Exploiting Hidden Morphophonemic Constraints for Finding the Underlying Forms of 'weak' Arabic Verbs

Allan Ramsay
School of Computer Science
University of Manchester
Manchester M13 9PL, UK
Allan.Ramsay@manchester.ac.uk

Hanady Mansour
Department of Arabic,
College of Arts and Sciences
Qatar University, State of Qatar
hanadyma@qu.edu.qa

Abstract

We present a treatment of Arabic morphology which allows us to deal with 'weak' verbs by paying attention to the underlying phonological process. This provides us with a very clean way of thinking about such verbs, and also makes maintenance of the lexicon very straightforward.

1 Introduction

It is well known that Arabic morphology is complex: the language uses a combination of concatenative and discontinuous processes, and the effects of these are obscured by the fact that many phonologically significant items (short vowels, geminations) are not written in modern Arabic.

We present a treatment of Arabic morphology which covers the standard cases, but which has two significant advantages. (i) We delay making decisions about the underlying form until we have the information that is necessary for getting the decision right. Unlike most attempts at diacriticisation, we do not enumerate all the possible forms and then try to choose between them. Instead we leave decisions on specific diacritics until we are in a position to get them right–*e.g.* we delay choosing between declarative and interrogative present tense prefixes for a verb until we know whether it is being used in a statement or a question. This enables us to weave morphological and syntactic processing together very efficiently, as described below. (ii) We can take account of the phonological processes that produce the varying forms of 'weak' verbs without having to declare these verbs as belonging to a special class. Weak verbs are in fact regular verbs whose spelling reflects a small set of phonological contractions. Our analysis allows us to obtain 'underlying forms' for the surface forms of weak verbs which show how they are related to their roots.

2 Basic Mechanisms

The basic problems of Arabic morphology are well known. A single word may have numerous forms, marking various syntactic features, where a form may have a combination of prefixes and affixes and the vowels at the heart of the word may vary. Thus كَتَبَ **(kataba)** . . . are all forms of a single verb, with a variety of prefixes and suffixes marking such things as tense, agreement and mode, and with each form involving different vowels between the consonants كْ?تْ?بْ **(k?t?b)**. The situation is made worse by the fact that the short vowels, and a number of other significant items, are not generally written. This means that the full forms كَتَبَ **(kataba)**, كُتِبَ **(kutiba)**, . . . are all written as كتب (*ktb*). To make things even worse, Arabic generally forms families of words around a single root. These are sometimes marked by derivational prefixes, but in many cases there is no visible prefix of this kind, so that the written form كتب (*ktb*) also corresponds to a plural noun (كُتُبْ **(kutub)**) and to two forms of two different verbs, كَتَبَ **(kataba)** and كُتِبَ **(kutiba)** (active and passive of '*to write*') كَتَّبَ **(kattaba)** and كُتِّبَ **(kuttiba)** (active and passive of '*to make write*'). Thus we have three sets of interlinked problems: different forms of the same word may be written quite differently, different forms of the same word may be written the same but have different underlying sets of vowels, and different words may be written the same (and may or may not have different underlying sets of vowels).

We follow fairly standard practice by describing a word in terms of a template and a set of fillers (*e.g.* (McCarthy and Prince, 1990)); we use a categorial description of the way roots and affixes combine (Bauer, 1983); in order to improve the efficiency of the process of lexical lookup, we store the lexicon as a lexical trie; and then we add a set of spelling rules to account for the variations in surface forms that are observed under various

Proceedings of Recent Advances in Natural Language Processing, pages 448–454,
Hissar, Bulgaria, 12-14 September 2011.

conditions.

2.1 Characters

We represent the graphemes that make up a word as bundles of information that we will refer to as 'characters'. A character has a number of properties. It has (usually) a written form; it has an 'underlying' form, which might be a diacritic mark (and hence unwritten in normal text) and which can be used to derive the phonetic transcription; it can be classified as being a consonant or a vowel, and in the latter case it can be either long or short; and it has various other features, which we will introduce as they become relevant. Thus the semivowel وْ (**w**) is represented as in Fig. 1 (note that this item marked as being both a consonant and a vowel, since it has the properties of both).

*character(char(*وْ **(w)**),
underlying("w"),
vc(+*vowel*, +*consonant*, +*long*))

Figure 1: The character وْ **(w)**

To save space we will sometimes simply write a character like the one in Fig. 1 as #وْ **(w)**, but whenever you see something of this form you should try to remember that it is just a shorthand for a complex object of the kind shown in Fig. 1.

2.2 Templatic Specification of Lexical Items

In order to know what forms a word may take, you need to know three things: what are the consonants in the root, what are the vowels that fill the gaps between those consonants under different conditions, and are the consonants geminated?

We therefore represent a root by providing a template, as in Fig. 2.

history(diacritics(choices(actvPres(["*o*","*u*"]),
actvPast(["*a*","*a*"]),
psvPast(["*u*","*i*"]),
psvPres(["*o*","*a*"]))),
actualVowels(*A*)))
consonants(*targetConsonants*(*B*),
actualConsonants(*B*)))

Figure 2: Template for one sense of كْتُبْ **(k?t?b)**

This template specifies the vowels that are to be used for filling the gaps in the root for different tense/voice combinations. The slot for the 'actual vowels' will be bound to one of the options, once the tense and voice are actually known. The template further specified that the underlying consonants are the same as the ones that appear in the written form–we will see examples where this is not so below.

2.3 Categorial Treatment of Inflectional Morphology

In addition to describing how the vowels and consonants of the root change in the underlying form depending on the tense and mood (for verbs) and the number and gender (for nouns), we have to specify the patterns of affixes that a given root takes. We do this using a categorial description of the affixes that a given item requires in order to complete itself. We make two assumptions: (i) we assume that an open-class word will typically be obtained from an underlying root via a derivational suffix. Thus we assume that استكْتُبْ (*āstk?t?b*), مكْتُبْ (*mk?t?b*), اكْتُبْ (*āk?t?b*) and so on are all obtained by adding a derivational prefix to the root كْتُبْ (*k?t?b*). For consistency we further assume that forms with no visible derivational affix are nonetheless obtained from the root by adding an empty prefix. (ii) We assume that each individual affix specifies what further affixes are required, using the extended categorial rules in Fig. 3 to process words strictly from right to left, as proposed by (Ades and Steedman, 1982) for handling syntactic relations. Allowing each affix to specify what else is required allows roots to require variable numbers of affixes, *e.g.* the derivational affix است (*āst*) which obtains a verb from كْتُبْ (*k?t?b*) starts a different chain of affixes from the prefix م (*m*) which obtains a noun from this root. This provides a more flexible approach to describing the structure of a word than using a context-free grammar, as suggested by (Kiraz, 2001).

$$G / H \Longrightarrow G / I,\ I / H$$
$$G \setminus H \Longrightarrow G / I,\ I \setminus H$$

Figure 3: Combinatory categorial rules

Consider the written form يستكتب (*ystktb*). This has two possible readings, as an active transitive verb or as the passive form of that verb. In both cases it is made out of a number of pieces, as shown in Fig. 4.

يَ+اِيسْتَ+كْتِبْ+٠+؟ **(ya+īsta+kotib+0+?)**,
يُ+اُوسْتُ+كْتَبْ+٠+؟ **(yu+ūstu+kotab+0+?)**

Figure 4: Possible structure for يستكتب (*ystktb*)

Fig. 4 shows that يستكتب (*ystktb*) is actually made out of five pieces–the root كْتُبْ (*k?t?b*), the derivational prefix ايست (*īst*), a tense circumfix consisting of the prefix ي؟ (*y?*) and an empty suffix, and an agreement marker whose form cannot be decided out of context. The slot fillers in the root and the tense prefix vary depending on

whether the verb is active or passive, with the phonological consequence that the prefix would be pronounced '*yaasta*' in the active and '*yuustu*' in the passive. The underlying form of the agreement marker cannot in fact be determined until the mood of the verb is known–it will be ـُ (**u**) if the verb is used in a statement or a question, ـَ (**a**) if it is being used in a context where a subjunctive is required, and ـْ () if a jussive is intended.

2.4 The Lexicon as a Trie

We store the lexicon as a trie. This is a well-known technique for managing dictionaries, since it facilitates lexical lookup. The only slightly odd thing about the trie we use is that it contains arcs with unknown items, to mark the fact that roots have holes in them which can be filled in a variety of ways. In particular, a single hole may be filled by either an (unwritten) short vowel or a (written) long vowel, depending on fine-grained syntactic factors. This makes the normal processing of traversing the trie more complex, but is unavoidable: how we deal with this is discussed in Section 2.5.

2.5 Spelling Rules

In most languages, phonological processes and other quirks of the writing system mean that there are a range of 'boundary effects' where elements of a word are joined together. The prefix *im-* on the English words '*impossible*' and '*imperfect*', for example, is a variant on the negation prefix *in-* that appears on '*incorrect*' and '*indecisive*' which arises because it is easier to get from saying '*m*' to '*p*' (because they both involve closing your lips) than to get from '*n*' to '*p*'.

Phenomena of this kind are generally dealt with by specifying 'spelling rules', often in the form of finite-state automata of some kind. We will write such rules using the format $/L/\boxed{P}/R/\Longrightarrow Q$, meaning that if P occurs in a context where it is preceded by L and followed by R then it should be replaced by Q, as suggested by (Chomsky and Halle, 1968). We will use *c0, c1, ...* to denote arbitrary consonants, *v0, v1, ...* to denote vowels and *x0, x1, ...* to denote arbitrary consonants, and we will add specific features by including them in square brackets $[\ldots]$. If the context is unimportant then we will write /???/. It is important to note that we are using these rules in the reverse of the standard direction: morphophonemic rules are usually used to describe what the surface form would be given a particular set of constituents. We are using these rules to recover the underlying form from the surface forms. This should be borne in mind when reading the rules below.

There are numerous such cases in Arabic: we will illustrate the form of our rules by considering the feminine agreement marker, which is pronounced differently depending on whether it is the last element of the word to which it is attached.

We assume that the canonical form of this item is the one that appears at the end of a word, which is pronounced '*ha*', and is written as ة (*t*). We then have a spelling rule that says that if you see a ت (*t*) in the middle of a word, it might actually be this item, having undergone a change in the way it is written to reflect the fact that it is easier to say '*ta*' than '*ha*' in the middle of a word. The rule in Fig. 5 says that if the written form of a word contains a consonant $\#c0$, where this consonant is not a slot filler ($-$*query*), and this is followed by an ordinary ت (*t*) and another character $\#x0$, then maybe the underlying item was the feminine marker ة (*t*), which has been replaced by ت (*t*) to reflect the change in pronunciation of this item when it appears in the middle of a word (note that this character carries the marker $+taa$ to indicate that it is not just the normal character #ت (*t*)). Thus application of this rule to the word دارستان (*dārstān*) produces the underlying form دَارِسْةَانْ (**dārist**ةَ**ān**)[1]

/c0:$[-$*query*$]/\boxed{\#$ تْ **(t)**$}/$*x0/* $\Longrightarrow$ #ةْ **(t):**$[+$*taa*$]$

Figure 5: Rule for tamarbuta replacement

Application of spelling rules is interwoven with the search through the lexical trie. You cannot search the trie effectively without being aware of the potential application of these rules, but it is unrealistic either to apply the rules to lexical entries before constructing the tree (since this would lead to an explosion in the size of the trie) or to apply them blindly to the surface string (because this would again lead to the construction of an exponentially large number of forms, many of which have no correspondents in the lexicon). Our strategy is to apply rules as they become relevant during traversal of the trie. That way we do not apply rules to strings that have no counterparts in the trie, but we do apply them as soon as their effect would lead to exploration of a branch. The left-hand con-

[1]This might look a little odd, since it has the word final version of the feminine marker appearing in the middle of the word, but that's the whole point of this rule: the item in question *is* the feminine marker, but because it is in a word-internal position it has undergone a phonological change.

text of the antecedent of a rule is thus the route that has been followed so far through the trie, the right-hand context is the currently unconsumed portion of the input string, and the consequent of the rule is the branch of the trie that is to be explored.

We also use rules of this kind to insert the 'fillers' into the templatic descriptions of roots. As noted above, we include the gaps between the consonants in a root as arcs in the trie. These gaps can in general be filled either by short (unwritten) vowels or by long (written) ones. In any particular form of a given root, the way that they are to be filled is determined, but since you do not know which form you have until you have looked the word up, and possibly not until you have examined its syntactic role, you have to allow for all possible ways of traversing these arcs. To do this we make use of the two rules in Fig. 6.

/*c1*:[−*query*]/[*c2*:[−*query*, −*taa*]] /???/
⟹*?*: [+*vowel*, −*long*, +*query*, +*inserted*]

/*c1*:[−*query*]/[*v1*:[+*long*, −*query*]:*B*] /???/
⟹*?*:[+*long*, +*inserted*, −*multiple*]:*C*
if [*underlying*]@*B*↔ [*underlying*]@*C*

Figure 6: Rules for slot-fillers

The first of these rules says that if you've just traversed a consonant *c1*, and the next character is another consonant *c2*, where neither *c1* nor *c2* is itself a query or the tarmabuta, then you might try inserting an unspecified short vowel, i.e. an item whose surface form is *?*, so that it can be used to traverse an arc for a slot filler. The second rule in Fig. 6 says that if you have just traversed a consonant *c1*, and the next character is a long vowel *B*, then you can try replacing the long vowel by a *?* which is marked as being long, and which shares the same *underlying* character as the original long vowel.

Between them these rules allow us to account for the slot-and-filler structure of Arabic nouns and verbs, since we simply introduce *?*s at appropriate points, marking them as corresponding to short or long vowels appropriately, and in the case of long vowels remember what the actual underlying form of the long vowel was.

There are a number of other spelling rules, which can be used to account for a range of phenomena from fairly trivial things (such as the fact that the hamza can be omitted on word-initial characters) to more interesting cases such as the deletion of the second occurrence of a repeated consonant after a sukun (*e.g.* obtaining the underlying form جَدُدْ (**ğadod**) from the written form جد (*ğd*)).

2.6 Delayed Decisions about Underlying Forms

In general, just looking at a word will not tell you what the short vowels in its underlying form are. Consider, for instance, the word يدرس (*ydrs*). This has a number of interpretations, as the active and passive forms of verbs meaning 'study' and 'teach', but even if we consider just one of these, say the active form of the version meaning 'study', we see that there are a number of possibilities. In particular, it could occur in a context requiring an indicative form, *e.g.* as the main verb of a declarative sentence, or in one requiring a subjunctive form (*e.g.* after certain complementisers and modifiers), or one requiring a jussive form. The final agreement marker takes different forms in the different kinds of context, as shown in Fig. 7.

(1) a. يدرس الولد الدرس. (*ydrs ālwld āldrs.*)
b. نل يدرس الولد الدرس. (*nl ydrs ālwld āldrs.*)

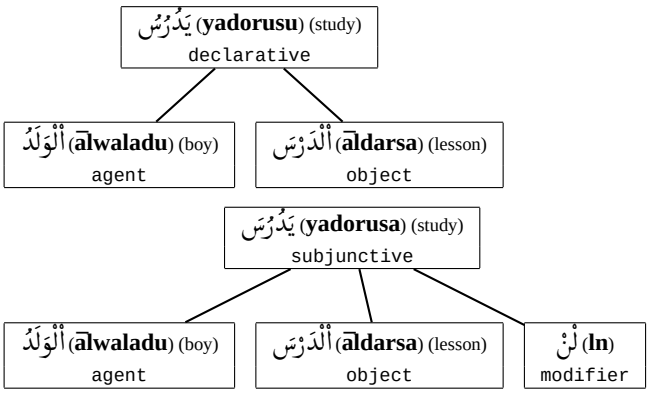

Figure 7: indicative/subjunctive forms of يدرس (*ydrs*)

Generating both these versions as soon as you saw the written form would be a major problem for any system that was going to attempt to parse the input text, since it would double the number of analyses that needed to be explored.

There are plenty of similar instances. Even in Fig. 7, for instance, the nouns الولد (*ālwld*) and الدرس (*āldrs*) have different case markers (اْلْوَلَدُ (**ā-lwalad-u**) and اْلْدَرْسَ (**āldars-a**)), because الولد (*ālwld*) is the subject and الدرس (*āldrs*) is the object. But these case markers cannot be determined until the syntactic role of these items is known (and, indeed, not until the context in which the verb itself

	past actv	past psv	present actv	present psv
1st pers sing	وَقَفْتُ (**waqaftu**)	وُقِفْتُ (**wuqiftu**)	أَقِفُ (**ʔa'qifu**)	أَوْقَفُ (**ʔawqafu**)
1st pers not sing	وَقَفْنَا (**waqafnā**)	وُقِفْنَا (**wuqifnā**)	نَقِفُ (**naqifu**)	نَوْقَفُ (**nwqafu**)
2nd pers sing masc	وَقَفْتَ (**waqafta**)	وُقِفْتَ (**wuqifta**)	تَقِفُ (**taqifu**)	تَوْقَفُ (**twqafu**)
2nd pers sing fem	وَقَفْتِ (**waqafti**)	وُقِفْتِ (**wuqifti**)	تَقِفْيْنَ (**taqifyna**)	تَوْقَفْيْنَ (**twqafyna**)
2nd pers dual	وَقَفْتُمَا (**waqaftumā**)	وُقِفْتُمَا (**wuqiftumā**)	تَقِفَانِ (**taqifāni**)	تَوْقَفَانِ (**twqafāni**)
2nd pers plural masc	وَقَفْتُمْ (**waqaftum**)	وُقِفْتُمْ (**wuqiftum**)	تَقِفُوْنَ (**taqifwna**)	تَوْقَفُوْنَ (**twqafwna**)
2nd pers plural fem	وَقَفْتُنَّ (**waqaftuna**)	وُقِفْتُنَّ (**wuqiftuna**)	تَقِفْنَ (**taqifna**)	تَوْقَفْنَ (**twqafna**)
3rd pers sing masc	وَقَفَ (**waqafa**)	وُقِفَ (**wuqifa**)	يَقِفُ (**yaqifu**)	يَوْقَفُ (**ywqafu**)
3rd pers sing fem	وَقَفَتْ (**waqafat**)	وُقِفَتْ (**wuqifat**)	تَقِفُ (**taqifu**)	تَوْقَفُ (**twqafu**)
3rd pers dual masc	وَقَفَا (**waqafā**)	وُقِفَا (**wuqifā**)	يَقِفَانِ (**yaqifāni**)	يَوْقَفَانِ (**ywqafāni**)
3rd pers dual fem	وَقَفَتَا (**waqafatā**)	وُقِفَتَا (**wuqifatā**)	تَقِفَانِ (**taqifāni**)	تَوْقَفَانِ (**twqafāni**)
3rd pers plural masc	وَقَفُوَا (**waqafwā**)	وُقِفُوَا (**wuqifwā**)	يَقِفُوْنَ (**yaqifwna**)	يَوْقَفُوْنَ (**ywqafwna**)
3rd pers plural fem	وَقَفْنَ (**waqafna**)	وُقِفْنَ (**wuqifna**)	يَقِفْنَ (**yaqifna**)	يَوْقَفْنَ (**ywqafna**)

Figure 8: Full conjugation for وقف (*wqf*) (attested by (Khwask, 1992; El-Dahdah, 1991))

appears is known, because in some contexts subjects are marked as being accusative). Again, generating all the possibilities at the point when you look the word up will multiply the options that a parser would have to explore: if we had generated the nominative, accusative and genetive forms of the two nouns, and the indicative, subjunctive and jussive forms of the verb, when we looked up the words in (1)(a) then we would have had to look at potentially 27 times as many possibilities.

To cope with this, we use 'just-in-time constraints' (similar to Hewitt (1971)'s 'if-added demons', or to 'watched literals' in theorem proving (Moskewicz et al., 2001)) to leave an unspecified item in the underlying form, to be filled in when the required information becomes available. Thus the forms that are produced when we first look up the words يدرس (*ydrs*), الولد (*ālwld*) and الدرس (*āldrs*) are يَدْرُسْ؟ (**yadrus?**), اْلْوَلَدْ؟ (**ālwalad?**) and اْلْدَرْسْ؟ (**āldars?**), where the ?s indicate that there is some element of the word which is not yet known, because the contextual information that would fix it is not yet available.

3 Weak Verbs

So far so good. We can produce fine-grained diacriticisations using a combination of slot-and-filler templates, a lexical trie and a set of spelling rules, and we can delay decisions about the underlying form until relevant syntactic information turns up. We now turn to the question of 'weak' verbs.

These are verbs whose root contains a semivowel (usually و (*w*), ي (*y*) or ا (*ā*)), which sometimes appears in the written form and sometimes goes missing or changes its form. These words do not appear to fit the normal slot-and-filler pattern, since the set of consonants in the written form appears to vary, so it does not look as though you can set a single template and fill in the slots. A typical example is the verb وقف (*wqf*), whose conjugation is given in Fig. 8.

The awkward thing about Fig. 8 is that most of the table looks as though it corresponds to a verb whose root is وقف (*wqf*), but in the column for the active present tense the initial و (*w*) is missing.

Why is it missing here and nowhere else? Crucially, why is it missing in the column for the present active but not the column for the present passive?

The only differences between the active and passive are that the *underlying* forms of the prefixes are different–the active prefix is يَ (**ya-**), the passive one is يُ (**yu-**)–and that the diacritics that fill in the slots may be different.

It is hard to see what the diacritics for the active present of وقف (*wqf*) would be. Because the initial consonant has disappeared there is no obvious trace of a vowel following the position where it would have been, but it seems reasonable to assume that it is a ـَ (**a**), since the cases where we can see the diacritics seem fairly regular, and a ـَ (**a**) for the first diacritic in the active present is common for regular verbs. It therefore looks as though the initial وْ (**w**) disappears if it is preceded by a ـَ (**a**) in the underlying form. There is, of course, no trace of this in the written form, and there is indeed no trace of it in the phonetic form, but the underlying process is that the awkwardness of pronouncing اَوَ (**awa**) has led to the deletion of the وَ (**wa**)

We therefore introduce a spelling rule which says that if you have just traversed a consonant and an (unwritten) ـَ (**a**), and the next item is a consonant, you should consider the possibility that a وْ (**w**) has been deleted from the surface form. This

	past actv	past psv	present actv	present psv
1st pers sing	شَكَوْتُ (**šakawtu**)	شُكِيتُ (**šukiytu**)	أَشْكُو (**ʔa’škuw**)	أَشْكَى (**ʔa’škaā**)
1st pers not sing	شَكَوْنَا (**šakawnā**)	شُكِينَا (**šukiynā**)	نَشْكُو (**naškuw**)	نُشْكَى (**nuškaā**)
2nd pers sing masc	شَكَوْتَ (**šakawta**)	شُكِيتَ (**šukiyta**)	تَشْكُو (**taškuw**)	تُشْكَى (**tuškaā**)
2nd pers sing fem	شَكَوْتِ (**šakawti**)	شُكِيتِ (**šukiyti**)	تَشْكِينَ (**taškiyna**)	تُشْكَيْنَ (**tuškayna**)
2nd pers dual	شَكَوْتُمَا (**šakawtumā**)	شُكِيتُمَا (**šukiytumā**)	تَشْكُوانِ (**taškuwāni**)	تُشْكَيَانِ (**tuškayāni**)
2nd pers plural masc	شَكَوْتُمْ (**šakawtum**)	شُكِيتُمْ (**šukiytum**)	تَشْكُونَ (**taškuwna**)	تُشْكَيْنَ (**tuškayna**)
2nd pers plural fem	شَكَوْتُنَّ (**šakawtuna**)	شُكِيتُنَّ (**šukiytuna**)	تَشْكُونَ (**taškuwna**)	تُشْكَيْنَ (**tuškayna**)
3rd pers sing masc	شَكَا (**šakā**)	شُكِيَ (**šukiya**)	يَشْكُو (**yaškuw**)	يُشْكَى (**yuškaā**)
3rd pers sing fem	شَكَتْ (**šakat**)	شُكِيَتْ (**šukiyat**)	تَشْكُو (**taškuw**)	تُشْكَى (**tuškaā**)
3rd pers dual masc	شَكَوَا (**šakawā**)	شُكِيَا (**šukiyā**)	يَشْكُوانِ (**yaškuwāni**)	يُشْكَيَانِ (**yuškayāni**)
3rd pers dual fem	شَكَتَا (**šakatā**)	شُكِيَتَا (**šukiyatā**)	تَشْكُوانِ (**taškuwāni**)	تُشْكَيَانِ (**tuškayāni**)
3rd pers plural masc	شَكَوَا (**šakawā**)	شُكُوا (**šukuwā**)	يَشْكُونَ (**yaškuwna**)	يُشْكَوْنَ (**yuškawna**)
3rd pers plural fem	شَكَوْنَ (**šakawna**)	شُكِينَ (**šukiyna**)	يَشْكُونَ (**yaškuwna**)	يُشْكَيْنَ (**yuškayna**)

Figure 9: Conjugation for شكو (*škw*) (attested by (Khwask, 1992; El-Dahdah, 1991))

rule only applies if the وْ (**w**) is also followed by an unwritten ـَ (**a**), so we will insert this as well. Note that the item being rewritten here is in fact the empty string: this rule just inserts وْ (**w**), ـَ (**a**) between *c0*, ـَ (**a**) and *c1*.

/c0, #ـَ (**a**)/Ø/*c1*/$\Longrightarrow$ #وْ (**w**), #ـَ (**a**)

Figure 10: Spelling rule for missing وْ (**w**)

Fig. 10 says that if you have just traversed arcs corresponding to a consonant *c0* and a gap which was filled by an unwritten vowel whose underlying form was ـَ (**a**), and the next character to be scanned is another consonant *c1*, then you could try inserting a وْ (**w**) and a following gap-filler which also has underlying form ـَ (**a**).

This rule allows us to spot that the surface form تقف (*tqf*) corresponds to an underlying diacriticisation تَوَقُّفَ (**tawaqufa**), but not to the passive form تُوُقِفَ (**tuwuqifa**) because in the latter case the underlying vowel in the prefix was ـُ (**u**), which does not trigger the rule.

Now consider the conjugation of شكو (*škw*), as shown in Fig. 9. Much of this can be accounted for by assuming that the diacritics for the four tense/mood combinations for this verb are *actvPast=["a", "a"]*, *psvPast=["u", "i"]*, *actvPres=["o", "u"]*, *psvPres=["o", "a"]*. The past active column is accounted for by the rule in Fig. 10: the third singular masculine and feminine and third dual feminine forms have suffixes which begin with ـَ (**a**) added to them, so Fig. 10 deletes the final اَوْ (**aw**) from the root to produce شكا (*škā*), شَكْتْ (**škt**) and شَكْوَا (**škwā**) as the surface forms. The other cases are slightly more complex. Most of the present active column leaves the end of the root unchanged as اُو (**uw**), most of the passive past column is produced by a rule of the form /???/Ø/#ـِ (**i**)#يْ (**y**)/ $\Longrightarrow$ #ـُ (**u**)#وْ (**w**) and most of the passive present is produced by /#ـَ (**a**)/#ىٰ (**ā**)/v0[-long]/$\Longrightarrow$ #وْ (**w**) and /#ـَ (**a**),#يْ (**y**)/Ø/???/$\Longrightarrow$#وْ (**w**).

In each of these columns, however, there are cases that do not fit the main pattern. Why, for instance, is the 2nd person singular feminine active present tense تَشْكِينَ (**taškiyna**) when every other entry in this column has شْكُو (**škuw**) as its root? Inspection of the components of this item show that it is made up of تُ+شْكُو+ِينَ (**tu+škuw+iyna**). But in that case the rule /???/Ø/#ـِ (**i**)#يْ (**y**)/ $\Longrightarrow$ #ـُ (**u**)#وْ (**w**) that we introduced to cover the past passive forms applies here also, producing the observed form. Similarly, the presence of يُشْكَوْنَ (**yuškawna**) as the 3rd person plural masculine passive present appears odd in the passive present column, where most of the time شْكَوْ (**škaw**) has been turned to شْكِي (**šky**); but again consideration of the components of يُشْكَوْنَ (**yuškawna**) as يُ+شْكَوْ+وْنَ (**yu+škaw+wna**) shows that the relevant rule here is /#وْ (**w**)/Ø/???/ $\Longrightarrow$ #وْ (**w**) (i.e. the وْ (**w**) at the end of the stem is deleted in the surface form) rather than /#ـَ (**a**),#يْ (**y**)/Ø/???/$\Longrightarrow$#وْ (**w**).

Thus the vast majority of the cases in Fig. 9 arise very straightforwardly by applying spelling rules which reflect simple phonological processes. If you look only at the surface forms, these rules are hard to spot, but looking at the full underlying forms they become much more apparent. The case of the 3rd person masculine plural passive past, however, requires a little more attention. The basic building blocks here are شُكِوْ+وَا (**šukiw+wā**).

Applying produces شُكِيوَا (**šukiy+wā**). But then the sequence اِيوْ (**iy+w**) is itself awkward, so a subsequent rule /???/$\boxed{\text{\#ـِ (i), \#يْ (y)}}$/\#وْ (w)/ $\Longrightarrow$ \#ـُ (u) comes into play, leading finally to شُكُوا (**šukuwā**).

We thus have the rules in Fig. 11:

/*c0*, #ـَ (a)/$\boxed{\emptyset}$/*c1*/$\Longrightarrow$#وْ (w), #ـَ (a)
/???/$\boxed{\emptyset}$/#ـِ (i)#يْ (y) / $\Longrightarrow$ #ـُ (u)#وْ (w)
/#ـَ (a)/$\boxed{\text{\#ىٰ (ā)}}$/v0[-long]/$\Longrightarrow$ #وْ (w)
/#ـَ (a)#يْ (y)/$\boxed{\emptyset}$/???/$\Longrightarrow$#وْ (w)
/#وْ (w)/$\boxed{\emptyset}$/???/ $\Longrightarrow$ #وْ (w)
/???/$\boxed{\text{\#ـِ (i), \#يْ (y)}}$/\#وْ (w)/ $\Longrightarrow$ #ـُ (u)

Figure 11: Spelling rules for وْ (**w**)

These rules are phonologically plausible, in that they all reflect changes in pronunciation that arise from awkward combinations of phonemes. Applying them allows us to reconstruct the underlying forms from the surface forms, without having to put complex descriptions in the lexicon. We can simply say that شكو (*škw*) is a regular verb, with the slot fillers given above, rather than having to list all the forms of the stem and assigning very precise sets of affixes to them, as in for instance the Buckwalter analyser (Buckwalter, 2004).

Lexicons that require multiple specifications for a single item are hard to maintain, since you have to know a great deal about the meanings of the tags that say what affixes will attach to a given item (see (Algihaad and Abdelfatah, 2009) for a similar approach). It is much easier to simply say that شكو (*škw*) is a regular verb that takes *actvPast=["a", "a"], psvPast=["u", "i"], actvPres=["o", "u"], psvPres=["o", "a"]* as its diacritics, and to let the spelling rules look after the surface appearance. Indeed, the Buckwalter analyser misses out a number of the forms in Fig. 9, notably several of the passive forms (and some cases which have both active and passive readings, e.g. يشكون (*yškwn*)). The output of this analyser also relies on the sense tagging (given as the English gloss) to link the different forms of a single word. The morphological analysis of شكت (*škt*), for instance, is given as شْكْ+اَتْ (**šk+at**). The only way to ascertain that this is a form of the same word as the others in Fig. 9 is by noting that they have same English gloss–there is nothing in the structure that makes the link clear.

4 Conclusions

We have shown how using phonologically motivated spelling rules allows us to treat Arabic weak-initial and weak-final verbs in exactly the same way as other verbs, specifying a template and a set of slot fillers for the various tense/mood combinations (the same approach also works for weak-middle verbs, but there was no space to discuss these here). This has two major advantages: it provides a very clear separation between the cause of the apparent irregularity of these verbs and their actual adherence to the usual slot-and-filler pattern of Arabic verbs; and by providing this separation, it makes it easy to maintain the lexicon. Comparison with a small number of examples shows that this approach provides correct analyses for several cases which the Buckwalter analyser misses.

References

Ades, A E and M J Steedman. 1982. On the order of words. *Linguistics and Philosophy*, 4:517–558.

Algihaad, A and A Abdelfatah. 2009. Morphological analyizer for arabic verbs. In *3rd IEEE International Conference on Arabic Language Processing (CITAL'09)*, Rabat, Morocco. IEEE.

Bauer, L. 1983. *English Word Formation*. CUP, Cambridge.

Buckwalter, T. 2004. Buckwalter Arabic morphological analyzer version 2.0. Linguistic Data Consortium.

Chomsky, N and M Halle. 1968. *The sound pattern of English*. MIT Press, Cambridge, Mass.

El-Dahdah, A. 1991. *Dictionary of Arabic verb conjugation*. Liprairie du Liban.

Hewitt, C. 1971. Planner: a language for proving theorems in robots. In *2nd International Joint Conference on Artificial Intelligence*.

Khwask, Z. 1992. *Lessons in Syntax and Morphology*. Dar Almarefah Alghamayah, Alexandria, Egypt.

Kiraz, G. 2001. *Computational Nonlinear Morphology: with emphasis on Semitic languages*. Cambridge University Press, Cambridge.

McCarthy, J and A Prince. 1990. Prosodic morphology and templatic morphology. In Eid, M and J McCarthy, editors, *Perspectives on Arabic linguistics II: papers from the second annual symposium on Arabic linguistics*, pages 1–54, Amsterdam. Benjamins.

Moskewicz, M, C Madigan, Y Zhao, L Zhang, and S Malik. 2001. Chaff: Engineering an efficient SAT solver. In *39th Design Automation Conference*, Las Vegas.