Arabic Morphology Generation Using a Concatenative Strategy

Violetta Cavalli-Sforza
Carnegie Technology Education
4615 Forbes Avenue
Pittsburgh, PA, 15213
violetta@cs.cmu.edu

Abdelhadi Soudi
Computer Science Department
Ecole Nationale de L'Industrie Minerale
Rabat, Morocco
asoudi@enim.ac.ma

Teruko Mitamura
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213
teruko@cs.cmu.edu

Abstract

Arabic inflectional morphology requires infixation, prefixation and suffixation, giving rise to a large space of morphological variation. In this paper we describe an approach to reducing the complexity of Arabic morphology generation using discrimination trees and transformational rules. By decoupling the problem of stem changes from that of prefixes and suffixes, we gain a significant reduction in the number of rules required, as much as a factor of three for certain verb types. We focus on hollow verbs but discuss the wider applicability of the approach.

Introduction

Morphologically, Arabic is a non-concatenative language. The basic problem with generating Arabic verbal morphology is the large number of variants that must be generated. Verbal stems are based on triliteral or quadriliteral roots (3- or 4-radicals). Stems are formed by a derivational combination of a root morpheme and a vowel melody; the two are arranged according to canonical patterns. Roots are said to interdigitate with patterns to form stems. For example, the Arabic stem katab (he wrote) is composed of the morpheme ktb (notion of writing) and the vowel melody morpheme 'a-a'. The two are coordinated according to the pattern CVCVC (C=consonant, V=vowel).

There are 15 triliteral patterns, of which at least 9 are in common use, and 4 much rarer quadriliteral patterns. All these patterns undergo some stem changes with respect to voweling in the 2 tenses (perfect and imperfect), the 2 voices (active and passive), and the 5 moods (indicative, subjunctive, jussive, imperative and energetic).[1] The stem used in the conjugation of the verb may differ depending on the person, number, gender, tense, mood, and the presence of certain root consonants. Stem changes combine with suffixes in the perfect indicative (e.g., katab-naa 'we wrote', kutib-a 'it was written') and the imperative (e.g. uktub-uu 'write', plural), and with both prefixes and suffixes for the imperfect tense in the indicative, subjunctive, and jussive moods (e.g. ya-ktub-na 'they write, feminine plural') and in the energetic mood (e.g. ya-ktub-unna or ya-ktub-un 'he certainly writes'). There are a total of 13 person-number-gender combinations. Distinct prefixes are used in the active and passive voices in the imperfect, although in most cases this results in a change in the written form only if diacritic marks are used.[2]

Most previous computational treatments of Arabic morphology are based on linguistic models that describe Arabic in a non-concatenative way and focus primarily on analysis. Beesley (1991) describes a system that analyzes Arabic words based on Koskenniemi's

[1] The jussive is used in specific constructions, for example, negation in the past with the negative particle lam (e.g., lam aktub 'I didn't write'). The energetic expresses corroboration of an action taking place. The indicative is common to both perfect and imperfect tenses, but the subjunctive and the jussive are restricted to the imperfect tense. The imperative has a special form, and the energetic can be derived from either the imperfect or the imperative.

[2] Diacritic marks are used in Arabic language textbooks and occasionally in regular texts to resolve ambiguous words (e.g. to mark a passive verb use).

(1983) two-level morphology. In Beesley (1996) the system is reworked into a finite-state lexical transducer to perform analysis and generation. In two-level systems, the lexical level includes short vowels that are typically not realized on the the surface level. Kiraz (1994) presents an analysis of Arabic morphology based on the CV-, moraic-, and affixational models. He introduces a multi-tape two-level model and a formalism where three tapes are used for the lexical level (root, pattern, and vocalization) and one tape for the surface level.

In this paper, we propose a computational approach that applies a concatenative treatment to Arabic morphology generation by separating the issue of infixation from other inflectional variations. We are developing an Arabic morphological generator using MORPHE (Leavitt, 1994), a tool for modeling morphology based on discrimination trees and regular expressions. MORPHE is part of a suite of tools developed at the Language Technologies Institute, Carnegie Mellon University, for knowledge-based machine translation. Large systems for MT from English to Spanish, French, German, Portuguese and a prototype for Italian have already been developed. Within this framework, we are exploring English to Arabic translation and Arabic generation for pedagogical purposes. We generate Arabic words including short vowels and diacritic marks, since they are pedagogically useful and can always be stripped before display.

Our approach seeks to reduce the number of rules for generating morphological variants of Arabic verbs by breaking the problem into two parts. We observe that, with the exception of a few verb types, there is very little interaction between stem changes and the processes of prefixation and suffixation. It is therefore possible to decouple, in large part, the problem of stem changes from that of prefixes and suffixes. The gain is a significant reduction in the size number of transformational rules, as much as a factor of three for certain verb classes. This improves the space efficiency of the system and its maintainability by reducing duplication of rules, and simplifies the rules by isolating different types of changes.

To illustrate our approach, we focus on a particular type of verbs, termed *hollow* verbs, and show how we integrate their treatment with that of more regular verbs. We also discuss how the approach can be extended to other classes of verbs and other parts of speech.

1 Arabic Verbal Morphology

Verb roots in Arabic can be classified as shown in Figure 1.[3] A primary distinction is made between weak and strong verbs. Weak verbs have a weak consonant ('w' or 'y') as one or more of their radicals; strong verbs do not have any weak radicals.

Strong verbs undergo systematic changes in stem voweling from the perfect to the imperfect. The first radical vowel disappears in the imperfect. Verbs whose middle radical vowel in the perfect is 'a' can change it to 'a' (e.g., qaTa`a 'he cut' -> yaqTa`u 'he cuts'),[4] 'i' (e.g., Daraba 'he hit' -> yaDribu 'he hits'), or 'u' (e.g., kataba 'he wrote' -> yaktubu 'he writes') in the imperfect. Verbs whose middle radical vowel in the perfect is 'i' can only change it to 'a' (e.g., shariba 'he drank' -> yashrabu 'he drinks') or 'i' (e.g., Hasiba 'he supposed' -> yaHsibu 'he supposes'). Verbs with middle radical vowel 'u' in the perfect do not change it in the imperfect (e.g., Hasuna 'he was beautiful' -> yaHsunu 'he is beautiful'). For strong verbs, neither perfect nor imperfect stems change with person, gender, or number.

Hollow verbs are those with a weak middle radical. In both perfect and imperfect tenses, the underlying stem is realized by two characteristic allomorphs, one short and one long, whose use depends on the person, number and gender.

[3] Grammars of Arabic are not uniform in their classification of "hamzated" verbs, verbs containing the glottal stop as one of the radicals (e.g. [sa?al] 'to ask'). Wright (1968) includes them as weak verbs, but Cowan (1964) doesn't. Hamzated verbs change the written 'seat' of the hamza from 'alif' to 'waaw' or 'yaa?', depending on the phonetic context.

[4] In the Arabic transcription capital letters indicate emphatic consonants; 'H' is the voiceless pharyngeal fricative ; '`' the voiced pharyngeal fricative ; '?' is the glottal stop 'hamza'.

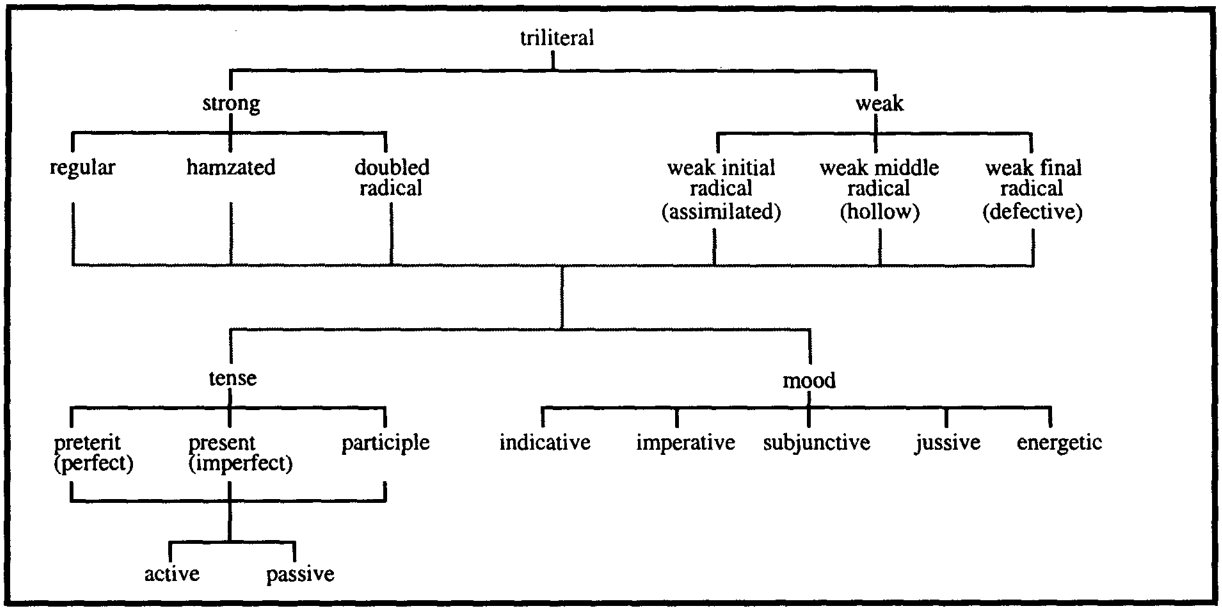

Figure 1: Classification of Arabic Verbal Roots and Mood Tense System

Hollow verbs fall into four classes:

1. Verbs of the pattern CawaC or CawuC (e.g. [Tawul] 'to be long'), where the middle radical is 'w'. Their characteristic is a long 'uu' between the first and last radical in the imperfect. E.g.,

   From the underlying root [zawar]:
   zaara 'he visited' and yazuuru 'he visits'

   Stem allomorphs:
   Perfect: -zur- and -zaar-
   Imperfect: -zur- and -zuur-

2. Verbs of the pattern CawiC, where the middle radical is 'w'. Their characteristic is a long 'aa' between the first and last radical in the imperfect. E.g.,

   From the underlying root [nawim]:
   naama 'he slept and yanaamu 'he sleeps'

   Stem allomorphs :
   Perfect: -nim- and -naam-
   Imperfect: -nam- and -naam-

3. Verbs of the pattern CayaC, where the middle radical is 'y'. Their characteristic is a long 'ii' before the first and last radical in the imperfect. E.g.,

   From the underlying root [baya`]:
   baa`a 'he sold' and yabii`u 'he sells'

   Stem allomorphs :
   Perfect: -bi`- and -baa`-
   Imperfect: and -bi`- and -bii`-

4. Verbs of the pattern CayiC, where middle radical is 'y'. E.g.,

   From the underlying root [hayib]:
   haaba 'he feared' and yahaabu 'he fears'

   Stem allomorphs :
   Perfect: -hib- and -haab-
   Imperfect: -hab- and -haab-

In the relevant literature (e.g., Beesley, 1998; Kiraz, 1994), verbs belonging to the above classes are all assumed to have the pattern CVCVC. The pattern does not show the verb conjugation class and makes it difficult to predict the type of stem allomorph to use. To avoid these problems, we keep information on the middle radical and vowel in the base form of the verb. In generation, classes 2 and 4 of the verb can be handled as one because they have the same perfect and imperfect stems.[5]

[5] The only exception is the passive participle. Verbs of classes 1 and 2 behave the same (e.g. Class 1: [zawar]: mazuwr 'visited'; Class 2 [nawil] → manuwl 'obtained'), as do verbs of classes 3 and 4 (e.g. Class 3: [baya`] → mabii` 'sold', Class 4: [hayib] → mahiib 'feared').

We describe our approach to modeling strong and hollow verbs below, following a description of the implementation framework.

2 The MORPHE System

MORPHE (Leavitt, 1994) is a tool that compiles morphological transformation rules into either a word parsing program or a word generation program.[6] In this paper we will focus on the use of MORPHE in generation.

Input and Output. MORPHE's output is simply a string. Input is a feature structure (FS) which describes the item that MORPHE must transform. A FS is implemented as a recursive Lisp list. Each element of the FS is a feature-value pair (FVP), where the value can be atomic or complex. A complex value is itself a FS. For example, the FS for generating the Arabic **zurtu** 'I visited' would be:

```
((ROOT "zawar")
 (CAT V)(PAT CVCVC)(VOW HOL)
 (TENSE PERF)(MOOD IND)
 (VOICE ACT)
 (NUMBER SG)(PERSON 1))
```

The choice of feature names and values, other than ROOT, which identifies the lexical item to be transformed, is entirely up to the user. The FVPs in a FS come from one of two sources. Static features, such as CAT (part of speech) and ROOT, come from the syntactic lexicon, which, in addition to the base form of words, can contain morphological and syntactic features. Dynamic features, such as TENSE and NUMBER, are set by MORPHE's caller.

The Morphological Form Hierarchy. MORPHE is based on the notion of a morphological form hierarchy (MFH) or tree. Each internal node of the tree specifies a piece of the FS that is common to that entire subtree. The root of the tree is a special node that simply binds all subtrees together. The leaf nodes of the tree correspond to distinct morphological forms in the language. Each node in the tree below the root is built by specifying the parent of the node and the conjunction or disjunction of FVPs that define the node. Portions of the Arabic MFH are shown in Figures 2-4.

Transformational Rules. A rule attached to each leaf node of the MFH effects the desired morphological transformations for that node. A rule consists of one or more mutually exclusive clauses. The 'if' part of a clause is a regular expression pattern, which is matched against the value of the feature ROOT (a string). The 'then' part includes one or more operators, applied in the given order. Operators include addition, deletion, and replacement of prefixes, infixes, and suffixes. The output of the transformation is the transformed ROOT string. An example of a rule attached to a node in the MFH is given in Section 3.1 below.

Process Logic. In generation, the MFH acts as a discrimination network. The specified FS is matched against the features defining each subtree until a leaf is reached. At that point, MORPHE first checks in the irregular forms lexicon for an entry indexed by the name of the leaf node (i.e., the MF) and the value of the ROOT feature in the FS. If an irregular form is not found, the transformation rule attached to the leaf node is tried. If no rule is found or none of the clauses of the applicable rule match, MORPHE returns the value of ROOT unchanged.

3 Handling Arabic Verbal Morphology in MORPHE

Figure 2 sketches the basic MFH and the division of the verb subtree into stem changes and prefix/suffix additions.[7] The inflected verb is generated in two steps. MORPHE is first called with the feature CHG set to STEM. The required stem is returned and temporarily substituted for the value of the ROOT feature.

[6] MORPHE is written in Common Lisp and the compiled MFH and transformation rules are themselves a set of Common Lisp functions.

[7] The use of two parts of the same tree for the two problems is a constraint of MORPHE's implementation, which does not permit multiple trees with separate roots.

The second call to MORPHE, with feature CHG set to PSFIX, adds the necessary prefix and/or suffix and returns the fully inflected verb.

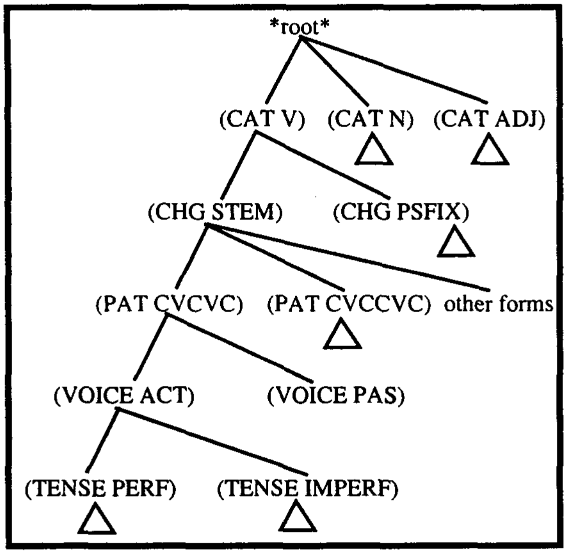

Figure 2 : The Basic Verb Hierarchy

Figure 2 also shows some of the features used to traverse the discrimination tree. The feature PAT is used in conjunction with the ROOT feature to select the appropriate affixes. Knowing the underlying root and its voweling is crucial for the determination of hollow verb stems, as described in Section 1. Knowing the pattern is also important in cases where it is unclear. For example, verbs of pattern CtVCVC insert a 't' after the first radical (e.g. ntaqal 'to move, change location', intransitive). With some consonants as first radicals, in order to facilitate pronunciation, the 't' undergoes a process of assimilation whose effects differ depending on the preceding consonant. For example, the pattern CtVCVC verb from zaHam 'to shove' instead of *ztaHam is zdaHam 'to team'. It is also difficult to determine from just the string ntaqal whether this is pattern nCVCVC of the verb *taqal (if it existed) or pattern CtVCVC of naqal 'to transport, move', transitive).

3.1 Handling Strong and Hollow Verb Morphology in MORPHE

As a demonstration of our approach, we discuss the case of hollow verbs, whose characteristics were described in Section 1. Figure 3 shows the MFH for strong and hollow verbs of pattern CVCVC in the perfect tense, active voice. We use the feature VOW to carry information about the voweling of the verb in the imperfect (discussed below) and overload it to distinguish hollow and other kinds of verbs.

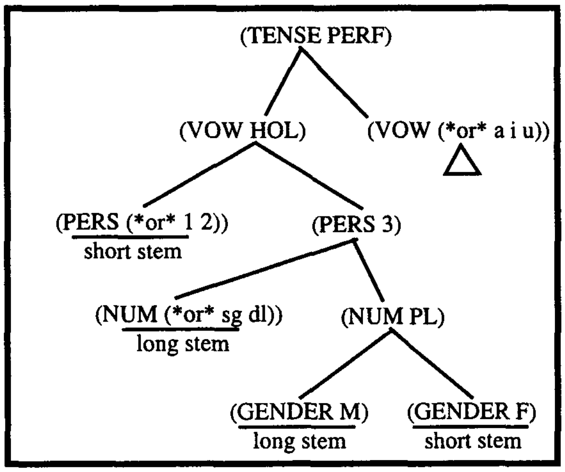

Figure 3: The Perfect Stem Change Subtree for Strong and Hollow Verbs of Pattern CVCVC

In the perfect active voice, regular strong verbs do not undergo any stem changes, but doubled radical verbs do. Rules effecting these changes are attached to the node labeled with the FVP (VOW (*or* a i u)).[8] The hollow verbs, on the other hand, use a long stem with a middle 'alif' (e.g. [daam] 'to last') for third person singular and dual (masculine and feminine) and for third person plural masculine. The remaining person-number-gender combinations take a short stem whose voweling depends on the underlying root of the verb, as specified earlier. Transformation rules attached to the leaf nodes perform the conversion of the ROOT feature value to the short and long stem.

Inside the stem change rules, the four different classes of hollow verbs are treated as three separate conditions (classes 2 and 4 can be merged, as described in Section 1) by matching on the middle radical and the adjacent vowels and replacing them with the appropriate vowel.

[8] Hamzated verbs changes are due to interactions with specific suffixes and are best dealt with in the prefixation and suffixation subtree.

An example of such a rule, which changes the perfect stem to a short one for persons 1 and 2 both singular and plural, follows.

```
(morph-rule v-stem-f1-act-perf-12
 ("^%{cons}(awa)%{cons}$"
   (ri *1* "u"))
 ("^%{cons}(a[wy]i)%{cons}$"
   (ri *1* "i"))
 ("^%{cons}(aya)%{cons}$"
   (ri *1* "i")))
```

The syntax `%{var}` is used to indicate variables with a given set of values. Enclosing a string in parenthesis associates it with a numbered register, so the replace infix (ri) operator can access it for substitution.

Figure 4 shows the imperfect subtree for strong and hollow verbs. Strong verbs are treated efficiently by three rules branching on the middle radical vowel, given as the value of VOW. The consonant-vowel pattern of the computed stem is shown (e.g. for kataba 'he wrote', the imperfect stem would be -ktub- in the pattern CCuC). As described in Section 1, the possible vowel in the imperfect is restricted but not always determined by the perfect vowel and so must be stored in the syntactic lexicon.[9] Separating stem changes from the addition of prefixes and suffixes significantly reduces the number of transformation rules that must be written by eliminating much repetition of prefix and suffix addition for different stem changes. For strong verbs of pattern CVCVC, there is at least a three-fold reduction in the number of rules for active voice (recall the different kinds of vowel changes for these verbs from perfect to imperfect described in Section 1). Other patterns and the passive of pattern CVCVC verbs show less variation in stem voweling but more variation in prefix and suffix voweling. Since some of the patterns share the same prefix and suffix voweling, once the stem has been determined, the prefixation and suffixation rules can be shared by pattern groups.

The hollow verb subtree is not as small for the imperfect as it is for the perfect, since the stem depends not only on the mood but also on the person, gender, and number. It is still advantageous to decouple stem changes from prefixation and suffixation. Suffixes differ in the indicative and subjunctive moods; if the two types of changes were merged, the stem transformations would have to be repeated in each of the two moods and for each person-number-gender combination. The same observation applies to stem changes in the passive voice as well. Significant replication of transformational rules that include stem changes makes the system bigger and harder to maintain in case of changes, particularly because each transformational rule needs to take into consideration the four different classes of hollow verbs.

3.2 An Example of Generation

Consider again the example verb form zurtu 'I visited' and the feature structure (FS) given in Section 2. During generation, the feature-value pair (CHG STEM) is added to the FS before the first call to MORPHE. Traversing the MFH shown in Figure 2, MORPHE finds the rule `v-stem-f1-act-perf-12` given in Section 3.1 above. The first clause fires, replacing the 'awa' with 'u' and MORPHE returns the stem -zur-. This stem is substituted as the value of the ROOT feature in the FS and the feature-value pair (CHG PSFIX) is changed to (CHG PSFIX) before the second call to MORPHE. This time MORPHE traverses a different subtree and reaches the rule:

```
(morph-rule v-psfix-perf-1-sg
  (""
     (+s "otu")))
```

This rule, currently simply appends "otu" to the string, and MORPHE returns the string "zurotu", where the 'o' denotes the diacritic "sukuun" or absence of vowel. This is the desired form for zurtu 'I visited'.

[9] In the presence of certain second and third radicals, the middle radical vowel is more precisely determined. This information can be incorporated into the syntactic lexicon as it is being built.

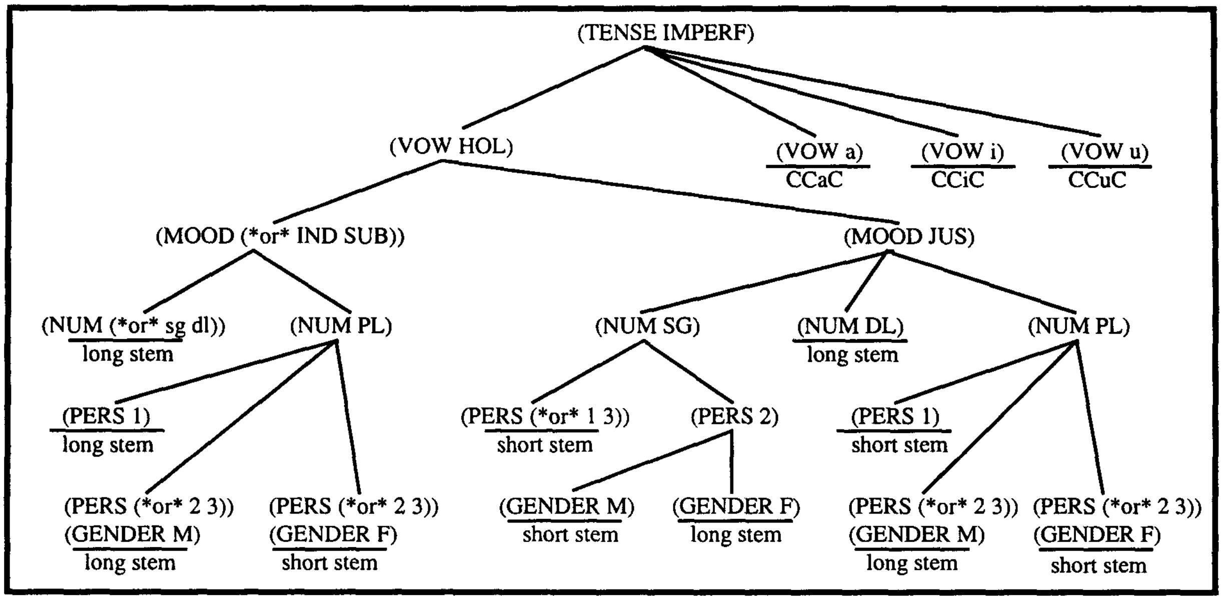

Figure 4: The Imperfect Stem Change Subtree for Strong and Hollow Verbs of Pattern CvCvC

4 Extensions

In this paper so far we have focused on regular and hollow verbs of the pattern CVCVC. Here we examine how our approach applies to other verb types and other parts of speech.

4.1 Extending the Approach to Other Verb Types

The two-step treatment of verbal inflection described in this paper is easily extended to the passive, to doubled radical and hamzated verbs, and to different patterns of strong and hollow verbs. In fact, since not all higher patterns are affected by the presence of a middle or weak radical (e.g. patterns CVCCV, CaaCVC, taCVCCVC and others), the subtrees for these patterns will be significantly less bushy than for pattern CVCVC. The two-step treatment also covers verbs with a weak first radical, especially the radical 'w', which is normally dropped in the active imperfect (e.g. perfect stem warad 'to come', imperfect stem -rid-).[10] Alternatively, it can be placed in the irregular lexicon, which MORPHE consults when it reaches a leaf node, prior to applying any of the transformational rules.

Verbs with a weak third radical, including doubly or trebly weak verbs, are the most problematic since the stem changes interact heavily with the inflectional suffixes, and less is gained by trying to modify the stem separately. We are currently investigating this issue and the best way to treat it in MORPHE.

4.2 Extending the Approach to Other Parts of Speech

The two-step approach to generating verbal morphology also presents advantages for the inflectional morphology of nouns and adjectives. In Arabic, the plural of many nouns, especially masculine nouns, is not formed regularly by suffixation. Instead, the stem itself undergoes changes according to a complex set of patterns (e.g. rajul 'man' pluralizes as rijaal 'men'), giving rise to so-called "broken plurals". The inflection of broken plurals according to case (nominative, genitive, accusative) and definiteness, however, is basically the same as the inflection

[10] Exceptions to this rule exist (e.g. the verb wajil 'to be afraid'), with imperfect stem – wjal-) but are rare and can be handled in MORPHE by placing the irregular stem in the syntactic lexicon and checking for it prior to calling MORPHE for stem changes.

The radical 'y' is largely not dropped or changed.

of most masculine or feminine singular nouns. The same holds true for adjectives.

Finally we note that our two-step approach can also be used to combine derivational and inflectional morphology for nouns and adjectives. Deverbal nouns and present and past participles can be derived regularly from each verb pattern (with the exception of deverbal nouns from pattern CVCVC). Relational or "nisba" adjectives are derived, with small variations, from nouns. Since these parts of speech are inflected as normal nouns and adjectives, we can perform derivational and inflectional morphology in two calls to MORPHE, much as we do stem change and prefix/suffix addition.

Conclusion

We have presented a computational model that handles Arabic morphology generation concatenatively by separating the infixation changes undergone by an Arabic stem from the processes of prefixation and suffixation. Our approach was motivated by practical concerns. We sought to make efficient use of a morphological generation tool that is part of our standard environment for developing machine translation systems. The two-step approach significantly reduces the number of morphological transformation rules that must be written, allowing the Arabic generator to be smaller, simpler, and easier to maintain.

The current implementation has been tested on a subset of verbal morphology including hollow verbs and various types of strong verbs. We are currently working on the other kinds of weak verbs: defective and assimilated verbs. Other categories of words can be handled in a similar manner, and we will turn our attention to them next.

References

K. Beesley. 1990. Finite-State Description of Arabic Morphology. In *Proceedings of the Second Cambridge Conference: Bilingual Computing in Arabic and English.*

K. Beesley. 1991. Computer Analysis of Arabic: A Two-Level Approach with Detours. In B. Comrie and M. Eid, editors, *Perspectives on Arabic Linguistics III: Papers from the Third Annual Symposium on Arabic Linguistics*. Benjamins, Amsterdam, pages 155-172.

K. Beesley. 1996. Arabic Finite-State Morphological Analysis and Generation. In *Proceedings COLING'96*, Vol. 1, pages 89-94.

K. Beesley. 1998. Consonant Spreading in Arabic Stems. In *Proceedings of COLING'98.*

D. Cowan. 1964. *An introduction to modern literary Arabic.* Cambridge University Press, Cambridge.

G. Hudson. 1986. Arabic Root and Pattern Morphology without Tiers. *Journal of Linguistics*, 22:85-122.

G. Kiraz. 1994. Multi-tape Two-level Morphology: A Case study in Semitic Non-Linear Morphology. In *Proceedings of COLING-94*, Vol. 1, pages 180-186.

K. Koskenniemi. 1983. *Two-level morphology: A General Computational Model for Word-Form Recognition and Production.* PhD thesis, University of Helsinki.

A. Lavie, A. Itai, U. Ornan, and M. Rimon. 1988. On the Applicability of Two Level Morphology to the Inflection of Hebrew Verbs. In *Proceedings of the Association of Literary and Linguistic Computing Conference.*

J.R. Leavitt. 1994. MORPHE: A Morphological Rule Compiler. Technical Report, CMU-CMT-94-MEMO.

J. McCarthy and A. Prince. 1990. Foot and Word in Prosodic Morphology: The Arabic Broken Plural. *Natural Language and Linguistics Theory*, 8: 209-283.

J. McCarthy and A. Prince. 1993. *Template in Prosodic Morphology.* In Stvan, L. et al., editors, *Papers from the Third Annual Formal Linguistics Society of Midamerica Conference,.* Bloomington, Indiana. Indiana University Linguistics Club, pages 187-218.

G. Ritchie. 1992. Languages Generated by Two-Level Morphological Rules. *Computational Linguistics*, 18(1), pages 41-59.

R. Sproat. 1992. *Morphology and Computation.* MIT Press, Cambridge, Mass.

H. Wehr. 1971. *A Dictionary of Modern Written Arabic*, J.M. Cowan, editor. Spoken Language Services, Ithaca, NY, fourth edition.

W. Wright. 1988. *A Grammar of the Arabic Language.* Cambridge University Press, Cambridge, third edition.