Using Augmented Transition Network for Morphological Processing of Arabic

Kouninef Belkacem
Institut National des Télécommunications
et des TIC Oran
ALGERIA

Saidi Abderrahmane
Institut National des Télécommunications
et des TIC Oran
ALGERIA

ABSTRACT

By its morphological, syntactic and phonetic properties, the Arabic language is considered as being one of the languages that are difficult to apprehend in the field of automatic processing of written and spoken language. This paper presents a more effective analysis morphologically (or morphosyntactic) an Arabic word voweled. From this method we can define and determine the type of word and its morphosyntactic whatever the word is (simple or composed).

The construction of an Arabic word is different from a word in French or in English; it can mean a sentence in French, which explains the difficulty of morphological analysis. Thanks to these characteristics we can do morphological analysis by the use of one of the methods used in the parsing of French or English. This method is based on the automaton and called ATN (Augmented Transition Network).

Keywords

Morphology, Morphosyntactic analysis, ATN (Augmented Transition Network), Automate.

1. INTRODUCTION

Our objective is to identify the category of a given word and assign a morphosyntactic label (verb unaccomplished, accusative name, etc…) The major problem that arises is the elements that are bound affixes to the word (suffixes and / or prefixes) and (proclitics and enclitic). The principle of division of a word is based on the use of formalism of augmented transition network (ATN)[ ]. Any system of morphosyntactic analysis consists of two main parts namely, a lexical database that contains and describes the resources and a lexical analyzer.

2. STRUCTURE OF WORD

In Arabic a word can mean a sentence thanks to its compound structure, which is an agglutination of elements of grammar; the following representation outlines a possible structure of a word. Note that the reading and writing of a word is from right to left [5].

Enclitic + suffix + Schematic body + Prefix + proclitic

- Proclitics are prepositions or conjunctions.

- Prefixes and suffixes express grammatical features and indicate the functions: event name, mode of the verb and the modalities (number, gender, person ...).

- Enclitics are personal pronouns.

Example:

أ تَتَذْكَّرُونَنَا

This word expresses the phrase in English: "Do you remember us?" [5] The segmentation of this word gives the following constituents:

أ +تَ+ تَذْكَّرُ+ ونَ+ نَا

Proclitic : أ conjunction of interrogation

Prefix : تَ verbal prefix time of unaccomplished.

Schematic body: تَذكَّر derived from the root: ذكر according to the schema

Suffix : ون verbal suffix expressing the plural

Enclitic : نَا pronoun suffix complement to the name

3. CONCEPTION OF THE LEXICAL BASE

It is clear that such a system relies on the use and handling of linguistic data of the Arabic language, representing its knowledge base. For that, we were interested in this work to design a language database containing all the primitive morphology of the Arabic language [15].
The realization of such a database requires not only the collection of linguistic data but also their organization according to techniques well suited to facilitate their exploitation.

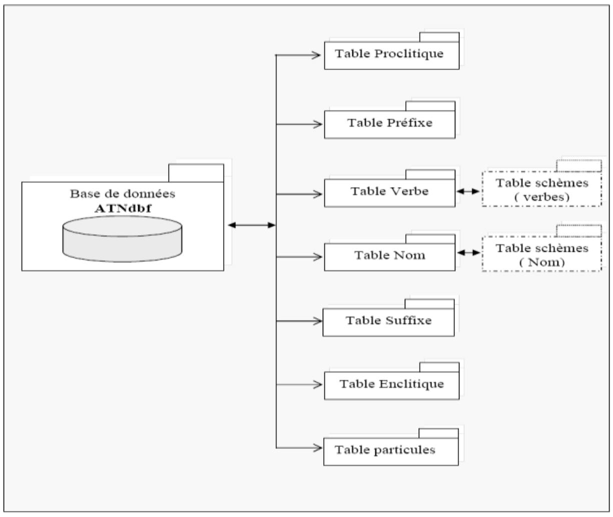

Fig 1: Architecture of the lexical database.

3.1 The pack of names

The morphological system of Arabic names distinguishes between two categories of names. The first category includes all derived nouns (names obtained by using the derivation rules). A derived name is completely characterized by its morphological representation « root – schema » [11].

The second category includes all the special names that does not respect any rule of derivation. Names are listed in their pack so that their ending are without vowels ' HaraKa حركة ', the reason why the analysis of affixes of inflection is easily treated.

We quote that the character schema for names, is a virtual pack, as it will be called where it is useful, on the one hand, not to obstruct the natural progress of the application, and on the other hand for not increase the size of the physical database.

3.2 The pack of verbs

The morphological system of Arabic verbs is very particular. Indeed it is on the one hand robust and totally regular in the case of healthy verbs " الأفعال الصحيحة " based on the representation "roots - schemes", and on the other hand, irregular as a rule in the case of non-healthy verbs " الأفعال المعتلة "or " defective verbs "(a family of verbs are distinguished by the syntax and the inability to take certain forms of conjugation, hence the reason for nomination" incomplete"). The verbs are listed in their pack so that their ending is without vowels ' HaraKa حركة ', the reason why the analysis of affixes of conjugation is easily treated [13].

In our system we used only healthy verbs.

The simple verbs are used in the database, on the other hand the augmented are derived from simple verbs, and when the program inspect the totalitty of the simple verbs without having any result, it moves from one schema to another by applying to every simple verb a specific function for each schema.

3.3 The pack of proclitic

This pack contains a reduced number of this kind of objects. The following table shows us some proclitic used in our application [6].

Table 1. The main used proclitics

ال	وَ	فَـ	وَال	فَال	أ	أفَ	أوَ	سَ	فَسَ
وَسَ		بال			فَبال	وَبال	لِ	فَلِ	وَلِ
كَ	وَكَ	فَكَ	كَال	وَكَال	فَكَال	أسَ	أفَس	أوَس	أبِ
أكَ	ألِ	أبال	أكَال	ألِل	أفَلِل	أوَلِل	لِلْ	فَلِلْ	وَلِل
أفَبِ	أوَبِ	أفَكَ	أوَكَ	أفَلِ	أوَلِ	أفَبال	أوَبال	أفَكَال	أوَكَال
بِ	فَبِ	وَبِ		لَ	فَلَ				

Another feature that characterizes this kind of lexical items is that they do not have the same kind of link, either with verbs,names, particles or particulates and names.

- Boxes ▒ indicate that these proclitics agree with the names, verbs and particles.
- Boxes ▨ indicate an agreement with names only
- Boxes □ specify the link with verbs only.
- Boxes ▣ specify the link with the particles and names.

3.4 The pack of enclitic

Also for this kind of object we used the ones that appear in the following table [6]:

Table 2. The main used enclitic

نـي	ي	نا	كَ	كِ	كُمَا	كُمْ	كُنَّ	هُ
هِ	هَا	ه	هُمَا	هِمَا	هُمْ	هِمْ	هُنَّ	هِنَّ

For these objects, the question of relationship is as following:

- Boxes ▨ indicate that this enclitic agree with the names, verbs and particles.
- Boxes ▒ indicate an agreement with names only.
- Boxes □ specify the link with verbs and particles.

3.5 The pack of prefix

Note that these particles are placed at the beginning of verbs in the imperfective (the present continuous (المضارع) and the imperative (الأمر)). The principle implemented prefixes are [6]:

Table 3. Major used prefixes

أ	أ	يَ	يُ	تَ	تُ	نَ	نُ	إ	أ	ا

.

3.6 The pack suffix

The main implemented suffixes are [6]:

Table 4. Major used suffixes.

َ	ُ	ِ	أً	ٌ	ٍ	ةَ	ةٌ	ةِ	ةً
ةٌ	ةٍ	اً	اتُ	اتِ	اتٌ	اتٍ	تْ	تْ	تِ
تْ	انِ	تَا	تَان	تَيْ	تَيْن	يْ	يْنَ	يْن	وُ
وا	ونَ	ي	ينَ	تُمَا	تُمْ	تُنَّ	نْ	نَا	نَّ

- The boxes ▨ indicate that these suffixes agree with names.
- The boxes ▒ indicate agreement with verbs only.
- The boxes □ specify the link with verbs and nouns.

3.7 The pack of particles [3]

The particle is an invariable word that accompanies a noun or a verb, and cannot convey any sense when it is isolated. Among the particles, some are working with names, some with verbs, and others with the name and verb.

In the study of words of Arabic language we distinguish particles of (they are voweled in the system):

- Supplementation (حروف العطف): و, ف, ثم... etc.

- Interrogative (حروف الاستفهام): هل, أ, ماذا, متى... etc
- Preposition say (حروف الجر): في, إلى,, ب... etc
- unaccomplished 'mansoub' (حروف النصب): لن, أن... etc..
- Determination (حروف التعريف): ال.

and others.

In our system we used one of these particles as proclitics (in the pack of proclitics, Table1) and others in the pack of particles, such as: preposition (على, عن, من ..), interrogative (ماذا, متى ...)

And also in this pack we find the indeclinable names (غير المتصرفة) as demonstrative (أسماء الإشارة) (هذا, هذه...) linked (الأسماء الموصولة) (الذي, الذين...)

4. MORPHO-SYNTACTIC ANALYSIS AND TRANSITION NETWORKS

4.1 Transition networks

The transition network is an extension of finite state automaton. As we noted earlier, there is equivalence between finite state automaton and regular grammar. However, regular grammar is insufficient to address issues related to natural language. By integrating some extensions, the transiton networks have increased the power of automaton by making them equivalent to context-free grammar and even in some cases of contextual grammar.

The type of automaton associated with regular grammars is a finite automaton [4]

4.1.1- The RTNs (Recursive Transition Network)

Recursive transition networks are automaton in which a transition arc A may be tagged by the identifier of another network R.

This means that the transition A will be completed only if the sub-network R is traversed from its initial state to one of its final states.

This first extension allows the improvement of the modularity of the formalism. It becomes possible to organize networks based on syntactic categories they represent. However it does not increase its power. The second extension allows an arc to call the network to which it belongs (hence the term recursive). It is also possible to have an indirect recursion, ie an arc A1 in the network R1 can call a network R2, and an arc A2 in the network R2 may in turn call the network R1. This second extension makes RTNs equivalent to context-free grammars.

4.1.2- The ATN (augmented transition networks) [7, 8]

ATN are extensions of finite state automaton used primarily for language processing and specification of man-machine interfaces.

Augmented transition network (Augmented Transition Network - ATN) is a recursive transition network in which is added extensions that give it a higher descriptive power than the one of a type 2 grammar. These extensions are three in number. This is to add register to the networks of transitions, and to impose conditions on transitions, and assign actions to the transitions made.

The RTNS allow having a formalism whose power is equivalent to context-free grammars. However this is still insufficient to deal with problems in which context plays a decisive role. This Is Why augmented transition networks were introduced by Woods. These networks can increase the power of RTNs by combining to arcs of transition conditions that will allow to restrict the circumstances under which it is possible to pass over an arc, and actions [12] that will permit to realise semantic processing.

4.2 The analysis steps morpho-syntactic

We come now to the most important part of this work, based on the application of the technique of transition network on the Arabic language (morphology). The goal is to determine the category of the word (particle, noun, verb) and to produce their morpho syntactic characteristics [1][2].

Augmented Transition Netwoks, ATN are recursive nature strengthened by a set of conditions and actions. Their simplest form is characterized by arches that are having an initial extremity end and another one which is final. The ends represent nodes that interpret the state of transition, and each network must have necessarily a global initial state and one final, and a set of records in each arc crossing.

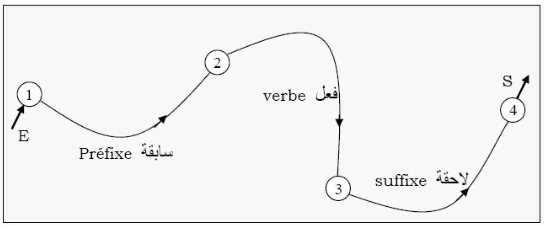

Fig 2: ATN simple verbs.

Taking the word "يذهبون" can be seen from the network as shown in the figure below:

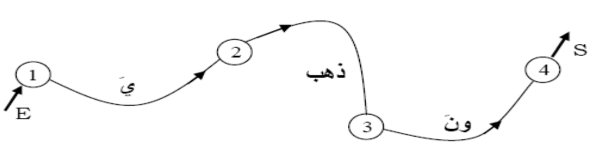

Fig 3: Example of the word "يذهبون.

The particle "ي" belongs to the prefixes, "ذهب" to verbs, and "ون" to suffixes. If the network has been covered from the initial state to the final state, then the word is practically identified [16]. However, the word is incorrect, if and only if, the path is completed by a non-final state, while the word is completely covered, or where it is impossible to move to another state.

The specification of this method is to identify one of the three main categories that the word belongs to (noun, verb, and particle). These three categories brought into play, will allow us an initial analysis phase of the word.

Now, we present the Arabic word for ATN:

We have three networks in our case, we start with the particles network [10,14] then the names network and finally the verbs network.

The network of particles is thus schematized as follows:

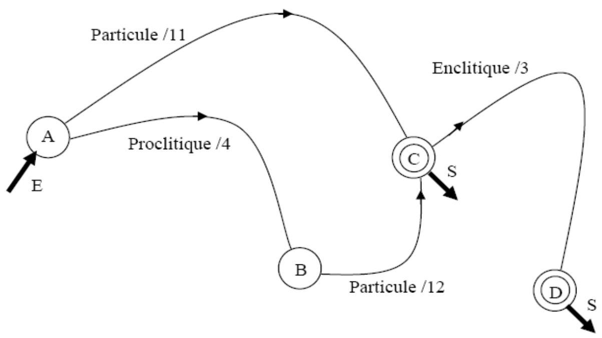

Fig 4: ATN for the particule.

In this case, if the path starts from the initial state A to the final state D in the network of particles, the word will be identified as a particle. In case of failure on the recognition particle, the analysis will be directed to the Network names shown as follows:

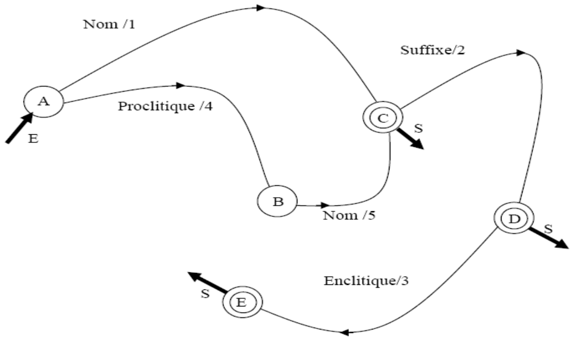

Fig 5. ATN for the name

Similarly, if the course was able to traverse the network of names since its initial state upon arrival at one of these final states, then the word is classified into the category of names, and a third choice will be made automatically, continuing analyzing while traversing the network of verbs, as illustrated below:

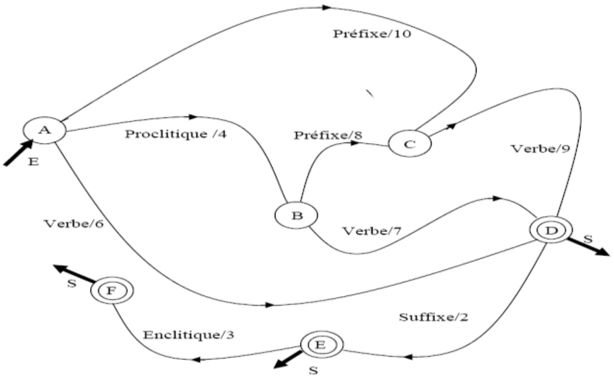

Fig 6: ATN for the verb

Of course, transitions from one state to another are governed by a set of conditions and actions to perform, with registration statements already borrowed in records with flags. This is the strength of such networks. If the arc of prefixes has been marked, it is a condition that interprets a class of unfulfilled verbs (المضارع) or imperatives verbs (الأمر), whereas if the arc 6 is marked, the category of the verb is the past.

There is another idea of synthesis that brings together all networks mentioned in a unified network as shown in Figure 7:

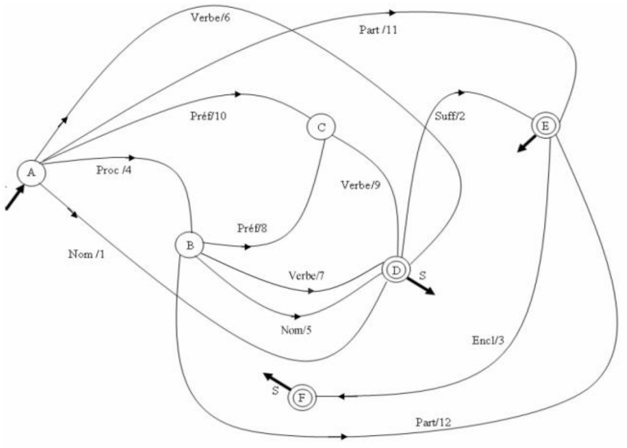

Fig 7 : ATN synthesis.

Example: Use the word " سَنُعَلِّمُهُمْ " :

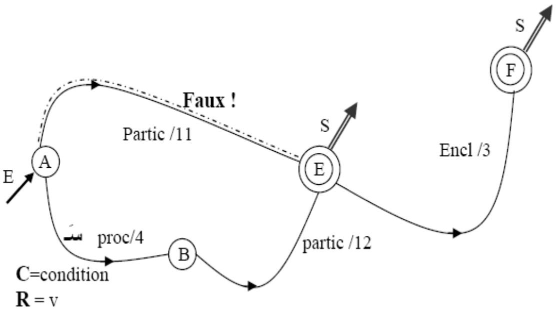

Fig 8. Example "سنعلمهم" on ATN of particule.

We begin by verifying if the word belongs to the particles, if it doesn't we try into names, authorwise, we look into the network of verbs.

In the network of particles, the arc 11 will be marked to test whether the word belongs to the whole package of the particles, if not, the last letter of the word will be decided "سنعلمهـ”. While continuing the test into the package of particles, the resulting word may have potential suffixes or enclitic, until we are sure the word, or part it is not into the packing of particles. Then, moving towards the arc 4 of the same network, so that the first four letters of the word "سنعل" could be tested if they belong to the package of proclitics, and if it doesn't, then we subtract the last letter of the word, and so on, until that the letter "س" belongs to the package of proclitic [17,18].

Arriving at this result, an action will be triggered to change the analyzed word to the word "نعلمهم. At this point, we store the value V (flag) into a specific register R1, which means an obligation of succession by a verb, because the letter "سـ" combines with verbs in their right side. In case we get the letter "ال" instead of letter "س", the register R1 is activated by a value N (flag) Coming to the arcs that characterize the particles and the names (12 / and 5 /).

For example in order to continue the path from the arc 5 / (of names), we must have checked the condition attached to this arc:

R1 (Si (R=N) ou (R=NV) ou (R=VNP))

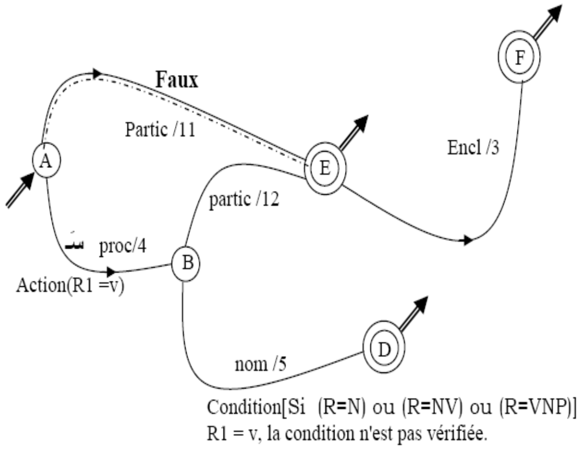
Faux
Partic /11
F
E
Encl /3
A
partic /12
سـَ proc/4
B
Action(R1 =v)
D
nom /5
Condition[Si (R=N) ou (R=NV) ou (R=VNP)]
R1 = v, la condition n'est pas vérifiée.

Fig 9: Action and Condition in ATN.

We have the value V (flag) in our registry, so there will be no progress towards the following arcs in the network of names. The connection will be oriented to the network of verbs.

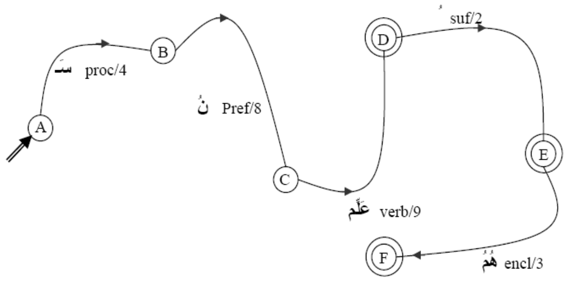

Fig 10 : Example"سنعلمهم" in ATN of verb

We noted earlier that the letter "س" combines with verbs of imperfective, so the move will be towards the arc of prefixes (pref / 8) and the letter "ن" shall be cut, giving the word "علمهم" witch naturally is a verb.

5. CONCLUSION

The work we have presented is a preliminary step for the automatic processing of Arabic morphology. This method facilitates the parsing and syntax errors decreases. Our analytical method can be exploited to morphosyntactic realization of different processing systems of the Arabic language such as machine translation systems, speech synthesis from text [9], systems AutoCorrect text.

Our contribution in this system by applying several examples of Arabic words compounds, demonstrated the effectiveness of this method in the morphological analysis and can be used in the parsing of Arabic with vowels and non-vowels.

6. REFERENCES

[1] Ahmed a. rafea , khaled f. shaalan"Lexical Analysis of Inflected Arabic Words using Exhaustive Search of an Augmented Transition Network"software-practice and experience, vol. 23(6), 567–588 (june 2003).

[2] Ahmed Farouk Ahmed, Developing an Arabic parser in a multilingual machine translation system" , Thesis, Cairo University, 2009.

[3] Aïda KHEMAKHEM, "ArabicLDB : une base lexical normalisée pour la langue arabe", mémoire master, Tunis, 2006.

[4] Faure C. Grumbach A. Likforman-Sulem Sigelle M, méthodes structurelles et neuronales école national supérieur des télécommunications, 2005-2006.

[5] Fouad Soufiane Douzidia,"Résumé automatique de texte arabe",M.Sc Université de Montréal ,2004.

[6] Ramzi Abbes, "la conception et la réalisation d'un concordancier électronique pour l'arabe", thèse de doctorat, Lyon, 2006.

[7] Stuart C. Shapiro, « generalized augmented transition network grammars for generation from semantic networks», Department of Computer Science, SUNY at Buffalo.

[8] Woods, William A. Woods. «Transition Network Grammars for Natural Language Analysis in Communication » ACM, October 1970, Vol.13, n°10, p.591-606.

[9] Z. Zemirli, S. Khabet , « TAGGAR : Un analyseur morphosyntaxique destiné à la synthèse vocale de textes arabes voyellés » Institut National d'Informatique, Algérie, 2006.

[10] Ahmad Al Taani: An Adaptive Parser for Arabic Language Processing , International Journal of Computer Processing Of Languages (IJCPOL) Volume: 23, Issue: 1 (2011) pp. 67-80 DOI: 10.1142/S1793840611002218

[11] Abdelhadi Soudi, Antal van den Bosch and G¨unter Neumann, editors. Arabic Computational Morphology. Knowledge-Based and Empirical Methods. Dordrecht, The Netherlands: Springer. 2007. ISBN 978-1- 4020-6045-8 DOI:10.1017/S1351324908004828

[12] Attia M., Foster J., Hogan D., Le Roux J., Tounsi L., van Genabith J. 2010. "Handling Unknown Words in Statistical Latent-Variable Parsing Models for Arabic, English and French". Proceedings of the NAACL HLT 2010, Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2010), Los Angeles, CA.

[13] Ruhi Sarikaya, Mohamed Afify, Yonggang Deng, Hakan Erdogan, and Yuqing Gao Joint Morphological-Lexical Language Modeling for Processing Morphologically Rich Languages With Application to Dialectal Arabic, IEEE Transaction on audio, speech and language processing, Vol 16 n°7, September 2008.

[14] R. Sarikaya and Y. Deng, "Joint morphological-lexical modeling for machine translation," in Proc. HLT/NAACL'07, Rochester, NY, 2007, pp. 145–148.

[15] B. Xiang, K. Nguyen, L. Nguyen, R. Schwartz, and J. Makhoul, "Morphological decomposition for Arabic broadcast news transcription," in Proc. ICASSP'06, Toulouse, France, 2006, pp. I-1089–I-1092.

[16] N. Habash, O. Rambow, and R. Roth, 'Mada+tokan: A toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization', in Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR), Cairo, Egypt, (2009).

[17] Imad A. Al-Sughaiyer and Ibrahim A. Al-Kharashi. 2004. Arabic morphological analysis techniques: A comprehensive survey. *Journal of the American Society for Information Science and Technology*, 55(3):189–213.

[18] Nizar Habash. 2004. Large scale lexeme based arabic morphological generation. In Proceedings of Traitement Automatique du Langage Naturel (TALN-04).Fez, Morocco.