*A fully inflected Arabic verb resource constructed constructed from a lexicon of lemmas by using finite-state transducers**

ALEXIS Amid Neme

Laboratoire d'informatique Gaspard-Monge – LIGM, Université Paris-Est
77454 Marne-la-Vallée Cedex 2, France

alexis.neme@gmail.com

Abstract: *We describe a fully inflected lexicon of 2.5 million verbal forms generated by using finite-state transducers. The lexicon is constituted of 15 400 verbal entries or lemmas. The lexicon of Arabic verbs is constructed on the basis of Semitic patterns and used in a resource-based method of morphological annotation of written Arabic text. An enhanced FST implementation for Semitic languages was created. This system is adapted also for generating inflected forms. The language resources can be easily updated. We propose an inflectional taxonomy that increases the lexicon readability and maintainability for Arabic speakers and linguists. Traditional grammar defines inflectional verbal classes by using verbal pattern-classes and root-classes, related to the nature of each of the triliteral root-consonants. Verbal pattern-classes are clearly defined but root-classes are complex. In our taxonomy, traditional pattern-classes are reused and root-classes are simply redefined. Our taxonomy provides a straightforward encoding scheme for inflectional variations and orthographic adjustments due to assimilation and agglutination. We have tested and evaluated our resource against 10 000 diacriticized verb occurrences in the Nemlar corpus and compared it to Buckwalter resources. The lexical coverage is 99.9 %. A laptop needs two minutes in order to generate and compress the 2.5 million form lexicon into 4 Megabytes for fast retrieval. The analysis of a verb takes 0.5 millisecond.*

Résumé : *Nous décrivons un lexique complètement fléchi de 2,5 millions de formes verbales générées par des transducteurs à états finis. Le lexique est constitué de 15 400 entrées ou lemmes. Le lexique de ces verbes arabes est construit sur la base des schèmes de la grammaire traditionnelle. Cette ressource verbale est ensuite utilisée par un logiciel d'annotation morphologique du texte écrit en arabe. Un ajustement de l'implémentation de ces transducteurs a été spécialement crée afin de traiter les langues sémitiques. Ce système est également adapté pour générer des formes fléchies. Les ressources linguistiques peuvent être facilement mis-à-jour. Nous proposons une taxonomie de la flexion verbale qui augmente la lisibilité du lexique et la maintenabilité pour les locuteurs et linguistes arabes. La grammaire traditionnelle définit des classes de flexion verbales en utilisant des classes de schèmes et des classes de racines, liées à la nature de chacune des consonnes d'une racine trilitères. Les classes de schèmes verbaux sont clairement définies alors que les classes de racines sont complexes. Dans notre taxonomie, les classes de schèmes traditionnelles sont réutilisées et les classes de racines sont redéfinies de façon plus simple. Notre taxonomie fournit un schéma de codage simple des variations flexionnelles et des ajustements orthographiques dus à l'assimilation ou à l'agglutination d'une particule grammaticale. Nous avons testé et évalué notre ressource sur 10 000 occurrences voyellées de verbes extraites du corpus Nemlar et nous l'avons comparé à la ressource de Buckwalter. La couverture lexicale est de 99,9%. Un ordinateur portable a besoin de deux minutes pour générer et compresser les 2,5 millions de formes fléchies en 4 Méga-octets pour une recherche rapide. L'analyse d'un verbe prend 0,5 milliseconde.*

Keywords: *Arabic, Natural Language Processing, Semitic morphology, POS tagging, root and pattern.*
Mots clés *: Langue arabe, traitement automatique du langage naturel, morphologie sémitique, étiquetage grammatical, racine et schème.*

**Une ressource de verbes arabes entièrement fléchie constituée à partir d'un dictionnaire de lemmes à l'aide de transducteurs finis.*

© Revue RIST, vol. 20, n° 2, p.7-19

1 Introduction

No dictionary suitable for Arabic NLP is currently available.'Arabic spell checking is an active area of research since results are not satisfactory' (Shaalan et al., 2003:240). 'Although many research projects have focused on the problem of Arabic morphological analysis using different techniques and approaches, very few have addressed the issue of generation of fully inflected words for the purpose of text authoring' (Shaalan et al., 2012:719). 'The need for incorporating linguistic knowledge is a major challenge in Arabic Data-driven MT. Recent attempts to build data-driven systems to translate from and to Arabic have demonstrated that the complexity of word and syntactic structure in this language prompts the need for integrating some linguistic knowledge and with a minimum cost since the amount of linguistic resources added has consequences for computational complexity and portability' (Zbib, Soudi, 2012:2).

Arabic morphology can be described by many formal representations. However, Semitic morphologyor *root-and-pattern* morphology (Kiraz, 2004) is a natural representation for Arabic[1]. The *root* represents a morphemic abstraction, usually for a verb a sequence of three consonants, like *ktb*. A *pattern is* a template of characters surrounding the root consonants, and in which the slots for the root consonants are shown by indices. The combination of a root with a pattern produces a surface form. For example, *kataba* and *yakotubu* are represented by the root *ktb* and the patterns *1a2a3a* or *ya1o2u3u*.

Root-and-pattern morphology is standard in Arabic and is learned in grammar text books. Arabic linguists use *root-and-pattern* representation in order to list verbal entries and related inflected forms. On the other hand, FSTs have shown their simplicity and efficiency in inflectional morphology for western languages. Computer scientists appoint FSTs as standard devices for inflection.

Various formal representations for Arabic morphology have been created by computer scientists to avoid root-and-pattern representation. The point that motivated this trend is that FSTs formalism would not be fitted for root-and-pattern morphology since FSTs are concatenative whereas root-and-pattern morphology is not. In concatenative representation, the root-and-pattern representation is replaced by a stem- or lexeme-based representation. For these formalisms, a stem is a basic morpheme that undergoes affixations with other morphemes in order to form larger morphological or syntactic units. For root-and pattern morphology, a stem derives from a root and a particular pattern and subsequently undergoes affixations.

At the operational level, the lexical representation of the concatenative model is entirely concatenative in order to compel with the *[prefix][stem][suffix]* representation. However, these representations imply a manual stem precompilation based on a root-and-pattern representation. The concatenative models are generally composed of three components: lexicon, rewrite rules, and morphotactics. The lexicon consists of multiple sublexica, generally *prefix*, stem, and *suffix*. The rewrite rules map the multiple lexical representations to a surface representation. The morphotactics component aims with a subjacent representation to generate or to parse the surface form *[prefix][stem][suffix]* and performs alternation rules at morpheme boundaries such as deletion, epenthesis, and assimilation.

1 We would like to thank Eric Laporte and Sébastien Paumier for helpful discussions, contributions and for the adaptation of Unitex to Arabic. Unitex is an open source multilingual corpus processor. http://www-igm.univ-mlv.fr/~unitex.

Any formal representation that is not adapted to root-and-pattern morphology will be rejected by the majority of Arabic-speaking linguists. When linguists work in a newly created formalism, they continue to work with *root-and-pattern* representation on paper and subsequently, they unfold their descriptions for a specific formalism. Their contribution for updating and correcting lexical resources is complex and time-consuming, and therefore error-prone.

Our approach resorts to classical techniques of lexicon compression and lookup in an inflected full-form dictionary that includes orthographic variations related to morpheme agglutination. The formalization of all possible verbal tokens requires complex and interdependent rules. For these issues, we define a taxonomy for Arabic verbs composed of 460 inflectional classes. We demonstrate that FSTs are compatible with root-and-pattern representation. Our taxonomy encodes simultaneously in the lexical representation three variations at the surface level:

- inflectional classes of a lemma;
- inflectional subclasses related to morphophonemic assimilation;
- orthographic adjustments related to the agglutination of a pronoun.

In our orthographic representation, we use a fully diacriticized lexicon and we take advantage of the clear boundary, already defined in traditional grammar, between verbal inflection and verbal agglutination to describe these two levels independently. In order to satisfy both computer scientists and Arabic linguists, we have created in Unitex an enhanced version of FSTs adapted to root-and-pattern representation.

In Section 2, we outline the state-of-the-art approaches to Arabic morphological annotation. Section 3 describes the methodology and particularly the inflectional verbal taxonomy. Section 4 describes agglutination as morpheme combinatorics. Section 5 reports the construction of a fully inflected verb resource. Section 6 reports the evaluation of this resource. A conclusion and perspectives are presented in Section 7.

2 State of the Art

Several morphological annotators of Arabic are available. Beesley's (1996, 2001) system for Arabic inflection formalizes the traditional version of the root-and-pattern model and classifies in the root/pattern/rule approach. Its rules deal with root alternations, morphophonological alternations and spelling adjustments. They are encoded in the form of finite automata and compiled with the dictionary into a finite transducer. For morphological analysis, these rules are applied regressively, i.e. they take surface forms as input and they output deep forms.

The verbal lexical coverage is medium: 4 930 roots producing 90 000 stems. Nonetheless, this number of stems does not measure the number of entries because the formal model of the system does not include the notion of lexical entry (Beesley, 2001:7).

This system faces several challenges. One of them is that of analysis speed: 'the finite-state transducers (FSTs) tend to become extremely large, causing a significant deterioration in response time' (Altantawy et al., 2011:116).By the way, this wasthe main motivation for devising the multi-stem approach such as Buckwalter Arabic Morphological Analyzer (2002).

The MAGEAD system (Habash, Rambow, 2006; Altantawy et al., 2010, 2011) is close to Beesley's (2001) in its design: 'We use "deep" morphemes throughout, i.e., our system includes both a model of roots, patterns, and morphophonemic/orthographic rules, and a

complete functional account of morphology' (Altantawy et al., 2010:851); the rules are also compiled with the lexicon into a finite transducer. The lexicon is derived from Buckwalter's (Habash, Rambow, 2006:686). The project has an on-going part for nouns, including broken plurals (Altantawy et al., 2010).

MAGEAD improves upon Beesley (2001) in several ways. The notion of lexical entry is represented. The notion of inflectional class is adopted for patterns, but not for root alternations (Habash, Rambow, 2006:683): each lexical entry is assigned a code that identifies the patterns it admits. There are 41 classes for verbs (Habash, Rambow, 2006:684). Thus, verbal inflectional information is shared at class level, reducing redundancy between entries. This facilitates dictionary checking, update and extension, reducing the cost of management of the dictionary: when an error is detected in the patterns of a class, the correction of the error affects all the class; when a new class is found and encoded, it can be shared by all the future members of the class through a simple code assignment.

However, MAGEAD still faces many problems:

- The resources of MAGEAD-Express compile in 48 h, and the analysis of a verb takes 6.8 ms (Altantawy et al., 2011:123);
- The analysis opts for deep roots, complexifying the computation of the root from the surface form;
- Root alternations are not taken into account in inflectional classes, but controlled by a single set of rules for all entries. Encoding such rules is a challenge: 'we also exclude all analyses involving non-triliteral roots and non-templatic word stems since we do not even attempt to handle them in the current version of our rules' (Altantawy et al., 2010:856).

In addition, the lexical coverage is still limited. The lexical data are borrowed from Buckwalter (2002): 8 960 verbs (Altantawy et al., 2011:122) and 32 000 nouns, including those with suffixal plural (Altantawy et al., 2010:854), but the rules are compatible only with triliteral broken plural nouns.

The open-source Alkhalil morphological analyser[2] (Boudlal et al., 2010) is used in various projects and won the first prize at a competition by the Arab League Educational, Cultural Scientific Organization (ALESCO) in 2010. We counted that Alkhalil's lexical resources cover 97% of the verb occurrences of our test sample (cf. section 6), which is comparable to the coverage of Buckwalter (2002). The patterns are scripted in Arabic. The system includes broken plural (BP). As in Beesley (2001), the output of the analyser does not identify lexical entries: nothing connects a noun in the BP to its singular. This deficiency in a definition of lexical entries in Alkhalil hinders, among others, the lexicon readability and maintainability for Arabic speakers and linguists.

For a complete survey of morphological parsers, readers are invited toconsider Al-Sughaiyer & Al-Kharashi (2004), and Habash (2010).

[2] http://sourceforge.net/projects/alkhalil/

3 Method of description

3.1 A taxonomy for verb inflection

Our method is based on a precompiled diacriticized full-form dictionary with all possible inflected forms and their orthographic variations due to morphophonemic alternations. We exclude from this inflectional representation agglutinated prefixes and suffixes such as conjunctions and pronouns. We associate morphosyntactic feature values to each entry in the generated list of 2.48 million surface forms. In order to obtain this list, we provide a list of lemmas manually associated to codes defined by a taxonomy, each code representing a transducer. The full-form list is produced after inflecting each lemma by applying the encoded transducer (Silberztein, 1998).

Arabic and other Semitic languages have long been described in terms of a root interwoven with a pattern. The root is a sequence of consonants. Each Arabic verb contains 3 or 4 consonants that remain generally unchanged in all conjugated forms and make up the consonantal root; all the remaining information on a conjugated form is called 'pattern'. For example, Beesley (1996) representsyakotubuwna by [ktb & ya1o2u3uwna] through the interdigitationof the root ktb with the patternof active-Perfect-3rdperson-masculine-plural-indicative ya1o2u3uwna. Below some precisions:

- Some root consonants change. They are the glottal stop, noted *h* in the taxonomy, and glides, noted *w*, *y*; those that never change are written in patterns in the form of their position 1, 2, 3 or 4.
- At the surface level, the orthographic representation of glottal stop and glides can change. The glottal stop is represented by six allographs depending on the context. At the phonological level, the glides become short vowels */i, u/* or long vowels */a:, i:, u:/* or are omitted and transcripted as *zero-vowel, o*[3] (see also footnote 4).
- A pattern indicates the position of its letters relative to the root consonants. Generally, these letters are vowels and/or affixes related to a derived verb form such as *IisotakotabuwA = [ktb & Iisota1o2a3uwA]*. The surface form may also be subdivided in *[prefix] [stem] [sufix]*. The *stem pattern* formalizes all infixation operations such as *kotub = [ktb & 1o2u3]*. Inflectional prefixes and suffixes can be concatenated subsequently to the stem form ya*kotubuwna = [ya] [ktb & 1o2u3] [uwna]*.
- The third root consonant can be identical to the second one. In the root, it is represented by a gemination mark *G,* and in the pattern, by *2,* such as *madadota = [mdG & 1a2a2ota]*.
- By convention, the perfect-3rd person-masculine-singular is the form used as lemma. The corresponding pattern is called the canonical pattern. All patterns are defined in function of the canonical pattern.

Verbal pattern classes are clearly defined in Arabic grammar but root-classes are intricate and involve a complex terminology. Root-classes are defined according to the nature of some of the root consonants: regular, weak, geminated, with glottal stop, and to their position 1, 2, 3 or 4. In this terminology, *qaAla/yaquwlu* قال "say" is a *hollow verb of w kind*, with a weak consonant *w* at the second position; whereas *baAEa/yabiyEu* باع "sell" is a *hollow verb of y kind*. Moreover, two or three special values of the root consonants can

[3] The zero-vowel marks the silent-vowel or the absence of vowel between two consonants.

appear at the same time. A verb like *OataY/yaOotiy* أتى "arrive" has a glottal stop at the first position and a weak consonant *y* at the third position. A classification with nature/position criteria and each with 4 sub-criteria yields to an intricate terminology and is not consensual in Arabic grammar.

Our classification is bi-dimensional like the traditional one and based on the traditional pattern-classes which are reused and root-classes which are redefined more simply. Traditional grammar defines an inflectional verbal class by a pattern-class and a root-class. Triliteral verbs are compatible with 16 possible canonical patterns and quadrilateral verbs with 4 canonical patterns. Our classification defines 31 root-classes. The root classes are defined according to the nature of the root consonants. The special values for the consonants are *w, y* and the glottal stop (h). An irregular root is a root with at least one special value in its consonants. The inflected forms of a verb are easily predictable on the basis of the features of the root. We revisited and simplified, with no loss of information, the root-based traditional classification by using three consonantic slots, noted *123,* except for special values: glottal stop (*h*), *w*, *y*, for each slot; and when the 3rd root consonant is identical to the 2nd, the slots are noted *122*. Thereby, the lemma *ktb* will be encoded *$V3au-123* where:

$ is the Semitic mode for FST which means the root consonants interdigitate into the pattern: *[ktb & ya1o2u3u]= yakotubu;*

V is the verbal POS;

3au is the class of triliteral verbs used with the patterns *1a2a3/ya1o2u3* for Perfect/Imperfect;

123 is the class of roots in which no slot is occupied by a special value.

Each root/canonical-pattern pair corresponds to a lemma. This representation seems well-founded and also well-established in Arabic morphology. Above all, it is ubiquitous in the Arabic-speaking world. Below, some examples from the entries of the lemma-based lexicon:

```
/Lemma,encoding/ canonical-patt. Special values
-----------------------------------------------
/ simple forms
نقض,$V3au-123    / 1a2a3a/ya1o2u3u   no special values
جرَ,$V3au-122     /          third root identical to second
عاد,$V3au-1w3     /          with waw as a second root
غفا,$V3au-12w     /          with waw as a third root
فتح,$V3aa-123     / 1a2a3a/ya1o2a3u
لمز,$V3ai-123     / 1a2a3a/ya1o2i1u
حاك,$V3ai-1y3     /          with yeh as a second root
سرى,$V3ai-12y     /          with yeh as a third root
أوى,$V3ai-hwy     /          with hamza, waw and yeh
علم,$V3ia-123     / 1a2i3a/ya1o2a3u
وطئ,$V3ia-w2h     /          waw and hamza as 1rst and 3rd
كرُم,$V3uu-123    / 1a2u3a/ya1o2u3u
حسب,$V3ii-123     / 1a2i3a/ya1o2i3u
/ Derived forms
أقبل,$V61-123     / Aa1o2a3a
دشّن,$V62-123     / 1a2Ga3a
د ا هم,$V63-123    / 1aA2a3a
إنشغل,$V64-123    / Iino1a2a3a
إنطلى,$V64-12y    /          with yeh as a third root
إختنق,$V65-123    / Ii1ota2a3a
إزهرّ,$V66-123    / Ii1o2a3Ga
تهاجن,$V67-123     / ta1aA2a3a
```

```
تـآكل,$V67-h23    /           with  hamza as a first root
تـحدّد,$V68-122   / ta1a2Ga2a with identical 3rd root
تـلكّأ,$V68-12h   /            with hamza as a third root
إستبسل,$V69-123  / Iisota12a3a
اعشوشب,$V70-123  / Ii1o2aw2a3a
```

The format of the lexicon is a list of lemma entries. In our format, the string before comma transcribes plain letters and the gemination mark but no short vowel diacritics. The pattern includes the encoding of short vowels (*a, i, u).* This transcript choice is consistent with usual practice in traditional paper dictionaries.

Our full-form lexicon is produced by FSTs. The FST output format is *surface-form,lemma.V:feature-values* such as *:*

```
تَـكْـتُـبُ,كـتب.V:aI3fsN   /active-Imperfect-3rd pers-fem-sing-
iNdicative (marfuuE)
```

The *feature values* are:

- Voice: active (a), passive (b);
- Tense: Perfect, Imperfect, Imperative (Y);
- Person: 1, 2, 3;
- Gender: masculine, feminine;
- Number: singular, dual, plural;
- Mode: indicative (N), Subjunctive, Jussive, Energetic.

The FSTs generate a huge list of surface forms. Below, some fragments of the 2.48 million lines representing the inflected lexicon:

```
/full-form,Lemma.V:inflectional-features
-----------------------------------------------
تَـكْـتُـبُ,كـتب.V:aI3fsN
تَـكْـتُـبَ,كـتب.V:aI3fsS
تَـكْـتُـبْ,كـتب.V:aI3fsJ
تَـكْـتُـبَنِّ,كـتب.V:aI3fsE
تَـكْـتُـبُنَّ,كـتب.V:aI2mpE
تَـكْـتُـبُوا,كـتب.V+nopro:aI2mpS
تَـكْـتُـبُو,كـتب.V+pro:aI2mpS
تَـكْـتُـبُوا,كـتب.V+nopro:aI2mpJ
تَـكْـتُـبُو,كـتب.V+pro:aI2mpJ
إكْـتُـبَنِّ,كـتب.V:Y2msE
إكْـتُـبِي,كـتب.V:Y2fsJ
أكْـتُـبُوا,كـتب.V+nopro:Y2mpJ
أكْـتُـبُو,كـتب.V+pro:Y2mpJ
أكْـتُـبْنَ,كـتب.V:Y2fp
ثُـكْـتَـبُ,كـتب.V:bI3fsN
ثُـكْـتَـبُو,كـتب.V+pro:bI2mpJ
ثُـكْـتَـبُونَ,كـتب.V:bI2mpN
كَـتَـبْتُ,كـتب.V:aP1s
كَـتَـبْنَا,كـتب.V:aP1p
كَـتَـبَتِ,كـتب.V:aP3fs
كُـتِـبَ,كـتب.V:bP3ms
كُـتِـبَا,كـتب.V:bP3md
كُـتِـبُوا,كـتب.V+nopro:bP3mp
كُـتِـبُو,كـتب.V+pro:bP3mp
...
يُـهَـيِّـجُوا,هيّج.V+nopro:aI3mpS
يُـهَـيِّـجُو,هيّج.V+pro:aI3mpS
يُـهَـيِّـجُوا,هيّج.V+nopro:aI3mpJ
يُـهَـيِّـجُو,هيّج.V+pro:aI3mpJ
يُـهَـيِّـجُونَ,هيّج.V:aI3mpN
يُـهَـيِّـجْنَ,هيّج.V:aI3fp
```

```
....
إستقرأ,تَسْتَقْرَأُ.V+nopro:aI3fsN
إستقرأ,تَسْتَقْرَأَ.V:aI3fsS
إستقرأ,تَسْتَقْرَأُ.V:aI3fsJ
إستقرأ,تَسْتَقْرَأَنَّ.V:aI3fsE
إستقرأ,تَسْتَقْرَؤُ.V+pro:aI3fsN
...
```

In the following sub-section, we present the inflectional transducers.

3.2 The inflection transducers

An inflection transducer specifies the inflectional variations of a word. It is shared by the class of words that inflect in the same way. The input parts of the transducer encode the modifications that have to be applied to the canonical forms. The corresponding output parts contain the codes for the inflectional features. A transducer is represented by a graph and can include subgraphs. The transducers are displayed in Unitex style, i.e. input parts are displayed in the nodes, and output parts below the nodes (Neme, 2011 - Fig. 2). A Buckwalter transliteration is used as a standard to map Arabic characters into Latin ones. An XML-friendly version of this transliteration was created in order to handle this format. We create a new XML-friendly version where all special characters such as (', |, * , $, ~) are respectively replaced by (c, C, J, M, G) [4]. Many systems use special characters in a special way.

In order to generate the full-form dictionary, the following steps are accomplished.

- The lemma lexicon is transliterated.
- The FSTs are applied to the list and produces a transliterated full-form dictionary output.
- The output is transliterated into Arabic script.
- So, both the lemma lexicon and the full-form dictionary are in Arabic script which is handier to read for Arabic linguists.

For example, the lexical entry *ktb,$V3au-123* is processed by the transducer named *V3au-123* in order to get all inflected forms. The main graph contains five subgraphs referring to the five voice-tense variations (Neme, 2011 - Fig. 1). In turn, each subgraph contains suffixes of Person, Gender, Number for the perfect and Person, Gender, Number, Mode for the Imperfect (Neme, 2011 - Fig. 2).

4 Agglutination and omission of diacritics

4.1 Orthographic adjustments and agglutination

In Arabic, a token delimited by spaces or punctuation symbols is composed of a sequence of segments. Each segment in a token is a morpheme. In Unitex, this segmentation is formalized via a morphological dictionary graph. Such graphs introduce morphological analyses in the text automaton (Fig. 1) where dashed lines connect segments.

The combination of a sequence of morphemes obeys a number of constraints. Checking these constraints is necessary to discard wrong segmentations. In Arabic, a verbal token is composed by one morpheme <*V*> or the concatenation of up to 4 morphemes such as:

[4] This transliteration is called the Buckwalter-Neme code. and adopted in Unitex to map Arabic <=> Latin: ء, c; آ, C; أ, O; ؤ, W; إ, I; ئ, e; ا, A; ب, B; ة, p; ت, T; ث, V; ج, J; ح, H; خ, x; د, d; ذ, J; ر, r; ز, z; س, s; ش, M; ص, S; ض, D; ط, T; ظ, Z; ع, E; غ, g; ف, f; ق, q; ك, k; ل, l; م, m; ن, n; ه, h; و, w; ى, Y; ي, y; ً, F; ٌ, N; ٍ, K; َ, a; ُ, u; ِ, i; ّ, G; ْ, o.

<CONJC> <CONJS> <V> <PRO+Accusative>

Where <CONJC> is a coordinating conjunction, <CONJS> is a subordinating conjunction and *<PRO+Accusative>* an agglutinated object pronoun.

<CONJC> combines freely with any inflected verb. The *<CONJS>* constrains the verb to the Imperfect Subjunctive or Jussive. Finally, an inflected verb form is often insensitive to the agglutinated pronoun but some forms are sensitive like verbs in the active-Perfect-3rd person-masculine-plural and forms with a glottal stop as the third root consonant.

The subgraph selects only V+pro variants from the full-form dictionary (Fig. 1). When followed by a pronoun, a verbal segment may have an orthographic adjustment. This is often the case when the verbal segment ends with a long */a:/ A,* its allograph *Y*, or a glottal stop. The glottal stop has 6 allographs depending on its position and the surrounding vowels. For verbs, the roots with a glottal stop as the third consonant change their graphemic representation. A suffix subgraph related to classes Vpp-rrh represents the orthographic variations of an ending glottal stop due to pronoun agglutination.

In the middle box of Fig. 1, *ItGhm* إِتَّهم identifies the lemma of the inflected form. The morphological dictionary graph restricts the selection to V+pro, which is the agglutinated variant *IitGahamuw* إِتَّهَمُو, without ending Alif. The V+nopro variant is with ending Alif, *IitGahamuwA* إِتَّهَمُوا. The *aP3mp* code means active-Perfect-3rd person-masculine-plural.

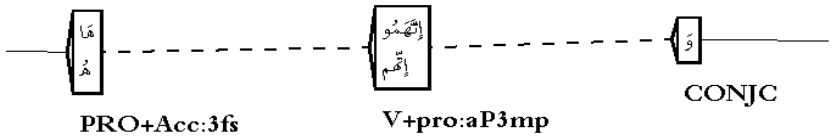

Figure 1: A morphological analysis of *wa_IitGahamuw-haA* وَاتَّهَمُوهَا *(and_suspect-they_her) "and_they_suspect_ her"*.
Text automaton as output of the application of a graph dictionary.
Dashed lines connect segments in the same token.

The generation of the agglutinable variants of an inflected verb is performed directly with a lexicon of words, which is another way to implement a rule. In fact, the dictionary graph links each morphological variant to the correct context, which also expresses a rule. The variants are generated during the compilation of the resources, not at analysis time as in rule-based systems in which a rule should compute each morphological variant at run time, then link each variant to the correct context. The advantage of our method is that it simplifies and speeds up the process of annotation.

4.2 Diacritics

Diacritics are often omitted in Arabic written text. According to our corpus study of 6930 tokens from Annahar newspaper, 209 tokens (3%) include at least a diacritic. 140 tokens (2 %) are with the *F* diacritic *(–an)*. 57 (0,8 %) are with gemination mark *G*, in which 49 (0.7 %) are related to a verbal form. 9 are with the short vowel *u*. For the *u* diacritic, 7/9 involve a passive verbal form. For the gemination diacritic, 49/57 involve a verbal form and are distributed as follows:

- 41 belong to class V62 and are *1a2**G**a3a* derived forms (فَعَّلَ).
- 5 to V68 and are *ta1a2**G**a3a* derived forms (تَفَعَّلَ).
- 2 to V65G and are *Ii1**G**a2a3a* derived forms (اِفْتَعَلَ).
- 1 to V3au and is *ya1o2**u**lu,* a triliteral simple form (فعَل يَفعُل).

Editors generally display diacritics for unusual forms such as passive verb forms. When displayed, they exclude misinterpretations. For verbs, diacritics are the short vowels *(a, i, u)* or the gemination mark followed by a short vowel. Arabic verbs can include a sequence of two diacritics: the gemination mark followed by a short vowel. In the case of two diacritics, diacritics omission is not totally free. One can omit the two diacritics or the last diacritic but never the gemination mark alone.

Consequently, processing written Arabic text should take into account undiacriticized and partially diacriticized text. A lookup procedure in Unitex[5] has been adjusted to deal with omission of diacritics in Arabic. This procedure finds in the diacriticized full-form dictionary all possible diacriticized candidate forms compatible with a given undiacriticized or partially diacriticized form. When a diacritic is present in a surface form, the lookup procedure excludes the candidates in the lexicon which do not have that diacritic at the same position.

5 Some figures

Our lexicon is composed of 15 400 entries. Each entry is inflected into 144 surface forms and in average 158 forms if we include orthographic variations due to agglutination. The size of the full-form dictionary is 2.48 million surface forms. The size of the full-form dictionary in plain text is 132 Megabytes in Unicode little Endian and it is compressed and minimized into 4 Megabytes which are loaded to memory for fast retrieval; and the analysis of a verb takes 0.5 millisecond (*versus* 6.8 ms for MAGEAD-Express). The generation, compression and minimization of the full-form lexicon lasts two minutes on a Windows laptop (*versus* 48 hours for MAGEAD-Express).

The number of main inflectional graphs is 460. Each main graph is composed of 5 subgraphs for voice-tense features variations, that is 2300 subgraphs. These subgraphs use also 540 suffix subgraphs related to person-gender-number-mode features. In all, the number of graphs and subgraphs is 3300 (460+2300+540), to be compared with nearly 100 graphs and subgraphs dedicated to the verbal inflection system for Brazilian Portuguese constructed also for Unitex (Muniz et al. 2005).

6 Testing and evaluation

6.1 Testing the lexical coverage

We have chosen the NEMLAR Arabic Written Corpus (Attia et al., 2005), first to improve our lexicon of verbs, and then to constitute our test collection. The Nemlar data consists of about 500 000 words of Arabic text from 13 genres. The text is provided in 4 versions: raw text, fully diacriticized text, text with Arabic lexical analysis, and text with Arabic POS-tags. The database was produced and annotated by RDI, Egypt, for the Nemlar Consortium.

The extraction of occurrences of verbs from "text with Arabic POS-tags" provided 50 000 occurrences of verbs. These occurrences were split in two disjoint parts: nearly 40 000 token occurrences (11 050 token types) for correcting the resource and a test

[5] The lookup procedure and the compression algorithm was adjusted for Semitic by Sébastien Paumier.

collection of 10 000 token occurrences (5 222 token types) for testing it after the correction stage. The test collection shows that 10 verb lemmas were missing in our lexicon[6]. Hence, the fault rate of the resource is 0.1% in this corpus.

6.2 Testing the morphological annotation

In order to test our lexicon on real texts, we selected also three documents totaling 3 550 tokens (about 10 pages) from the Nemlar Corpus and containing popular science about three topics: pollution and fishing in Egypt, earthquakes in the world, and quality of water. We used the documents in the fully diacritized version. Below, three concordances with morphological annotation.

The first concordance locates "all inflected forms of the lemma <Istxdm> (to use)". It has been produced by submitting the lexical mask "<إستخدم>", in Arabic script, to Unitex:

يَعْرِفُ أَنَّ الْكُلُورْ غَازٌ سَامٌّ اسْتُخْدِمَ كَأَحَدِ الْأَسْلِحَةِ الْكِيمْيَائِيَّةِ
وَنَفْتَحُ الصَّنَابِيرَ فِي بُيُوتِنَا لِنَسْتَخْدِمَهُ هَنِيئًا مَرِيئًا ، فَمَاءُ الصَّنَابِيرِ
تَنْظِيفِ أَدَوَاتِنَا. أَيْضًا نَسْتَخْدِمُ الْمَاءَ فِي الطَّبْخِ، فَكُلُّ الْأَكْلِ
وَالْفِيتَامِينَاتِ. وَنَحْنُ لَا نَسْتَخْدِمُ الْمَاءَ فِي الشُّرْبِ فَقَطْ فَالْمَاءُ
نَاصِرِ؛ وَبِسَبَبِ هٰذِهِ الْقُدْرَةِ نَسْتَخْدِمُ الْمَاءَ فِي التَّنْظِيفِ، سَوَاءً تَنْظِي

The second concordance locates "all occurrences of inflected form of verb in the subjunctive (*mansuub*) preceded by the subordinating conjunction *li-*". It has been produced by submitting the lexical mask "لِ<V:S>" to Unitex:

أَمْوَاجُ تِسُونَامِي شَرْقًا عَبْرَ الْمُحِيطِ الْهَادِئِ لِتَصِلَ إِلَى جَزِيرَةِ هِيلُو فِي هَاوَايْ، ثُمَّ
لْأَنَابِيبِ. وَنَفْتَحُ الصَّنَابِيرَ فِي بُيُوتِنَا لِنَسْتَخْدِمَهُ هَنِيئًا مَرِيئًا ، فَمَاءُ الصَّنَابِيرِ
بِأَنَّ التَّغْيِيرَ الْإِيجَابِيَّ لَمْ يَكُنْ لِيَتَحَقَّقَ بِدُونَ حُدُوثِ طَفْرَةٍ هَائِلَةٍ فِي
لَا يَكْفِي، فَمَا زَالَ فِي جَعْبَةِ الْكُلُورِ الْكَثِيرُ لِيُقَدِّمَهُ لَنَا. قَدْ يَقُولُ الْبَعْضُ

The third concordance locates "all occurrences of verb in the plural followed by an agglutinated pronoun". It has been produced by submitting the lexical mask "<V:p><PRO+Acc>" to Unitex:

أَجْسَادِنَا، وَمِنَ الْمَاءِ الَّذِي نَتَنَاوَلُهُ نَسْتَمِدُّ الْعَنَاصِرَ الْغِذَائِيَّةَ
الطَّبْخِ، فَكُلُّ الْأَكْلِ الَّذِي نَأْكُلُهُ يَدْخُلُ فِيهِ الْمَاءُ. بِاخْتِصَارٍ:
يِبِعُونَ الْمَاءَ وَكَأَنَّهُمْ صَنَعُوهُ !!وَلِأَنَّ الْمَاءَ أَحَدُ الْمَوَارِدِ
لَهُ أَمْلَاحًا مَعْدِنِيَّةً ثُمَّ يَبِيعُونَهُ عَلَى أَنَّهُ مَاءٌ مُعَبَّأٌ مِنَ الْآبَا
وَنَفْتَحُ الصَّنَابِيرَ فِي بُيُوتِنَا لِنَسْتَخْدِمَهُ هَنِيئًا مَرِيئًا ، فَمَاءُ الصَّنَابِيرِ
تَعِيدَ الْمَاءَ الصِّحِّيَّ النَّقِيَّ وَنَجْعَلَهُ رَخِيصًا -وَمَجَّانًا إِنْ أَمْكَنَ- لَا أَن
الْحُصُولَ عَلَى هَوَاءٍ نَقِيٍّ نَتَنَفَّسُهُ أَيْضًا؟؟ سُؤَالٌ قَدْ يُجِيبُ عَنْهُ الْ

Summing up, by using the fully inflected verb resource and a word-internal grammar that models agglutination (cf. section 4.1), Unitex is able to identify occurrences of all inflected forms of a verb, with their specific inflectional features, with and without agglutination.

[6] jzm,$V32-123; qrGZ,$V62-123; thrGb,$V68-123; rDb,$V33-123; kfl,$V34-123; tnAqM,$V67-123; sAb,$V32-1y3; zEq,$V33-123; DnG,$V32-1nn; tAh,$V32-1y3

7 A conclusion and perspectives

With our model for Arabic verbs, we constructed a fully inflected verbal resource of 2.48 million forms with the following features. A detailed and simple taxonomy is based on root-and-pattern representation. Lemma-based verbs are used as entries in the lexicon. FSTs are used to produce inflected forms. Agglutination is described independently from inflection. Our experimentation shows that the method outperforms state-of-the-art systems of Arabic morphological annotation.

We made language resources the central point of the problem. All complex operations were integrated among resource management operations. Morphological annotation of Arabic text is performed directly with a lexicon of words and without morphological rules, which simplifies and speeds up the process. The undiacriticized, partially and fully diacriticized Arabic text can be annotated excluding incompatible analyses.

All forms are stored in the resources, including spelling variants; roots and patterns are handled at surface level. The dictionary is compiled by finite transducers that combine roots, patterns and inflectional suffixes. Each of the 460 inflectional classes is assigned one of the transducers, which ensures that the management of classes is mutually independent. The encoding of a new verb amounts to assigning it an inflectional code.

We reuse traditional Semitic patterns and we provide a clear scheme for root-class encoding by avoiding intricate terms. Root-and-pattern representation facilitates our task in encoding the lexicon since it is a standard but also it helps to debug our transducers quickly which is not the case of a rule-based system.

This system shows a concern with the comfort and efficiency of human encoding, checking and update of dictionaries. NLP companies need easy procedures for dictionary management, because most projects involve a specific domain with a particular vocabulary, and terminology evolves constantly; in addition, dialects show lexical differences, which are relevant to speech processing if not for written text processing; finally, the main advantage of dictionary-based analysers is that they provide a way of controlling the evolution of their accuracy by updating the dictionaries. None of the other authors surveyed above mentions the objective of facilitating manual dictionary management, and we reported the weak points of their analysers in this regard. We identify the problem as belonging not only to computation and morphology, but also to NLP dictionary management, and consider language resources as the key point, as Huh & Laporte (2005). Our dictionaries are constructed and managed with the dictionary tools of the open-source Unitex system (Paumier, 2011).

This work opens the perspective to extend our methodology to inflection of nouns and adjectives, mainly to encode singular and the broken plural under the same lemma entry using Semitic patterns (Neme, Laporte, forthcoming). This extension could address, among others, ‘the issue of generation of fully inflected words for the purpose of text authoring’ (Shaalan et al., 2012).

8 References

Attia., M., Yaseen., M., Choukri., K. (2005). Specifications of the Arabic Written Corpus produced within the NEMLAR project, www.NEMLAR.org.

Altantawy, Mohamed; Habash, Nizar; Rambow, Owen; Saleh, Ibrahim (2010). Morphological Analysis and Generation of Arabic Nouns: A Morphemic Functional Approach. In Proceedings of the Language Resource and Evaluation Conference (LREC), Malta, pages 851-858.

Altantawy, Mohamed; Habash, Nizar; Rambow, Owen (2011). Fast Yet Rich Morphological Analysis. In Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing (FSMNLP), pages 116-124.

Beesley, Kenneth R. (1996). Arabic finite state morphological analysis and generation. In COLING'96, volume 1, pages 89– 94, Copenhagen, August 5-9. Center for Sprogteknologi. The 16th International Conference on Computational Linguistics, 1996.

Beesley, Kenneth R. (2001). Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001. In Proceedings of the ACL/EACL Workshop 'Arabic Language Processing: Status and Prospects', pages 1-8.

Boudlal, Abderrahim; Lakhouaja, Abdelhak; Mazroui, Azzeddine; Meziane, Abdelouafi (2010). Alkhalil Morpho SYS1: A Morphosyntactic Analysis System for Arabic Texts. International Arab Conference on Information Technology (ACIT).

Buckwalter, T. (2004). Issues in Arabic Orthography and Morphology Analysis. In Proceedings of theCOLING 2004. Workshop on Computational Approaches to Arabic Script-based Languages, pages 31–34.

Buckwalter Arabic Morphological Analyzer Version 1.0. (2002). LDC Catalog No.: LDC2002349.

Habash, N., Rambow, O. (2005). Arabic Tokenization, Morphological Analysis, and Part-of-Speech Tagging in One Fell Swoop. In Proceedings of the Conference of American Association for Computational Linguistics (ACL05).

Habash, N., Rambow, O. (2006). MAGEAD: A Morphological Analyzer and Generator for the Arabic Dialects. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pages 681–688, Sydney, Australia, July.

Habash, N. (2010). Introduction to Arabic Natural Language Processing. Morgan & Claypoll Publishers.

Huh, H.-G. Laporte É. (2005). A resource-based Korean morphological annotation system. In Proc. Int. Joint Conf. on Natural Language Processing, Jeju, Korea, 2005.

Kiraz,A.(2004):http://www.scribd.com/doc/46443095/Computational-Nonlinear-Morphology-With-Emphasis-on-Semitic-Languages-Studies-in-Natural-Language-Processing-9780521631969-41686

Muniz, Marcelo C.M., Maria das Graças V. Nunes, and Éric Laporte (2005). UNITEX-PB, a set of flexible language resources for Brazilian Portuguese. Workshop TIL'05. pp. 2059–2068.

Neme, Alexis (2011). A lexicon of Arabic verbs constructed on the basis of Semitic taxonomy and using finite-state transducers. In Proceedings of the International Workshop on Lexical Resources (WoLeR) at ESSLLI.

Neme, Alexis, Laporte Éric (2013). Pattern-and-root inflectional morphology: the Arabic broken plural.In Language SciencesVolume 40, November 2013, Pages 221–250.

Paumier, Sébastien (2011). Unitex 3.0 – User manual, University of Marne-la-Vallée.

Shaalan, Khaled, Allam Allam, Gomah AbdAllah(2003). Towards automatic spellchecking for Arabic. Conference on Language Engineering.

Shaalan, Khaled, Samih, Younes, Attia, Mohammed, Pecina, Pavel, &van Genabith, Josef (2012).Arabic Word Generation and Modelling for Spell Checking. Language Resources and Evaluation (LREC). Istanbul, Turkey. Pages: 719-725.

Silberztein, Max. (1998). INTEX: An integrated FST toolbox, in Derick WOOD, Sheng YU (éd.), Automata Implementation, p. 185-197, Lecture Notes in Computer Science, vol. 1436. Second International Workshop on Implementing Automata, Berlin/Heidelberg: Springer.

Al-Sughaiyer, Imad A., Al-Kharashi, Ibrahim A. (2004). Arabic Morphological Analysis Techniques: A Comprehensive Survey. In Journal of the American Society for Information Science and Technology, 55(3):189–213.

Zbib, Rabih; Soudi, Abdelhadi (2012). Introduction. Challenges for Arabic machine translation.In Soudi, Abdelhadi; Farghaly, Ali; Neumann, Günter; Zbib, Rabih (eds.), Challenges for Arabic Machine Translation, Natural Language Processing, 9, Amsterdam: Benjamins, p. 1-13.