An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks

Mohammed A. Attia
School of Informatics,
The University of Manchester
mohammed.attia@postgrad.manchester.ac.uk

Morphological ambiguity is a major concern for syntactic parsers, POS taggers and other NLP tools. For example, the greater the number of morphological analyses given for a lexical entry, the longer a parser takes in analyzing a sentence, and the greater the number of parses it produces. Xerox Arabic Finite State Morphology and Buckwalter Arabic Morphological Analyzer are two of the best known, well documented, morphological analyzers for Modern Standard Arabic (MSA). Yet there are significant problems with both systems in design as well as coverage that increase the ambiguity rate. This paper shows how an ambiguity-controlled morphological analyzer for Arabic is built in a rule-based system that takes the stem as the base form using finite state technology. The paper also points out sources of legal and illegal ambiguities in MSA, and how ambiguity in the new system is reduced without compromising precision. At the end, an evaluation of Xerox, Buckwalter, and our system is conducted, and the performance is compared and analyzed.

Keywords: Arabic morphology, morphological ambiguity, Arabic morphological analyzer, Arabic finite state transducer

1. INTRODUCTION

Morphological ambiguity in Arabic is a notorious problem that has not been sufficiently addressed (Kiraz 1998). This ambiguity represents hurdles in the way of POS taggers (Freeman 2001) syntactic parsers, and machine translation. Overcoming ambiguity is the major challenge for NLP in Arabic (Kamir et al 2002). Xerox Arabic Finite State Morphology and Buckwalter Arabic Morphological Analyzer are two of the best known, well documented, morphological analyzers for Modern Standard Arabic (MSA). Yet there are significant problems with both systems in design as well as coverage that increase the ambiguity rate. Xerox morphology is root based, yet it has uncurbed generative power that makes it produce forms that are unknown in the language. Buckwalter's morphology is a stem-based database that lacks the generality and power of a rule-based system. Both systems include a large number classical entries that are not part of MSA and do not occur in contemporary Arabic texts, the matter that leads to an increased number of ambiguities.

Ambiguity is also increased by the inappropriate application of spelling relaxation rules and by overlooking rules that combine words with clitics and affixes (grammar-lexis specifications). Another source of confusion is whether to allow Arabic verbs to inflect for the imperative mood and the passive voice or not. Xerox adopted the overgeneralization that all verbs inflect for the imperative and the passive, leading it to overgenerate. Buckwalter's morphology, on the other hand allowed only some verbs to have these inflections. Yet, because it did not follow a unique criteria or a systematic

approach, the analysis is either underspecified or superfluous. This paper shows how an ambiguity-controlled morphological analyzer for Arabic is built in a rule-based system that takes the stem as the base form using finite state technology. The paper also points out sources of legal and illegal ambiguities in MSA, and how ambiguity in our system is reduced without compromising precision. The system is based on a contemporary corpus of news articles to ensure that the scope of the lexicon is restricted to MSA. Our morphology emphasizes the idea that inflecting all verbs in the passive and the imperative is semantically and pragmatically incorrect. Moreover, a set of broadly-defined criteria is devised to select which verbs can have a passive voice and which verbs can occur in the imperative.

In this introduction we discuss sources of acceptable ambiguity in Arabic, and propose the ambiguity pyramid hypothesis in which we claim that ambiguity decreases with the build-up of words by adding affixes and clitics. Then we explain the main strategies followed in developing Arabic morphologies and analyse two of the well-known systems.

In Section 2 we explain our system design and how we managed to make it less ambiguous. In the last section, an evaluation of Xerox, Buckwalter, and our system is conducted, and the performance is compared and analyzed.

1.1 Sources of legal morphological ambiguity in Arabic

Many words in Arabic are homographic: they have the same orthographic form, though the pronunciation is different. There are many recurrent factors that contributed to this problem. Among these factors are:

1. Orthographic alternation operations (such as deletion and assimilation) frequently produce inflected forms that can belong to two or more different lemmas. Example (1) is an extreme case of a surface form that can be interpreted as belonging to five different stems.

   (1 يعد y'd

يعِد (أعاد)	يعُد (عاد)	يعِد (وعد)	يَعُدّ (عد)	يُعِدّ (أعد)
yu'id	ya'ud (‘aada)	ya'id (wa'ada)	ya'udd (‘adda)	yu'idd (a'adda)
(‘a'aada)	[return]	[promise]	[count]	[prepare]
[bring back]				

2. Some lemmas are different only in that one of them has a doubled sound which is not explicit in writing. Arabic Form I and Form II are different only in that Form II has the middle sound doubled.

   (2 علم ‘alm

علِم	علّم
‘alima	‘allama
(know)	(teach)

3. Many inflectional operation underlie a slight change in pronunciation without any explicit orthographical effect due to lack of short vowels (diacritics). An example is the recurring ambiguity of active vs. passive vs. imperative forms

   (3 أرسل ‘rsl

أرسَل	أرسِلُ	أرسِل
‘arsala	‘ursila	‘arsil
(sent)	(was sent)	(send [imperative])

4. Some prefixes and suffixes can be homographic with each other. The prefix *t-* can indicate 3rd person feminine or 2nd person masculine.

(4 تكتب ta-ktub تكتب ta-ktub

(you.m write) (she writes)

Another recurring ambiguity is the person suffix *–t* which is shared by four features.

(5 كتبت ktbt

كتبتُ	كتبتَ	كتبتِ	كتبتْ
katabtu	katabta	katabti	katabat
(I wrote)	(you.m wrote)	(you.f wrote)	(she wrote)

Similarly, the dual is always confused with the plural in the accusative case.

(6 أمريكيين

أمريكيَين	أمريكيين
‘amriikiyain	‘amriikiyiin
(Americans [dual])	(Americans [plural])

5. Prefixes and suffixes can accidentally produce a form that is homographic with another full form word. This is termed “coincidental identity” (Kamir et al 2002).

(7 أسد asd

أسُدّ) أ+سد)	أسد
‘asuddu	‘asad
(I block)	(lion)

6. Similarly, clitics can accidentally produce a form that is homographic with another full word.

(8 علمي

علمي	علمي (علم + ي)
‘ilmi	‘ilm-i
(scientific)	(my knowledge)

7. There are also the usual homographs of uninflected words with/without the same pronunciation, which have different meanings and usually different POS’s.

(9 ذهب

ذهب	ذهب
dhahab	dhahaba
(gold)	(go)

1.2 The Ambiguity Pyramid Hypothesis

The ambiguity pyramid hypothesis assumes that the rich and complex system of Arabic inflection and concatenation helps to reduce ambiguity rather than increase it. Unmarked stems are usually ambiguous but when they are inflected and/or when clitics are added, ambiguity is reduced, as shown in (10).

(10

stem:	كتب	ktb	books / wrote / was-written
inflected:	يكتب	ya-ktb	writes / is-written
clticized:	يكتبه	ya-ktb-hu	[he]-writes-it

Words from a few randomly selected sentences were morphologically analyzed at different levels. First they were analyzed as whole words, then they were analyzed after separating words from clitics, and at last they were analyzed after separating clitics and stripping off all inflectional prefixes and suffixes, that is using the base stem. The highest rate of ambiguity appeared in the stem level. The rate decreased with inflection, and decreased even further with the addition of clitics. Figure 1 illustrates that ambiguity rates decrease, on average, with the increase in word build-up.

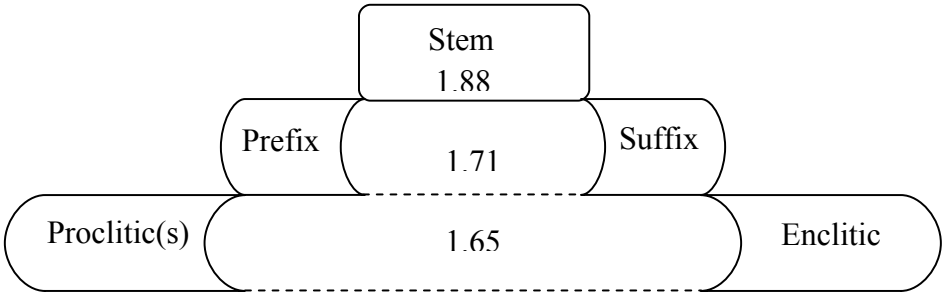

FIGURE 1: The ambiguity pyramid hypothesis

This is still a hypothesis that still needs to be verified. Further testing with some other sentences contradicted these assumptions, and large scale testing on a large number of words is not possible. For example, a list of 30,000 full form words was reduced to 15,000 unique words after stripping off clitics. Comparing the ambiguity rates for two different numbers is not indicative, as the same transducer will usually give different ambiguity rates when it is fed different ranges. So in order to verify this hypothesis, testing needs to be done on several hundred sentences, rather than words. This may not even be very meaningful, as a sentence containing 30 full form words will break down into about 50 tokens and break down further into 70 base forms. So comparing the rates at these different numbers cannot constitute strong evidence. It is also found that words with the highest scores are inflected forms.

1.3 Development Strategies of Arabic Morphology

Arabic is known for its morphological richness and complexity (McCarthy 1985; Azmi 1988; Beesley 1998b; Ratcliffe 1998; Ibrahim 2002). Arabic morphology has always been a challenge in computational processing and a hard testing ground for morphological analysis technologies. There are mainly two strategies for the development of Arabic morphologies depending on the level of analysis:

1. Stem-based morphologies: analyzing Arabic at the stem level and using regular concatenation. A stem is the least marked form of a word, that is the uninflected word without suffixes, prefixes, proclitics or enclitics. In Arabic, this is usually the perfective, 3rd person, singular verb, and in the case of nouns and adjectives they are in the singular indefinite form.
2. Root-based morphologies: analyzing Arabic words as composed of roots and patterns in addition to concatenations. A root is a sequence of three (rarely two or four) consonants which are called *radicals*, and the pattern is a template of vowels, or a combination of consonants and vowels, with slots into which the radicals of the root are inserted as shown in Figure 2. This process of insertion is usually called *interdigitation* (Beesley 2001).

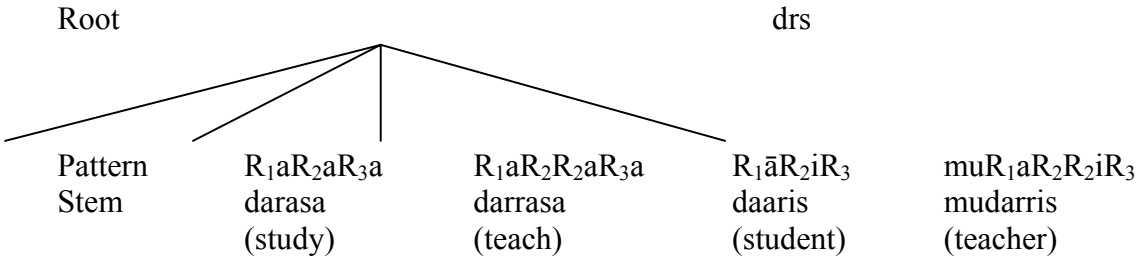

FIGURE 2: Root and Pattern Interdigitation

There has been an intense contest between proponents and opponents of using the root as the base form. Beesley (2001) defended the "linguistic reality of Semitic roots" and cited, as a practical motivation, that traditional dictionaries are indexed by roots. It has even been maintained that "the use of Arabic roots as indexing terms substantially improves the [information] retrieval effectiveness over the use of stems" (Darwish 2002).

However, several researchers criticized this approach. Kamir et al (2002) assumed that the stem is the lemma, or the basic grammatical unit, in Arabic, and argued that the root is an abstract "super-lemma" that groups all the words that share a semantic field. They also maintained that the role of a root appears in word formation, or derivational morphology, while the stem is the actual manifestation of the root, and it is the stem that takes part in inflectional morphology. Dichy and Fargaly (2003) dedicated a lengthy paper to the subject and maintained that a root-and-pattern system included "huge numbers of rule-generated word-forms, which do not actually appear in the language" and that morpho-syntactic and semantic information need to be added to lexical entries at the stem level.

In our implementation we adopted the idea that a root is an abstract form that does not belong to a specific POS, but it plays a crucial part in stem formation. So using the stem as base form is far less complex in developing and maintaining, less ambiguous, and more suitable for syntactic parsers that aim at translation. The effectiveness of a root-and-pattern system in information retrieval is even doubted as some verbs like أمِنَ amina (to be safe), أمُنَ amuna (to be honest) and آمَن aamana (to believe) have the same root but each has a different pattern and different semantic field (examples adapted from (Dichy and Fargaly 2003)). So أمان amaan (safety), أمانة amaanah (honesty) and إيمان ʻiimaan (believe) should not be made related in IR.

1.4 Existing Arabic Morphological Systems

There are many morphological analyzers for Arabic, some of them are available for research and evaluation while the rest are proprietary commercial applications. Among those known in the literature are Xerox Arabic Morphological Analysis and Generation (Beesley 1998a; Beesley 2001), Buckwalter Arabic Morphological Analyzer (Tim Buckwalter. 2002), Diinar (Dichy and Hassoun 1998), Sakhr (Chalabi 2004), and Morfix (Kamir et al 2002). The first two are the best known and most quoted in literature, and they are well documented and available for evaluation.

1.4.1 Buckwalter Arabic Morphological Analyzer

Buckwalter Morphology is well-known in the literature and has even been considered as the "most respected lexical resource of its kind" (Hajič et al 2005). It contains 38,600

lemmas, and is used in LDC Arabic POS-tagger, Penn Arabic Treebank, and the Prague Arabic Dependency Treebank. It is designed as a main database of word forms interacting with other concatenation databases. Every word form is entered separately. It takes the stem as the base form, and information on the root is also provided. Buckwalter's morphology reconstructs vowel marks and provides English glossary, and it is less ambiguous than Xerox's. The disadvantages, however, are:

1. Not rule-based. All word forms are entered manually. After each entry, all forms that belong to that specific entry at different inflectional levels are listed. So it does not capture generalities, and it increases the cost of maintenance.
2. The system is not suited for generation.
3. Underspecification in the clitic question morpheme which can be prefixed to verbs and nouns. This was perhaps intended to reduce ambiguity, but, still, it limits coverage.

   (11 أأقول 'a'aqul (do I say) – not found
   أمحمد 'Mohammed (is Mohammed) – not found
4. Underspecification in imperative forms: Out of 9198 verbs, only 22 verbs (%0.002) have imperative forms. This is far less than the %37 allowed in our morphology. This restricts Buckwalter's from dealing with instruction manuals, for example. No imperative senses are associated with verbs in (12).

   (12 حاول haawil try
   انتظر intazhir wait
   اضرب idrib hit
5. Underspecification in the passive morphology. Out of 9198 verbs, only 1404 verbs (%15) are allowed to have a passive form. In our system, %41 of verbs can have a passive form. Buckwalter's passive forms are also restricted by tense. Only 110 of them have a passive form in the past (perfective) tense. There are passive forms for verbs with low probability such as in (13).

   (13 يمات yumat be made to die
   يعاش yu'ash be lived

   While other verbs with high probability are not allowed in the passive, such as those in (14).

   (14 قابل qabal meet
   استعمل ista'mala use
6. It accounts for the classical affirmative clitics ل "la" (indeed) which is prefixed to nouns. This makes it ambiguous with the preposition which has the same form.

   (15 لأحزاب la-ahzab indeed + parties
7. Some proper names are associated with senses that are no longer used in the language

   (16 حسام Husam / sword
   حنيفة Hanifah / orthodox
8. No handling of multiword expressions (MWEs). MWEs have high frequency in texts and when they are identified and analyzed correctly they add a sense of certitude to the analysis and reduce ambiguity. However, when MWEs are analyzed compositionally, they lose their meaning and add to the ambiguity problem, as component parts may be individually ambiguous.

   (17 أبي أسعد
   abi as'ad
   my father / proud happier / make happy
   (Abu As'ad [proper name])

9. Inclusion of classical entries. Every entry added to the lexicon of a morphological analyzer is very costly in terms of ambiguity, so terms should be extracted from contemporary data, rather than from traditional dictionaries, if they are meant to handle modern texts. There are many hints that Buckwalter and Xerox took Hans Wehr's Arabic English Dictionary of Modern Written Arabic (Wehr 1979) as the backbone reference. However, in the very introduction, Hans Wehr stated that the dictionary "not only lists classical words and phrases of elegant rhetorical style side by side with new coinages that conform to the demands of the purists, it also contains neologisms, loan translations, foreign loans, and colloquialisms which may not be to the linguistic taste of many educated Arabs" (Wehr 1979). Buckwalter includes some roots that are totally no longer in use, such as those in (18).

(18 قف qaffa (to be dry)
أبد abada (be untamed)
أب abba (desire)

Some forms are fossilized in contemporary usage, as their usage is limited to expressions in a certain syntactic and morphological context.

(19 لا يأبه la ya'bah (not care) [he does not care]
Root: أبه abaha (be interested)

All the above forms are homographic in some way with other forms that are in contemporary usage. Still, it can be proven statistically that Buckwalter included classical terms by showing the Google score for some selected entries found Buckwalter's morphology in Table 1.

#	Word	Transliteration	Meaning	Google
1	قلعط	qal'at	sully	8
2	قلفط	qalfat	caulk	9
3	استكد	istakadda	wear	4
4	غملج	ghamlaj	fickle	7
5	ائتكال	i'tikal	erosion	7

TABLE 1: Google score for entries from Buckwalter morphology

10. Improper spelling relaxation rules. Buckwalter justified the inclusion of these relaxation rules by the fact that they are common in the data analyzed (Buckwalter 2004). We reckon however, that this is not a solid justification because, firstly, we should take into account that Arabic electronic texts are relatively recent, and that not so many authors are well trained in using proofing tools. Secondly, misspelled words should be handled as special cases, or apply rules when the form fails to receive an analysis. Applying the rules globally in this case led to a massive increase in the ambiguity level for correctly spelled words, as shown in (20). Thirdly, misspelling is even common in English. The Google score for the misspelled word "arround", for example, is 2,530,000 and for "vedio" is 2,150,000, and this will not be deemed as a plausible ground for including these misspelled words in an English morphological analyzer.

(20 فاشل fashil (failure)
-> فأشل fa'ashul (then I paralyze)
واقف waqif (standing)
-> وأقف wa'aqif (and I stand)

11. Incomprehensive treatment of the rules that govern the combination of words with clitics, or grammar-lexis specification (Dichy 2001; Dichy and Fargaly 2003; Abbès

et al 2004). As clitics are syntactic units, syntactic rules should apply when they combine with words. For example, when a preposition precedes a noun, the noun must be in the genitive case. Similarly, while it is acceptable for the noun to be followed by possessive pronouns, this is not acceptable for adjectives, which is not observed by Buckwalter, as shown in (21).

(21 معادي mu'adi (hostile/anti- + my)
معدي mu'di (contagious/infectious + my)

Another wrong analysis is shown in (22) where a verbal noun derived from an intransitive verb is attached to an accusative pronoun clitic.

(22 مصري mussirr-i (determined/insistent + my)

Similarly, names of places are usually followed by relative suffixes, not possessive pronouns, the rule which is ignored in (23).

(23 عراقي 'raqi (Iraq + my), should be Iraqi
إيراني irani (Iran + my), should be Iranian

1.4.2 Xerox Arabic Morphological Analysis and Generation

Xerox Morphology is "based on solid and innovative finite-state technology" (Dichy and Fargaly 2003). It adopts the root-and-pattern approach. It includes 4,930 roots and 400 patterns, effectively generating 90,000 stems. The advantages are that it is rule based with large coverage. It also reconstructs vowel marks and provides an English glossary for each word. The system inherited many disadvantages from Buckwalter's morphology such as the lack of specifications for MWEs, and improper spelling relaxation rules. It even includes more classical entries, and lacks more grammar-lexis specifications. Example (24) shows an extreme case which violates the syntactic rule that a pronoun must be free within its binding domain, or "co-reference of the subject and of the object" (Dichy 2001).

(24 نضربنا nadribuna (we hit us).

Additional disadvantages of Xerox morphology are:

1. Overgeneration in word derivation. The distribution of patterns for roots is not even, and although each root was hand-coded in the system to select from among the 400 patterns, the task is understandably tedious and prone to mistakes.

word	transliteration	root	meaning
قال	qaal	qwl	say (verb)
		qlw	fry (active participle)
		qll	decrease (active participle)

TABLE 2: Overgeneration of illegal stems

The first analysis is valid, while the other two are illegal derivations that have no place in the language, and not mentioned in classical dictionaries.

2. Underspecification in POS classification, which makes it unsuited for serving a syntactic parser. Words are only classified into:
   - Verbs
   - Nouns, which include adjectives and adverbs.
   - Participles
   - Function words, which include prepositions, conjunctions, subordinating conjunctions, articles, negative particles, and all other particles.

– Increased rate of ambiguity. Due to the above-mentioned factors, the system suffers from a very high level of ambiguity, as it provides so many analyses (many of them spurious) for most words, as shown in (25).

(25 مصري misri Egyptian
Xerox (22 solutions), Buckwalter (10 solutions), Attia (2 solutions)

2. SYSTEM DESCRIPTION

Our system is built using finite state technology (Attia 2005), and it is suitable for both analysis and generation. It is based on contemporary data (a corpus of news articles of 4.5 million words), and takes the stem as the base form. It contains 9741 lemmas and 2826 multiword expressions. The core system provides a full and efficient coverage of MSA for its specific domain (news articles). The system is available for research and evaluation at www.attiaspace.com, along with a set of relevant finite state tools: a tokenizer, a white space normalizer and a morphological guesser. The system is rule based; there is only one entry for each stem, and all inflection operations and orthographical changes are handled through xfst alternation rules. This helps in separating the task of the developer and the linguist. As adding new terms to the lexicon in a morphological transducer is a never ending process, the lexicographer's job is made clearer and easier.

A point of strength in the system that gives it an advantage over other morphological analyzers is the coverage of multiword expressions (Attia 2006). The system can efficiently handle compound names of people, places, and organizations, as shown in (26), in addition to more complex expressions which can undergo inflections and lexical variations.

(26 أبو عمار
abu 'ammaar (lit. father of 'Ammar)
Abu 'Ammar
بيت لحم
bait lehem (lit. house of meat)
Bethlehem
مجلس الأمن
majlis al-amn
Security Council

A disadvantage of the system, however, is its limited coverage. Between Buckwalter's 38,600 and Attia's 12,500 lemmas, a good coverage, general-domain morphology is expected to be around 20,000 lemmas including MWEs. Our system does not handle diacritized texts. The decision to ignore diacritics was taken after examining a corpus of 4.5 million Arabic words, where only 54 words were found to carry meaningful diacritic marks, which is statistically insignificant. Other disadvantages are that it does not reconstruct diacritics, or provide English glossaries. These limitations do not affect the functionality of the morphology especially when the target is to feed a syntactic parser, yet it has been customary in Arabic morphology to provide diacritics and glossaries for illustration and pedagogical purposes.

2.1 Finite State Technology

Finite state technology has successfully been used in developing morphologies for many languages, including Semitic languages. There are a number of advantages of this

technology that makes it specially attractive in dealing with human language morphologies, among these advantages are:

- Handling concatenative and non-concatenative morphotactics (Beesley 1998b).
- The technology is fast and efficient. It can handle very huge automata of lexicons with their inflections. Compiling large networks that include several millions of paths is only a matter of seconds in a finite state calculus. Moreover, these large networks can be easily combined together to give even larger networks.
- Unicode support, which enables developers to accommodate native scripts that use non-Latin alphabets.
- Multi-platform support. Xerox finite state tools work under Windows, Linux, UNIX and Mac OS, which means that a morphological transducer developed using Xerox finite state compilers can serve applications under any of these platforms.
- A finite state system is fully reversible. So it can be used for analysis as well as generation.
- The regular expressions used in finite state closely resemble standard linguistic notations (Yona and Wintner 2007) so the rules are reasonably readable and intelligible.

In a standard finite state system, lexical entries along with all possible affixes and clitics are encoded in the lexc language which is a right recursive phrase structure grammar (Beesley 2001; Beesley and Karttunen 2003). A lexc file contains a number of lexicons connected through what is known as "continuation classes" which determine the path of concatenation. In example (27) the lexicon *Proclitic* has a form *a* which has a continuation class *Prefix*. This means that the forms in *Prefix* will be appended to the right of *a*. The lexicon *Proclitic* has also an empty string, which means that *Proclitic* is optional and that the path can proceed without it. The bulk of all lexical entries are presumably listed under *Root* in the example.

(27
LEXICON Proclitic
a Prefix;
Prefix;
LEXICON Prefix
c Root;
LEXICON Root
efg Suffix;
hij Suffix;
LEXICON Suffix
k Enclitic;
LEXICON Enclitic
l #;

In a natural language, it usually happens that an affix or a clitic requires or forbids the existence of another affix or clitic. This is what is termed as "separated dependencies" or "long distance dependencies" which constrain the co-occurrence of morphemes within words (Beesley and Karttunen 2003). So Flag Diacritics were introduced as an extension to Xerox finite state implementation to serve as filters on possible concatenations to a stem. The most common form of Flag Diacritics is the unification type. Suppose that we want to prevent the *Proclitic* and *Enclitic* lexicons from co-

occurring. We can add a Flag Diacritic to each of them with the same feature name, but with different value, as shown in (28).

(28
…
a@U.Clitic.On@ Prefix;
…
l@U.Clitic.Off@ #;
…

With inflections and concatenations, words usually become subject to changes or alternations in their forms. Alternations are the discrepancies between underlying strings and their surface realization (Beesley 1998b), and alternation rules are the rules that relate the surface forms to the underlying forms. In Arabic, long vowels, glides and the glottal stop are the subject of a great deal of phonological (and consequently orthographical) alternations like assimilation and deletion. Most of the trouble a morphological analyzer faces is related to handling these issues. In our system there are about 130 replace rules composed on the bottom of the lexicon to handle alternations that affect verbs, nouns, adjectives and function words when they undergo inflections or are attached to affixes and clitics. Alternation rules are expressed in finite state using XFST replace rules of the general form:

(29 a -> b || L _ R

This means that the string *a* is replaced with the string *b* when *a* occurs between the left context *L* and the right context *R*. When no context is specified the replacement operates globally, and the special symbol '.#.' can be used instead of *L* to indicate a left boundary, meaning when the string *a* occurs at the beginning of a word. When '.#.' is used instead of *R*, it indicates a right boundary, meaning when the string *a* occurs at the end of a word. These replace rules can be composed one over the other, so that the output of one rule can be the input for another rule. This can effectively account for multi phonological and orthographical processes.

At the end we obtain a transducer with a binary relation between two sets of strings. The first set of strings is conventionally known as the lower language and contains the surface forms, and the second set of strings is the upper language and contains the lexical forms, or the analysis, as shown in (30) for the verb يشكرون yashkurun ([they] thank).

(30 Upper Language: شكر+masculine+present+plural+3rdPerson
Lower Language: يشكرون

2.2 Handling Arabic Morphotactics

Morphotactics is the study of how morphemes combine together to form words (Beesley 1998b). These can be concatenative with morphemes either prefixed or suffixed to stems or non-concatenative, with stems themselves undergoing alternations to convey morphosyntactic information.

It seems that Arabic traditional grammarians (Ibrahim 2002) have been persuaded by morphology to classify words into only three types: verbs, nouns and particles. Adjectives take almost all the morphological forms of nouns. Adjectives, for example, can be definite, and are inflected for case, number and gender.

Arabic traditional grammarians have also classified tense into imperfective (present), perfective (past) and imperative. This, as well, is influenced by the fact that verbs in Arabic are inflected for imperfective, perfective and imperative. Moreover, both the perfective and imperfective have two forms: the active form and the passive form. To summarize, verbs are inflected to provide five forms: active perfective, passive perfective, active imperfective, passive imperfective and imperative. The base form of the verb is the perfective tense, 3rd person, singular. There are a number of indicators that tell how the base form would be inflected to give the other forms. Among these indicators are the number of letters of the base form and its template. A template (Beesley and Karttunen 2003) is a kind of vocalization mould in which a verb fits. Vocalism is a major factor in template shaping. Although diacritics (the manifestation of vocalism) are not present in modern writing, we still need to worry about them as they trigger other phonological and orthographical processes, such as assimilation and deletion and the re-separation (or spreading) of doubled letters.

2.2.1 Verbs

Possible concatenations and inflections in Arabic verbs are shown in Table 3. All elements are optional except the stem, and they can be connected together in a series of concatenations.

Flag Diacritics are used to handle long distance morphotactic restrictions or what is termed "separated dependencies" for Arabic verbs. These restrictions can be considered as grammatical constraints, or grammar-lexis specifications, that govern the morphological process. These can be summarized as follows:

- The yes-no-question article أ "a" (does) cannot co-occur with imperatives or with the accusative case.
- The complementizer ل "li" (to) cannot co-occur with the nominative case.
- Cliticized object pronouns do not occur either with passive or with intransitive verbs.
- Affixes indicating person and number in the present tense come in two parts one preceding and one following the verb and each prefix can co-occur only with certain suffixes.
- The imperfective, perfective and imperative have each a range of prefixes or suffixes or both which must be precisely constrained.
- A first person object pronoun cannot co-occur with a first person prefix (to account for the rule that a pronoun must be free within its binding domain), and similarly a second person object pronoun cannot co-occur with a second person prefix.

–

Proclitics		Prefix	Stem	Suffix	Enclitic
Conjunction/ question article	Complementizer	Tense/mood – number/gender	Verb	Tense/mood – number/gender	Object pronoun
Conjunctions و "wa" (and) or ف "fa" (then)	ل "li" (to)	Imperfective tense (5)	Stem	Imperfective tense (10)	First person (2)
				Perfective tense (12)	Second person (5)
Question word أ "a" (does or did)	س "sa" (will)	Perfective tense (1)			Third person (5)
	ل "la" (then)	Imperative (2)		Imperative (5)	

TABLE 3: Possible concatenations in Arabic verbs

The maximum number concatenations in Arabic verbs as shown by Table 3 is six; one stem plus 5 other bound morphemes representing affixes and clitics. Statistically, concatenations in Table 3 can give as much as 33,696 forms. In real constrained examples, some verbs, such as شكر "shakar" (to thank), can generate up to 2,552 valid forms. This considerable amount of form variations is a good indication of the richness and complexity of Arabic morphology.

2.2.2 Nouns

Possible concatenations and inflections in Arabic nouns are shown in Table 4 below. The maximum number of concatenations in Arabic nouns is five; one stem plus 4 other bound morphemes representing suffixes and clitics, bearing in mind that the genitive pronoun and the definite article do not co-occur.

Flag Diacritics are also used to handle separated dependencies for nouns. These can be summarized as follows:

- The definite article ال "al" (the) cannot co-occur with a genitive pronoun.
- The definite article cannot co-occur with an indefinite noun marking (*nuun* with the dual and plural or *tanween* with the singular).
- The cliticized genitive pronoun cannot co-occur with an indefinite noun marking.
- Prepositions cannot co-occur with nominative or accusative case markings.

Proclitics			Stem	Suffix	Enclitic
Conjunction/ question article	Preposition	Definite article	Noun	Gender/Number	Genitive pronoun
Conjunctions و "wa" (and) or ف "fa" (then)	ب "bi" (with), ك "ka" (as) or ل "li" (to)	ال "al" (the)	Stem	Masc Dual (4)	First person (2)
				Fem Dual (4)	
Question word أ "a" (does or did)				Masculine regular plural (4)	Second person (5)
				Feminine regular plural (1)	Third person (5)
				Feminine Mark (1)	

TABLE 4: Possible concatenations in Arabic nouns

Statistically, uncontrolled concatenations in Table 4 can give 6,240 forms. In real examples some nouns, such as معلم "mu'allim" (teacher), can generate up to 519 valid forms.

Another problem with nouns is the issue of broken plurals (Ratcliffe 1998; Ibrahim 2002), which is the traditional grammarians' term for describing the process of non-concatenative plural formation. The term was chosen to indicate that the base form of the nouns is broken either by removing one or more letters, adding one or more letters, changing vocalization or a combination of these. Arabic singular nouns have 30 templates served by 39 broken plural templates. Some templates of singular nouns can select from up to seven broken plural templates. The different plural templates were historically meant to indicate some meaning differences, such as whether the number of the plural is below or above ten, whether the noun describes a profession or an attribute, and whether the attribute is static or transient. These subtle meaning differences are no longer recognized even by well-educated native speakers.

These broken plural forms are, to a great extent, fossilized, i.e., they are not productive any more. So, the system relies only on the lexicographer's knowledge to tell whether a particular noun is to have a regular or broken plural form. Trying to rely on the system to guess the broken plural form will make the transducer overgenerate excessively and needlessly.

2.2.3 Alternation Rules

As verbs are the category most affected by alternation operations, we focus here on the main conditions that trigger orthographical changes in verbs. Arabic verbs are generally classified (regarding the number of letters of the base form) into three-, four-, five- and six-letter verbs. Furthermore, trilateral verbs are traditionally classified into:

A. Strong verbs. These are the verbs that contain no weak letters. They are further classified into three categories:

1. Regular verbs. These are the verbs whose formative letters do not contain either a hamzated, doubled or weak letter.

2. Hamzated verbs. These are the verbs that contain a hamza (glottal stop) among its formative letters.
3. Doubled verbs. These are the verbs that are composed of two letters and the second is doubled.

B. Weak verbs. These are the verbs that contain a weak letter. A weak letter is one of three letters representing either long vowels or glides. They are ا (alif) for the long vowel *aa* (which can also be represented orthographically by the letter ى (alif maqsuura). The second weak letter is و (waw) for the glide *w*. The third weak letter is ي (yaa) for the glide *y*. Weak verbs are also classified into three categories:
   1. Assimilated (mithal). A verb that contains an initial weak letter.
   2. Hollow (ajwaf). A verb that contains a middle weak letter.
   3. Defective (naqis). A verb that contains a final weak letter.

We can extend this notion of weak and strong verbs into the four-, five- and six-letter verbs. This classification is of crucial importance in writing alternation rules. Strong regular verbs are generally not so much affected, orthographically, by inflection. The verb in (31) undergoes one alternation operation that is the deletion of the first letter, when inflected into the imperfective.

(31 استخرج istakhraja (extracted) -> يستخرج yastakhrij (extract)

However, more attention should be given to verbs that contain a weak, hamzated, or doubled letter at any position, as this usually requires more orthographical alterations during inflection. The verb in (32) undergoes two operations: deletion of the fist letter and assimilation of the pre-final letter from ا "aa" into ى "ii".

(32 استقال istaqaala (resigned) -> يستقيل yastaqiil (resign)

In our lexc file, the start and end of stems are marked to provide information needed in conducting alternation operations, as shown by Figure 3. The tags are meant to provide the following information:

1. Start and end of verb stem. The multi-character symbol "^ss^" stands for stem start and "^se^" for stem end.
2. Which letter is doubled in the linear order, as the entries from 4 to 8 in Figure 3 show. The mark "^dbl2^dbl", for example means that the second letter is doubled.
3. If there is a long vowel that undergoes assimilation, the assimilated form needs to be explicitly stated. This is represented by the entries from 10 to 13 in Figure 3. In traditional terms the origin of ا "aa" in قال "qaal" (said) is و "uu". which means that "aa" changes to "uu" when the verb is inflected into the imperfective
4. The flag diacritic "@D.V.P@" means "disallow the passive voice", and "@D.M.I@" means "disallow the imperative mood.

These markings are considered an intermediate language which is removed in the final stage, so that only surface strings are left on the bottom and analysis strings (or lexical strings) are left on the top of the network (Beesley 1996).

```
1  LEXICON Verbs
2  ^ss^شكر^se^                                  Transitive;
3  ^ss^فرح^se^@D.V.P@                           Intransitive;
4  ^ss^رد^se^^dbl2^dbl@D.V.P@                   Transitive;
5  ^ss^أمر^se^^dbl2^dbl@D.M.I@                  Transitive;
6  ^ss^أضر^se^^dbl3^dbl@D.V.P@@D.M.I@           Intransitive;
7  ^ss^امتد^se^^dbl4^dbl@D.V.P@@D.M.I@          Intransitive;
8  ^ss^تمخض^se^^dbl3^dbl@D.V.P@@D.M.I@          Intransitive;
9  ^ss^استقر^se^^dbl5^dbl@D.V.P@@D.M.I@         Intransitive;
10 ^ss^باع^se^^origي^orig                       Transitive;
11 ^ss^قال^se^^origو^orig                       Intransitive;
12 ^ss^غزا^se^^origو^orig@D.V.P@@D.M.I@         Transitive;
13 ^ss^رمى^se^^origي^orig                       Transitive;
```

FIGURE 3: Verb stem entries

2.3 Techniques followed in limiting ambiguities

In order to make the system produce only valid solutions and avoid spurious solutions, the following considerations and techniques were followed.

1. Using the stem as the base form. Automatic derivation from the root can be risky as it may create stems not used in the language.

2. Non-inclusion of classical words or word senses, as they add only to the size of the lexicon and the level of ambiguity.

3. Observation of the rules governing the combination of words with affixes and clitics, or grammar-lexis specifications, which work as filters for spurious ambiguities (Dichy 2001; Dichy and Fargaly 2003; Dichy and Hassoun 1998; Abbès et al 2004). For example, adjectives, names of places, verbal nouns do not combine with possessive pronouns. Also verbal nouns derived from intransitive verbs do not combine with accusative pronouns. Yet more can be done regarding the filtering of human objects from verbs that allow only non-human objects (Dichy and Fargaly 2003) such as (33), which is still accepted by our system.

   (33 قرأتهم
   qara'tu-hum
   I read them

   There are also nouns that semantically do not allow the affixation of genitive pronouns, such as (34) which is still not properly handled by our system.

   (34 كيميائي
   kimia'i
   my chemistry

4. Specifying which verbs can have the passive forms. From 1297 verbs, only 544 verbs (%41) are allowed to have passive forms (500 transitive verbs, and 44 intransitive verbs). Initially all transitive verbs were allowed to have a passive form and all intransitive verbs were not. Then all verbs were reviewed manually for acceptability according to the author's judgment. A sum of 198 transitive verbs was not allowed to have a passive form, while some intransitive verbs are allowed. (Levin 1993) stated that intransitive, prepositional verbs can have passive constructions under constraints on the semantic roles of the arguments. In our

system, verbs in the 1st and 2nd person are not allowed to have a passive form. The 1st and 2nd persons are deemed as highly unlikely forms, first, because MSA is a formal written language, and these persons are mostly used in conversations or autobiographies. Second, these persons have orthographical shapes that are identical with other forms, and writers will tend to use other syntactically equivalent structures for expressing the passive in this case. Another good idea for limiting the use of the passive would be to constrain it according to tense, as done in the Buckwalter's system.

5. Specifying which verbs can have imperative forms. Out of 1297 verbs, only 483 verbs (%37) are allowed to have an imperative form (324 Transitive verbs, 159 Intransitive verbs). According to (Levin 1993), the imperative construction does not appear with verbs of perception and admire-type psych-verbs. It does not also appear with verbs of entity-specific change of state. These are the "verbs that describe changes of state that are specific to particular entities", such as bloom, erode, corrode, decay, dry, stagnate, blossom, wither, tarnish, swell.

3. EVALUATION

Our aim is to evaluate Xerox Arabic Morphological Analysis and Generation, Buckwalter Arabic Morphological Analyzer and Attia's Arabic Morphological Transducer with respect to ambiguity. Due to the fact that a gold standard annotated corpus for Arabic is not yet available (to our knowledge), a large scale, automatic evaluation is not possible. Therefore we conducted a small-scale manual evaluation experiment to test the ambiguity rate of the three morphologies on one hand and to test precision of the two morphologies with the least ambiguity rate on the other hand.

We selected five recent documents from Al-Jazeera web site containing a total of 950 unique words and 67 MWEs. We tested these words on each of the target morphologies, and then we conducted a detailed analysis for the two morphologies with the least ambiguity rate to see how accurate they were in obtaining the correct set of analyses and avoiding spurious ambiguities. We first show the precision evaluation in Table 5. A "complete" analysis is a precise one that neither contains a spurious ambiguity nor lacks a plausible solution. An "over-specified" analysis is one that contains all plausible solutions beside one or more spurious ambiguities. It must be noted here that a spurious ambiguity is an illegal ambiguity that falls outside the domain of the language, not a context or subject related ambiguity. An "underspecified" analysis is one that fails to account for one or more plausible solutions among the list of solutions. "Over-&under-specified" analysis denotes those solutions which contain spurious ambiguities and at the same time do not include one or more plausible solutions. "Irrelevant" words are misspelled words or those that do not occur alone, but usually occur as part of a MWE. Buckwalter's precision score is %64, while Attia's Morphology achieved %79. Although Attia's morphology is almost a quarter of the size of Buckwalter, it does not contain too many underspecified analyses. As Attia and Buckwalter achieved a relatively high score of precision at a low rate of solutions per word, it can be easily deduced that Xerox, with its high number of solutions, is over-specified for most words, and so no breakdown was perceived to be needed.

Criteria	Buckwalter	Xerox	Attia
Complete	617	-	756
Over-specified	247	-	67
Underspecified	40	-	75
Wrong Analysis	1	-	10
Over-&under-specified	20	-	5
Irrelevant	5	5	5
Not found	20	39	32
Total Solutions for 895 words (after excluding 55 not found)	2332	3871	1574

TABLE 5: Breakdown of evaluation results

English ambiguity rate is tested using XLE morphological transducer (Butt et al 2002) on 979 words, in order to be used as a baseline, and received 1732 solutions, giving an ambiguity rate of 1.76.

In order to measure the ambiguity rate in the three morphologies in our experiment, all words that were not known to any of the morphologies (that was a total of 55 words) were removed from the test list, which was reduced to 895. The ambiguity rates for the three morphologies are shown in Figure 4. A total of 67 MWEs were excluded from overall evaluation, as they are not supported on Buckwalter or Xerox. Attia, however, recognized 25 MWEs, that is 37% coverage.

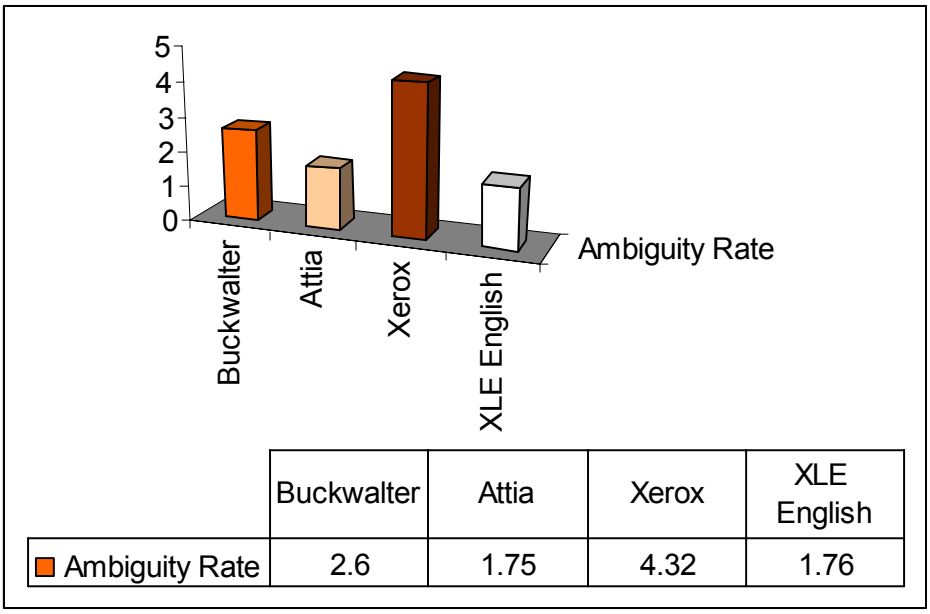

FIGURE 4: Comparison of the ambiguity rates in three morphologies

Error review shows that the sources of illegal ambiguities in Buckwalter and Xerox morphologies are summarized in the following three main points:

1. Inclusion of classical terms.
2. Incompliance with grammar-lexis relations rules.
3. Improper application of spelling relaxation rules.

4. CONCLUSION

The rich and complex morphology of Arabic does not automatically mean that it is highly ambiguous. The analysis and evaluation conducted in this paper shows that most of the ambiguities produced by Xerox Arabic Finite State Morphology and Buckwalter Arabic Morphological Analyzer are spurious ambiguities caused by the inclusion of classical entries, rule-created overgenerated stems with no actual place in the language, overlooking word-clitic combination rules (or grammar-lexis specifications), and overdoing spelling relaxation rules. By avoiding these pitfalls a more focused, less ambiguous morphological analyzer can be developed.

REFERENCES

Abbès R, Dichy J, Hassoun M (2004): The Architecture of a Standard Arabic lexical database: some figures, ratios and categories from the DIINAR.1 source program, *The Workshop on Computational Approaches to Arabic Script-based Languages, COLING 2004*. Geneva, Switzerland.

Attia MA (2005): Developing a Robust Arabic Morphological Transducer Using Finite State Technology, *8th Annual CLUK Research Colloquium*. Manchester, UK.

Attia MA (2006): Accommodating Multiword Expressions in an Arabic LFG Grammar. In Salakoski T, Ginter F, Pyysalo S (eds), *FinTAL 2006*, Vol 4139. Turku, Finland: Springer-Verlag Berlin Heidelberg, pp 87 - 98.

Azmi M (1988): *Arabic Morphology: A Study in the System of Conjugation*. Hyderabad: Hasan Publishers.

Beesley KR (1996): Arabic Finite-State Morphological Analysis and Generation, *Proceedings of the 16th conference on Computational linguistics*, Vol 1. Copenhagen, Denmark: Association for Computational Linguistics, pp 89-94.

Beesley KR (1998a): Arabic Morphological Analysis on the Internet, *The 6th International Conference on Multilingual Computing*. Cambridge.

Beesley KR (1998b): Arabic Morphology Using Only Finite-State Operations, *Proceedings of the Workshop on Computational Approaches to Semitic languages*. Montreal, Quebec, pp 50-57.

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

Beesley KR, Karttunen L (2003): *Finite State Morphology*. Stanford, Calif.: Csli.

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

Butt M, Dyvik H, King TH, Masuichi H, Rohrer C (2002): The Parallel Grammar Project, *COLING-2002 Workshop on Grammar Engineering and Evaluation*. Taipei, Taiwan.

Chalabi A (2004): Sakhr Arabic Lexicon, *NEMLAR International Conference on Arabic Language Resources and Tools*. Cairo, Egypt.

Darwish K (2002): Building a Shallow Morphological Analyzer in One Day, *Proceedings of the workshop on Computational Approaches to Semitic Languages in the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*. Philadelphia, PA, USA.

Dichy J (2001): On lemmatization in Arabic. A formal definition of the Arabic entries of multilingual lexical databases, *ACL 39th Annual Meeting. Workshop on Arabic Language Processing; Status and Prospect*. Toulouse, pp 23-30.

Dichy J, Fargaly A (2003): Roots & Patterns vs. Stems plus Grammar-Lexis Specifications: on what basis should a multilingual lexical database centred on Arabic be built?, *Proceedings of the MT-Summit IX workshop on Machine Translation for Semitic Languages*. New-Orleans.

Dichy J, Hassoun M (1998): Some aspects of the DIINAR-MBC research programme, *Proceedings of 6th ICEMCO, International Conference and Exhibition on Multi-lingual Computing*. Cambridge, England.

Freeman A (2001): Brill's POS tagger and a Morphology parser for Arabic, *39th Annual Meeting of Association for Computational Linguistics & 10th Conference of the European Chapter, Workshop on Arabic Language Processing: Status and Prospects*. Toulouse, France.

Hajic J, Smrž O, Buckwalter T, Jin H (2005): Feature-Based Tagger of Approximations of Functional Arabic Morphology, *The Fourth Workshop on Treebanks and Linguistic Theories*. Universitat de Barcelona.

Ibrahim K (2002): *Al-Murshid fi Qawa'id Al-Nahw wa Al-Sarf [The Guide in Syntax and Morphology Rules]*. Amman, Jordan: Al-Ahliyyah for Publishing and Distribution.

Kamir D, Soreq N, Neeman Y (2002): A Comprehensive NLP System for Modern Standard Arabic and Modern Hebrew, *Proceedings of the workshop on Computational Approaches to Semitic Languages in the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*. Philadelphia, PA, USA.

Kiraz GA (1998): Arabic Computational Morphology in the West, *Proceedings of the 6th International Conference and Exhibition on Multi-lingual Computing*. Cambridge, UK.

Levin B (1993): *English Verb Classes and Alternations: A Preliminary Investigation*: University of Chicago Press.

McCarthy JJ (1985): *Formal Problems in Semitic Phonology and Morphology*. New York ; London: Garland.

Ratcliffe RR (1998): *The Broken Plural Problem in Arabic and Comparative Semitic : Allomorphy and Analogy in Non-concatenative Morphology*. Amsterdam ; Philadelphia: J. Benjamins.

Wehr H (1979): *A Dictionary of Modern Written Arabic*, 4th Edition ed. Ithaca, NY: Spoken Language Services, Inc.

Yona S, Wintner S (2007): A finite-state morphological grammar of Hebrew. *Natural Language Engineering*.