous in at least three ways. It can be a SUBJPRON (they); OBJPRON (them) or a POSSESSIVE-PRON (their). It could also be a lexical word NOUN (worry) or a VERB (initiated). However, it is possible to disambiguate a word like هم by grammatical rules. If it is attached to a verb, it is an object pronoun, whereas if it is attached to a noun it must be a pronoun. If attached to a complementizer such as أن *anna* or a conjunction like و *wa* or ف *fa*, it is a subject pronoun.

4.4 Arabic as a Pro-Drop Language

Another linguistic feature that complicates NLP systems for Arabic is due to the fact that Arabic is a pro-drop language. In Arabic, subject pronouns [Farghaly 1982] may be freely dropped subject to the Recoverability of Deletion Condition [Chomsky 1965]. The property of dropping the subject pronoun and allowing "subjectless sentences" is not limited to the Arabic language, as Italian, Spanish, and Korean are a few of the languages that permit subjectless sentences.

In Chomsky's theory of Principles and Parameters [Chomsky 1981, 1982], it is assumed that Universal Grammar (UG) consists of principles and parameters that have different values. A child's innate UG will fix the setting of the parameters and principles based on his early linguistic experience. The pro-drop is a parameter within UG; so an Arabic speaking child, based on his exposure to his native language will set it to the value "positive." As a result, s/he can comprehend and produce subjectless sentences whereas an English speaking child would set it to "negative" and not allow dropped subject pronouns.

Of course it is more challenging to process languages that incorporate the pro-drop feature. For example, [Dell'Orletta et al. 2005] point out that in

Italian as a pro-drop language, the sequence of Verb-Noun could be either Verb-Object with a dropped subject or it could be a sequence of Verb-post verbal subject, which results in a case of syntactic ambiguity; this observation may be applied to Arabic and possibly all pro-drop languages.

4.5 Solutions

There are several resources available for the morphological analysis of Arabic. The Xerox Arabic Morphological Analyzer Generator was developed in the 1990's by Ken Beesley at Xerox Research Center in Europe.[1] The implementation uses finite state technology. It recovers Arabic roots and performs both analysis and generation. Arabic words can be entered using the Arabic script. Another resource for Arabic morphology is Tim Buckwalter's morphological analyzer.[2] It differs from the Xerox morphological system in that it is stem-based and used a transliteration system that maps Arabic characters to Latin-based representation. A third resource is "Sarf" which is an engine that can generate Arabic verbs, nouns, gerunds, adjectives from their roots.[3]

5. SYNTACTIC STRUCTURE OF ARABIC

Arabic is a relatively free word order language. While the primary word order in Classical Arabic and Modern Standard Arabic is verb-subject-object (VSO), they also allow subject-verb-object (SVO) and object-verb-subject (OVS). It is common to use the SVO in newspapers headlines. Arabic dialects exhibit the SVO order. All varieties of Arabic allow subjectless sentences when the subject is recoverable. Like Russian, all varieties of Arabic allow equational sentences without explicit use of the equivalent of verb "to be" in English. So, "I a student" meaning "I am a student" is perfectly grammatical in Arabic.

Constituent questions in Arabic are formed by placing the corresponding wh-phrase at the beginning of the question as in "من قابلت أمس" "who did you meet yesterday." However, Egyptian Arabic does not front the question phrase but keeps it in place as in "امبارح إنت شفت مين" "who did you see yesterday."

Like Hebrew and many other languages, Arabic retains a resumptive pronoun referring to the lexical head in relative clauses. A typical Arabic relative clause will look like this literal translation "I met the woman who you talked to her last night."

Arabic has a very rich and complex agreement system. A noun and its modifiers have to agree in number, gender, case, and definiteness. In SVO structures, a verb must agree with its subject in gender, number, and person. However, in VSO sentences the verb is always in the singular even when its subject is dual or plural. The feature definiteness plays an important role in constituent formation. For example a noun construct usually begins with an indefinite noun followed by either a definite or indefinite noun as in "مدير البنك" "the manager of the bank." The first term of the Arabic construct "مدير"

[1]http://www.arabic-morphology.com
[2]http://www.qamus.org
[3]http://sourceforge.net/projects/sarf

"manager" is indefinite whereas the second noun "البنك" "the bank" is definite. In noun phrases consisting of a quantifier and a noun, the quantifier and the noun must disagree in gender. For example "ثلاث رجال" "three-masc men-masc" is ungrammatical because the word "three" is masculine, and so is "men". The phrase "ثلاثة رجال" "three-fem men-masc" is a grammatical Arabic phrase because the quantifier and the noun disagree in gender.

The only complete grammar available to ANLP researchers is that of Classical Arabic which was developed in from the 8th to the 10th centuries. Classical Arabic grammar was written to account for a closed corpus; that of the Quran and the Hadith which represents the Prophet's sayings. Traditional Arabic grammarians clearly defined their goal which was to protect the purity of the Arabic language which was threatened when large number of people converted to Islam after the Islamic conquests and when the features of the native languages of these new converts to Islam interfered with their ability to learn Arabic. So as a result, they made errors while they were reciting the Quran. In the absence of a Classical Arabic reference grammar and lexicons it would not be clear which usage was correct. Arab grammarians undertook the task of standardizing the correct usage of Arabic and in particular the correct pronunciation "التجويد" (recitation) of the Quran as well as its correct interpretation. Correspondingly, the Quran is always written with full representation of the short vowels and with explicit case marking. Thus Arab grammarians had to account for case markings in their grammar texts.

Most contemporary texts such as newspapers, academic papers, and modern books which represent much of the input to ANLP programs are written in MSA and as such they do not show short vowels nor do they have explicit representation of most case markings. In processing such texts, syntactic constituency is crucial for the correct analysis of Arabic and for identifying constituent boundaries. Traditional Arabic grammar does not provide ANLP developers of MSA texts with the type of grammar they need. For example, traditional Arabic grammar would analyze the following three phrases as *idaafa* or a noun construct:

1. مدير البنك "manager of the bank"
2. حاد الذكاء "with sharp intelligence"
3. فوق المنزل "on top of the house"

Arabic traditional grammarians noted that in all the above sentences, the second term is governed by the first and is assigned a genitive case. So they grouped them together as *idaafa*. In an Arabic machine translation system for example, we must be able to distinguish these three phrases by one of the following analyses: (1) is a noun phrase, (2) is an adjectival phrase, and (3) is a prepositional phrases. Such an analysis is more relevant and useful in ANLP applications than the traditional analysis which treats them as having one structure based on case ending which is not applicable to texts in MSA.

ANLP would benefit greatly from surface-based grammars of MSA the same way NLP systems benefited from lexical-based grammar formalisms such

as Lexical Functional Grammar [Bresnan 2000] and Head Phrase Structure Grammar [Sag and Pollard 1994].

5.1 Solutions

Grammatical descriptions of Modern Standard Arabic has started to appear [Badawi et al. 2004; Ryding 2005]. Such descriptions are very useful in the processing of contemporary Arabic although they were not written from a computational viewpoint. The annotated Arabic corpora that have been developed at the LDC is extremely valuable for ANLP applications. Modern Standard Arabic texts have been analyzed with insights from traditional Arabic grammar as well as from modern linguistic theories. The LDC has also compiled corpora for some Arabic dialects and Arabic-English parallel corpora that are very useful for machine translation. Recently, as we mentioned earlier, the LDC released an annotated entity extraction corpus for Arabic. Another important resource is the Prague Arabic Dependency Treebank which implements a functional approach to the analysis of Modern Standard Arabic. There are also resources for Arabic dialects such as the Arabic Treebank at Columbia University and the Arabic dialects corpora at the LDC.

6. CONCLUSION

There are Arabic language features that are inherently challenging for ANLP researchers and developers. These features include the nonconcatenative nature of Arabic morphology, the absence of the orthographic representation of Arabic short vowels from contemporary Arabic texts, the need for an explicit grammar of MSA that defines linguistic constituency in the absence of case marking. The new grammar also must describe important aspects such as anaphoric relations, the subjectless sentences, and discourse analysis. In spite of these challenges, significant work has been done in ANLP in applications such as entity extraction [Shaalan and Raza 2009], machine translation [Farghaly and Senellart 2003; Shaalan et al. 2004; Fraser and Wong 2009; Sawaf 2009; Abdel Monem et al. 2009], and sentiment analysis [Almas and Ahmed 2007].

REFERENCES

ABDEL MONEM, A., SHAALAN, K., RAFEA, A., AND BARAKA, H. 2009. *Generating Arabic Text in Multilingual Speech-to-Speech Machine Translation Framework, Machine Translation.* Springer.

ALMAS, Y. AND AHMED, K. 2007. A note on extracting "sentiments" in financial news in English, Arabic, and Urdu. In *Proceedings of the 2nd Workshop on Computational Approaches to Arabic Script-based Languages (CAASL'07)*. 1–12.

AL-SUGHAIYER, I. AND AL-KHARASHI, I. 2004. Arabic morphological analysis techniques: A comprehensive survey. *J. Amer. Soc. Inform. Sci. Technol. 55*, 3, 189–213.

ATTIA, M. 1999. A large scale computational processor of Arabic morphology and applications. Master's Dissertation, Computer Engineering, Cairo University, Egypt.

ATTIA, M. 2007. Arabic tokenization system. In *Proceedings of the Association of Computational Linguistics (ACL'07)*.

ATTIA, M. 2008. Handling Arabic morphological and syntactic ambiguities within the LFG framework with a view to machine translation. PhD Dissertation, University of Manchester.

BADAWI, E., CARTER, M. G., AND GULLY, A. 2004. *Modern Written Arabic: A Comprehensive Grammar*. Routledge, London.

BAKALLA, M. H. 2002. *Arabic Language Through Its Language and Literature*. Kegan Paul, London.

BEESLEY, K. 1996. Arabic finite-state morphological analysis and generation. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*. 89–94.

BEESLEY, K. 2001. Finite-state morphological analysis of Arabic at Xerox Research: Status and plans in 2001. In *Proceedings of the Workshop on Arabic Natural Language Processing at the 39th Annual Meeting of the Association for Computational Linguistics (ACL'01)*. 1–8.

BENJAJIBA, Y., DIAB, M., AND RASSO P. 2008. Arabic named entity recognition using optimized feature sets. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP'08)*. 284–293.

BRESNAN, J. 2000. *Lexical Functional Syntax*. Blackwell Publishers Inc., Malden, MA.

BUCKWALTER, T. 2002. Arabic transliteration. http://www.qamus.org/aramorph/.

BUCKWALTER, T. 2004. Issues in Arabic orthography and morphology analysis. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages (CAASL'04)*. 31–34.

CHOMSKY, N. 1965. *Aspects of the theory of syntax*. MIT Press, Cambridge, MA.

CHOMSKY, N. 1981. *Lectures on Government and Binding*. Foris Publications, Dordrecht.

CHOMSKY, N. 1982. Some concepts and consequences of the theory of government and binding. MIT Press, Cambridge, MA.

CHOUKRI, K. 2009. MEDAR: Mediterranean Arabic language and speech technology: Inventory of the HLT products, players, projects and language resources. In *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR'09)*.

CAVALLI-SFORZA V., SOUDI, A., AND MITAMURA, T. 2000. Arabic morphology generation using a concatenative strategy. In *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP'00)*. 86–93.

CIA. 2008. *CIA Word Fact Book*. Central Intelligence Agency, Washington, D.C.

DELL'ORLETTA, F., LENCI, A., MONTEMAGNI, S., AND PIRRELLI, V. 2005. Climbing the path to grammar: A maximum entropy model of subject/object learning. In *Proceedings of the 2nd Workshop on Psycho-Computational Models of Human Language Acquisition, Association for Computational Linguistics (ACL'05)*. 72–81.

DIAB, M. AND HABASH, N. 2007. Arabic dialect tutorial. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'07)*.

FARGHALY, A. 1982. Subject pronoun deletion rule. In *Proceedings of the 2nd English Language Symposium on Discourse Analysis (LSDA'82)*. 110–117.

FARGHALY, A. 1987. Three level morphology for Arabic. In *Proceedings of the Arabic Morphology Workshop (AMW'87)*.

FARGHALY, A. 1999. Arabic diglossia and Arabic identity in the information age. *Al-Fikr Al-Arabi, March-April*.

FARGHALY, A. 2005. A case for inter-Arabic Grammar. In Eligbali, A., Ed., *Investigating Arabic: Current Parameters in Analysis and Learning*. Brill, Boston.

FARGHALY, A. 2007. Information retrieval and the Arabic noun construct. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages (CAASl'07)*.

FARGHALY, A. 2008. Arabic NLP: Overview, the state of the art: Challenges and opportunities. In *Proceedings of the International Arab Conference on Information Technology (ACIT'08)*.

FARGHALY, A. 2010. *Introduction in Arabic computational linguistics*. CSLI Publications, Stanford, CA.

FARGHALY, A. AND SENELLART, J. 2003. Intuitive coding of the Arabic lexicon. In *Proceedings of the MT Summit IX, the Association for Machine Translation in the Americas (AMTA'03)*.

FERGUSON, C. 1959. Diglossia. *WORD, 15* 3, 325–340.

FERGUSON, C. 1996. Epilogue: Diglossia revisited. In *Contemporary Arabic Linguistics in Honor of El-Said Badawi*. The American University in Cairo.

FRASER, A. AND WONG, W. 2009. The language weaver Arabic to English statistical machine translation system. To appear in Farghaly, A., Ed., *Arabic Computational Linguistics*. CSLI Publications. To appear.

GRISHMAN R. AND SUNDHEIM, B. 1996. Message understanding conference (MUC–6): A brief history. In *Proceedings of the 16th International Conference on Computational Linguistics (ICCL'96)*. 466–471.

HABASH, N. 2004. Large-scale lexeme based Arabic morphological generation. In *Proceedings of Traitement Automatique du Langage Naturel (TALN'04)*.

HABASH, N., OWEN, R., AND GEORGE, K. 2005. Morphological analysis and generation for Arabic dialects. In *Proceedings of the Association for Computational Linguistics (ACL'05)*.

HABASH, N. AND OWEN, R. 2005. Arabic tokenization, part of speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of the Association for Computational Linguistics (ACL'05)*.

HLAL, Y. 1985. Morphological analysis of Arabic speech. In *Proceedings of the 2nd Conference on Computer Processing of the Arabic Language (CPAL'85)*.

HOSNY, A., SHAALAN, K., AND FAHMY, A. 2008. Automatic morphological rule induction for Arabic. In *Proceedings of the Workshop on Human Language Translation and Natural Language Processing within the Arabic World (LREC'08)*. 97–101.

LARKEY, L. AND CONNELL, M. E. 2001. Arabic information retrieval at UMASS in TREC-10. In *Proceedings of the 10th Text Retrieval Conference (TREC'01)*.

MAAMOURI, M. AND BIES, A. 2010. The Penn Arabic Treebank. In Farghaly, A., Ed., *Arabic Computational Linguistics*. CSLI Publications, Stanford, CA.

MCCARTHY, J. 1981. A prosodic theory of nonconcatenative morphology. *Linguistic Inquiry. 12*, 373–418.

RAFEA, A. AND SHAALAN K. 1993. Lexical analysis of inflected Arabic words using exhaustive search of an augmented transition network. *Softw. Prac. Exper. 23*, 6, 567–588.

RYDING K. 2005. *Reference grammar of modern standard Arabic*. Cambridge University Press, Cambridge, UK.

SAG I. AND POLLARD, C. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press, Chicago, IL.

SAWAF, H. 2009. The AppTek hybrid machine translation system. In Farghaly, Ali, Ed., *Arabic Computational Linguistics*. CSLI Publications. To appear.

SHAALAN K. 2005a. An intelligent computer-assisted language learning system for Arabic learners. *J. Int. Comput. Assist. Lang. Learn. 18*, 1/2, 81–108.

SHAALAN, K. 2005b. *Arabic GramCheck: A Grammar Checker for Arabic, Software Practice and Experience*. John Wiley & Sons, Ltd., 643–665.

SHAALAN K., RAFEA, A., ABDEL MONEM, A., AND BARAKA, H. 2004. Machine translation of English noun phrases into Arabic. *Int. J. Comput. Proc. Oriental Lang. 17*, 2, 121–134.

SHAALAN, K., ABDEL MONEM, A., AND RAFEA, A. 2006. Arabic morphological generation from Interlingua: A rule-based approach. In *Intelligent Information Processing III*, Z. Shi, K. Shimohara, and D. Feng, Eds. Springer, 441–451.

SHAALAN, K., ABO BAKR, H., AND ZIEDAN, I. 2007. Transferring Egyptian colloquial into modern standard Arabic. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'07)*. 525–529.

SHAALAN, K. AND RAZA H. 2008. Arabic named entity recognition from diverse text types. In *Proceedings of the 6th International Conference on Natural Language Processing (GoTAL'08)*. B. Nordström, and A. Ranta, Eds.

SHAALAN, K. AND RAZA, H. 2009. NERA: Named entity recognition for Arabic. *J. Amer. Soc. Inform. Sci. Technol. 60*, 7, 1–12.

SOUDI, A., BOSCH, A., AND GÜNTER, N., eds. 2007. *Arabic Computational Morphology: Knowledge-Based and Empirical Methods (Text, Speech, and Language Technology)*, Springer.

VERSTEEGH, K. 1997. *The Arabic Language*. Columbia University Press, New York.

Received June 2009; revised September 2009; accepted October 2009