

# Arabic morphology: a categorial approach

Allan Ramsay\* & Hanady Mansur<sup>+</sup>

\*Dept of Computation & <sup>+</sup>Dept of Language Engineering  
UMIST, PO Box 88, Manchester M60 1QD, UK

## Abstract

We provide an account of Arabic morphology from a categorial viewpoint which enables us to reconstruct the diacritics that are missing from written Modern Standard Arabic but which would be essential for any text-to-speech system for MSA. This treatment provides a partial solution to this problem, but there are written forms which cannot be uniquely resolved in this way without reference to the surrounding syntactic and semantic context. The implementation described here is embedded within a larger framework for syntactic and semantic analysis, but the current paper concentrates on purely morphological issues.

## 1 Outline

Written Modern Standard Arabic lacks overt information about short vowels. The written form *'yktbwn'*<sup>1</sup>, for instance, omits the diacritic markers that indicate its pronunciation, where the full written form *'yaktubwan'* with diacritics makes this clear. Arabic speakers, however, have very little difficulty in reconstructing this information. Any Arabic speaker can read an MSA text aloud without any difficulty. How can they do this? Because they can put the vowels back in. The information they use comes in various forms:

<sup>1</sup>We are using Roman equivalents of Arabic letters here, and writing words from left→right, simply for convenience: not much rides on this, though there are some useful distinctions in the written Arabic form that are missing from our transliterations.

- the root *'k\*t\*b\*'* belongs to a group of words that share the same consonants in the same linguistic contexts. The \*'s in this stem mark places where diacritics are required in the full phonologically marked form.
- the tense of the verb, marked partly by its affixes and partly by its syntactic context, constrains the surface form.
- morphotactic rules describe how vowels mutate in specific phonological contexts.

In the approach described below, we provide a categorial account of the structure of complex Arabic words such as verbs and nouns (see [Bauer, 1983] for a general description of categorial morphology, and [Schulze and Ramsay, under review, 2000] for an application of this approach to German). This is used to assign the fine syntactic properties of particular surface forms (e.g. to note that *'kataba'* must be the third singular past tense form, whereas *'kutub'* is a plural nominal form). In most cases the pattern of affixes determines the interpretation, but in situations where the written form is ambiguous between a number of readings the surrounding context will normally lead to a single choice which makes global sense (e.g. the form *'drs'* is locally ambiguous between a noun and a verb; but within the context of a sentence only one of these will generally lead to a well-formed analysis).

We propose a three phase approach to the task of converting short-form MSA words to full-form lexical items with the required diacritics:

1. Compare the sequence of written characters one at a time with the branches of a letter trie (a representation of the lexicon in which words whose initial letters are identical are found down the same branch). This letter trie contains both roots and affixes, so that at the end of the lookup process a sequence of word-parts will have been found.
2. Use the rules of categorial grammar to combine these word parts as appropriate.
3. Apply morphotactic rules to the resulting morpheme sequence to reflect any changes that result when particular sequences of diacritics are found adjacent to one another.

## 2 Categorial Morphology

We use the notation of categorial grammar to describe ‘incomplete’ lexical items. To take a simple English example, we can describe ‘dog’ as being of type  $\{\text{cat}=\text{noun}, \text{affixes}=[\{\text{cat}=\text{noun}, +\text{affix}, +\text{num}\}]\}$  and the affix ‘s’ as being of type  $\{\text{cat}=\text{noun}, +\text{affix}, +\text{num}, \text{dir}=\text{right}\}$ <sup>2</sup>, so that if we decompose ‘dogs’ as ‘dog+s’ then we have a set of objects looking like

```
{form="dog", cat=noun,
  affixes=[{cat=noun,
            +affix,
            +num}]}
+ {form="s", cat=noun, +affix, +num,
  num=pl, dir=right}}.
```

The basic cancellation rule of categorial grammar converts this to  $\{\text{cat}=\text{noun}, \text{num}=\text{pl}, \text{affixes}=[]\}$ . The general form of this rule is

```
{form=F1, cat=X,
  affixes=[{cat=Y, +affix} | A]}
{form=F2, cat=Y, +affix, dir=right}
==>
{form=F1+F2, cat=X, affixes=A}
```

<sup>2</sup>This way of writing it makes the order and direction in which cancellation will take place clearer than the classical noun/num notation, and also allows us to be non-specific about the direction in which to look for a given affix.

```
{form=F2, cat=Y, +affix, dir=left}
{form=F1, cat=X,
  affixes=[{cat=Y, +affix} | A]}
==>
{form=F2+F1, cat=X, affixes=A}
```

where we unify the descriptions of the first affix on the list and the item which is supposed to match it (just as we do when ticking off elements of the subcat list in parsing). For inflectional morphology, X and Y will be identical, whereas for derivational morphology they may not be.

This notation can easily be extended to cover stems which need several affixes, so that a French adjective like ‘blond’ might be described as being of type

```
{form="blond", cat=adj,
  affixes=[{cat=adj, +affix, +gender},
           {cat=adj, +affix, +num}]}
```

If we decomposed ‘blondes’ as ‘blond+e+s’ then we would have

```
{form="blond", cat=adj,
  affixes=[{cat=adj, +affix, +gender},
           {cat=adj, +affix, +num}]}
+{form="e", cat=adj,
  +affix, +gender, gender=f, dir=right}
+{form="s", cat=adj,
  +affix, +num, num=pl, dir=right}
```

Combining the first two elements of this would lead to

```
{form="blond"+"e", cat=adj,
  affixes=[{cat=adj, +affix, +num}],
  gender=f}
+{form="s", cat=adj,
  +affix, +num, num=pl, dir=right}
```

which would in turn reduce to

```
{form=("blond"+"e"+"s", cat=adj,
  affixes=[]},
  gender=f,
  num=pl}
```

This example can, however, be dealt with slightly differently. We could say that ‘blond’, ‘e’ and ‘s’ were as follows:

```
{form="blond", cat=adj,
  affixes=[{cat=adj, +affix, +gender}]}
```

```
{form="e", cat=adj,
+affix, +gender, gender=f, dir=right}
affixes=[{cat=adj, +affix, +num}]}
```

```
{form="s", cat=adj,
+affix, +num, num=pl, dir=right}
```

Here the adjective just requires a gender marker, but the gender marker itself requires a number marker. With this description of the lexical and sublexical items the sequence *'blond+e+s'* would reduce either by combining 'e' and 's' to form something of type {cat=adj, +affix, +gender, gender=f, num=pl, affixes=[]}, which could then be combined with *'blond'* to get a fully inflected adjective, or by a version of the categorial 'cancellation' rule  $X/Z \Rightarrow X/Y, Y/Z$ . We choose to take the latter route, since it enables us to block garden path decompositions early.

This latter approach provides considerable flexibility in cases where the number of affixes required varies, as happens for instance with verbs in languages where finite forms require person and number markers but non-finite ones don't. We exploit this flexibility in Arabic by allowing a written form such as *'drs'*, which can lead to the production of either verbal or nominal forms, to simply require a 'first affix'. The lexical entry for this item looks like

```
{form="drs", cat=X,
affixes=[{cat=X, +affix, +first}]}
```

Consider the following affixes:

```
{form="y", cat=verb, +affix,
+first, dir=left, tns=present,
affixes=[{cat=verb, +affix, +agr}]}
```

and

```
{form="m", cat=noun, +affix,
+first, dir=left,
affixes=[{cat=noun, +affix, +agr}]}
```

If 'y' appears to the left of *'drs'* they will combine to make something of the form

```
{form="y"+"drs", cat=verb, tns=present,
affixes=[{cat=verb, +affix, +agr}]}
```

i.e. a verb that needs an agreement marker. The constraint between the cat of *'drs'* and that of its first affix (expressed by the repetition of the variable X in the lexical entry) marks the whole thing as being a verb. The fact that the affix 'y' says that it is the kind of thing that appears to the left of the item looking for it means that the appropriate version of the categorial rule is invoked (since the affix and the entry on the list of affixes for *'drs'* are unified, and hence the entry for *'drs'* matches the second version of the rule of combination).

If we have *'wn'* as an item of type

```
{form="wn", cat=verb, +affix, +agr,
person=3, num=plural, dir=right}
```

then this will be picked up by *'ydrs'* to make a fully inflected verb<sup>3</sup>:

```
{form=("y"+"drs"+"wn", cat=verb,
tns=present,
person=3, num=plural,
affixes=[]]}
```

If, on the other hand, *'drs'* had been preceded by 'm' then it would have accepted this as its first affix, producing

```
{form="m"+"drs", cat=noun,
affixes=[{cat=noun, +affix, +agr}]}
```

Suppose that we described *'aan'* (which is a nominal +agr suffix) as

```
{form="aan", cat=noun, +affix, +agr,
person=3, num=dual, gender=m,
dir=right}
```

then the final result would be

```
{form=("m"+"drs"+"aan", cat=noun,
person=3, num=dual, gender=m,
```

Note that the lexical entry for *'drs'* did *not* say that it was a noun or a verb. There is a single lexical entry whose category is *unspecified*, but which needs an initial affix whose syntactic properties it is going to share. The prefixes 'y' and 'm' do say that they should attach to a verb and a noun respectively:

<sup>3</sup>*'ydrs'* does not need an overt agreement marker to complete it. We will discuss zero/invisible affixes below, but to keep the presentation clear we will concentrate initially on cases involving overt affixes.

in each case, the resulting item requires an agreement marker, but since ‘*ydrs*’ is a verb it will only accept verbal agreement markers, and likewise ‘*mdrs*’ will only accept nominal agreement markers.

### 3 Diacritic Placement

The categorial approach above will enable us to analyse surface forms into their constituent parts. But that, of course, is only part of the task. Once we have done that, we need to work out what the diacritics would have been if they had been present.

The only place we can record information about the expected diacritics is in the lexicon. The first move, then, is to make sure that the lexical entries are in fact full form sequences. So for affixes like ‘*y*’ and ‘*m*’ we simply note that the underlying form of these items is actually ‘*ya*’ and ‘*mu*’. These items are relatively fixed, and where there are changes they are due to morphotactic processes that happen late on in the analysis, so we can afford to put them into the dictionary as complete (sub-)lexical items.

For roots like ‘*drs*’, however, the diacritics vary. That is, after all, the whole problem. So the best we can do is to put place-holders into the lexical entry marking points where a vowel is required<sup>4</sup>. The lexical entry for ‘*drs*’ is therefore ‘*d\*r\*s\**’.

With these changes, then, the analysis in Section 2 would enable us to treat forms that looked like ‘*mud\*r\*s\*aan*’, which is hardly something that anyone would write.

We assume, in general, that surface forms and underlying strings are linked by morphotactic rules which indicate when a given surface form may correspond to an underlying concatenation of morphemes – rules like

[c0,v0,c1] ==> [c0,e,+,v0,c1]

(this is the one that spots the deleted ‘*e*’ in ‘*changing*’ in English). These rules can be implemented using the technology of 2-level FSTs [Koskiennemi, 1985]. More importantly,

<sup>4</sup>This vowel could be a long vowel, with a graphemic representation, or a short vowel, to be indicated by a diacritic, or even a zero.

they can be interwoven with the process of following a branch in a letter trie so that they do not involve exploration of long spurious paths. We can therefore introduce a rule for Arabic which says

[c0,c1] ==> [c0,v0,c1]

where v0 is an arbitrary vowel. In other words, we will investigate the consequences of inserting any vowel between any pair of consonants in a surface form.

This is not an expensive option, because if a vowel is *not* required at this point we will immediately spot that there is no appropriate route through the letter trie. But if we treat \* as a vowel then we can easily see that ‘*mdrsaan*’ is the surface form for ‘*mud\*r\*s\*aan*’ (we do not insert word breaks using this rule, since it applies to internal vowels as well as to affixes). And as we saw in Section 2, we can analyse this as a masculine dual noun.

So we now know that the short vowel ‘*u*’ was missing in the original surface form, and that there are three other vowels to be inserted. What are they?

The particular configuration of vowels in a nominal form is determined by the nominal prefix. ‘*mu*’ wants all three to be ‘*a*’ (‘*mu-darasataan*’), whereas ‘*ma*’ wants the first to be empty and the second and third to be ‘*a*’ (‘*madrasataan*’). To accommodate this, we extend the entries for ‘*mu*’ and ‘*ma*’ as follows:

```
{form="mu",
  diacritics="a+a+a",
  cat=noun, +affix,
  +first, dir=left,
  affixes=[{cat=noun, +affix, +agr,
            dir=right}]}
```

```
{form="ma",
  diacritics="0+a+a",
  cat=noun, +affix,
  +first, dir=left,
  affixes=[{cat=noun, +affix, +agr,
            dir=right}]}
```

The analysis of ‘*mdrsaaat*’ will then give rise to two analyses, namely

```
{form=("mu"+"d*r*s*")+ "aat", cat=noun,
  diacritics="a+a+a",
  person=3, num=pl, gender=f},
```

and

```
{form=("ma"+"d*r*s*")+ "aat", cat=noun,
  diacritics="0+a+a",
  person=3, num=pl, gender=f},
```

The first of these, *'mudarasaat'*, means teachers, and the second, *'madrasaat'*, means schools. How to determine which of these is meant in a given context is beyond the scope of the current paper (since they are both feminine plural nouns, the surrounding syntactic context will not settle the matter, and we will be exploiting the ideas proposed by [Ramsay, 1999b, Wedekind, 1996] for this task). But once we have chosen which form we want, it is straightforward to replace the place-holders in *'mad\*r\*s\*aat'* by the entries in *'0+a+a'* to obtain *'ma+d0rasa+aat'*. Subsequent inverse application of the morphotactic rules [c0, c1] ==> [c0, 0, c1] and [v1] ==> [d0, +, v1] (where d0 is a short vowel) then eliminates the 0 item and also removes the short vowel 'a' at the end of the stem because of the following long vowel 'aa' to produce the final full written form *'madrasaat'*.

That's fine for nouns, where the nominal affix largely determines the pattern of diacritics. The situation is less straightforward with verbs. It looks as though the information is stored in two places:

- the stem knows what diacritics it requires in different tenses, so that the third singular past tense of *'drs'* is *'darasa'*, the third singular masculine present is *'yadrasu'*, and the third singular future is *'syadrasu'*.
- the tense affix knows what the tense is.

We therefore need to make the tense marker select the appropriate option. To do this, we extend the entries for *'d\*r\*s\*'* and *'ya'* as follows:

```
{form="d*r*s*", cat=X,
  doptions={past="a+a+a",
```

```
  present="0+a+u",
  future="0+a+u"}
  affixes=[{cat=X, +affix, +first}]}
```

```
{form="ya", cat=verb, +affix,
  +first, dir=left, tns=present,
  doptions={past=_,
            present=D,
            future=_},
  diacritics=D,
  affixes=[{cat=verb, +affix, +agr,
            dir=right}]}
```

Unifying these will produce

```
{form="ya+d*r*s*", cat=X,
  doptions={past="a+a+a",
            present="0+a+u",
            future="0+a+u"},
  tns=present,
  diacritics="0+a+u",
  affixes=[{cat=verb, +affix, +agr,
            dir=right}]}
```

The verb provided three options for the diacritics: the present tense marker specified that the diacritics in any item which contained it would have to be the 'present' option, which in this case was *'0+a+u'*. For verbs belonging to other classes, the option selected by the affix would be different, but it would still be the one that verbs of that class took in the present tense. This item still requires an agreement marker, and only when that has also been found and the appropriate morphotactic rules applied will we have the final written form.

We are now in a position where we can recover the full form of a written noun or verb, at least in cases where the appropriate affixes are all overtly present and where everything goes smoothly. We do have to allow for empty affixes – in particular, we have to allow for an empty past tense marker, and we also have to allow empty nominal prefixes. The introduction of an empty inflectional affix is not all that radical – any categorial approach to morphology for English or German or French or Spanish or ... will require empty markers for singular NPs and other similar cases (a categorial approach to

the morphology of French adjectives, for instance, requires an empty masculine marker and an empty singular marker). Introducing what looks like an empty derivational prefix for obtaining a noun such as ‘*daarasataan*’ from ‘*d\*r\*s\**’ looks more dubious; but note that even this affix is not phonologically null, since it imposes a particular choice of infix items<sup>5</sup> on the full form, as follows:

```
{form="",
  diacritics="aa+a+0",
  cat=noun, +affix,
  +first, dir=left,
  affixes=[{cat=noun, +affix, +agr,
            dir=right}]}
```

#### 4 Conclusions

The treatment of Arabic morphology outlined above covers a substantial range of examples. There are, clearly, *irregular* forms that cannot be neatly captured within this framework, but irregular forms cannot, by definition, be dealt with using regular rules. More importantly there are a number of *regularities* within the various diacritic patterns (in particular, the fact that the vowels within a given pattern tend to be in ‘harmony’) which require further analysis. We noted above that we require a final round of morphotactic rules to capture effects at morpheme boundaries (reducing the written form ‘*ma+d0rasa+aat*’ to ‘*madrasaat*’, for instance) and we expect to be able to deal with at least some of these regularities at this point.

It is clear that written forms are highly ambiguous: the omission of the diacritics means that numerous distinct words have identical written forms, particularly in cases where there is no overt prefix indicating whether the item is a noun or a verb. There is nothing to be done about this at the lexical level: if the written form ‘*drs*’ could be resolved to a 3rd singular past active masculine verb or two a passive verb or two either of two nouns, then no amount of staring at it will resolve

<sup>5</sup>which may be long vowels, which will be realised graphemically on the standard written form, or short vowels or other phonetic elements for which we are attempting to find the appropriate diacritic

the issue. In a *sentence*, however, only one of these interpretations will lead to a well-formed analysis. Our treatment of morphology is embedded within a general framework for syntactic and semantic analysis which has been successfully applied to a range of European languages (English, German, Spanish, Greek) and which is currently being adapted for Arabic. The syntactic analysis is carried out within the HPSG-like framework reported in [Ramsay, 1999a, Ramsay and Seville, 2000]. The semantic analysis has been reported in numerous places, but it has not yet been applied to Arabic.

#### References

- L Bauer. *English Word Formation*. CUP, Cambridge, 1983.
- K Koskiennemi. A general two-level computational model for word-form recognition and production. In *COLING-84*, pages 178–181, 1985.
- A M Ramsay. Direct parsing with discontinuous phrases. *Natural Language Engineering*, 5(3):271–300, 1999a.
- A M Ramsay. Dynamic and underspecified semantics without dynamic and underspecified logic. In H Bunt, L Kievit, R Muskens, and M Verlinden, editors, *Computing Meaning*, volume 1, pages 208–220, Dordrecht, 1999b. Kluwer Academic Publishers (SLAP 73).
- A M Ramsay and H Seville. Unscrambling English word order. In M Kay, editor, *Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000)*, pages 656–662, Universität des Saarlandes, July 2000.
- M Schulze and A M Ramsay. Die struktur deutscher lexeme. *German Linguistic and Cultural Studies*, under review, 2000.
- J Wedekind. On inference-based procedures for lexical disambiguation. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, pages 980–985, Copenhagen, 1996.