The Use of the Relational Concept in the Arabic Morphological Analysis

Said Iazzi[1], Abderrazak Iazzi[2]
LRIT Associated Unit to the CNRST-URAC29, Faculty of Sciences, Mohammed V University
Rabat, Morocco

Saida Laaroussi[3]
Laboratoire des Sciences de l'Ingénieur, ENSA, IbnTofail University, Kenitra, Morocco

Abdellah Yousfi[4]
Team ERADIASS
FSJES, Mohammed V University
Rabat, Morocco

***Abstract*—The Arabic language differs from other natural languages in its structures and compositions. In this article we have developed an Arabic morphological analyzer. For this, we have used the relational concept in the database to build our Arabic morphological analyzer. This analyzer uses a set of tables which are linked together by relationships. These relations model certain numbers of compatibility rules between different affixes. Our morphological analysis have been trained and tested on the same databases. The tests of our new approach have given good results and the numbers obtained are very close to those of existing analyzer.**

Keywords—Arabic language rules; morphology; morphological analyzers; database; relational concept

I. Introduction

Morphological analysis is a central task in language processing. It consists in detecting the different morphological entities of an input word and provides a morphological representation of it. Morphological analysis has been and remains the focus of researchers in the automated processing of the Arabic language [1][6] [9][10][18].

Studies on Arabic morphology at the computer level have received great attention from computer engineers and linguists since the early eighties. A large number of morphological analyzers are designed for use in various applications. The attention is due to the richness and the complexity of Arabic morphologies, the importance also appears for the morphological analyzers in the main applications to facilitate and provide solutions in the fields of machine translation, information search and information retrieval [3] [4][8] [21].

Automatic language processing requires several efforts in the development level of all advanced computer methods and techniques [5] [7] [15] [23]. For many automatic language processing tasks, a complete and rich lexical database is essential, even a simple word list can often be an invaluable source of information. One of the most difficult problems with lexicons is that of non-vocabulary words, especially for languages that have a richer morphology like Arabic. To evaluate our morphological analysis systems, we need a body adapted to Arabic morphological analyzers that facilitate data processing tasks and have efficient results. Corpus analysis is focused on the experimental, while interpretation can be qualitative or quantitative.

We also describe the construction and the methodology of the necessary linguistic resources, a morphological dictionary and an adapted morphological corpus, and assess the effect of resource size on the accuracy of the analysis, showing what results can be obtained with limited linguistic resources [12].

II. Approaches Used in Morphological Analysis

There are different ways to build morphological analyzers. There was research to set the rules of grammar, morphology, grammar and spelling to build morphological analyzer systems. The development requires the study of the properties of the data words by essentially raising issues concerning the morphological analysis and presentation of Arabic words [9][10][14][15].

The methods used in the construction of morphological analyzers are quite varied. Indeed, some researchers have developed methods based on finding diacritic symbols at the character level, others have exploited these methods to identify diacritics at the word level. A group of researchers has developed hybrid methods coupling approaches to improve these methods. Darwish [15] suggested classifying the approaches into the symbolic approach, the statistical approach and the hybrid approach.

Researchers in the field of morphological analyzer use several methods to analyze a word:

A. Approaches based on Linguistic Rules

For Arabic morphological analyzers several researchers use approaches based on linguistic rules. They use a knowledge base of rules written by linguists to assign solutions to different morphological attributes of Arabic words. The approach based on linguistic rules uses algorithms purely based on the morphological knowledge of the language. It requires rules to cover all morphological shapes. These rules are often classified into grammatical, structural and logical categories. This consists of using criteria and linguistic properties in the form of rules expressing the functioning of the natural language used.

The linguistic approach requires a large number of lists and tables. To develop a set of rules to find the appropriate decomposition, this approach is based on a thorough morphological analysis of the Arabic language [2][14].

The subjective linguistic approach simulates the process used by a linguistic expert. It consists of removing affixes by comparison with predefined lists and transforming what remains, the stem into the root, after a possible alteration by the addition, deletion or modification of some of its letters.

B. *Dictionaries based Approach*

The resources used for a morphological analyzer are a dictionary of root words that has been created manually using different resources. In addition, a morphological dictionary is used for both a morphological analyzer and a morphological generator, depending on the direction in which it is read by the system [14].

C. *Approach based on Patterns of Words*

The use of morphological patterns, depending on morphological affixes in all its forms, there are patterns for verbs, patterns for names, patterns for adjectives, etc. There are some common patterns between these types.

To apply this type of analysis, it was necessary to determine the morphological patterns. By counting the morphs that enter them, and the morphemes included therein and the list between them and the grammatical affixes they share with them at the beginning or at the end of the words [10], [19], [22].

D. *Approach based on Graph*

The graph approach has been dominant since the 1980s. The finite state approach for morphological analysis was initially studied at Xerox and the first practical application was due to Koskenniemi [20][23]; this has been used to develop wide coverage morphological analyzers for several languages.

In this type of approach, the morphotactic and spelling rules are programmed in a finite state transducer (FST), they require too much manual processing to state rules in an FST and not to analyze words that do not appear in Arabic dictionaries [11][3], [24].

Other analyzers use graphs to perform the morphological analysis of Arabic words [16][17].

E. *Statistical Approaches*

This approach uses the probability of succession of certain morphemes to perform the morphological analysis of a given word. Statistical data obtained by a corpus allowing to acquire knowledge on the morphology of the language, it learns prefixes, suffixes and patterns from a corpus or a list of words in the target language without any human intervention. This approach uses a list of prefixes, suffixes and patterns to transform from stem to root. The possible prefix-pattern-suffix combinations are constructed for a word in order to obtain the possible roots.

F. *Hybrid Approach*

Hybrid methods are algorithms that combine several approaches already mentioned. For example in the case of the Buckwalter analyzer, the latter combines linguistic rules (when it introduces the notion of compatibility between prefix, suffix and stem) and dictionaries (when it uses the dictionary of Arabic stems) [14]. Other works use this type of approach [13][15].

III. Presentation of the Relational Database Approach

Our new approach is based on the relational concept between tables to perform the morphological analysis of a word.

The database used in this approach uses several tables:

- The tables of proclitics and prefixes of Arabic words.
- The tables of suffixes and enclitics of Arabic words.
- The table of surface patterns of Arabic words.
- The table of Arabic stems.

A. *The Tables of Proclitics and Prefixes*

The two tables are composed of the Arabic prefixes and the proclitics list. These two tables are related by links which model the prefixes and the proclitics combination rules.

The computability between proclitics and prefixes are going to be explained in Fig. 1.

For example: the prefix "ال" does not combine with the proclitic "س".

These Compatibility Tables generates a query (Query1) of prefixes attached to the proclitics which correspond to them, with other information, such as the word type, the pronoun, … (Table I).

B. *The Suffix and Enclitic Tables*

The two tables are composed of the Arabic suffix and enclitic list. These two tables are related with links which model the rules of combination between suffixes and enclitics.

The computability between suffix and enclitic are going to be explained in Fig. 2.

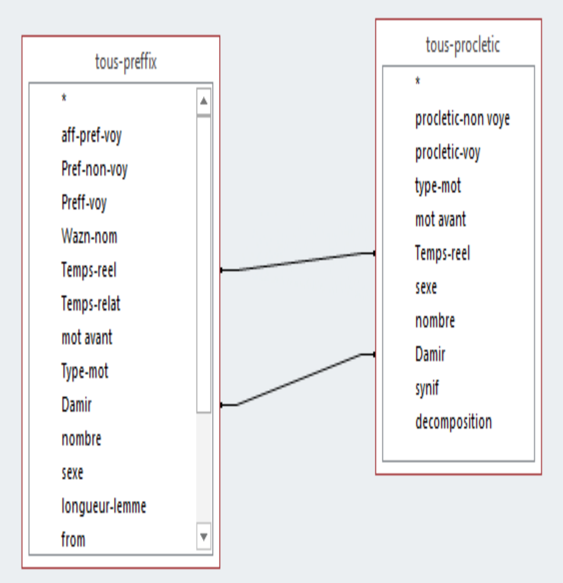


Fig. 1. Compatibility Tables between Proclitic and Prefixes.

TABLE I. EXTRACT OF THE REQUEST CREATED BETWEEN THE PREFIXES AND PROCLITICS

Prefix + proclitic	Proclitic	Prefix	Word-type	Gender of a word	Pronoun
فَلِيَ	فَلِ	يَ	المضارع المنصوب	فعل	هو
أَفَتُ	أَفَ	تُ	المضارع المجهول	فعل	أنت
.	.	.	.	.	.
فَسَنُ	فَسَ	نُ	المضارع المنصوب	فعل	نحن

For example: the enclitic "كَ" does not combine with the suffix "تَ".

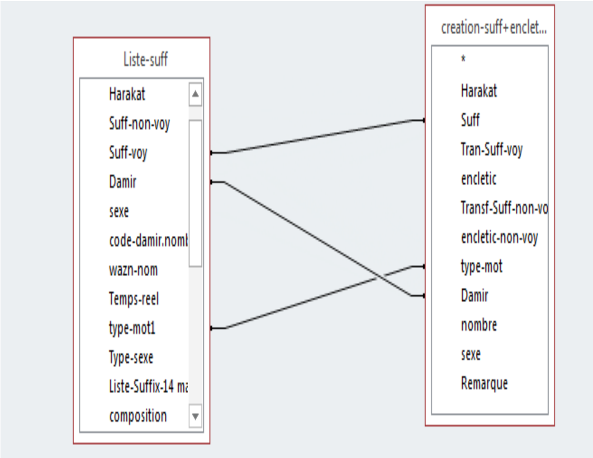


Fig. 2. Compatibility Tables between Suffixes and Enclitics.

This database generates a query (Query2) which contains several information, such as word type, gender, pronoun and numbers. To explain more, the following Table II is going to explain the request created between the suffixes and enclitics.

TABLE II. EXTRACT OF THE REQUEST CREATED BETWEEN THE SUFFIXES AND ENCLITICS

Suffix + enclitic	Encletic	Suffix	Pronoun	numbers	Genre	Type of Word
تَاهُ	هُ	تَا	هُمَا(مؤنث)	مثنى	مؤنث	فعل معلوم
تُكُنَّ	كُنَّ	ةُ	-	مفرد	مؤنث	اسم
.	.	.	.	.	.	.
يْنِهِنَّ	هِنَّ	يْنِ	-	مثنى	مذكر	اسم

C. *The Tables of Surface Patterns and Stems of Non-derived Words*

This table is composed of surface patterns stems and stems of non-derived Arabic names. Each record is composed of several information, such as lemma, stem of the surface patterns or word, type of word, gender, number, class of the surface patterns, etc.

D. *Morphological Analysis using the Relational Model*

Our approach uses the relation concept used in databases, to connect the query and the tables described previously: Query1, Query 2, and Table III. The resulting query is noted Query-main.

TABLE III. EXTRACT OF THE QUERY CREATED BETWEEN SURFACE PATTERNS AND STEMS OF NON-DERIVED WORDS

lemma	Stem-voy	Damir	number	sex	Type of Word
جَعَرَ	جْعَر	أَنَا	مفرد	مذكر	المضارع المجهول
جَاعَ	جُوعُ	أنت	مفرد	مذكر	المضارع المعلوم
الله	الله	-	-	-	اسم جلالة
نـاقَة	نَاقَة		مفرد	مؤنث	اسم حيوان
جعفر	جعفر		مفرد	مذكر	اسم علم

So the Relations between queries and tables are going to be explained in Fig. 3.

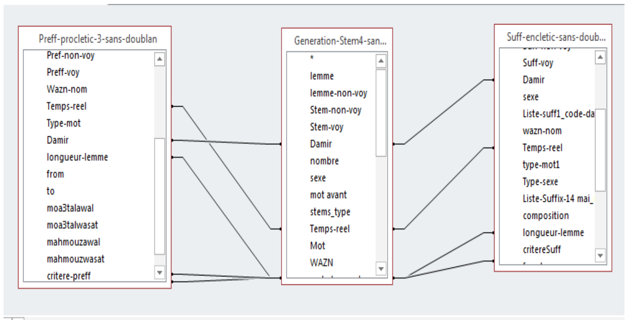
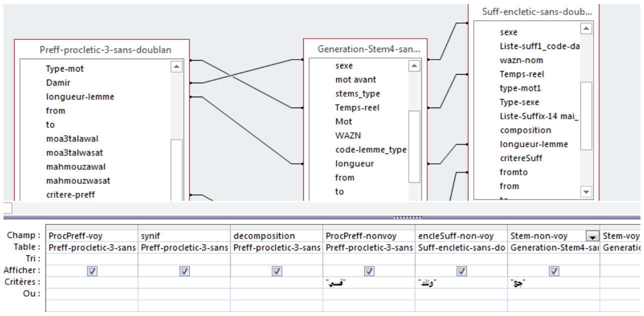


Fig. 3. Relations between Queries and Tables.

These query and table are linked by several links modeling the compatibility rules between prefix-proclitic stem and suffix-enclitic stem.

To analyze a given word, the system goes through the following steps:

- Partitioning of the word to be analyzed into a set of proclitics, prefixes, stems, suffixes and enclitics.
- Find all the surface patterns associated with each stem.
- Create a query from this information.
- For example, for the word to analyze "فدخلت", the pattern of the lemma is "جعر" associated with the proclitic "ف" and the suffix "ت", therefore, the stem calculated from the pattern of the lemma is "دخل"(Fig. 4).
- Reconstruction of solution lemmas from surface patterns of lemmas obtained in this query.
- Verification of the set of lemmas obtained, with the basis of the lemmas.



Champ :	ProcPreff-voy	synif	decomposition	ProcPreff-nonvoy	encleSuff-non-voy	Stem-non-voy	Stem-voy
Table :	Preff-procletic-3-sans	Preff-procletic-3-sans	Preff-procletic-3-sans	Preff-procletic-3-sans	Suff-encletic-sans-do	Generation-Stem4-sa	Generatio
Tri :							
Afficher :	☑	☑	☑	☑	☑	☑	
Critères :				"فسي"	"ونك"	"جو"	
Ou :							

Fig. 4. Request from the Information.

Example:

For the word "أموته", we find among the solution lemmas "مات - يموت" with the prefix "أ", the enclitic "ه" and the time "فعل مضارع معلوم". But when we check in the database of the lemmas, we find that the verb "مات-يموت" does not admit a complement, which therefore implies that each verb derived from this lemma, does not admit any enclitic. This is why this solution will be excluded.

Fig. 5 illustrates the different layers and process of our analyzer.

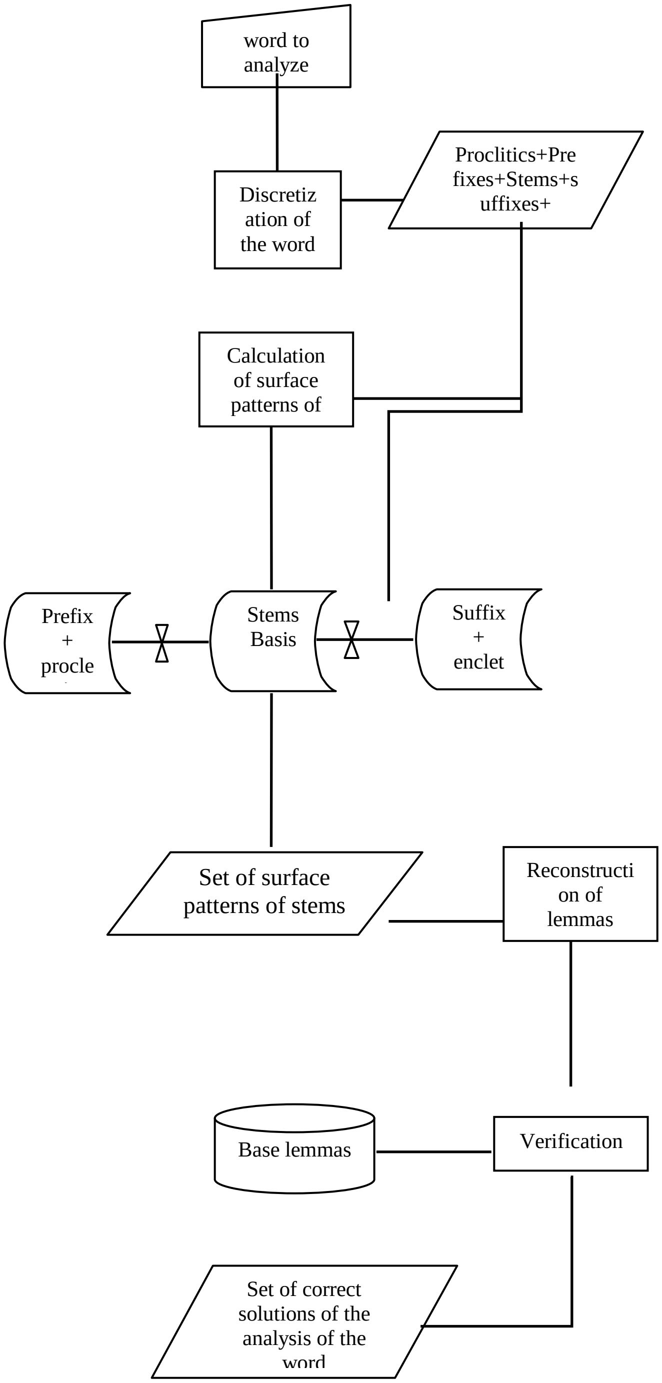

word to analyze
Discretiz ation of the word
Proclitics+Pre fixes+Stems+s uffixes+
Calculation of surface patterns of
Prefix + procle
Stems Basis
Suffix + enclet
Set of surface patterns of stems
Reconstructi on of lemmas
Base lemmas
Verification
Set of correct solutions of the analysis of the word

Fig. 5. Schema of the Morphological Analysis Process

IV. Experimentation and Results

A. Implementation

We used with the Java language to build our analyzer morphological. Our application contains three layers including, the presentation layer which processes the data reading part and the result display. The second layer contains the implementation of the algorithm where the rules and filters to apply to analyze a given word as input. The third layer deals with the communication with the MySQL database. The Fig. 6 illustrates the different layers of our analyzer.

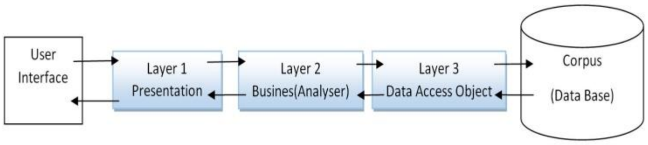
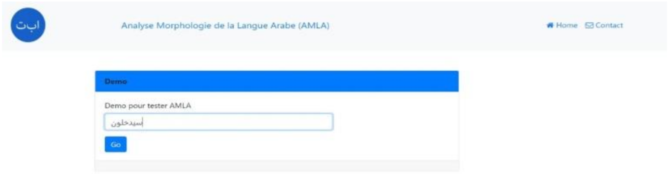


Fig. 6. Analyzer Implementation Layers.

B. Application

To see the results, we have developed a demo as a web interface that allows a user to analyze a word entered as input.

Fig. 7 shows the demo page, which contains a button to start, a text area for entering a word to analyze. The text area accepts only Arabic letters.



Demo pour tester AMLA

أسيدخلون

Fig. 7. User interface of our Analyzer.

Example:

For the word "أموته", our system goes through the steps:

- the possible discretizations are showed on Table IV:
- Creation of the table (noted table-sol) of the surface patterns of stems and stems, result is showed on Table V.
- The creation of a query from the table-sol as shown in Table VI.

TABLE IV. Possible Discretization

proclitic	prefix	Stem	suffix	enclitic
		أموته		
	أ	موته		
	أ	موت		ه
		أمو	ت	ه
أ		موته		
أ		موت		ه
أ		مو	ت	ه

TABLE V. SURFACE PATTERNS OF STEMS AND STEMS

Proclitic	prefix	Stem+pattern de stem	suffix	Enclitic
		فعللل		
	أ	فعلل		
	أ	فعل		ه
	أ	فول		ه
		فعل	ت	ه
		فعو	ت	ه
أ		فعلل		
أ		فعل		ه
أ		فول		ه
أ		فع	ت	ه
أ		فو	ت	ه
		أموته		
	أ	موته		
	أ	موت		ه
		أمو	ت	ه
أ		موته		
أ		موت		ه
أ		مو	ت	ه

TABLE VI. RESULT OF QUERY

Proclitic	prefix	Stem+pattern de stem	Root	suffix	enclitic
	أ	فول	موت		ه
أ		فعل	موت		ه
أ		موت	موت		ه
	أ	موت	موت		ه

C. Validation and Comparison

To validate our proposed analyzer, we performed significant experiments on the database discussed in the previous section. Our test database contains 20000 words manually constructed from the prefix, suffix and infix. These test words are validated by linguistic experts.

Several tests were performed to evaluate the recognition rate of the analyzer based on the number of words including false words and valid ones. So the Table VII is going to explain this recognition rate.

These results show a significant validation rate. Our analyzer extracts valid words with a rate of 98%. The robustness of our analyzer is demonstrated by the number of possible solutions found.

TABLE VII. THE RECOGNITION RATE

Number of test words	Invalid words	Possible solutions	Valid solutions	Validation rate
20000	1000	100000	98000	98%
15000	0	85000	84000	98.8%

The error rate of our analyzer does not exceed 2%. Most of the found errors are related to insufficient corpus used, which means that our approach is robust against possible false solutions. This robustness is validated by our corpus and the number of criteria used to filter the invalid solutions.

V. CONCLUSION

The approach presented in this article, uses the relational concept relative to databases for making the morphological analysis of Arabic words.

In this approach, the Arabic morphological rules are modeled by links between the different tables used in the main database. The main advantage of this approach is its simplicity of implementation. Moreover, all the variations and the morphological rules are included in the relations between tables. The results obtained are satisfactory and show the importance of the proposed approach.

As future work, we will focus on adding new rules to improve results. Also, we are interested to upgrade our database in order to challenge the performance of our method.

REFERENCES

[1] Alexia Blanchard, Analyse morphologique des réponses d'apprenants en environnement d'Apprentissage Assisté par Ordinateur. Mémoire de Master, Université Stendhal-Grenoble III,UFR des Sciences du Langage, 2006.

[2] Al-Fedaghi, S. S., and A1-Anzi, F. S. 1989. A New Algorithm to Generate Arabic Root-Pattern Forms, Proceedings of the 1 lth National Computer Conference and Exhibition, March, Dharan, Saudi Arabia, 391-400.

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

[4] Audebert C, Jaccarini A. (1988). De la reconnaissance des mots outils et des tokens. Annales islamologiques 24, Institut francais d'archeologie orientale du Caire.

[5] Azmi, Aqil, Reham S Almajed. 2015. A survey of automatic Arabic diacritization techniques. Natural Language Engineering, 21, pp 477–495. doi:10.1017/S1351324913000284.

[6] Gaubert C., « Analyse morphologique d'un texte par ordinateur – Résultats et évaluation », AnIsl 29 (1996), IFAO, p. 283-311.

[7] Jaafar, Y., Bouzoubaa, K., Yousfi, A., Tajmout, R., & Khamar, H. (2016). Improving Arabic morphological analyzers benchmark. International Journal of Speech Technology, 19(2), 259–267. doi:10.1007/s10772-016-9340-x.

[8] Goldsmith and John.A (2001). Unsupervised learning of the morphology of a natural language. Computational Linguistics, 27(2), 153-198.

[9] Hilal, Yahiah, 1985 : Morphological analysis of Arabic speech. In: Computer processing of the Arabic language.

[10] Beesley.KR (1998). Arabic Morphology Using Only Finate-State Operations, Proceedings of the Workshop on Computational Approaches to Semetic languages. Montreal, Quebec, pp 50-57.

[11] K. R. Beesley and L. Karttunen, Finite State Morphology, CLSI Studies in Computional Linguistics, vol.509, 2003.

[12] Soudi, A., Violetta Cavalli-Sforza (2001). A Computational Lex-eme-Based Treatment of Arabic Morphology. In Proceedings of the Association for Computational Linguistics, Arabic Processing Workshop, Toulouse, July 2001, France.

[13] Boudchiche. M,Mazrouia. A. al 2017 . AlKhalil Morpho Sys 2: A robust Arabic morpho-syntactic analyze. Journal of King Saud University - Computer and Information Sciences Volume 29, Issue 2, April 2017, Pages 141-146.

[14] Buckwalter.T (2002). Buckwalter Arabic Morphological Analyzer. Version 1.0. Linguistic Data Consrtium, catalog. Number LDC2002L49 and ISBN 1-58563-257-0.

[15] Darwish.K (2002). “Building a Shallow Morphological Analyzer in One Day”. Proceedingsof the workshop on Computational Approaches to Semitic Languages in the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02). Philadelphia, PA, USA.

[16] S.Iazzi, A.Yousfi, M.Bellafkih, D.Aboutajdine 2018 : Arabic Morphological Analysis Based On Graphs And Correspondence tables Between Affixes And Root. ISIVC 2018: 318-322.

[17] S.Iazzi, A.Yousfi, M.Bellafkih. 2020: Comparison between the morphological analyzers based on graph and the one based on surface patterns. SITA 2020: 26:1-26:4.

[18] Iazzi, S, Yousfi, A, Bellafkih, M, Aboutajdine, D. Graph-based morphological analysis. Journal of Computer Science and Engineering Volume 19, Issue 2 June 2013.

[19] Iazzi, S, Yousfi, A, Bellafkih, M, Aboutajdine, D. Morphological Analyzer of Arabic Words Using the Surface Pattern. IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 2, No 1, March 2013, ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784.

[20] Koskenniemi and Kimmo (1983). Two Level Morpology. A General Computational Model for Word-form Recognition and Production. Publication No. 11, Dep. of General Linguistics, University of Helsinki, Helsinki.

[21] Yousfi.A, Iazzi.S : "نحو محلل صرفي عربي يعتمد على أوزان الكلمة". 7th International Computing Conference in Arabic (ICCA'11). Riyadh, Saudi Arabia (May 31- June 2, 2011).

[22] Yousfi.A (2010). The morphological analysis of Arabic verbs by using the surface patterns. IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No 11, May 2010.

[23] G. D. Forney, “The Viterbi Algorithm,” Proceedings of IEEE, Vol. 61, No. 3, 1973, pp. 268-278.

[24] Dichy J. & Fargaly A. (2003), Roots & Patterns vs. Stems plus Grammar-Lexis Specifications: on what basis should a multilingual lexical database centred on Arabic be built?, Proceedings of the MTSummit IX workshop on Machine Translation for Semitic Languages, New-Orleans.