Morphological Analysis of the Qur'an

Judith Dror

Department of Arabic Language and Literature

University of Haifa, Mount Carmel, 31905 Haifa, Israel

judith.dror@hotmail.com

Dudu Shaharabani

Department of Computer Science

University of Haifa, Mount Carmel, 31905 Haifa, Israel

dudushr@yahoo.com

Rafi Talmon

Department of Arabic Language and Literature

University of Haifa, Mount Carmel, 31905 Haifa, Israel

rstalmon@research.haifa.ac.il

Shuly Wintner*

Department of Computer Science

University of Haifa, Mount Carmel, 31905 Haifa, Israel

shuly@cs.haifa.ac.il

*Corresponding author

Abstract

We present a computational system for morphological analysis and annotation of the Qur'an, for research and teaching purposes. The system facilitates a variety of queries on the Qur'anic text that make reference not only to the words but also to their linguistic attributes. The core of the system is a set of finite-state based rules which describe the morpho-phonological and morpho-syntactic phenomena of the Qur'anic language. Using a finite-state toolbox we apply the rules to the Qur'anic text and obtain full morphological analysis of its words. The results of the analysis are stored in an efficient database and are accessed through a graphical user interface which facilitates the presentation of complex queries. The system is currently being used for teaching and research purposes; we exemplify its usefulness for investigating several morphological, syntactic, semantic and stylistic aspects of the Qur'anic text.

1 Introduction

We present a system for morphological analysis and annotation of the Qur'an, for research and teaching purposes. It provides a tool by which queries can be made which enable search of intricate syntactic (but also, to some extent, semantic and stylistic) relations in the Qur'an. The system is currently being used for teaching and research purposes; we exemplify its usefulness for investigating several morphological, syntactic, semantic and stylistic aspects of the Qur'anic text.

The importance of this text in the history of the Arabic language and Islamic civilization needs no introduction. The Qur'an has the advantage of being a closed corpus in the following senses: First, it demonstrates a frequent repetition of structures, indeed of the same phrases, to the extent of what may be considered formulaic style. Second, the Qur'an is traditionally identified with one person, a specific region, and a certain period of time, and its volume is relatively restricted.[1] These two facts justify treatment of the Qur'an as an independent corpus which deserves an independent study of its language in general and syntax in particular (Talmon 2001).

The system we describe provides means for presenting a variety of queries on the Qur'anic text that make reference not only to the words but also to their linguistic attributes. Thus, users are able to extract from the text certain words; or word patterns, using features of the words (such as root, pattern, lexeme, gender, number, dependent pronouns, tense and aspect, etc.); or combinations of words which conform to a particular structure (such as a nominative noun followed immediately

[1] Of course, there is no communis opinio about this tri-partite identification. Contradictory theories are discussed, which deny it partly or even as a whole.

by a finite verb). This capability enables the linguist to access complex information that is unavailable in ordinary dictionaries, thesauri or concordances. Such information can be used for teaching and research purposes; it facilitates linguistic and literary analyses of the Qur'anic text, and is instrumental in exploring aspects of its syntax, semantics and style.

The core of the system consists in morphological analysis of the text. This is done automatically, using a finite-state based toolbox. The major task here is the stipulation of the morphophonemic and morphographemic rules of the corpus. The product of this phase is a database of morphological analyses associated with each word token in the corpus. On top of the database, a graphical user interface was implemented which enables users to access the database of the annotated Qur'an, present queries and collect information in a structured manner.

The contribution of this work is manifold:

- The system enables both scholars and students to upgrade their linguistic tools in the study of the structure of Classical Arabic and its leading literary texts.

- The model is applicable for computerized study of other corpora, in fact of the whole Classical Arabic literature (we are currently applying it to a medium-sized corpus of Classical Arabic poetry, see section 5). .

- The methodology we developed is in principle applicable for other, similar tasks. While the morpho-phonological rules are characteristic of Classical Arabic, at times even specific to the corpus we used, the same methodology can be used for investigating linguistic and literary aspects of other corpora.

- The grammatically annotated Qur'an facilitates study of other language aspects of this text, especially its style.

In the next section we discuss the motivation for this research and relate it to existing works. Section 3 describes the details of the system, and some results of its usage are listed in section 4. We conclude with suggestions for further research.

2 Motivation

2.1 Challenges

Any linguistic or literary investigation of texts can benefit from computational technology. Evidently, the use of computational dictionaries, concordances and indexes, augmented by sophisticated searching tools, can be extremely useful for scholars who are interested in such investigations. However, for languages with productive morphology such technology is usually insufficient. Dictionaries and concordances are indexed by lemma, whereas in a language such as Arabic a given lemma can be inflected in dozens, sometimes hundreds of different forms.

As an example, consider the noun *diin* "religion" (we refer to matters of transcribing Arabic presently). It occurs as *diin-u* in the nominative case, *diin-a* in the accusative and *diin-i* in the genitive. Each of these forms can be optionally extended by the indefinite marker *n*. Adding to the confusion is the fact that nouns can combine with dependent pronominal pronouns, such as *-i* for the first person singular one. Thus, the form *diin-i* is ambiguous as to whether the *-i* suffix is a case marker or a dependent pronoun.

Verbs are even more complicated. There are twelve major verbal stems in

Classical Arabic, each with its inflectional paradigm. Verbs inflect for person, number, gender, tense/aspect and mood and can additionally be associated with object clitics. For example, the verb *'akala* "eat, third person singular masculine perfect active" also occurs as *'akaluu* (plural) or *ya'kul* (imperfect); or as *'akala-hu* (with an attached pronominal clitic indicating the direct object of the verb).

On top of the morphological wealth of the language, traditional Arabic script dictates that certain particles, which in some languages are independent morphemes, attach to the words which immediately follow them. Such is the case with the definite article, the coordinating conjunction, many of the prepositions etc. Sequences of two, three and even four particles can combine with a single word, and the number of different prefix sequences is in the hundreds. Since these particles can in most cases also be analyzed as parts of lemmas, the orthography adds yet another dimension of ambiguity to the problem of analyzing Arabic words.

For example, each of the inflected forms of *diin* discussed above can be preceded by the definite article *l-*; the conjunction *wa-* "and"; or the prepositions *bi-* "in" or *li-* "to". Combinations of the prefix particles are also allowed, such as *bi-l-diin-i* "in the religion" or *wa-li-l-diin-i* "and to the religion". Verbs, too, can be preceded by particles such as *wa*, *fa* or *fal* (conjunctions), *la* (an affirmative particle) or the prepositions *bi-*, *li-* or *ka-*.

Thus, a simple search engine which uses the lemma to index word forms may fail to recognize several of the inflected and derived forms of the lemma in the text. More sophistication is required in order to provide appropriate search tools: in other words, full morphological analysis of the text is inevitable if one is interested in accurate results.

2.2 Related Work

Automatic morphological analysis of Arabic is not new; several such systems exist (Beesley 1996; Beesley 1998a; Beesley 2001; Kiraz 1998; Kiraz 2000; Al-Shalabi and Evens 1998; Berri, Zidoum, and Atif 2001; Darwish 2002). A major drawback of some systems is limited coverage; for example, Al-Shalabi and Evens (1998) only deal with verbs and deverbal forms. Berri, Zidoum, and Atif (2001) describe a system which does not use a lexicon; as the implementation of the system is reported to be "underway", and no examples of output are given, it is hard to assess its coverage. Darwish (2002) decidedly aims at producing only the root of the word as the result of the analysis; and he does so using automatically acquired rules and statistical measures. In contrast, our system is capable of providing full coverage of the corpus it is designed for, and extensions of the system to larger corpora is only a matter of extending the lexicon, as the system implements linguistically motivated rules.

Similarly to works of Beesley and Kiraz mentioned above, but in contrast to other systems, our system is based on linguistic rules. This is advantageous both on theoretical and on practical grounds. First, the design of the rules results in additional linguistic insights, as noted above. Second, rule-based systems are very easy to extend, maintain and modify. Finally, the reliance on finite-state technology guarantees extremely efficient processing. In our case, the entire corpus of some 80,000 words is analyzed in less than 20 seconds on stock hardware.

However, the major inadequacy of existing systems for our purposes stems from the fact that most of them deal with Modern Standard Arabic (MSA); the language of the Qur'an is Classical Arabic. In many respects it is unique, and its lex-

icon, morphology and syntax require dedicated attention.[2] As Berg (2001) notes, "Except for isolated efforts..., little has been done with computer-assisted analysis of the text... Thus, for the present, computer-assisted analysis of the Qur'ān remains an intriguing but unexplored field." The work we report on here is one step towards exploring this field.

2.3 Characteristics of the Transcription

Most systems for processing Arabic use the standard Arabic script, or a one-to-one transliteration thereof, as their input script. Our system, in contrast, uses a phonemic transcription of the text, in which some of the morphological and orthographic ambiguity is reduced. The transcription is based on pure ASCII notations, largely with single-symbol equivalents of the Arabic graphemes, and double letters expressing long vowels. The conventions of the Arabic orthography are basically retained, e.g., one-letter particles which are prefixed to the noun or verb are hyphenated to the following word (*wa-kaana* "and was"), as are pronominal and case/mood suffixes (*yas'al-u-nii* "he will ask-indicative-me"). In general, hyphenation serves to isolate noun bases from the various (inflectional and derivational) affixes. This process is sometimes inapplicable for Arabic verbal forms, whose complexity calls for creation of a detailed set of derivation rules instead.

[2]Classical Arabic is a major fundament of MSA. Both are identified as one register, termed *fuSHaa*, by native speakers of Arabic (who use both for cultural purposes, not for daily communication). But MSA is comprised also of several other layers of Arabic, such as influence of the modern dialects, influence of foreign (mainly Western) languages, and Middle Arabic. See, e.g., the delailed discussion in Blau (1973), Blau (1976), mainly for the domain of syntax. It is noteworthy that some scholars (notably W. Fischer) consider Qur'anic Arabic as pre-classical, see summary in Talmon (2001, p. 346)

As an example of the transcription, figure 1 lists the seven verses of the first suura. More details are given in Appendix A.

1.bi-sm-i llaah-i l-raHmaan-i l-raHiim-i

2.l-Hamd-u li-llaah-i rabb-i l-&aalam-iina

3.l-raHmaan-i l-raHiim-i

4.maalik-i yawm-i l-diin-i

5.'iyyaa-ka na&bud-u wa-'iyyaa-ka nasta&iin-u

6.hdi-naa l-SiraaT-a l-mustaqiim-a

7.SiraaT-a lla(dh)iina 'an&amta &alay-him gayr-i l-magDuub-i &alay-him wa-laa l-Daall-iina

Figure 1: Example of the Arabic transcription: *suurat-u l-faatiHat-i*

The use of a phonemic script facilitates a simpler, clearer stipulation of morphological rules. It also helps to reduce the degree of ambiguity in certain cases. Since our transcription is more informative than the standard script, it is possible to deterministically convert the former to the latter; see section 3.3.

3 Description of the System

In order to perform full morphological analysis one needs a complete lexicon and a complete stipulation of the morphological rules of the language at hand. We use a finite-state based toolbox (LEXC/XFST, Beesley and Karttunen (2003)) which facilitates the stipulation of the lexicon and the rules. This information is then compiled into finite-state transducers which constitute the morphological analyzer.

The use of the toolbox enabled us to avoid the bottleneck of having to annotate

the corpus manually. Furthermore, using a finite-state toolbox such as XFST (as opposed to implementing a dedicated morphological analyzer from scratch) has three additional advantages: first, the morphological analyzer is not a "black box" which outputs analyses when given a string. Rather, the rules which constitute the system make sense linguistically. The mere process of designing the rules yields new insights concerning the morpho-phonology of the language. Maintaining such a system is a relatively easy task, as the rules are available in a human-readable form. Second, as finite-state networks are inherently reversible, the system can be used both for analysis and for generation. The generation mode was extremely useful when the system was debugged: it enabled us to generate both arbitrary and manually crafted inflected forms, and test their plausibility. Finally, as XFST compiles its rules into finite-state networks, we can benefit from the computational efficiency of such systems, where analysis of a string takes time linear in the length of the string (Karttunen, Chanod, Grefenstette, and Schiller 1996).

In this setup, the main tasks of the computational linguist are the design and implementation of the lexicon; and the stipulation of morphological rules. This section describes the lexicon, the set of rules and a few additional programs that are used by the system.

3.1 Lexicon

We divided the lexicon of the Qur'an into three classes: closed-class words (including prepositions, pronouns, particles, conjunctions, adverbials, etc.); nominal bases; and verbal bases. We describe each of these classes below.

Using a concordance (Abd al-Baaqii 1987), we manually constructed a full list of the closed-class words (a few hundreds occur in the Qur'an). Closed-class

words are lexical items such as pronouns (personal, demonstrative, relative and interrogative), prepositions and particles. Examples include the pronouns *hum* ("they") or *naHnu* ("we"), the prepositions *&alaa* ("on") or *min* ("from") and particles such as *'iyyaa* (the accusative marker, denoted Acc). Note, however, that in Arabic such words inflect and can combine with other particles, so the lexicon accounts also for inflected forms such as *&alay-him* ("on+3pPlMasc") or *'iyyaa-ka* ("Acc+2pSgMasc"). Furthermore, certain particles are combined to words as prefixes, such as the conjunction *wa-* ("and"): *wa-naHnu* ("and-we"). The lexicon handles such cases by means of systematic rules which generate the inflected (and derived) forms from the basic word list. Some phenomena, however, such as phonetic rules which might apply, are dealt with by subsequent stages of processing; see section 3.2.

We also used the concordance to construct a complete list of all the nominal bases which occur in the corpus (approximately twenty-five hundred). The lexicon of noun bases is more complex; interesting phenomena include differences in the feminine and plural inflections, including the broken plural, and proper names. We solve the problems using brute-force encoding of the irregular forms in the lexicon. Again, since we are mostly concerned with a closed corpus here, this is a reasonable solution. It is worth mentioning that such phenomena *can* be handled by finite-state machinery (Beesley 1998b; Beesley and Karttunen 2000), but in our case such solutions were unnecessary.

The lexicon associates with each lexeme its root and pattern. The format of each entry consists of three parts: a *lexical string*, which typically consists of the root and patter information, but sometimes contains more information, especially for the idiosyncratic forms; a *surface form*, which is the actual representation of

the lexeme in our script; and a *continuation class*, which tersely specifies which suffixes the lexeme can combine with. The format, in the syntax of the LEXC finite-state toolbox, is:

```
lexical_string:surface_form continuation_class
```

We use the '+' symbol in lexical strings to separate morphemes and tags of the analysis from each other. For example, consider the following lexical entries:

```
swr+fu&lat:suurat                NounEndingFem;
Hmd+fa&l:Hamd                    NounEnding;
'insaan=nws+fa&l:naas            PluralNounEnding;
```

The first specifies that *suurat* ("Qur'an chapter") is a noun whose root is *s.w.r* and whose pattern is *fu&lat*. Furthermore, the continuation class `NounEndingFem` indicates that the lexeme can be suffixed by feminine nominal affixes. The second entry indicates that *Hamd* ("praise") is a regular (masculine) noun whose root is *H.m.d* and whose pattern is *fa&l*. The last example refers to *naas* ("men") and stipulates, in addition to the root *n.w.s* and the pattern *fa&l*, also the singular form *'insaan* "man". The continuation class here indicates that *naas* can combine with plural nominal suffixes.

As was the case with the previous class, certain aspects of noun inflection, such as concatenation of particles (prefixes), gender, number and case morphemes and dependent pronouns (suffixes), as well as definite and indefinite markers, are handled in the lexicon. Subsequent processing handles morpho-phonemic alternations. For example, all nouns can be suffixed by *-ii* to indicate a first person singular dependent pronoun (e.g., *&aduww-ii* "my enemy"). The lexicon will add

such suffixes to all regular nouns, including *bu(sh)raa* "good news". Only further processing will correct the resulting form to *bu(sh)raa-ya* ("good news+1pSg").

As another example, the lexicon generates all the combinations of nouns with the definite article *l-*, which is a prefix, and with the indefinite marker *n*, which is a suffix. This means that ungrammatical forms such as $*$*l-naas-un* are generated by the lexicon and will have to be pruned by the rules.

The verbs lexicon is the most complicated. While it was possible to manually construct a list of all noun bases occurring in the corpus, such a task would have been far more complex for the verbs. However, a list of the verbal *roots* and *stems* occurring in the Qur'an (including perfect/imperfect base variations in Stem 1) is available (Chouémi 1966; Ambros 1987); we automatically generated all possible instantiations of these roots in all the verbal patterns of Qur'anic Arabic. Of course, this leads to vast over-generation: our lists contain approximately 1000 roots and almost 100 verbal patterns. Of the 100,000 possible verb bases, only a small percentage is actually realized in Arabic. Furthermore, following the practice of noun bases and closed group words, we also generate all possible inflections of the verbal bases in the lexicon (again, deferring morpho-phonological alternation to subsequent processing). In theory, such over-generation could have led to unbearable morphological ambiguity; our experience, however, shows that this is not the case. As our objective here is limited to analysis of the Qur'anic text only, we were not obliged to consider word forms which do not occur in the Qur'an. Intersecting the huge number of inflected forms with the corpus, most of the artificial forms disappear and the remaining ones contribute only mildly to the degree of ambiguity.

Finally, it is important to note that the lexicon does not only generate surface

forms; it also associates with each surface form a lexical string which lists information about form's morphemes and morphological tags. The following example lists pairs of surface forms and their associated lexical strings as generated by the lexicon (and before any of the rules was applied):

```
l-naas-u    Def+'insaan=nws+fa&l+Noun+Triptotic+Masc+BrokenPl+Nom
l-naas-un   Def+'insaan=nws+fa&l+Noun+Triptotic+Masc+BrokenPl+Nom+Tanwiin
fa-'akalaa  fa+Particle+Conjunction+'kl+Verb+Stem1+Perf+Act+3P+Dual+Masc
fii-hum     fii+Prep+Pron+Dependent+3P+Pl+Masc
```

Note that the second example above is ungrammatical due to the occurrences of both the definite article and the indefinite marker; the last example is also ungrammatical, as a vowel harmony rule dictates that the dependent pronoun *hum* be realized as *him* when attached to a word ending in *ii*, *i* or *y*. These problems are corrected by the rule component.

3.2 Finite-State Rules

As noted above, the lexicon generates base forms, with additional affixed morphemes that represent particles such as the conjunction *wa-* ("and"), the definite article *l-* or the preposition *bi-* ("in"), morphological information pertaining to number, gender, case etc. such as the suffix *-u* "+Nominative", dependent pronouns such as the suffix *-ka* ("+2pSgMasc") etc. However, such affixes are simply concatenated to the bases they attach to, and morpho-phonological alternations are deferred to this stage of processing. Furthermore, the verb bases that are generated in the lexicon ignore completely the peculiarities of the weak paradigms. The rule component of the system has four tasks: to filter out redundancies of the lexicon; to handle morpho-phonological alternations; to take care of idiosyncrasies, such as

the verbal weak paradigms; and to implement purely phonetic rules. We exemplify each of these below.

As an example of redundancies in the lexicon, consider the prepositions *li-* ("to") and *ka-* ("as"). These prepositions can only attach to nouns in the genitive case. However, the lexicon will wrongly generate strings in which these prepositions combine with nominative or accusative nouns. A simple finite-state rule filters out analyses which contain both the preposition *li-* or *ka-* and a noun in accusative or nominative case. The examples below use the syntax of the XFST finite-state tool; while it is impossible to detail the syntax and semantics of the XFST language here, we note in passing that '`~?*`' denotes the empty set; '`<-`' is a replace rule, and the example below states that whenever a pattern which is captured by whatever is to the right of the '`<-`' is matched, it is replaced by the empty set (in other words, filtered out). The pattern to the right of the arrow matches occurrences of the *l* or *k* prepositions ('`|`' denotes union) when they are *in the context of* the beginning of a word (denoted by '`.#.`') on the left and either the tag '`+Acc`' or the tag '`+Nom`' on the right. The '`_`' separates the left context from the right context. Ignore the '`%`' symbols which are an XFST technicality.

```
~?* <-  [l %+Prep | k %+Prep] \/
        [.#.] _ [?* [%+Acc | %+Nom]];
```

In general, replace rules have five components: a pattern to be matched; a pattern for replacing it; a type of replacement; and, optionally, left and right contexts which constrain the replacement. As another example, a simple rule filters out analyses which contain both the definite article and an indefinite marker (*tanwiin*):

```
~?* <- %+Noun  \/ [Def%+ ?*] _ [?* %+Tanwiin];
```

Other rules of this kind filter out analyses of diptotic nouns whose pattern is *fa&laa'* or *'af&al* in the genitive case when they are not definite; or, similarly, indefinite tri-syllabic broken plurals in the genitive case.

As an example of a morpho-phonological alternation rule, consider the suffix *-uuna* ("Rectus"). When added to a noun which ends in *aa*, the long vowel is shortened and the suffix is contracted, so that *l-'a&laa+uuna* ("the supreme ones") becomes *l-'a&l-awna*. Similarly, the obliquus suffix *iina* is contracted to *ayna*. These phenomena are easily handled with finite-state rules:

```
[aa %- uu n a] -> [%- a w n a];
[a y %- uu n a] -> [%- a w n a];
[aa %- ii n a] -> [%- a y n a];
[a y %- ii n a] -> [%- a y n a];
```

Here, the format of the rule is simpler, as the replacement is unconditional (i.e., not constrained by the context). For example, any occurrence of the pattern *aa - uu n a*, wrongly generated in the lexicon, is replaced by the pattern *- a w n a*. Assimilation phenomena in the verb are handled similarly:

```
t a -> s || [%+Stem5 | %+Stem6 ] _ [s ];
z t -> z d || [%+Stem8] _ ;
```

In addition, certain rules handle idiosyncrasies such as 'frozen' nouns which are not marked for case etc. More interesting is the treatment of the weak verb paradigms. Most of the rules in the system are dedicated to weak verbs, handling phenomena such as breaking a tri-consonantal cluster with a vowel in the context of geminite roots: $R_1R_2V_1R_2V_2 \rightarrow R_1V_1R_2R_2V_2$; or the omission of the *w* in prima-w roots, as well as the other phenomena associated with this paradigm; etc.

Finally, finite-state rules also handle pure phonetic rules. For example, such a rule implements a vowel harmony phenomenon which changes the *u* vowel of the dependent pronouns *-hu, -hum, -humaa* and *-hunna* ("him, them-PlMasc, them-Dual, them-PlFem", respectively) to *i* when attached to words ending in *i* or *y*:

```
[%- h u m] -> [%- h i m] || [ii | i | y] _ [%- | .#.];
[%- h u m aa] -> [%- h i m aa] ||
                 [ii | i | y] _ [%- | .#.];
[%- h u n n a] -> [%- h i n n a] ||
                  [ii | i | y] _ [%- | .#.];
[%- h u] -> [%- h i] || [ii | i | y] _ [%- | .#.];
```

3.3 A Computational Morphological Analyzer

Once the lexicon and the finite-state rules have been finalized, we used an existing finite-state toolbox (Beesley and Karttunen 2003) to compile them to finite-state networks, implementing a full morphological analyzer of the corpus. Examples of analyses are provided in Figure 2.

Furthermore, we have implemented a finite-state transducer for converting our transcription to the standard Arabic script (represented in Unicode). While it is not incorporated into the graphical user interface (see below), this transducer provides an efficient conversion of the analysis results to the standard script, while the internals of the system can still be implemented using our phonemic transcription. Users who prefer to use the standard script can thus search the morphologically analyzed Qur'an in the traditional way (albeit without the aid of the graphical user interface).

```
suurat-u        swr+fu&lat+Noun+Triptotic+Fem+Sg+Nom
l-faatiHat-i    Def+ftH+Verb+Triptotic+Stem1+ActPart+Fem+Sg+Gen
bi-sm-i         b+Prep+sm+Noun+Triptotic+Masc+Sg+Gen
llaah-i         Def+llaah+ProperName+Gen
l-raHmaan-i     Def+rHm+fa&laan+Noun+Triptotic+Adjective+Masc+Sg+Gen
l-raHiim-i      Def+rHm+fa&iil+Noun+Triptotic+Adjective+Masc+Sg+Gen
l-Hamd-u        Def+Hmd+fa&l+Noun+Triptotic+Masc+Sg+Nom
li-llaah-i      l+Prep+Def+llaah+ProperName+Gen
rabb-i          rbb+fa&l+Noun+Triptotic+Masc+Sg+Pron+Dependent+1P+Sg
rabb-i          rbb+fa&l+Noun+Triptotic+Masc+Sg+Gen
l-&aalam-iina   Def+&lm+faa&al+Noun+Triptotic+Masc+Pl+Obliquus
l-raHmaan-i     Def+rHm+fa&laan+Noun+Triptotic+Adjective+Masc+Sg+Gen
l-raHiim-i      Def+rHm+fa&iil+Noun+Triptotic+Adjective+Masc+Sg+Gen
maalik-i        mlk+Verb+Triptotic+Stem1+ActPart+Masc+Sg+Gen
yawm-i          ywm+fa&l+Noun+Triptotic+Masc+Sg+Gen
l-diin-i        Def+dyn+fi&l+Noun+Triptotic+Masc+Sg+Gen
'iyyaa-ka       'iyyaa+Particle+Pron+Dependent+2P+Sg+Masc
na&bud-u        &bd+Verb+Stem1+Imp+Act+1P+Pl+Masc/Fem+NonEnergicus+Indic
wa-'iyyaa-ka    wa+Particle+Conjunction+'iyyaa+Particle+Pron+Dependent+2P+Sg+Masc
nasta&iin-u     &wn+Verb+Stem10+Imp+Act+1P+Pl+Masc/Fem+NonEnergicus+Indic
nasta&iin-u     &yn+Verb+Stem10+Imp+Act+1P+Pl+Masc/Fem+NonEnergicus+Indic
hdi-naa         hdy+Verb+Stem1+Imperative+2P+Sg+Masc+NonEnergicus+Pron+Dependent+1P+Pl
l-SiraaT-a      Def+SrT+fi&aal+Noun+Triptotic+Masc+Sg+Acc
l-mustaqiim-a   Def+qwm+Verb+Triptotic+Stem10+ActPart+Masc+Sg+Acc
```

Figure 2: Example analyses

The Qur'an consists of approximately 80,000 word forms (tokens). Our morphological analyzer is now capable of producing analyses for all of them (full coverage). Evidently, our system is currently incapable of performing (context-dependent) morphological disambiguation, and sometimes the number of analyses

per word can be rather high, especially in the verb, as is the case with *nasta&iin-u*, which is assigned four analyses here (but see section 5 below). However, the average number of analyses per word token in our corpus is only 1.37, and most of the tokens (70%) are assigned a unique analysis.

The results of the analysis are stored in a database in a form that encodes, for each analyzed word, its morphological features and their values. For example, an analysis such as:

```
swr+fu&lat+Noun+Triptotic+Fem+Sg+Nom
```

is converted to a record structure of the form:

root	swr
pattern	fu&lat
part of speech	Noun
case marking	triptotic
gender	Fem
number	Sg
case	Nom

We use a small program to parse the analyses and add the records to the database (we use MySQL (DuBois 1999) for database management). The database provides an efficient means for searching the analyzed corpus by a variety of keys, including the surface word, its root, its pattern but also key features such as part of speech etc. This facilitates complex queries as described below.

To demonstrate the efficiency of finite-state technology in general, and the Xerox tools (LEXC and XFST) we used in particular, for large-scale morphological analysis, we provide here some technical data regarding the system. The corpus

we deal with contains some 80,000 word tokens. The lexicon, expressed in *LEXC*, contains approximately 2500 noun forms and 100,000 verb bases, in addition to closed-class words. The number of rules, expressed in *XFST*, is approximately 50 for nouns and 300 for verbs. Both the lexicon and the rules are compiled into a finite-state network that is then minimized and stored compactly; the number of nodes in the network is 185,000, with more than 345,000 connecting arcs. The size of the network file is only 1Mb. We use the network to analyze the entire corpus; on an ordinary personal computer (with a 500MHz processor and 320Mb of memory), this takes less than 20 seconds.

3.4 Graphical User Interface

We designed a graphical user interface for accessing the information stored in the database. As users of the system are not expected to be proficient in SQL, a database query language, the GUI provides menus for easing the construction of rather complex queries (see Figure 3).

The top part of the GUI is used for expressing queries. Queries can refer to a single word in the corpus or to several words; in the latter case, each sub-query refers to a single word. Sub-queries can be combined with two operators: *followed immediately by*, or *followed by*, which refers to words following the word indicated by the previous sub-query, up to the end of a verse. In addition, two sub-queries can refer to the same word using the logical operators *and* and *or*.

Each sub-query (which refers to a single word) can be used to express information about the word's properties. The word can be given explicitly; or the user can ask for a certain root, or a certain pattern; or, additionally, users can constrain the values of morphological features such as number, gender, case, aspect etc. Further-

more, agreement phenomena can be queried by setting the value of some feature in a sub-query to a *variable*, and using the same variable as the value of the same feature in a different sub-query referring to a different word. The menus are dynamic: for example, checking the value *noun* for the feature *part of speech*, more options will pop up for constraining properties of nouns. Different options pop up when the user opts for *verb* as the part of speech.

For example, by checking the *verb* value for part of speech, a sub-menu listing the twelve verb stems would open. The user can then check *stem1*, for example. Checking, in addition, *masc* for gender, *sg* for number and *gen* for case results in the query depicted in Figure 3.

Once the specification of constraints is done, a button enables the generation of an SQL query. This query can be further edited manually by more sophisticated users. Finally, a button submits the query to the database; the result, which is presented on a dedicated window, is a list of all the occurrences in the corpus of words which satisfy all the constraints (in the example of Figure 3 there are 261 such analyses). Each occurrence is preceded by its reference: the suura, verse and word number; and is followed by its analysis. The user can now select any of the analyses by clicking on it; in the bottom frame, which constantly displays the Qur'anic text, the view will be shifted to the actual occurrence of the desired word, and the word will be highlighted. In the example, this is the word *bi-l-baaTil-i*.

4 Results

As noted above, our system performs a full morphological analysis of the entire Qur'an. We evaluated the accuracy of the system by manually annotating the eighth

Figure 3: The graphical user interface

suura, consisting of a subset of 1248 words. For this subset, the system produced 1440 analyses, with an average degree of ambiguity 1.15. Comparing the analyses of the system to the manually annotated subset, 69 of the analyses were deemed incorrect, 205 as possible (but perhaps contextually wrong) and 1162 as *the* correct analysis. These figures yield 93% recall, 80% precision and an f-measure of 0.86. We believe that these measures are representative of the entire corpus.

The system is now ready for research purposes and teaching of advanced students in Arabic departments. Its development was conceived to enhance a systematic syntactic analysis of the Qur'an, and therefore it creates a basis for, and an introduction to (future) operation of, a more comprehensive tool, that will offer a syntactic parsing of our corpus. In what follows we discuss the system's advantages in searching issues of syntactic, semantic and stylistic relevance. We also compare the usability of the system to existing tools, such as manual and computerized concordances.

4.1 Range of the Morphological Data in the Lexicon

The system provides a wealth of morphological information about the words in the corpus. Nominals are associated with their root, pattern, number, gender and case; but also with subcategory information, so that proper names, numerals and adjectives are explicitly marked as such. Furthermore, for the pattern *'af&al*, which is used to denote three different functions, the correct function of each lexeme is specified in the lexicon (and can be searched for using the interface). Verbs are specified for root, stem, aspect, mode, mood, person, number and gender. Closed class words are classified according to their function. For example, pronouns are listed as demonstrative, relative, interrogative etc.; other particles are classified as

conjunctions, interrogatives, conditionals etc.

Various other (syntactic and semantic) attributes of words may be added relatively easily. For example, verbs can be tagged as transitive/intransitive, and their occurrences in various stems can partly be identified according to established semantic subcategorization of these stems. We leave such extensions for future work.

4.2 Morphological Studies with Syntactic Implications

The morphologically annotated Qur'an is designed basically to facilitate retrieval of data which are significant for a systematic syntactic analysis of this corpus. Recent studies in the noun and verb structure of Semitic languages in general and Arabic in particular emphasize the interrelation between the two spheres. The benefit of a fully annotated inventory of Qur'anic nouns is obvious. For example, efficient retrieval of broken plural patterns in their immediate context is vital for the study of morphological stipulations on syntactic agreement between heads and their complements. Is there any rule in the selection of *bararat* (pattern *fa&alat*) vs. *'abraar* (pattern *'af&aal*) as plurals of *barr* "pious"? Does this choice affect the properties of nominal adjuncts and modifiers of the head? We have used the system for this type of investigations.

Another example is taken from the verb domain. Parallel use of different infinitive (nomen actionis) patterns of the same root (sometimes of the same stem, mostly of stem 1; cf., e.g., *kawn, kiyaan, kaynuunat* "being") calls for study of its possible stipulations. The given Qur'anic occurrences are easily traced with our program.

Nominalization of verbs in the Qur'an is either morphological or syntactic. As an example of a morphological process, consider *taSuum-uuna* "you will fast,

plural masculine indicative" → *Siyaam-u-kum* "fasting+2ndPlMasc". In contrast, consider the syntactic process exemplified by *taSuum-uuna* → *'an taSuum-uu* "that you-fast+PlMascSubjunctive". Comparison of such pairs which occur in the same corpus may explain the rules of preference for selection of each construction. Our system facilitates presentation of queries which yield the specified constructions in their context, thereby providing the researcher with the relevant data needed for such a comparison.

4.3 Efficient Retrieval of Syntactic Constructions

While the system annotates the corpus mostly with morphological information, it can still be used to uniquely retrieve selected syntactic phrases, or at least reduce drastically the possibility of non-relevant occurrences. Consider the prepositional phrase called *partitive min* (*min al-bayaan*). It typically follows a relative clause headed by a relative pronoun (*man, maa, lla(dh)ii*, etc.), e.g.,

maa	*fa&ala*	*min-a*	*l-'a&maal-i*	*l-(sh)ariirat-i*
Rel	did-he	of	the-doings-Gen	the-bad-SgFemGen

"what he did of evil doings" or "the evil doings he performed"

Now manual and computerized concordances of the Qur'an adduce upon request such data as the total number of occurrences of the preposition *min*, which is 3221. Considerable time is saved if we need to check only the occurrence of this preposition in such constructions as "*min*, followed by a plural noun which is prefixed by the definite article", which is typical of this partitive *min*; or if our query includes the restriction that this construction is preceded by a relative particle. In fact, such a query produces only 12 results, all of which are instances of the desired syntactic construction.

Such uses of the system can be demonstrated with another structure. Classical Arabic uses an equational sentence ("X is Y") of the pattern Demonstrative + Independent ("copular") pronoun + noun prefixed by an article or suffixed by dependent pronoun, e.g.,

'uulaa'ika *hum-u* *l-Saadiq-u*
Those are the-earnest-Nom
"these are the earnest ones (in their belief)"

Using a concordance for retrieving such constructions, one is overwhelmed by 368 occurrences of demonstratives. Accurate retrieval of the specific structure is enabled by the following query: "Show all Demonstratives followed by 3rd person independent pronouns, which are followed immediately by an article". This yields exactly 9 results, all correct.

4.4 Historical and Stylistic Investigations

The prophetic and political message which constitutes the Qur'an is traditionally divided by scholars according to the Meccan and Medinan phases of Muhammad's activity. Study of the language of the Qur'anic text, either as a whole or of selected parts of it, involves syntactic and stylistic distinctions of such sentence types as indicative, imperative, vocative, etc., which are characterized mainly by selection of different verb aspects (perfect, imperfect-indicative, imperfect-subjunctive, passive or active forms, etc.), or a combination of these and certain particles, typical phrases, etc. Critical students of the text, including historians of early Islam, need an exhaustive language analysis of this unique text according to these and similar parameters. Whereas stylistic investigation is as old as the first attempts to create a chronological ordering by modern scholars of the Qur'anic message, it neverthe-

less has developed but scant observations concerning the relations of the syntactic structures and their possible chronological distribution.

As an example, we demonstrate the advantages of our tools for an analysis of stylistic themes and consider the possible implications of this analysis for study of Qur'an's chronology. Let us focus on the summons *yaa 'ayy-u-haa lla(dh)iina 'aamanuu* "O the believers", so typically Qur'anic. Selection of vocatives is generally of great interest, because on the one hand it may reflect 'ritual' or fashionable formulae, but on the other it may be idiosyncratic and disclose a significant aspect of the speaker's discourse strategies. Note that the Prophet addresses his supporters. Here are some relevant statistics:

- It occurs in the Qur'an 89 times in 20 suuras.
- There are 149 occurrences of the vocative *'ayy-u-haa* in the Qur'an altogether, always preceded by *yaa*.
- In the 20 suuras in which this specific summons occurs, there are 28 other expressions opened with the *yaa 'ayy-u-haa* element.
- The other 32 occurrences of *yaa 'ayy-u-haa* expressions are found in 19 suuras.
- In all but one of its occurrences, the specific expression is in a verse-initial position, and may well pose as address-initial in all these cases.
- Another common Qur'anic expression with a general appellation is *yaa 'ayy-u-haa l-naas-u* "O people", which occurs 20 times, half of them in the 20 suuras that include the previous expression.

We now check the chronological distribution of the various expressions discussed above. Our reference is provided by three western scholars, who divide the 114 suuras of the Qur'an into four periods of the Prophet's activity. Three belong to Muhammad's 12 years prophecy in Mecca, and the last to his 10 years in Medina. The results are the following:

- 18 of the 20 suuras including our expression are of the Medinese period. Of the remaining two, one is identified as belonging to the last suuras of the third Meccan period, and the other is recognized by some as Medinese, and by others as similar in status to the first.

- Of the group of 19 suuras which include *yaa 'ayy-u-haa* expressions other than ours, 11 are of the late (third) Meccan, and one of the Medinese period, whereas two go back to the second, and six to the first Meccan period of Muhammad's mission.

- It is noteworthy that all the 20 occurrences of the general appellation *yaa 'ayy-u-haa l-naas-u* belong to the two later periods, namely the third Meccan and the Medinese.

- The group of 20 suuras is further distinguished from its earlier counterpart which includes five plurality terms of address, which are not utterance-initial, e.g. *(th)umma 'inn-a-kum 'ayy-u-haa l-Daall-una...* "Now you, o lost ones...". This type of vocative expressions is absent in the late Meccan and Medinese suuras of both groups.

An earlier, general observation on the pattern *yaa 'ayy-u-haa l-naas-u* is noted briefly by Böwering (2000, p. 324). Böwering's note does not include the sig-

nificant details so comfortably retrieved by our tools. It reflects the inconclusive presentation of many other syntactico-stylistic issues studied by earlier scholars.[3] To conclude, our automatic analysis of stylistic issues efficiently provides a large amount of linguistic structures for consideration of seminal issues in the study of the chronology of the Qur'anic text and the earliest history of Islam.

4.5 Teaching Uses

The system is designed for teaching of classes of Qur'anic studies, in which advanced students take first experience in independent analysis of Classical Arabic corpora (together with acquaintance with the standard linguistic literature), and in computational approaches to the grammatical study of the Qur'anic text. Students will be challenged with such tasks as composition of meaningful queries to morphological and syntactic topics of their choice, to provide insights unnoticed in systematic grammars and specific studies. The system is available for use over the Internet; it is used both in the classroom and, through remote access, by students at home. So far our experience with this aspect of the system's use is limited to a small number of graduate students, but we expect it to be used more extensively for teaching purposes in the future.

5 Conclusion

We described a system that uses state-of-the-art finite-state technology for morphological analysis of the Qur'an, and makes the results available, through an efficient database and a graphical user interface, for complex queries that involve not only

[3] For a survey of earlier studies on Qur'anic syntax, see Talmon (2001, pp. 359-367).

the Qur'anic text but also its morphological, and to some extent also syntactic and semantic, properties. The system is being used for teaching and research purposes and is publicly available on the Internet.[4]

This work demonstrates that the use of modern computational linguistics technology can facilitate the construction of computational tools for processing linguistic and literary texts, and in general aid in Humanities research and education. The benefits of the system are expressed in additional linguistic insights which were hard to obtain otherwise, as was demonstrated above.

While the system is already being used to actively investigate linguistic aspects of the Qur'an as demonstrated above, it is still under development. Current and future extensions of the system are focused on two major issues: improving the accuracy of the morphological annotation, in particular disambiguation; and extending the annotation to cover syntactic constructions.

As can be seen in Figure 2 above, the current annotation still results in unnecessary ambiguity, especially in the verb system. We are constantly working on reducing the degree of ambiguity. This is done by enriching the dictionary with additional information, and by refining the rules. However, in order to significantly reduce the degree of ambiguity, contextual information must be taken into account. We are currently implementing a cascade of finite-state transducers (Abney 1996) in order to enrich the morphological analysis with syntactic information and to reduce the morphological ambiguity using short-context rules. We do not intend to perform a full syntactic analysis of the Qur'an, but rather to use finite-state technology for shallow parsing, detection of phrase boundaries, recognizing certain

[4] http://www.cl.haifa.ac.il/projects/quran/index.html

syntactic relations etc. Once this endeavor is complete, the user interface will provide access to syntactic information such as clausal and phrasal boundaries, phrase function etc.

Furthermore, we are currently extending the coverage of the system from the Qur'anic corpus to other texts of Classical Arabic. We have a significant corpus of Classical Arabic poetry (roughly half the size of the Qur'an corpus) which is already transcribed in our notation. We develop a variant of the rules which is aimed at this corpus; in other words, some of the morphological rules are marked as valid for Arabic in general, while others are corpus specific. We can thus generate two morphological analyzers, one for each of the corpora. This endeavor will lead to linguistic investigations concerning the similarities and differences between the two corpora.

Another dimension of extensions which we plan to pursue in the future involves enrichment of the annotation, especially by semantic and extra-linguistic information. For example, tagging of verses, groups of verses, suuras, and whole parts of the Qur'an according to chronological order (there are at least four, basically similar, established scientific systems) will add an important dimension to statistical studies based on our system. Regular use of certain (groups of) language patterns is indicative of the relevance of suggested chronological divisions of the text according to Muhammad's detailed biography. A combination of lexical and morpho-syntactic elements with chronological information is a significant contribution to the critical study of the Qur'anic language as a mirror of Muhammad's life. We hope that these extension will be instrumental in investigating the stylistic structure of the Qur'an, the text's history and its syntactic intricacies, and will eventually contribute to our understanding of its contents.

acknowledgements

We are grateful to Gal Goldschmidt and Eden Orion for their technical support in setting up this system. We benefitted greatly from the support of Xerox Research Center Europe, and in particular from the help of Ken Beesley and Ágnes Sandor. The work of the two last authors was supported by the Israeli Science Foundation, grants no. 136/01 and 745/99, as well as by a grant from The Caesarea Edmond Benjamin de Rothschild Foundation Institute for Interdisciplinary Applications of Computer Science at the University of Haifa.

References

Abd al-Baaqii, M. F. (1987). *al-Mu&jam al-mufahras li-'alfaaZ al-qur'aan al-kariim*. Cairo: Dar wa-Matabi' al-Sha'b.

Abney, S. (1996). Partial parsing via finite-state cascades. In *Workshop on Robust Parsing, 8th European Summer School in Logic, Language and Information*, Prague, Czech Republic, pp. 8–15.

Al-Shalabi, R. and M. Evens (1998, August). A computational morphology system for Arabic. In M. Rosner (Ed.), *Proceedings of the Workshop on Computational Approaches to Semitic languages*, Montreal, Quebec, pp. 66–72. COLING-ACL'98.

Ambros, A. A. (1987). Eine lexikostatistik des verbs im Koran. *Wiener Zeitschrift für die Kunde des Morgenlandes 77*, 9–36.

Beesley, K. (1996). Arabic finite-state morphological analysis and generation. In *Proceedings of COLING-96, the 16th International Conference on Com-*

putational Linguistics, Copenhagen.

Beesley, K. (1998a, April). Arabic morphological analysis on the internet. In *Proceedings of the 6th International Conference and Exhibition on Multilingual Computing*, Cambridge.

Beesley, K. R. (1998b, August). Arabic morphology using only finite-state operations. In M. Rosner (Ed.), *Proceedings of the Workshop on Computational Approaches to Semitic languages*, Montreal, Quebec, pp. 50–57. COLING-ACL'98.

Beesley, K. R. (2001, July). Finite-state morphological analysis and generation of Arabic at Xerox Research: Status and plans in 2001. In *ACL Workshop on Arabic Language Processing: Status and Perspective.*, Toulouse, France, pp. 1–8.

Beesley, K. R. and L. Karttunen (2000, August). Finite-state non-concatenative morphotactics. In *Proceedings of the fifth workshop of the ACL special interest group in computational phonology, SIGPHON-2000*, Luxembourg.

Beesley, K. R. and L. Karttunen (2003). *Finite-State Morphology: Xerox Tools and Techniques*. Stanford: CSLI.

Berg, H. (2001). Computers and the Qur'ān. In J. D. McAuliffe (Ed.), *Encyclopaedia of the Qur'ān*, Volume One, pp. 391–395. Leiden–Boston–Köln: Brill.

Berri, J., H. Zidoum, and Y. Atif (2001). Web-based Arabic morphological analyzer. In A. Gelbukh (Ed.), *CICLing 2001*, Number 2004 in Lecture Notes in Computer Science, pp. 389–400. Berlin: Springer Verlag.

Blau, J. (1973). Remarks on some syntactic trends in Modern Standard Arabic. *Israel Oriental Studies 3*, 172–231.

Blau, J. (1976). Some additional observations on syntactic trends in Modern Standard Arabic. *Israel Oriental Studies 6*, 158–190.

Böwering, G. (2000). Chronology and the Qur'an. In J. D. McAuliffe (Ed.), *Encyclopaedia of the Qur'ān*, Volume I, pp. 316–335. Leiden–Boston–Köln: Brill.

Chouémi, M. (1966). *Le verbe dans la Coran*. Paris: Klincksieck.

Darwish, K. (2002, July). Building a shallow Arabic morphological analyzer in one day. In M. Rosner and S. Wintner (Eds.), *Computational Approaches to Semitic Languages, an ACL'02 Workshop*, Philadelphia, PA, pp. 47–54.

DuBois, P. (1999). *MySQL*. New Riders.

Karttunen, L., J.-P. Chanod, G. Grefenstette, and A. Schiller (1996). Regular expressions for language engineering. *Natural Language Engineering 2*(4), 305–328.

Kiraz, G. A. (1998). Arabic computational morphology in the West. In *Proceedings of the 6th International conference and Exhibition on Multi-Lingual Computing*, Cambridge.

Kiraz, G. A. (2000, March). Multitiered nonlinear morphology using multitape finite automata: a case study on Syriac and Arabic. *Computational Linguistics 26*(1), 77–105.

Talmon, R. (2001). Grammar and the Qur'ān. In J. D. McAuliffe (Ed.), *Encyclopaedia of the Qur'ān*, Volume II, pp. 345–369. Leiden–Boston–Köln:

Brill.

A Characteristics of the Transcription

Rather than use the standard Arabic script, our system uses a phonemic transcription of the text, in which some of the ambiguity is reduced. The transcription is based on pure ASCII notations, largely with single-symbol equivalents of the Arabic graphemes, and double letters expressing long vowels. A conversion table of the consonants and vowels is given in figure 4.

The conventions of the Arabic orthography are basically retained, e.g., one-letter particles which are prefixed to the noun or verb are hyphenated to the following word (*wa-kaana* "and was"), as are pronominal and case/mood suffixes (*yas'al-u-nii* "he will ask-indicative-me"). In general, hyphenation serves to isolate noun bases from the various affixes. This process is practically inapplicable for Arabic verbal forms, whose complexity calls for creation of a detailed set of derivation rules instead.

Our transcription is largely phonemic; e.g., we treat the article invariably as *l-*, regardless of its assimilation with the first consonant of the following noun (hence, *l-Daall-iina* "those who lost their way", not *D-Daall-iina*, etc.). Nevertheless, we record in our transcription major phonetic peculiarities, especially those reflected in the traditional script. Included are:

- Assimilatory and dissimilatory processes, e.g., *taSaddaqa* → *SSaddaqa* "gave alms"; *yahtadii* → *yahiddii* "behaves righteously"; *min-maa* → *mim-maa* "from which"; *'an-laa* → *'al-laa* "so that not"; *tata(sh)aqqaq-u* → *ta(sh)aqqaq-u* "it will split"; etc.;

’	ء	D	ض
b	ب	T	ط
t	ت	Z	ظ
(th)	ث	&	ع
j	ج	g	غ
H	ح	f	ف
x	خ	q	ق
d	د	k	ك
(dh)	ذ	l	ل
r	ر	m	م
z	ز	n	ن
s	س	h	ه
(sh)	ش	w	و
S	ص	y	ي
a	ـَ	aa	ـَا
u	ـُ	uu	ـُو
i	ـِ	ii	ـِي

Figure 4: Conversion table of the Arabic characters

- Several pausal forms (prolongation of vowel at verse-end), e.g., *l-Zunuun-a* $\rightarrow$ *l-Zunuun-aa* “the evil thoughts”;
- Shortening of long vowels in word endings, e.g., *fa-rhabuu-nii* $\rightarrow$ *fa-rhabuu-*

ni "so fear Me", *l-daa&ii* → *l-daa&i* "he who calls (God) in prayer", *sa-nad&uu* → *sa-nad&u* "we shall call".

- Other peculiarities, e.g., omission of the verb's final *i*, without restoration of the original *-hu* of the dependent pronoun: *wa-yattaqi-hi* → *wa-yattaq-hi* "and he fears Him"; *laakin 'anaa* → *laakin-naa* "but I".

We have not included in our transcription system the following phonetic distinctions: (1) third person masculine singular dependent pronoun differences of *uu* vs. *u*, *ii* vs. *i*; (2) assimilations, e.g. *-un* → *-u*+assimilated consonant (as in Q9:27: *gafuur-un raHiim-un*, not the assimilated *gafuur-ur raHiim-un*); (3) distinctions based on the many instructions found in the *tajwiid* literature, such as those referring to special quality of alophonic variations; (4) such oppositions as *'alif maqSuura bi-Suurat 'alif* vs. *'alif maqSuura bi-Sururat yaa'*. Note also that the last three points in the above list of major phonetic peculiarities concern specific, largely sporadic, exceptional cases. They are included in our system because they represent orthographic oddities of the official Qur'anic script, which may reflect morphologically exceptional constructions.

Morphophonemic variations, especially in pronouns, are also preserved, e.g., *-ii, -iy, -i, -ya* (1st person singular dependent pronoun) and even the exceptional *[yaa bn-a 'umm]-a* "O son of my mother".

The hyphen isolates bound morphemes. It serves as an effective device in the formulation of the morphological generation rules. Its integration in the verb system is restricted; it marks the verb's system of mood morphemes (including the energicus), which are the verbal correlatives of the nominal system of case morphemes (with or without nunation), especially according to the Arabic grammatical

tradition. Note that according to our categorial classification active and passive participles (as well as the nomen actionis/infinitive) are included in the verbal group.

Reflection of the case/mood bound morphemes was our guideline in the special case of nouns and verbs with *w/y* third radical (and consequently with several other patterns with similar behavior). Absence of nominal case- or verbal mood-distinction yielded forms without hyphen, e.g., the nominal **l-hudy-u/a/i* → *l-hudaa* "the right way", and the verbal (indicative-subjunctive moods) **yabqay-u/a* → *yabqaa* "he will remain". Opposition between these and their pairs yield: *l-hudaa* "the right way" vs. *huda-n* "a right way" (the nunated form), *yabqaa* (both indicative and subjunctive) "he will remain, he may remain" vs. *yabqa* (jussive) "let him remain"; and cf. the non-nunated *'iHdaa* "one (f.)", *yusraa* "left (f.)", *yataamaa* "orphans", *'awlaa* "prior", etc.

Hyphenation helps to avoid ambiguity, in such cases as *'alaa* (vocative) vs. *'a-laa* (negative-interrogative "isn't it?"), *ta't-iina* "you will come, 2nd person singular feminine" vs. *ta'tiina* "you will come, 2nd person plural feminine", *tasaa'al-uu* "that you inquire, imperfect 2nd person plural masculine" vs. non-hyphenated *tasaa'aluu* "perfect 3rd person plural masculine", also "imperative 2nd person plural masculine".