

Morphophonemic and orthographic rules in a multi-dialectal morphological analyzer and generator for Arabic verbs

Nizar Habash and Owen Rambow

Center for Computational Learning Systems, Columbia University, New York City, USA

Keywords :Morphological analysis, morphological generation, linguistic rules, finite-state.

Abstract MAGEAD is a morphological analyzer and generator for the Arabic language family, by which we mean both Modern Standard Arabic (MSA) and the spoken dialects. MAGEAD explicitly addresses the need for processing the morphology of the dialects as well. Furthermore, MAGEAD can model both phonology and orthography explicitly. In this paper, we present in detail the morphophonemic and the orthographic rules which we have encoded in MAGEAD for MSA verbs as well as for Levantine dialectal Arabic verbs.

97

Introduction

In previous papers (Habash et al. 2005, Habash and Rambow 2006), we presented a new morphological analyzer for Arabic and its dialects, called MAGEAD. MAGEAD had several design goals: we wanted MAGEAD to provide an analysis to the level of root-and-pattern as well as to the level of lexeme-and-features, and we wanted to allow for a system that, given the right information, can cover all variants of Arabic, by which we mean both Modern Standard Arabic (MSA) and the spoken dialects. To our knowledge, MAGEAD is the first morphological analyzer and generator for an Arabic dialect that includes a root-and-pattern analysis. The specific contribution of this paper is to discuss in some detail the representation of morphological rules (phonological/morphophonemic and orthographic) needed for the analysis or generation of Arabic verbs. When generating an inflected word form, the morphological rules apply after the creation of the word stem from the root and pattern, and after the affixation of prefixes, suffixes, and circumfixes. We present an important subset of the set of rules needed for MSA and Levantine verbs using general linguistic notation (the whole set will be available as a technical report, (Habash and Rambow, forthcoming)). These rules should be of use to anyone wanting to construct a morphological analyzer or generator, independently of the framework used.

The paper is structured as follows: we first summarize the relevant facts about Arabic morphology and introduce our terminology. We then briefly review related work by other researchers. We summarize how we use multi-tape finite state automata in MAGEAD, and explain our use of abstract and concrete morphemes. After these preliminaries, we discuss the morphological rule system in MAGEAD, listing all types that are needed for MSA. We then show which rules need to be changed in Levantine and conclude.

Arabic Morphology

Variants of Arabic

The Arabic-speaking world is characterized by diglossia (Ferguson, 1959). Modern Standard Arabic (MSA) is the shared written language from Morocco to the Gulf, but it is not a native language of anyone. It is spoken only in formal, scripted contexts (news, speeches). In addition, there is a continuum of spoken dialects (varying geographically, but also by social class, gender, etc.) which are native languages, but rarely written (except in very informal contexts: collections of folk tales, newsgroups, email, etc). We will refer to MSA and the dialects as variants of Arabic. Variants differ phonologically, lexically, morphologically, and syntactically from one another; many pairs of variants are mutually unintelligible. In unscripted situations where spoken MSA would normally be required (such as talk shows on TV), speakers usually resort to repeated code-switching between their dialect and MSA, as nearly all native speakers of Arabic are unable to produce sustained spontaneous discourse in MSA.

In this paper, we discuss MSA and Levantine, the dialect spoken in Syria, Lebanon, Jordan, and Palestine. The discussion in this section uses only examples from MSA, but all variants of Arabic show a combination of root-and-pattern and affixational morphology and similar examples could be found for Levantine.

Types of Arabic Morphemes

Arabic morphemes fall into three categories: templatic morphemes, affixational morphemes, and non-templatic word stems (NTWSs). NTWSs are word stems that are not constructed from a



root/pattern/vocalism combination. They tend to be foreign names and borrowed terms, e.g., واشنطن waAšinTun ‘Washington’. Verbs are never NTWSs; therefore, we do not discuss them further in this paper.

Templatic Morphemes

Templatic morphemes come in three types that are equally needed to create a word stem: roots, patterns and vocalisms. The root morpheme is a sequence of three, four, or five consonants (termed radicals) that signifies some abstract meaning shared by all its derivations. For example, the words كتب katab ‘to write’, كاتب kaAtib ‘writer’, and مكتوب maktuwb ‘written’ all share the root morpheme ك ت ب ktb ‘writing-related’. The pattern morpheme is an abstract template in which roots and vocalisms are inserted. The vocalism morpheme specifies which short vowels to use with a pattern. We will represent the pattern as a string made up of numbers to indicate radical position, of the symbol V to indicate the position of the vocalism, and of pattern consonants (if needed). For example, the pattern 1V22V3 indicates that the second root radical is to be doubled. A pattern can include letters for additional consonants and vowels, e.g., the verbal pattern V1tV2V3. In some discussions of Arabic morphology where the vocalism and pattern are not separated, word stems are constructed from complex patterns (i.e. pattern+vocalism) and roots only. Separating the vocalism as its own morpheme allows us to abstract certain features that consistently vary across complex patterns, such as voice (passive versus active). A word stem is constructed by interleaving the three types of templatic morphemes. For example, the word stem كتب katab ‘to write’ is constructed from the root ك ت ب ktb, the pattern 1V2V3 and the vocalism aa.

Affixational Morphemes

Arabic affixes can be prefixes such as س+ sa+ ‘will/[future]’, suffixes such as و+ +uwna ‘[masculine plural]’ or circumfixes such as ت+ن+ ta++na ‘[imperfective subject 2nd person fem. plural]’. Multiple affixes can appear in a word. For example, the word وسيتكونها wasayaktubuwnahA ‘and they will write it’ has two prefixes, one circumfix and one suffix:

wasayaktubuwnahA					
wa+	sa+	y+	aktub	+uwna	+hA
and	will	3person	write	masculine-plural	it

We analyze the imperfective word stem as including an initial short vowel, and leave a discussion of this analysis to future publications. Some of the affixes can be thought of as orthographic clitics, for example the conjunction و+ w+ ‘and’, the preposition ل+ l+ ‘to/for’ or the pronominal object clitic ها+ +hA. Others are bound morphemes.

Morphological Rewrite Rules

An Arabic word is constructed by first creating a word stem from templatic morphemes or by using a NTWS. Affixational morphemes are then added to this stem. A number of phonological, morphophonemic and orthographic rules then modify the form of the created word: it is not a simple interleaving or concatenation of its morphemes. We discuss these rules in much greater detail in a separate section below.

Morpheme Type and Morpheme Function

The type of morpheme is independent of the morphological function it is used for (derivational or inflectional). Although affixational morphemes tend to be inflectional and templatic morphemes derivational, there are many exceptions. For example, the plural of كتاب kitAb ‘book’ is not formed through affixation of the inflectional plural morphemes ات+ +At or ون+ +uwn, but rather with a broken plural, i.e., through the use of a different pattern, resulting in كتب kutub ‘books’. Conversely, the adjective كُتُبِي kutubiy~ ‘book-related’ is derived from the noun كتب kutub ‘books’ using affixational morphemes.

Dialect Morphology

Arabic dialect morphology shares with MSA morphology the root-and-pattern system. Additionally, each dialect shares with MSA some morphemes, and some (or most) of the morphological rules. Consider the following forms by way of example:

Egyptian: ماينولهاالكش
 mAbinÿuwlhAlakš
 mA+ bi+ n+ [‘wl + V12V3 + iu] +hA +lak +š

MSA: لا نقولها لك
 lA naquwluhA laka
 lA / n+ [qwl + V12V3 + au] +u +hA / la +ka

Morpho-
 phonemic
 and
 orthographic
 rules

Here, the Egyptian stem is formed from the same pattern as the MSA stem, but the initial radical, ق q in MSA, has become ء (همزة) Hamza, glottal stop) in Egyptian through regular sound change. The vocalism in Egyptian also differs from that in MSA. Then, we add the first person plural subject agreement marker, the prefix +ن n+ (which in MSA is the circumfix ُ++ ن++u) and the third person feminine singular object clitic +ها hA+ (same in MSA). In Egyptian, we add a second person masculine singular indirect object clitic +لك +lak, the present progressive prefix +ب b+, and the negation circumfix ما++ش mA++š. None of these affixes exist in MSA: their meaning is represented with separate words, or as a zero morpheme in the case of the present tense marker.

Previous Work

There has been a considerable amount of work on morphological analysis for MSA; for an excellent overview, see (Al-Sughayer and Al-Kharashi, 2004). We very briefly summarize some of the work most closely related to ours. To our knowledge, there has been no computational work on dialectal morphology.

99

Kataja and Koskeniemi (1988) present a system for handling Akkadian root-and-pattern morphology by adding an additional lexicon component to Koskeniemi's two-level morphology (Koskeniemi, 1983). The first large scale implementation of Arabic morphology within the constraints of finite-state methods is that of (Beesley et al., 1989) with a 'detouring' mechanism for access to multiple lexica, which gives rise to other works by Beesley (1998) and, independently, by Buckwalter (2004).

The work of McCarthy (1981) describing root-and-pattern morphology in the framework of autosegmental phonology has given rise to a number of computational proposals. Kay (1987) presents a framework with which each of the autosegmental tiers is assigned a tape in a multi-tape finite state machine, with an additional tape for the surface form. Kiraz (2000, 2001) extends Kay's approach and implements a small working multi-tape system for MSA and Syriac. We follow the approach of Kiraz in our work. Other auto-segmental approaches (described in more details in Kiraz (2001, Chapter 4)) include those of Kornai (1995), Bird and Ellison (1994), Pulman and Hepple (1993) (whose formalism Kiraz adopts), and others.

The MAGEAD System: Implementation

MAGEAD relates (bidirectionally) a lexeme and a set of linguistic features to a surface word form through a sequence of transformations. In a generation perspective, the features are translated to abstract morphemes which are then ordered, and expressed as concrete morphemes. The concrete templatic morphemes are interdigitated and affixes added, and finally morphological rewrite rules are applied. In this section, we briefly discuss our organization of linguistic knowledge and give some examples, while the translation to morphemes and the morphological rules are discussed in the following two sections of this paper.

We follow Kiraz (2000) in using a multitape representation. We extend the analysis of Kiraz by introducing a fifth tier. The five tiers are used as follows:

- Tier 1: pattern and affixational morphemes;
- Tier 2: root;
- Tier 3: vocalism;
- Tier 4: phonological representation;
- Tier 5: orthographic representation.

In the generation direction, tiers 1 through 3 are always input tiers. Tier 4 is first an output tier, and subsequently an input tier. Tier 5 is always an output tier. All tiers are read or written at the same time, so that the rules of the multi-tier automaton are rules which scan the input tiers and, depending on the state, write to the output tier. The introduction of two surface-like tiers is due to the fact that many dialects do not have a standard orthography.

We have implemented multi-tape finite state automata as a layer on top of the AT&T two-tape finite state transducers (Mohri et al., 1998). We have defined a specification language for the higher multitape level, the new Morphtools format. Specifications in the Morphtools format of different types of information (such as rules or context-free grammars for morpheme ordering) are compiled

to the appropriate Lextools format (an NLP-oriented extension of the AT&T toolkit for finite-state machines (Sproat 1995)). For reasons of space, we omit a further discussion of Morphtools. For details, see (Habash et al., 2005).

Abstract and Concrete Morphemes *Morphological Behavior Classes*

Morphological analyses are represented in terms of a lexeme and features. The list of features is variant-independent, though in fact some variants may express features others do not. For example, many dialects have lost the mood distinctions found in MSA. We define the lexeme to be a triple consisting of a root (or an NTWS), a meaning index, and a morphological behavior class (MBC). We do not deal with issues relating to word sense here and therefore do not further discuss the meaning index.

We use as our example the surface form ازدهرت Aizdaharat (Azdhrt without diacritics) ‘she/it flourished’. The lexeme-and-features representation of this word form is as follows:

Root:zhr MBC:verb-VIII POS:V PER:3 GEN:F NUM:SG ASPECT:PERF

An MBC maps sets of linguistic feature-value pairs to sets of abstract morphemes. For example, MBC verb-VIII for form VIII verbs (افتعل) maps the feature-value pair ASPECT:PERF to the abstract root morpheme [PAT_PV:VIII], which in MSA corresponds to the concrete root morpheme V1tV2V3, while the MBC verb-I for form I verbs (فعل) maps ASPECT:PERF to the abstract root morpheme [PAT_PV:I], which in MSA corresponds to the concrete root morpheme 1V2V3. We define MBCs using a hierarchical representation with non-monotonic inheritance. The hierarchy allows us to specify only once those feature-to-morpheme mappings for all MBCs which share them. For example, the root node of our MBC hierarchy is simply labeled “word”, and all Arabic words share certain mappings, such as that from the linguistic feature conj:w to the clitic +و w+. This means that all Arabic words can take a cliticized conjunction. Similarly, the object pronominal clitics are the same for all transitive verbs, no matter what their templatic pattern is. Our hypothesis is that the MBC hierarchy is variant-independent, though as more variants are added, some modifications may be needed. Our current MBC hierarchy specification for both MSA and Levantine, which covers only the verbs, comprises 66 classes, of which 25 are abstract, i.e., only used for organizing the inheritance hierarchy and never instantiated in a lexeme.

Ordering and Mapping Abstract and Concrete Morphemes

To keep the MBC hierarchy variant-independent, we have also chosen a variant-independent representation of the morphemes that the MBC hierarchy maps to. We refer to these morphemes as abstract morphemes (AMs). The AMs are then ordered into the surface order of the corresponding concrete morphemes. The ordering of AMs is specified in a variant-independent context-free grammar. At this point, our example (ازدهرت Aizdaharat) looks like this:

[Root:zhr][PAT_PV:VIII][VOC_PV:VIII-act] + [SUBJSUF_PV:3FS]

Note that as the root, pattern, and vocalism are not ordered with respect to each other, they are simply juxtaposed. The ‘+’ sign indicates the ordering of affixational morphemes. Only now are the AMs translated to concrete morphemes (CMs), which are concatenated in the specified order. Our example becomes:

<zhr,V1tV2V3,iaa> +at

The interdigitation of root, pattern and vocalism then yields the non-final form iztahar+at.

Since the concrete morpheme is represented phonologically not orthographically, long vowels and shadda are not represented as they appear on the surface. For example, the abstract morpheme for first person pronominal object, [OBJ:1S], is mapped to +nī (note the line over the i indicating length), not +niy. Similarly, the perfective subject suffix for second person feminine plural is mapped to +tunna not +tun~a.

Overall for verbs, there are 92 non-stem mappings from AMs to CMs: three conjunction mappings, six particle mappings, thirteen object pronominal clitics, thirteen perfective suffixes, 52 imperfective prefixes and suffixes and five imperative verb suffixes.

For each verb form, there are four entries corresponding to perfective/imperfective and passive/active voice variations. The passive/active voice is done with vocalism change. The following are the

2. Morphophonemic Lexical Rules

A. Verb Form VIII (افتعل) Rules

Three rules are used to model the phonological changes that occur in verb form VIII from interaction with root consonants. The first rule changes the t in the pattern to d when the first consonant is z, d or ḏ (written as ز, د, or ذ respectively). For example, iztaharat becomes izdaharat.

V	1	t	V	2	V	3	+	a	t
	z			h		r			
i			a		a				
i	z	d	a	h	a	r	+	a	t

The second rule similarly changes the t in the pattern to T when the first consonant is emphatic: S, T, D or Ḍ (written as ص, ط, ض or ظ respectively). For instance, iDTarab+at becomes iDTarab+at. The third rule transforms the phonological realization of the first root radical when it happens to be w into t, thus assimilating to the t of the pattern. For example, iwtakal+a becomes ittakal+a.

B. Rule for Weak C1

The phonological form of the first radical is mapped to the special symbol ε, which represents the empty string, when preceded by a short vowel and followed by the second radical. For example, the imperfective form of the verb waSal, y+awSil+u becomes y+aSil+u.

C. Rules for Weak C2 and C3

These rules are used for modeling the behavior of weak radicals (WR) in second and third root position. We generally follow the rules proposed in Holes (2004) with a few exceptions detailed below. The third radical case is similar to the second radical except that the following short vowel must come from a suffix vowel on the pattern tier rather than a vocalism vowel. For example, daçaw+a becomes daçā. Two other special contexts are that only C3 can be followed by a zero suffix and that only C2 can be preceded by a consonant. These special conditions are exploited in our rule writing to maximize the efficiency of the rule set. Most of the rules are applied regardless of the identity of the WR (w or y). There are two sets of rules that are applied in strict order to efficiently account for the different behavior of WRs in open and closed syllables. We list some of the rules by way of example. There are two sets: the WR Modification Rules are applied first, and then the WR Closed Syllable Rules.

a. WR Modification Rules

We give some examples from this class.

1. A sequence consisting of an a followed by a WR followed by any short vowel or an ū is turned into an ā. For example, bayaç becomes bāç ‘to sell’; qawul becomes qāl ‘to say’.
2. A sequence consisting of an i followed by a WR followed by an a or ā is turned into a y. For example, duçiw+ā becomes duçiy+ā.
3. A sequence consisting of a u or i followed by a WR followed by an i or ī is turned into an ī. For example: quwil become qīl.
4. A sequence consisting of a WR followed by an a (and not preceded by a vowel) is turned into an ā. For example, y+uqwal+u becomes y+uqāl+u.
5. A third WR is deleted when followed by a zero suffix: for example, y+adçuw+0 becomes y+adçu+0 (as in jussive imperfective case).

b. WR Closed Syllable Rules

These rules account for the different behavior of WRs in closed syllables. Since these rules are applied after the previous set of rules (WR Modification Rules), they refer to the modified context. In the case of a WR in second root position that is followed by a short vowel on the vocalism tier and that has been realized as a long vowel on the phonology tier, the long vowel on the phonology tier is turned into the short vowel appearing on the vocalism tier. This is a nice example of the power of multi-tier morphology. For example, qawul+tu is initially turned into qāl+tu but then, according to this rule, into qul+tu. Similarly, bayiç+nā is initially turned into bāç+nā and then according to this rule into biç+nā.

Note here that Holes explains such cases using his “Shortening Middle Theory” which says that the short vowel in the perfective is the same as the shortened vowel of the imperfective (so, qawal+tu is

qul+tu because the imperfective is ‘+aqūl+u). We assume differently that the underlying form for qul+tu is qawul+tu not qawal+tu and thus we do not need the additional complication of relating the imperfective stem vocalism to a perfective stem morphophonemic rule. Additionally, we do not believe Holes explains some additional cases such the form of xif+tu (with underlying xawif+tu and the imperfective ‘+axāf+u).

D. Geminate Rules

Geminate radicals rules are applied when second and third radicals have the same consonantal value, e.g., mdd. The geminate rules also apply in Form IX (افعل) where the third radical is repeated in the pattern, e.g., AiHmarar.

a. Geminate Rule 1

A vocalism short vowel is deleted on the phonology tier when preceded by a vocalism vowel and a geminate radical and followed by a geminate radical and a vowel suffix. For example, madad+a becomes madd+a and AiHmarar+a becomes AiHmarr+a. The case where the suffix is consonantal, no changes occur, e.g., madad+tu.

b. Geminate Rule 2

The realization of a geminate radical followed by a short vowel vocalism on the phonology tier is inverted when preceded by a consonant and followed by a consonant and a vocalic suffix. For example, y+amdud+u becomes y+amudd+u.

3. Orthographic Default Rules

Symbols on the phonology tier are mapped without any change to orthography tier.

4. Orthographic Lexical Rules

These rules operate on the orthographic tier (tier 5), but they refer to the earlier tiers.

A. Alif Maqsura Rule 1

A long ā on the orthography tier is turned into Alif Maqsura when it is generated from a short vowel following a third radical y. For example, ramay+a becomes ram+ā according to one of the phonological rules for weak non-initial consonants (Weak C2 and C3, Rule 2.C.a above); however it is orthographically turned into ram+ỵ.

B. Alif Maqsura Rule 2

This rule overwrites Alif Maqsura Rule 1 in the case of the suffix +at. In this case, the Alif Maqsura is turned into a short a again: ramay+at becomes ram+ỵt (Rule 4.A) then finally ram+at (Rule 4.B).

C. Alif Maqsura Rule 3

Alif Maqsura is spelled as Alif in medial positions. We also add a short vowel a before final Alif Maqsura. For example, ram+ỵ+hA become ram+A+hA. And ram+ỵ becomes ram+ạỵ.

The following three orthographic lexical hamza rules are applied here to avoid later orthographic rules treating them as regular hamzas.

D. Third Radical Hamza Rules

A Hamza in third radical position which is (initially) realized as Hamza on the phonology and orthography tiers is changed to Alif with Hamza Above \hat{A} when preceded by an a which comes from the vocalism tier. The same Hamza is converted into a Hamza on Yah \hat{y} when preceded by an i which comes from the vocalism tier.

E. First Person Singular Hamza Rule

The Hamza of the first person singular is always realized as an Alif with Hamza Above.

F. Hamza of Form IV (أفعل) Rule

The Hamza of form IV is always realized as Alif with Hamza Above.

G. +wA Suffix Rule

The $\text{وا}+$ +wA verbal suffix is represented on the phonological tier as a long vowel ū. It is written as $\text{وا}+$ +uw in medial word position (i.e., when followed by a pronominal object), and as $\text{وا}+$ +wA when in final word position.

5. Orthographic Non-Lexical Rules

These rules are purely orthographic rules that do not interact with any other tiers of information. They are the last step in generation mode and the first step in analysis mode. They are presented below in their order of application in generation mode.

A. Clean Zeros Rules

All empty characters are removed.

B. Clean Plus Rule

At this point, we remove all plus signs (which separate morphemes).

C. Long Vowel Spelling Rule

The long vowels are spelled using a short vowel and a glide: \bar{i} is written as iy and \bar{u} as uw .

D. Hard W spelling Rule

The special symbol hard W is spelled using a regular w at this point. Hard W is a root radical w that is not weak. It is necessary to explain some behaviors such as the presence of two verbs derived from root jwb and verb form X (استفعل): Aistajwab (with hard w) and AistajAb (with weak w).

E. Sukun Rule

A sukun is added between any two adjacent consonants or after a consonant at the end of a word.

F. Shadda Rule

The second of two repeated consonants (separated by a sukun) is replaced with a shadda. The sukun is deleted. For example, $madda$ becomes $mad\sim a$.

G. Hamza Rules

The phoneme Hamza is written using seven orthographic symbols. The correct symbol is dependent on the context of the Hamza. The following rules are applied. First a word-initial hamza is written with Alif Hamza Below (A) when followed by i and with Alif Hamza Above otherwise $\hat{\text{A}}$. The Hamza is written on a Yah when either followed or preceded by i . Alternatively, it is written on a Waw when either followed or preceded by u . Otherwise Hamza is written on Alif. Also, Alif Hamza Above followed by a long \bar{a} is rewritten as Alif Madda ($\hat{\text{A}}$).

H. Vowel initial Spelling Rule

Since initial vowel diacritics cannot appear on their own, we add an additional Alif at the beginning of words with initial vowels. At this point in rule application, our earlier example, $izdaharat$ is modified to its final form $Aizdaharat$.

From MSA to Levantine

We modified MAGEAD so that it accepts Levantine rather than MSA verbs. Our effort concentrated on the orthographic representation and we did not develop a full phonological representation. We used the MSA-like orthography for Levantine developed at the Linguistic Data Consortium (Maamouri et al., 2006). Changes were done only to the representations of linguistic knowledge at the four levels discussed above, not to the processing engine. We discuss the four levels in turn.

Morphological Behavior Classes: The MBCs are variant-independent, so in theory no changes needed to be implemented. However, as Levantine is our first dialect, we expand the MBCs to include two AMs not found in MSA: the aspectual particle $+b$ $b+$ and the postfix negation marker $+š$.

Abstract Morpheme Ordering: The context-free grammar representing the ordering of AMs need to be extended to order the two new AMs, which is straightforward.

Mapping Abstract to Concrete Morphemes: This step requires four types of changes to a table representing this mapping. In the first category, the new AMs require mapping to CMs. Second, those AMs which do not exist in Levantine need to be mapped to zero (or to an error value). These are dual number, and subjunctive and jussive moods. Third, in Levantine some AMs allow additional CMs in allomorphic variation with the same CMs as seen in MSA. This affects three object clitics; for example, the second person masculine plural, in addition to $+kum$ (also found in MSA), also can be $+kuwA$. Fourth, in five cases, the subject suffix in the imperfective is simply different for Levantine. For example, the second person feminine singular indicative imperfective suffix changes from $+iyina$ in MSA to $+iy$ in Levantine. Note that more changes in CMs would be required

were we completely modeling Levantine phonology (i.e., including the short vowels).

Morphological, Phonological, and Orthographic Rules: We need to change one rule, and add one. In MSA, the vowel between the second and third radical is deleted when they are identical only if the third radical is followed by a suffix starting with a vowel (Geminate Rule 1, Rule 2.D.a above). In Levantine, in contrast, gemination always happens, independently of the suffix. If the suffix starts with a consonant, a long [ē] (written as ay) is inserted after the third radical. For example, madad+t is turned into مديت maddayt not the MSA-like مددت madadt. The new rule deletes the first person singular subject prefix for the imperfective, +¹A+, when it is preceded by the aspectual marker b+. For example, b+ A+aktub is written بكتب baktub not باكتب bAaktub.

Conclusion

We have presented the morphophonemic and orthographic rules we use in MAGEAD, a generator and analyzer for Arabic and its dialects. Specifically, we have given an account of the rules needed for MSA and Levantine verbs. The complete set of rules can be found in (Habash and Rambow, forthcoming). The full system is available to interested researchers. Please contact the authors for a copy of MAGEAD.

Acknowledgements

This work was supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of DARPA. We would like to thank George Kiraz and Richard Sproat for help with this work.

Orthography		Phonology	Orthography		Phonology
Arabic	Transliteration		Arabic	Transliteration	
ء	'	'	ط	T	T
آ	Ā	'ā	ظ	Ḍ	Ḍ
أ	Ā	'	ع	Ṣ	Ṣ
ؤ	w̄	'	غ	γ	γ
إ	Ā	'	ف	f	f
ئ	ŷ	'	ق	q	q
ا	A	', ā	ك	k	k
ب	b	b	ل	l	l
هـ	h	a, t	م	m	m
ت	t	t	ن	n	n
ث	θ	θ	هـ	h	h
ج	j	j	و	w	w, ū
ح	H	H	ى	ý	a, y, ī
خ	x	x	ي	y	y, ī
د	d	d	أ	a	a
ذ	ð	ð	أ	u	u
ر	r	r	أ	i	i
ز	z	z	أ	ā	an
س	s	s	أ	ū	un
ش	š	š	أ	ī	in
ص	S	S	أ	~	prev. cons.
ض	D	D	أ	.	Ø

Table 1.
Guide to
orthographic
and
phonological
symbols used
(Habash et
al., 2007)

References

- Al-Sughaiyer, I. A. and Al-Kharashi, I. A. (2004) "Arabic morphological analysis techniques: A comprehensive survey", Journal of the American Society for Information Science and Technology, 55(3):189–213.
- Beesley, K., Buckwalter, T., and Newton, S. (1989) "Two level finite-state analysis of Arabic morphology", In Proceedings of the Seminar on Bilingual Computing in Arabic and English.

- Beesley, K. (1998) "Arabic morphology using only finite-state operations", In Rosner, M., editor, *Proceedings of the Workshop on Computational Approaches to Semitic Languages*, pages 50–7, Montreal.
- Bird, S., and Ellison, T. (1994) "One-level phonology", *Computational Linguistics*, 20(1):55–90.
- Buckwalter, T. (2004) *Buckwalter Arabic morphological analyzer version 2.0*.
- Darwish, K. (2003) "Building a shallow Arabic morphological analyser in one day", in *ACL02 Workshop on Computational Approaches to Semitic Languages*, Philadelphia, PA. Association for Computational Linguistics.
- Ferguson, C. F. (1959) "Diglossia", *Word*, 15(2):325–340.
- Habash, N. and Rambow, O. (2006) "MAGEAD: Morphological Analyzer and Generator for Arabic and its Dialects", in *Proceedings of the Association of Computational Linguistics*, Sydney, Australia.
- Habash, N. and Rambow, O. (forthcoming) "Morphophonemic and Orthographic Rules for Modern Standard Arabic Morphology in MAGEAD", Technical Report, Center for Computational Learning Systems, Columbia University, New York, NY, USA.
- Habash, N., Rambow, O. and Kiraz G. (2005) "Morphological Analysis and Generation for Arabic Dialects", in *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, Ann Arbor, MI.
- Habash, N. (2004), "Large Scale Lexeme Based Arabic Morphological Generation", in *Proceedings of Traitement Automatique du Langage Naturel (TALN-04)*. Fez, Morocco.
- Habash, N., Soudi, A., Buckwalter, T. (2007) "On Arabic Transliteration", in "Arabic Computational Morphology: Knowledge-based and Empirical Methods", Editors van den Bosch, A. and Soudi, A.
- Holes, C. (2004), "Modern Arabic: Structures, Functions, and Varieties", Second edition, Georgetown University Press, Washington DC, USA.
- Kataja, L. and Koskenniemi, K. (1988) "Finite State Description of Semitic Morphology", in *Proceedings of the 12th International Conference on Computational Linguistics (COLING-88)*, volume 1, pages 313–15.
- Kay, M. (1987) "Non-Concatenative Finite-State Morphology", in *Proceedings of the Third Conference of the European Chapter of the Association for Computational Linguistics*, pages 2–10.
- Kiraz, G. (2000) "Multi-Tiered Nonlinear Morphology using Multi-Tape Finite Automata: A Case Study on Syriac and Arabic", *Computational Linguistics*, 26(1):77–105.
- Kiraz, G. (2001) "Computational Nonlinear Morphology: With Emphasis on Semitic Languages", Cambridge University Press.
- Kornai, A. (1995) "Formal Phonology", Garland Publishing.
- Koskenniemi, K. (1983) "Two-Level Morphology", Ph.D. thesis, University of Helsinki.
- Maamouri, M., Bies, A., Buckwalter, T., Diab, M., Habash, N., Rambow, O., and Tabessi, D. (2006) "Developing and using a Pilot Dialectal Arabic Treebank", in *Proceedings of LREC*, Genoa, Italy.
- McCarthy, J. (1981), "A Prosodic Theory of Non-concatenative Morphology", *Linguistic Inquiry*, 12(3):373–418.
- Mohri, M., Pereira, F., and Riley, M. (1998) "A Rational Design for a Weighted Finite-state Transducer Library", in Wood, D. and Yu, S., editors, *Automata Implementation*, Lecture Notes in Computer Science 1436, pages 144–58. Springer.
- Pulman, S. and Hepple, M. (1993), "A Feature-based Formalism for Two-level Phonology: A Description and Implementation", *Computer Speech and Language*, 7:333–58.
- Sproat, R. (1995) "Lextools: Tools for Finite State Linguistic Analysis", Technical Report 11522- 951108-10TM, Bell Laboratories.