Morphological Tagging of the Qur'an

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

Abstract

We present a computational system for morphological tagging 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 tagging 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 Objectives and overview

We present a system for morphological tagging of the Qur'an, for research and teaching purposes. The system aims at providing the linguist interested in Qur'anic syntax a tool, by which queries can be made which enable search of intricate syntactic relations in the Qur'an.

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.

The system 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 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; for example, it facilitates linguistic and literary analyses of the Qur'anic text, and is instrumental in exploring aspects of its syntax, semantics and style.

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

The core of the system consists in morphological tagging of the text, automated using a finite-state based toolbox (Beesley and Karttunen, 2003). 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 tagged Qur'an, present queries and collect information in a structured manner. In the future, we intend to use the morphologically tagged corpus in order to construct a shallow parser for the Qur'an, again using finite-state technology, thus augmenting the database with syntactic annotations.

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 Arabic and its leading literary texts.
- The model is applicable for computerized study of other corpora, in fact of the whole Classical Arabic literature.
- 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 tagged Qur'an facilitates study of other language aspects of this text, especially its style.

In the next section we discuss the methodology we used. Section 3 describes the details of the system, and some results of its usage are listed in section 4. We discuss related work in section 5 and conclude with suggestions for further research.

2 Methodology

2.1 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. 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. As an example of the transcription, figure 1 lists the seven verses of the first suura.

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

2.2 Computational morphological analysis

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 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. Using a concordance (Abd al-Baaqii, 1987), we manually constructed full lists of the words in the first two classes: our lexicon contains a few hundreds closed-class words and approximately 2500 noun bases. As the number of verbal bases is substantially greater, we decided to generate all possible verb bases automatically by interdigitating all the verbal roots of the Qur'an with all the verbal patterns. The over-generation problem is solved by intersecting the result with the corpus (see section 3.1).

Once the lexicographic work is done, most of the effort lies in the specification of the morpho-phonemic rules. We use a finite-state based toolbox (XFST, Beesley and Karttunen (2003)) which facilitates the stipulation of the rules. The rules are then compiled into finite-state transducers which constitute the morphological analyzer.

The use of XFST enabled us to avoid the bottleneck of having to tag the corpus manually. Furthermore, using a finite-state toolbox such as XFST 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.

3 Description of the system

3.1 Lexicon

The lexicon consists of three parts: closed-class words, noun bases and verb bases. 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* ("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.

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 (because such forms are listed explicitly). The lexicon associates with each lexeme its root and pattern. Typical entries are:[2]

```
swr+fu&lat:suurat   NounEndingFem;
Hmd+fa&l:Hamd       NounEnding;
```

The former specifies that *suurat* ("Qur'an chapter") is a noun whose root is *s.w.r* and whose pattern is *fu&lat.* Furthermore, it can be suffixed with feminine noun ending affixes. The latter indicates that *Hamd* ("praise") is a regular (masculine) noun whose root is *H.m.d* and whose pattern is *fa&l.* As the root and the pattern are listed explicitly with each noun, our analyzer can provide this information in the output. The nouns lexicon contains approximately 2500 entries.

As was the case with the previous group, 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").

The verbs lexicon is the most complicated.

[2] The examples use the syntax of the LEXC toolbox.

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); 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 918 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). However, 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. Therefore, we simply generate all the potential verb bases, inflect them in all the possible inflections and eventually intersect the results with the actual word forms of the corpus. In this way, most of the artificial forms disappear and the remaining ones contribute only mildly to the degree of ambiguity.

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; these, too, are handled with the rule component of the system.

As an example of how such rules work, 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:[3]

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

Similarly, 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 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];
```

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: $R_2R_2C \rightarrow R_2VR_2C$; or the omission of the *w* in prima-w roots, as well as the other phenomena associated with this paradigm; etc.

[3] The examples use the syntax of the XFST toolbox.

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*:

```
{%-hum} -> {%-him} ||
    [ii | i | y] _ [%- | .#.];
[%- h u m aa] -> [%- h i m aa] ||
    [ii | i | y] _ [%- | .#.];
{%-hunna} -> {%-hinna} ||
    [ii | i | y] _ [%- | .#.];
{%-hu} -> {%-hi} ||
    [ii | i | y] _ [%- | .#.];
```

3.3 Morphological analysis

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.

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). Examples of analyses are provided in Figure 2. 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. However, the average number of analyses per word in our corpus is only 1.8, and many of the words are assigned a unique analysis.

The results of the analysis are stored in a database in a way that encodes, for each analyzed word, its morphological features and their values. 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 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 220,000, with more than 500,000 connecting arcs. The size of the network file is only 2Mb. We use the network to analyze the entire corpus; on an ordinary personal computer, this takes approximately 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.

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, sub-queries refer each 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, case, aspect etc. Furthermore, 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.

Once the specification of constraints is done, a

```
suurat-u        swr+fu&lat+Noun+Fem+Sg+Nom
l-faatiHat-i    Def+ftH+Verb+Stem1+ActPart+Fem+Pron+Dependent+1P+Sg
l-faatiHat-i    Def+ftH+Verb+Stem1+ActPart+Fem+Sg+Gen
bi-sm-i         b+Prep+sm+Noun+Masc+Sg+Pron+Dependent+1P+Sg
bi-sm-i         b+Prep+sm+Noun+Masc+Sg+Gen
llaah-i         llaah+ProperName+Gen
l-raHmaan-i     Def+rHm+fa&laan+Noun+Masc+Sg+Pron+Dependent+1P+Sg
l-raHmaan-i     Def+rHm+fa&laan+Noun+Masc+Sg+Gen
l-raHiim-i      Def+rHm+fa&iil+Noun+Masc+Sg+Pron+Dependent+1P+Sg
l-raHiim-i      Def+rHm+fa&iil+Noun+Masc+Sg+Gen
l-Hamd-u        Def+Hmd+fa&l+Noun+Masc+Sg+Nom
li-llaah-i      l+Prep+llaah+ProperName+Gen
rabb-i          rbb+fa&l+Noun+Masc+Sg+Pron+Dependent+1P+Sg
rabb-i          rbb+fa&l+Noun+Masc+Sg+Gen
l-&aalam-iina   Def+&lm+faa&al+Noun+Masc+Pl+Obliquus
maalik-i        mlk+Verb+Stem1+ActPart+Masc+Pron+Dependent+1P+Sg
maalik-i        mlk+Verb+Stem1+ActPart+Masc+Sg+Gen
maalik-i        mlk+Verb+Stem3+Imperative+2P+Sg+Masc+NonEnergicus+HelpingVowel
yawm-i          ywm+fa&l+Noun+Masc+Sg+Pron+Dependent+1P+Sg
yawm-i          ywm+fa&l+Noun+Masc+Sg+Gen
l-diin-i        Def+dyn+fi&l+Noun+Masc+Sg+Pron+Dependent+1P+Sg
l-diin-i        Def+dyn+fi&l+Noun+Masc+Sg+Gen
'iyyaa-ka       'iyyaa+Particle+Pron+Dependent+2P+Sg+Masc
na&bud-u        &bd+Verb+Stem1+Imp+Act+1P+Pl+Fem+NonEnergicus+Jussive+HelpingVowel
na&bud-u        &bd+Verb+Stem1+Imp+Act+1P+Pl+Masc+NonEnergicus+Jussive+HelpingVowel
na&bud-u        &bd+Verb+Stem1+Imp+Act+1P+Pl+Fem+NonEnergicus+Indic
na&bud-u        &bd+Verb+Stem1+Imp+Act+1P+Pl+Masc+NonEnergicus+Indic
wa-'iyyaa-ka    wa+Particle+Conjunction+'iyyaa+Particle+Pron+Dependent+2P+Sg+Masc
nasta&iin-u     &yn+Verb+Stem10+Imp+Act+1P+Pl+Fem+NonEnergicus+Indic
nasta&iin-u     &yn+Verb+Stem10+Imp+Act+1P+Pl+Masc+NonEnergicus+Indic
nasta&iin-u     &wn+Verb+Stem10+Imp+Act+1P+Pl+Fem+NonEnergicus+Indic
nasta&iin-u     &wn+Verb+Stem10+Imp+Act+1P+Pl+Masc+NonEnergicus+Indic
hdi-naa         hdy+Verb+Stem1+Imperative+2P+Sg+Masc+NonEnergicus+Pron+Dependent+1P+Pl
l-SiraaT-a      Def+SrT+fi&aal+Noun+Masc+Sg+Acc
l-mustaqiim-a   Def+qwm+Verb+Stem10+ActPart+Masc+Sg+Acc
```

Figure 2: Example analyses

button enables the generation of an SQL query, which can be further edited manually by sophisticated users. Finally, a button submits the query to the database; the result is a list of all the occurrences in the corpus of words which satisfy all the constraints. Each occurrence is preceded by its 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 a separate window, 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.

4 Results

Our system performs a full morphological analysis of the entire Qur'an. As we do not have a manually tagged and verified subset, we cannot evaluate the accuracy of our analyzer automatically. However, we conducted systematic manual verification of the analyses as the project developed. As the system is being used, more inaccuracies are detected, but our experience shows that they deal mainly with missing or erroneous lexical items (and, sometimes, spelling errors in the transcribed Qur'an), and very rarely with inaccurate rules. Of course, lexical omissions and modifications are easy to handle. As far as we know, our transcription of the Qur'an is now error-free.

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. We have used the system successfully for a variety of tasks, including:

Morphological studies with syntactic implications For example, a study of morphological stipulations of syntactic agreement between heads and their complements. Is there any rule in the selection of *kafarat* (pattern *fa&alat*) vs. *kuffaar* (pattern *fu&&aal*) as plurals of *kaafir* "apostate"? Does this choice affect the properties of nominal adjuncts and modifiers of the head? Other examples include parallel use of different infinitive (nomen actionis) patterns of the same root; and investigation of nominalization of verbs in the Qur'an (which is either morphological or syntactic.)

Efficient retrieval of syntactic constructions Our morphologically tagged Qur'an can retrieve selected syntactic phrases or drastically reduce the possibility of non-relevant occurrences. For example, Classical Arabic uses an equational sentence of the pattern Demonstrative + Independent ("copular") pronoun + noun prefixed by an article or suffixed by dependent pronoun, e.g.,

'uulaa'ika	*hum-u*	*l-kafarat-u*
Those	are	the-apostates-Nom

"these are the apostates"

Accurate retrieval of this specific structure is enabled by the following query: "Show all Demonstratives followed by 3rd person independent pronouns, which are followed immediately by an article".

Historical and stylistic investigations The prophetic and political message which constitutes the Qur'an is traditionally divided by scholars according to the presumed Meccan and Medinan phases of Muhammad's activity. Study of the language of the Qur'anic text 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. Such queries are readily supported by our system.

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.

5 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 et al., 2001). A major drawback of some systems is limited coverage; for example, Al-Shalabi and Evens (1998) only deal with verbs and deverbal forms. In contrast, our system is capable of providing full coverage of the corpus it is designed for, and extensions to larger corpora is mainly 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; the language of the Qur'an is Classical Arabic. In many respects it is unique, and its lexicon, morphology and syntax require dedicated attention. Furthermore, most systems for processing Arabic use the standard Arabic script, or a one-to-one transliteration thereof, as their input script, whereas our system uses a phonemic transcription of the text, in which some of the ambiguity is reduced. As Berg (2001) notes, "for the present, computer-assisted analysis of the Qur'¯an remains an intriguing but unexplored field." Our work is one step towards exploring this field.

6 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 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.

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.

Currently, the system only employs rules that refer to the morphology of the Qur'anic corpus. In the future, we intend to use the same technology and methodology in order to stipulate short-context and even more general syntactic rules, thus reducing the degree of morphological ambiguity and improving the quality of the output produced by the system. This will enable users to present queries which refer to syntactic notions, in addition to morphological ones. We hope that this 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 Dudu Shaharabani and Judith Dror for their extensive help in designing and implementing this system. Thanks are due to Gal Goldschmidt and Eden Orion for technical support. This work was supported by the Israeli Science Foundation (grants no. 136/01 and 745/99).

References

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

Riyad Al-Shalabi and Martha Evens. 1998. A computational morphology system for Arabic. In Michael Rosner, editor, *Proceedings of the Workshop on Computational Approaches to Semitic languages*, pages 66–72, Montreal, Quebec, August. COLING-ACL'98.

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

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite-State Morphology: Xerox Tools and Techniques*. Cambridge University Press.

Ken Beesley. 1996. Arabic finite-state morphological analysis and generation. In *Proceedings of COLING-96, the 16th International Conference on Computational Linguistics*, Copenhagen.

Ken Beesley. 1998a. Arabic morphological analysis on the internet. In *Proceedings of the 6th International Conference and Exhibition on Multi-lingual Computing*, Cambridge, April.

Kenneth R. Beesley. 1998b. Arabic morphology using only finite-state operations. In Michael Rosner, editor, *Proceedings of the Workshop on Computational Approaches to Semitic languages*, pages 50–57, Montreal, Quebec, August. COLING-ACL'98.

Kenneth R. Beesley. 2001. 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.*, pages 1–8, Toulouse, France, July.

Herbert Berg. 2001. Computers and the Qur'ān. In Jane Dammen McAuliffe, editor, *Encyclopaedia of the Qur'ān*, volume One, pages 391–395. Brill, Leiden–Boston–Köln.

Jawad Berri, Hamza Zidoum, and Yacine Atif. 2001. Web-based Arabic morphological analyzer. In A. Gelbukh, editor, *CICLing 2001*, number 2004 in Lecture Notes in Computer Science, pages 389–400. Springer Verlag, Berlin.

Mustapha Chouémi. 1966. *Le verbe dans la Coran*. Klincksieck, Paris.

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

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