Standard Arabic formalization and linguistic platform for its analysis

Slim MESFAR
LASELDI, Franche-Comté University, France
mesfarslim@yahoo.fr

This article describes the construction of a lexicon and a morphological description for standard Arabic. This system uses finite state technology to parse vowelled texts, as well as partially and not vowelled ones. It is based on large-coverage morphological grammars covering all grammatical rules.

Keywords: NLP, Arabic language, lexical analysis, lexicographic analysis, morphological analysis, morphological grammar.

1. INTRODUCTION

From the beginning of the sixties, and starting with the first automatic analyzer proposed by David Cohen, one of the first theorists of NLP [1], research has continued with natural language processing and especially the automatic treatment of the Arabic language. In 1983, with a minimalist morphological analysis, based on the theory that any Arabic form is generated using root and pattern, researchers developed the first two-level morphological analyzer for Arabic (Koskenniemi 1983); this work was included within the project ALPNET (Beesley and Buckwalter 1989) using finite-state technology allowing only the concatenation of morphemes in the morphotactics. Since 1996, the Xerox research centre has enhanced this system using an algorithm of automatic combination between roots and patterns generating stems; this research is based on the ALPNET's dictionaries which were, considerably rebuilt using the Xerox finite-state technology (Beesley 2001). This technology is computationally very efficient for natural-language-processing; it's used within the developmental environment NooJ (Silberztein 2006). The use of finite-state machines within NooJ was extremely attractive, they are used to generate and analyse several thousands of words per second. This linguistic platform will be described inside this paper as the tool used for vocabulary formalization and analysis of standard Arabic language.

2. ARABIC LANGUAGE DESCRIPTION

The Arabic language is a Semitic language showing two great characteristics which have been the subject of much research: agglutination and non vocalization.

In fact, most forms [2] in Arabic writing can correspond to a succession of one or more prefixes, a radical and one or more suffixes. Radicals themselves are forms which have been inflected or derived from lemmas. However, non vocalisation is due to a lack of short vowels in usual texts from which a high degree of ambiguity ensues. When they are present, short vowels are represented by diacritics which appear above or below the consonants that they follow. In theory, only the Coran, and children's book are fully vowelled; the automatic analysis of Arabic must be able to parse fully vowelled, partially vowelled and unvowelled texts.

To resolve these problems, we use finite state machines which we associate to dictionaries of lemmas.

3. NOOJ AND ARABIC ANALYSIS

NooJ is a linguistic developmental environment which can analyze texts of several million words in real time. It includes tools to construct, test and maintain large-coverage lexical resources, as well as morphological and syntactic grammars. Dictionaries and grammars are applied to texts in order to locate morphological, lexicological and syntactic patterns, remove ambiguities, and tag simple and compound words.

NooJ recognizes all Unicode encodings and the runtime code that applies lexical transducers to input strings is completely language independent. Thus the code that runs the Arabic morphological analyzer is exactly the same code that processes a dozen languages, including some Roman, Germanic, Slavic, Semitic and Asian languages, etc.

NooJ can build lemmatized concordances of large texts from Finite-State or Context-Free grammars, and can accordingly perform cascading transformation operations on texts, in order to annotate the text, or to generate paraphrases.

The NooJ lexical module that will be used throughout this paper relies on operators performing transformations inside strings, and morphological graphs describing grammatical rules for morphological analysis. Generally, transformations inside strings are based on use of some generic predefined commands:

- <B>: keyboard Backspace,
- <L>: keyboard Left arrow,
- <R>: keyboard Right arrow,
- <S>: delete/Suppress current char,
- <N> [3]: go to end of Next word form,
- <P>: go to end of Previous word form,
- <E>: Empty string

Although these commands are defined in advance into NooJ, we can make meaning restatement or add new commands; for Arabic, as a highly inflectional and derivational language, we had to define three new operators such as the <T> operator that checks if the last consonant within a noun is a "ة" (T - *Teh marbouta*) to replace it with a "ت" (t – *Teh maftouha*) in some inflexional or derivational descriptions. For example, when performing the dual form from "مَدْرَسَة" (madorasaTa – *a school*), we have to substitute the "ة" (T - *Teh marbouta*) with a "ت" (t – *Teh maftouha*) before suffixing with "يْنِ" (yoni) to obtain "مَدْرَسَتَيْنِ" (madorasatayoni – *two schools*). The two other added operators <M> and <Z> will be described in section 4.1

These instructions can be associated with two argument types; either a number (e.g. <L3>: go left 3 times) or a "W" (e.g. <LW>: go to beginning of word). These commands operate on a letter pile, they require a O(n) transformation time. So, they guarantee a correspondence between lemma and corresponding inflected form in a linear time.

Morphological graphs have a great interest in our researches; so we will use the NooJ graph editor to construct either morphological or syntactic grammar. These grammars are used to extract sequences of interest in texts, but also to describe various linguistic phenomena. They represent a group of input sequences and associate them with some output.

Grammars are depicted as finite-state transducers (FST) [4]. A morphological FST represents sequences of letters (that spell a word form), and then produces some kind of lexical information (part of speech, a set of semantic codes, etc.). A syntactic FST represents word sequences, and then produces some kind of linguistic information (its syntactic structure for example).

Considering that certain linguistic phenomena descriptions require the construction of very complicated grammars; they would be simplified and built using dozens of elementary graphs. At the same time, most of these elementary graphs ("local grammars") can be re-used in different contexts, for the description of many different linguistic phenomena. This re-use of constructed elementary graphs, by mentioning them in other graphs, will permit users to construct virtual libraries of graphs, and in so doing, will cover increasingly complex phenomena from morphology to syntax.

NooJ [5] is used as a linguistic platform, an information retrieval system, to teach second languages, as a terminological extractor, as well as to teach computational linguistics to students. (Silberztein 2005_1) and (Silberztein 2005_2).

Given the explosion of Arabic resources, especially on-line, with more than 20 000 Arabic sites on the Web and more than 300 million users, we recognized the necessity of developing an Arabic component for NooJ platform, which would allow us to process and take advantage of this readily available data. We started building Arabic NooJ module with the purpose of providing automatic analysis of texts written in standard Arabic. This module will be used to a better understanding of the Arabic language based on description of its vocabulary and its transformational syntax according to the theory of Chomsky (1971) and Harris (1985).

4. ARABIC LEXICON FORMALISATION

The linguistic analysis must go through a first step of lexical analysis, which consists in testing membership of each word of the text to the Arabic vocabulary (Revuz 1991). So, we must begin on formalization of the Arabic vocabulary. This study started with the formalization of three sets: verbs, nouns and particles.

4.1 Verbs

The dictionary of verbs contains 10 000 fully vowelled entries [6]. Since automatic combination between roots and patterns leads to virtual lemmas generation or leaves a large number of lexical entries unrepresented and considering that each Arabic root can combine, legally, with a subset of the potential patterns (Dichy 2001), we chose to build a lemma dictionary to avoid such problems evoked within the Xerox lexical analyser. In our case, each entry represents a third person, singular, masculine, perfect verb. These verbs are associated to an inflectional description (among 130 hand-encoded inflexional paradigms for the totality of the verbs) [7].

By inflectional description we refer to the set of all possible transformations which allow us to obtain, from a lexical entry, all inflected forms. These inflexional descriptions include the mood (indicative, subjunctive, jussive or imperative), the voice (active or passive), the gender (masculine or feminine), the number (singular, plural or dual) and the person (first, second or third). On average, there are 122 inflected forms per lexical entry.

Example

"**كَلَّمَ**", V+Tr+FLX [8] = V_kallama (kallama – *to speak with someone*)

Among the 122 inflexional transformations which are described in the flexional paradigm "V_kallama", here is one: (<LW> يُ <R4><S> ِ <R><S> ُ A+P+3+m+s) [9].

This NooJ transformation means: position the cursor (|) at the beginning of the form (<LW>) (|kallama), insert "يُ" (yu) into the head of the form (yu|kallama), skip four

letters (<R4>) (yukall|ama), erase a letter (<S>) (yukall|ma), insert the vowel "ِ " (i) (yukalli|ma), skip a letter (<R>) (yukallim|a), delete of the following letter (<S>) (yukallim|)and finally insert the final vowel "ُ " (u) (yukallimu|).
These operations, applied in succession, generate the form: "يُكَلِّمُ" (yukallimu – he speaks with someone). It will be labelled with inflexional information: A + P + 3 + m + s, i.e. active voice (A), indicative present (P), third person (3), masculine (m) and singular (s).
The case of "weak" roots, which contains one of the letters "و " (ŵ), "ي " (y) or hamza ("] ("إ","أ","ا10], requires a particular study since they introduce some inflexional irregularities.

Example
The verb "وَبَأَ" (ŵabaǍa – *to have a catastrophe*) can be associated to two inflexional descriptions to give the two possible forms:

- "يَوْبَأُ" (yaŵobaǍa) with simple concatenation of "يَ" (ya) at the head of verb
- "يَبَأُ" (yabaǍa) with concatenation of "يَ" (ya) at the head of the verb and the suppression of "وَ" (ŵa)

The first version of inflexional descriptions included more than 130 paradigms to describe all verbal inflexional classes, but the possibility of defining new morphological operators (<Z> and <M> operators) enabled us to gather some descriptions where there wasn't a great inflexional difference. So, these operators allow us to reduce the number of inflexional descriptions to have only 130.

- The morphological <Z> operator is used to describe the two verbs "كَتَبَ" (kataba – *to write*) and "ثَبَتَ" (Ťabata – *to be proved*) within the same paradigm in spite of requiring two different transformations when giving the past tense. In fact, the first verb needs to delete the last vowel and add the letter succession "ْتُ" (otu) to obtain "كَتَبْتُ" (katabotu – *I wrote*) and the second verb needs to add Arabic shadda [11] "ّ " (w) and the final vowel "ُ" (u) after deleting the last vowel to obtain "ثَبَتُّ" (Ťabatwu – *I was proved*) ; this was the issue given by the definition of a specific transformational morphological command <Z> which verify if the last consonant of the verb is a "ت" (t) to process inflection in the right way.
- The <M> morphological command is used when gathering the two verbs "كَتَبَ" (kataba – *to write*) and "دَهَنَ" (dahana – *to smear*) in spite of having two different transformations for past form in the first person, plural when producing "كَتَبْنَا" (katabonA – *we wrote*) and "دَهَنَّا" (dahannA – *we smeared*).

4.2 Nouns

The nouns are described in three different ways:

- We built a dictionary which contains 15 000 entries in the form of primitive nouns [12], such as "كُرْسِيّ" (korsiyy – *a chair*). Each entry represents a singular noun form deprived of its final vowel.
- We added into the same dictionary some plural forms which do not have a singular correspondent form used, such as "مَخَاوف" (MakhAŵif – *dangers, perils*).
- We associated derivational descriptions to verbs described above. Generated forms represent the whole deverbals [13] such as IsmFaîl (i.e. active participle), IsmMafoûl (i.e. passive participle) or Masdar (i.e. infinitive form) (Dichy 2003).

Example
For the verbal entry of the dictionary:

"دَرَّسَ", V+DRV [14] = D_darrasa (darrasa – *to teach*)
Transformation "DRV=D_darrasa" produces, particularly, the two following transformations:

- "<LW> مُ <R4><S>ِ<R><S>/N [15]"→ "مُدَرِّس" (mudarris - *a teacher*): this form is obtained from the verb darrasa using the following transformations: position the cursor (|) at the beginning of the verb (<LW>) (|darrasa), insert "مُ" (mu) at the head of the form (mu|darrasa), skip four letters (<R4>) (mudarr|asa), remove a letter (<S>) (mudarr|sa), insert the vowel "ِ " (i) (mudarri|sa), skip a letter (<R>) (mudarris|a) and, finally, remove the last letter (<S>) (mudarris|/N).
- "<B><LW> مُ /N" → "مُدَرَّس" (mudarras - *a pupil*): this form is obtained by deleting the last vowel (<B>) (darras|), positioning the cursor (|) at the beginning of the verb (<LW>) (|darras), and concatenating "مُ " (mu) (mu|darras), thus we obtain (mudarras/N).

These nouns, deprived of their final vowel, are associated to inflexional descriptions to generate all inflected nominal forms labelled with some linguistic information such as gender (masculine, feminine or neuter), number (singular, dual and plural) and case (nominative, accusative or genitive). In addition, to perform plural forms from nominal entries, we had to develop about 125 paradigms when describing masculine regular plural, feminine regular plural and broken plural. These paradigms were carefully developed in order to treat certain specificities of Arabic plural such as the difference between plurals of small number and collective plurals such as the case of the form "شَهْر" (chahor – *a month*) which can have two plural forms: "أَشْهُر" (Ăachohur – *less than 12 months*) and "شُهُور" (chuhUr – *12 months and more*).
The formalized inflection of verbs, primitive nouns and deverbals allows recognition of all the corresponding inflected terms; the lookup algorithm of NooJ uses finite-state machines, which make possible simultaneous recognition (i.e. without additional computing) both of vowelled, partially vowelled or unvowelled forms.
In fact, the omission of diacritics in a written form can lead to numerous distinct fully vowelled words. For example, the unvowelled form "ktb" is supposed to have multiple vocalised annotations, our lookup algorithm based on finite state machines is able to return, at the same time, fifteen fully vowelled forms :

- "كُتُب" (kutub – *books*) with the five possible final vowels,
- "كَتْب" (katob – *a writing*) with the five possible final vowels,
- "كَتَبَ" (kataba – *he writes*) ,
- "كُتِبَ" (kutiba – *it was written*) ,
- "كَتَّبَ" (kattaba – *he makes write*) ,
- "كُتِّبَ" (kuttiba – *it was made write*) ,
- "كَتِّبْ" (kattibo – *make write*).

Moreover, each recognized form is associated by the lookup algorithm of NooJ a set of linguistic information: lemma, grammatical category, gender and number, syntactic information (e.g. +Transitive) and distributional information (e.g. +Human).

4.3 Particles

We listed about 580 vowelled particles. These particles include prepositions, adverbs, conjunctions, interjections, answers, negations, exceptions, etc.

5. MORPHOLOGICAL ANALYSIS AND GRAMMATICAL RULES DEFINITION

The Arabic language is a strongly agglutinant language; its morphological analyzer should separate and identify the component morphemes of the input word, labelling them somehow with sufficient information to be useful for the tasks at hand.

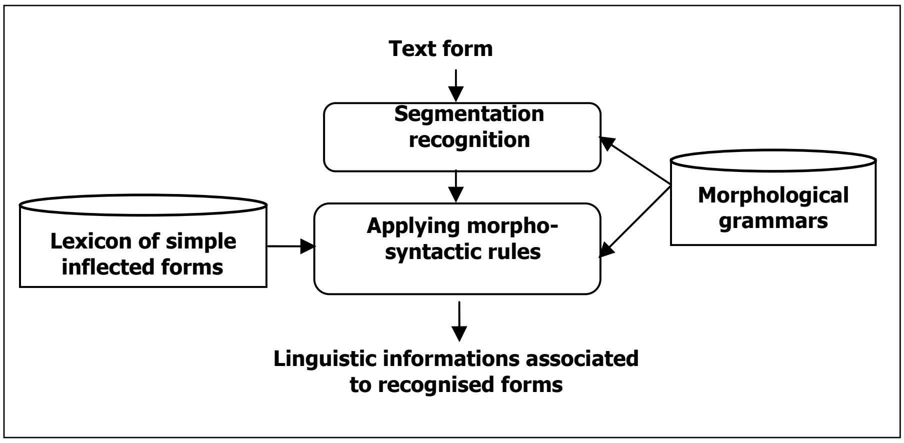
Text form
Segmentation recognition
Morphological grammars
Lexicon of simple inflected forms
Applying morpho-syntactic rules
Linguistic informations associated to recognised forms

FIGURE 1: Chain of a text form morphological analysis

We start our analysis with application of a decomposition system, implemented via a NooJ morphological grammar, to each word of the text to identify its radical and affixes. In the second step, grammars (finite-state transducers) produce lexical constraints checking the validity of segmentation thanks to a dictionary lookup. So, these grammars associate the recognition of a word to lexical constraints, working only with valid combinations of the various components of the form. Typically there are several output strings, each representing a possible analysis of the input word.

We continue the description of the morphological analysis of Arabic within the linguistic platform NooJ by detailing the different lexical constraints implemented inside morphological grammars.

5.1 Morphological constraints

Morphological constraints follow from the distortion of some radicals by agglutination with prefixes or suffixes. They enact restoration of the initial form such as it appears in the lexicon. We proceed by the application of some morphological transformations (addition of letters, deletion, substitution, etc.)

5.1.1 Letter addition

The morphological analysis of the form "كَتَبُوهُ" (katabuhu – *They wrote it*) requires the addition of a final letter before dictionary lookup:

- 1st step: Form segmentation into verb + suffix: "هُ + كَتَبُو" (katabu + hu)
- 2nd step: Addition of the long vowel "ا" (alef) + access to the dictionary: " + هُ كَتَبُوا " (katabul + hu) with "كَتَبُوا " (katabul – *They wrote*) is the inflected form of the third person, plural, masculine, perfect verb and "هُ " (hu - *it*) is a personal pronoun.

5.1.2 Letter substitution

The morphological analysis of the form "سَمَّانِي" (sammAniy – *He appointed me*) requires the substitution of final letter before dictionary lookup:

- 1st step: Form segmentation into verb + suffix: " سَمَّا + نِي" (sammA + niy)
- 2nd step: Substitution of the final long vowel "ا" (alef) for the long vowel "ى" (Y – *alef maksura*) + access to the dictionary: "سَمَّى +نِي" (sammaY + niy) with "سَمَّى" (sammaY – *to appoint someone*) is the inflected form of the third person, singular, masculine, perfect verb and " نِي" (niy - *me*) is a personal pronoun.

5.1.3 Letter deletion

The morphological analysis of the form "الدَّلْوُ" (alddalowu – *The bucket*) requires the deletion of Arabic shadda “ ّ ”, which implies deletion of the duplicated letter "d" in transliterated form, before dictionary lookup:

- 1st step: Form segmentation into prefix + verb: " ال + دَّلْوُ" (al + ddalowu)
- 2nd step: Deletion of the duplicated letter " ّ " (corresponding to the second letter “d” inside the transliterated form) + access to the dictionary: " ال + دَلْوُ " (al + dalowu) with " ال" (al - *The*) is a definite article and " دَلْوُ" (dalowu - *bucket*) is the inflected form of the singular nominative noun.

In fact, there are 14 Arabic consonant among the 28 letters in Arabic alphabet that requires fitting Arabic shadda “ ّ ” into forms when defining them. When present at the beginning of the form, these cases need the same handling as described processing.

The described transformations (letter addition, substitution or deletion) can be combined together to deal with more complex morphological phenomena.

This first type of constraints considers morphological incompatibilities which would have to be generated from a direct decomposition.

5.2 Constraints on the syntactic properties of verbs

These constraints verify the mark "+Transitive" of verbs in the dictionary. Indeed, the transitivity of verbs enables us, generally, to decide possibility of verb suffixation. Such agglutination will be only permitted for direct transitive verbs and indirect transitive ones conjugated at the singular third person (Achour 1998).

Example

- the verb "كَتَبَ " (kataba - *to write*) is direct-transitive → agglutination "kataba + hu" is accepted
- the verb "مَاتَ " (mAta - *to die*) is non-transitive → agglutination "mAta + hu" is prohibited
- the verb "آلَ " (éla - *to succeed*) is indirect transitive:
    - Inflexion at the singular third person → agglutination "ilta + hu" is accepted
    - Inflexion at the singular first person → agglutination "iltu + hu" is prohibited

The next simplified graph (FIGURE 2) shows that each term formed by letter succession (<L>) stored in a variable ($Verbe), followed by a personal pronoun (PRONPERS1, PRONPERS2, PRONPERS3) [16] will be recognized such as an accepted agglutination only if this variable represents a transitive direct verb (+Tr); this lexical constraint is represented in the <$Verbe=V+Tr> expression. In addition, an indirect transitive verb (+TrInd) inflected on the masculine (m), singular (s) third person (3); this constraint is represented in the <$Verbe=V+TrInd+3+m+s> expression.

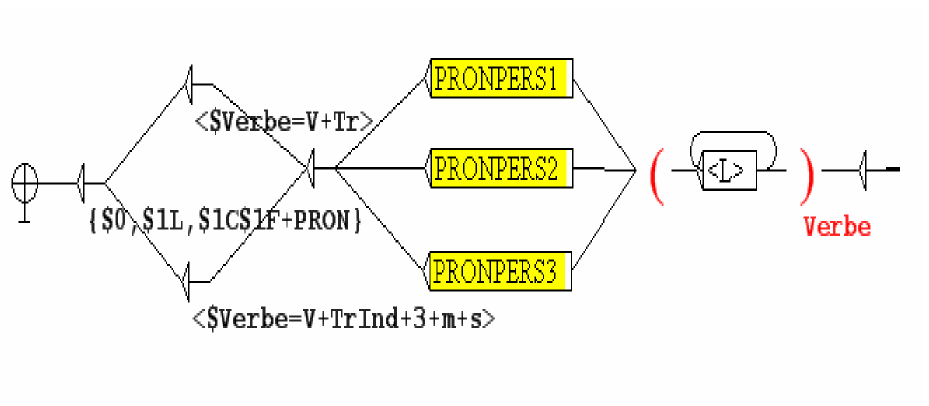

FIGURE 2: Transitivity verb NooJ graph

5.3 Orthographical constraints

These constraints look at letters which have orthographical variation during agglutination. We can expose the case of the letter "T" which can be written in two different orthographies with the same pronunciation. According to the nature of the form (noun, adjective or verb) and the position of the letter in this form (at the beginning, the middle or the end of the word), we can find either of the two spelling forms "ت" (t- *Teh maftouha*) or "ة" (T - *Teh marbouta*). This second spelling form can never be found in a verbal form and in case of its appearance at the end of a noun or an adjective it has to be converted into a "ت" (t- *Teh maftouha*) during word suffixation; thus the reverse operation is required before the dictionary lookup to associate corresponding linguistic information.

Example

The word "مَدْرَسَتِهِ" (madrasatihi – *his school*) is broken down into "مَدْرَسَتِ + هِ" (madrasati + hi). Before consulting the dictionary, we must substitute the letter "ت" (t) with the letter "ة " (T) to have the correct segmentation "مَدْرَسَةِ+ هِ " (madrasaTi + hi) with " مَدْرَسَةِ" (madrasaTi – *school*) is the genitive case of the noun and "هِ " (hi - *his*) is a personal pronoun.
We have similar problems with the alifs which have five different orthographies (" ,"ء ئ" ,"ؤ" ,"إ" ,"أ"") for the same pronunciation.

5.4 Phonological constraints

These constraints, generally combined with morphological ones, maintain a consonance inside agglutinated forms. They deal with the compatibility of the declension of radical and attached suffix.

Example

For morphological analysis of the word "كِتَابِهِ" (kitAbihi - *his book*), we start with a first step of segmentation which gives "كِتَابِ " (kitAbi - *book in the genitive case*) + hi (personal pronoun in the genitive case), in a second step, we apply grammatical rules to validate agglutination of the noun (kitAb – *book*) to the personal pronoun since they are both declined in the genitive case.

However, the concatenation of the same nominal form (kitAb – *book*) declined in the accusative case with the same personal pronoun, in the genitive case, makes a prohibited agglutination. So, the agglutinated form "كِتَابَهِ" (kitAba + *hi*) is refused because of the declension incompatibility between the accusative case of the noun "كِتَاب" (kitAb - *book*) and the genitive case of the pronoun as a suffix "هِ" (hi - *his*).

6. TEST AND EVALUATION

The test of the lexical coverage of our Arabic module is evaluated on lexical analysis of the corpora of the LASELDI collected from Internet. These corpora are composed of journalistic articles of the newspaper "Le monde diplomatique" [17] for five years (2001 – 2005), which include about 150 000 different terms. The lexical analysis, of these corpora, shows coverage of about 93% by our lexical and morphological resources. The unrecognized forms are classified in 4 subsets:

- 7 000 transliterated named entities : proper names of person such as “شِيرَاك” (chIrAk - *Chirac*) with some derived forms such as “شِيرَاكِيَّة” (chIrAkiyya - *Chiraquism*), cities such as “مَرْسِيلِيَا” (marsIliyA - *Marseille*) and organisations such as “مَايكْرُوسُوفِتْ” (mAyicrUsUfit - *Microsoft*),
- About 2000 borrowing terms such as “مِيتَافِيزِيقَا” (mItAphIzIkA - *metaphysics*),
- 1 400 spelling mistakes. The most common mistake is the misplacement of the last vowel such as the form “أَبَدَاً” (Ǎabadan – *never*) where the last vowel “ ً ” (an) is misplaced,
- Some irregular plural forms not yet covered by our study.

Actually, we are working on coverage enhancement; on the one hand, the majority of the unrecognized forms are named entities (names of people, organizations or localities) so we are implementing a module of named entity recognition using the NooJ syntactic module (Mesfar 2006); on the other hand, we noticed that some spelling mistakes are caused by badly placed vowels so we are applying morphological grammars with low priority (i.e. applied just in case of an unrecognised form) to start removing the last vowel of the form and if the form remains unrecognized we remove all vowels. We are also checking that the description of the recognized forms is correct using local grammars built with the syntactic module of NooJ.

ENDNOTES

[1] NLP: Natural Language Processing

[2] A form is a sequence of graphemes delimited with two white spaces or punctuations in a text. The terms form and word will be used interchangeably.

[3] <N> and <P> commands are used for compound-words transformations

[4] We give an example of a grammar represented as a finite state transducer in FIGURE 2 (subsection 5.2. Constraints on syntactic properties of verbs)

[5] More informations about the linguistic platform NooJ are available, on-line, at: www.nooj4nlp.net

[6] The list of verbs was built by Slim Mesfar and Ibtihal Farawi during their researches on Arabic in the LASELDI.

[7] The classification of verbs on flexional models is described in the book of Abou Il Azm (2003).

[8] FLX: introduces the functionality which describes all the potential verbal forms from a verbal lemma

[9] <LW>, <R> and <S> are commands, among ten, preset within NooJ

[10] People tend to not to write correctly hamzas, the bare alef could be either "أ" , "ا" or "إ"

[11] Arabic shadda is a consonant following sign which show gemination. It corresponds to consonant duplication in Latin languages (e.g. channel).

[12] A primitive noun indicates a noun which doesn't derive from a verb

[13] A deverbal is a noun which is derived from a verb

[14] DRV: indicates the functionality which allows derivation of nouns and adjectives from a verbal lemma

[15] "/N" indicates the fact that we obtain a noun after a derivation of a verb

[16] Coloured nodes (PRONPERS1, PRONPERS2, and PRONPERS3) represent sub-graphs containing all personal pronouns with the first, second and third person.

[17] The corpus was downloaded, in major part, from: www.mondiploar.com

REFERENCES

Abou Il Azm, A.(2003) Tasrif Moojim il afál: 10 000 verbs. Dar Ittawhidi, Rabat.

Abdeli, A., Cowie, J., Soliman, H. (2004) Arabic Information Retrieval Perspectives. JEP - TALN Session on the automatic treatment of the written and spoken Arabic language, Morocco.

Achour, H. (1998) Contribution à l'étude du problème de la voyellation automatique de l'arabe. PhD Thesis, Paris 7 University.

Beesley, K., Buckwalter, A. (1989) Two-level Finite State Analysis for Arabic Morphology. In Proceedings of the Seminar on Bilingual Computing in Arabic and English. The Literary and Linguistic Computing Centre, Cambridge.

Beesley, K. (1996) Arabic Finite-State Morphological Analysis and Generation. In Proceedings of COLING-96, the16th International Conference on Computational Linguistics, Copenhagen.

Beesley, K. (1998) Arabic Morphology Using Only Finite-State Operations. In M. Rosner, editor, Proceedings of the Workshop on Computational Approaches to Semitic Languages, Montreal.

Beesley, K. (2001) Arabic Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001. In Proceedings of ACL/EACL 2001 Workshop, ARABIC Language Processing: Status and Prospects, Toulouse.

Chomsky, N. (1971) Aspects of the Theory of Syntax. Trad. J.-C. Milner.

Debili, F. (2001) Traitement automatique de l'arabe voyellé ou non. Correspondances – IRMC.

Debili, F., Achour, H., and Soussi, E. (2002) La langue arabe et l'ordinateur : de l'étiquetage grammatical à la voyellation automatique. Correspondances – IRMC.

Dichy, J., and Farghaly, A. (2003) Roots & Patterns vs. Stems plus Grammar-Lexis Specifications: there what basis should is multilingual lexical database centred on Arabic be built? Workshop on Machine Translation for Semitic Languages, New Orleans, USA.

Dichy, J. (2001) On lemmatization in Arabic. A formal definition of the Arabic entries of multilingual lexical databases. In proceedings of ACL/ EACL 2001.

Harris, Z. S. (1985) Transformational Theory. Language, vol. 41, No. 9, pp. 363-401.

Khoja, S., Garside, R., and Knowles, G. (2001) A tagset for the morpho-syntactic tagging of Arabic. Paper given At the Corpus Linguistics 2001 conference, Lancaster.

Koskenniemi, K. (1983) Two-level morphology: a general computational model for word-form recognition and publication. University of Helsinki.

McCarthy, J. (1993) Template form in prosodic morphology. Paper from the Third Annual Formal Linguistics Society of Midamerica Conference, pages 187-218, Bloomington, Indiana. Indiana University Linguistics Club.

Mesfar, S. (2006) La reconnaissance des entités nommées (NER) en arabe standard. The 9th INTEX/NooJ conference, The Faculty of Mathematics, University of Belgrade.

Revuz, D. (1991) Dictionnaires et lexiques : méthodes et algorithmes. PhD Thesis, Paris 7 University.

Silberztein, M. (2005_1) NooJ's Dictionnaries. In proceedings of LTC 2005, Pozman University.

Silberztein, M. (2005_2) NooJ : The Lexical Module In NooJ pour le Traitement Automatique des Langues. S. Koeva, D. Maurel, M. Silberztein Eds, MSH Ledoux. Franche-Comté Academic Presses, 2005.

Silberztein, M. (2006) NooJ Manual. Download from "http://www.nooj4nlp.net"

Ouersighni, R. (2001) A major offshot of the DIINAR-MBC project: AraParse, a morpho-syntactic analyzer for unvowelled Arabic texts. In proceedings of ACL/EACL 2001.

Tuerlinckx, L. (2004) La lemmatisation de l'arabe non classique. The 7th international Days of Statistical Analysis of Textual Data, JADT.