# Towards detecting anomalies in the content of standardized LMF-dictionaries

**Wafa WALI**
MIR@CL Laboratory,
FSEGS,Sfax,Tunisia
wafa.wali@fesgs.rnu.tn

**Bilel GARGOURI**
MIR@CL Laboratory,
FSEGS,Sfax,Tunisia
bilel.gargouri@fesgs.rnu.tn

**Abdelmajid BEN HAMADOU**
MIR@CL Laboratory,
ISIMS,Sfax,Tunisia
abdelmajid.benhamadou@isimsf.rnu.tn

## Abstract

Dictionaries are reference resources for learning and diffusing natural languages. Their contents must be enriched carefully due to their importance. However, such contents might contain errors and inconsistencies that are hard to detect manually. Several researches have been made in recent years in order to perform this step automatically. However, they have dealt with the problem in a superficial way. The present paper deals with the detection of anomalies in the content of LMF-standardized dictionaries that covers lexical knowledge at the morphological, syntactic and semantic levels. Thus, we are proposing an approach based on a typological study of the potential anomalies that can occur in editorial dictionaries in general. This approach takes advantage of the LMF fine structure that highlights all kinds of relationships between entries' knowledge and distinguishes the role of each available text such as giving definitions and examples. An experiment of the proposed approach was carried out on an available LMF-standardized dictionary of the Arabic language. This experiment has been related to the morphological and syntactic levels.

## 1 Introduction

Dictionaries are important linguistic resources for learning and diffusing natural languages. They can be used for several purposes such as to find the meaning, the translation, the synonym or antonym of a word. Moreover, they can help to check the spelling or to find out grammatical information about a word.

For ages, editorial dictionaries (for human use) have been developed in paper versions for many natural languages. With the advent of the computer science, several editorial electronic dictionaries have been constructed to be released from the constraint of their paper versions. Thus, the use of the electronic dictionaries has been expanded to meet the NLP (Natural Language Processing) needs. Then, several models have been proposed to represent the dictionary knowledge. In addition, some projects have suggested a common representation of dictionaries such as TEI (Veronis and Ide, 1996), GENELEX (Antoni-Lay et al., 1994), EAGLES (Calzolari et al., 1996) and ISLE (Calzolari et al., 2003). Moreover, an ISO standard has been proposed for modeling lexical resources and electronic dictionaries accordingly. This standard, named Lexical Markup Framework (LMF: ISO 24613), provides a finely structured representation of large and common lexical knowledge (Francopoulo et al., 2008).

On the other hand, a good dictionary must contain accurate knowledge to give the right answers for any use. Thus, it is very important to assess the quality of dictionaries' contents, which is expensive to perform manually and requires high linguistic expertise (Fersoe and Morachina, 2004). In this context, a few works have been devoted to the evaluation of electronic dictionaries for many Latin and bilingual dictionaries (Zagic et al., 2011), (Rodrigues et al., 2011). For some other languages such Arabic, the published works still deal with paper versions (Alkhatib, 1967), (Alchidyâq, 1899), and (Hamzaoui, 1986). Thus, we can qualify the evaluation of dictionaries content as very important, notably with an automatic process.

In this paper, we are dealing with the automatic detection of anomalies in the content of standardized LMF dictionaries starting from a typological study of pertinent anomalies. In fact, we propose an approach that takes advantage of the fine structure of LMF. Indeed, LMF highlights all kinds of relationships between entries knowledge and distinguishes the role of each available text such as giving definitions and ex-



*Proceedings of Recent Advances in Natural Language Processing*, pages 719–726,
Hissar, Bulgaria, 7-13 September 2013.

amples. In order to experiment the proposed approach, we applied it on an available standardized dictionary for the Arabic language (Khamekhem et al., 2012). This experiment is related to the morphological and syntactic levels.

We are going to start with presenting some works related to the evaluation of dictionaries' contents. Then, we are reporting a typological study on the pertinent anomalies in the standardized dictionaries. Thereafter, we are describing the proposed approach. Finally, we are detailing the experiment that we carried out and we are giving the obtained results.

## 2 Related works

In this section, we have presented the most rele--vant works related to the evaluation of dictionaries. Some works are proposed to evaluate content of monolingual and bilingual dictionaries in paper versions. For monolingual dictionaries, most of the works focused on problems such as false derivation, incoherence of definition and incoherence between the example and the definition. These works deal with paper versions of dictionaries and are relatively old such as (A. Alkhatib, 1967), (A.F.alchidyâq, 1899), (I.Ben Mrad, 1987) and (M.Hamzaoui, 1986) that are dedicated for the Arabic dictionaries. Other works (M.Asfour, 2003), (M.Khoury, 1996), (A.Kasimi, 1998) dealt with the evaluation of bilingual dictionaries. They specially deal with translation problems.

Moreover, a few efforts are made to detect anomalies for electronic dictionaries as (Zagic et al., 2011) and (Rodrigues et al., 2011). The authors elaborated methods for detecting and correcting OCR problems in Urdu- English digital dictionaries using Dictionary Language Modeling (DML). However, these dictionaries are poorly structured resulting in the digitalization of paper versions. Furthermore, this situation generates a handicap for the evaluation of electronic dictionaries that require fine structure of the dictionary entries.

Finally, we believe that the lack of works on automatic detection of anomalies in the contents of dictionaries can be explained by the complexity of this task.

## 3 Study of anomalies in LMF standardized dictionaries

Based on subtle, powerful, universal LMF metamodel and applied to all natural languages, the present study was carried out on LMF standardized models of dictionaries for three languages used in the world (English, French and Arabic).

The dictionaries that we will evaluate, resulting from the conversion a paper dictionary in electronic version or went through a strict acquisition system. In this section, we will aim to give an overview of the standard LMF and to identify and classify pertinent anomalies in such dictionaries. We focused mainly on the morphological, syntactic and semantic linguistic levels.

### 3.1 Lexical Markup Framework-ISO 24613

The Lexical Markup Framework (LMF) (Francopoulo et al., 2008) provides a generic metamodel that can be applied for most natural Languages. It is composed of a core and several optional extensions as indicated in Figure 1 given below. The core and the extensions contain several classes detailing all lexical knowledge and the relationships between them. We can select the extensions and/or the classes with respect to a specific need to construct a dictionary. The selected model will be decorated by data categories from the DCR (Data Categories Registry) standardized with respect to the ISO 12620 standard.

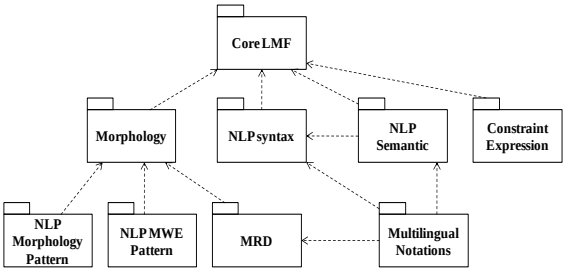


Figure 1: The LMF core and its extensions

### 3.2 Morphological anomalies

In the morphological model, each lexical entry has one lemma, many word forms that represent their inflected forms and morphological features (grammatical number, grammatical gender, person…) and many ordered stems. Indeed, each root or derived form in separate lexical entry are connected them by the class RelatedForm which has a Data Category (DC) type. This DC allows us to specify the type of relationship between the lexical entries whether it has a stem or a root. Thus, two kinds of anomalies can occur. The first one has something to do with false values of properties as shown in Figure 2.

Normally, the inflected form Muslims "مُسْلِمُون" is the plural of word Muslim "مُسْلِمٌ" as described in figure2. But it can find the anomaly mentioned in figure 3 such as the value of the attribute

"grammatical number" of the inflected form is singular.

| | | |
|---|---|---|
| Lemma | مُسْلِمٌ | Muslim |
| Inflected form in the plural | مُسْلِمُون | Muslims |

Figure 2: Example of Muslims"**مُسْلِمُون** "

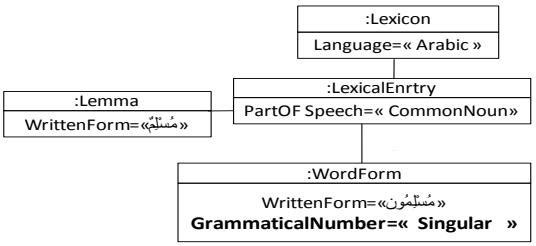


Figure 3: Illustration of the anomalies in proprieties values

The second anomaly is related to false morphologic links like incoherence between stem and lemma or incoherence between root and lemma. The Arabic word" مَكْتَبٌ - bureau" has a root "كَتَبَ - write" like the one presented in figure 4. Although, it can induce an anomaly as shown in figure 5 such as the root of the word" مَكْتَبٌ - bureau" is "كبت - inhibit".

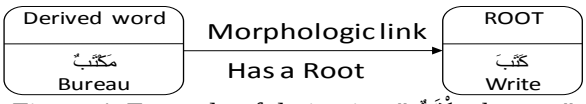


Figure 4: Example of derivation " **مَكْتَبٌ** -bureau"

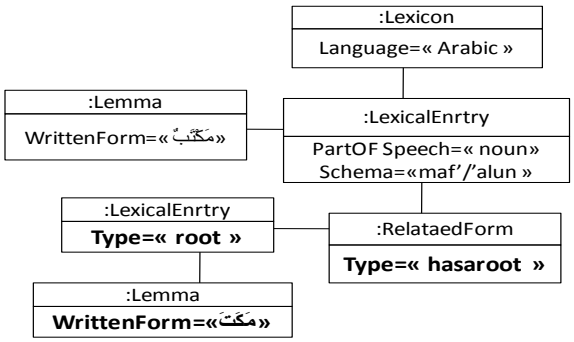


Figure 5: Illustration of anomalies in morphologic links

### 3.3 Syntactic anomalies

The syntactic model presents the syntax of sentences through sub-categorization frames. Then, it specifies the possible frames of a LexicalEntry (LE) and for each frame it specifies the various senses of the LE. The main class of syntactic model is SubcategorizationFrames that is a syntactic behavior of LE. This class is composed of a set of Syntactic Arguments and a LexemeProperty that include the characteristics of the central node of this frame.

In this syntactic model, we can find two types of anomalies like the incoherence between syntactic behavior and example. Indeed, the example" أَخَذَ الوَلَدُ الكِتَابَ-the boy takes the book" given in figure 6 has a syntactic behavior "verb subject object (VSO)". However, it can cause an error as indicated in figure 7 and present the syntactic behavior of the example like "subject verb (SV)".

| | | |
|---|---|---|
| Example | أَخَذَ الوَلَدُ الكِتَابَ | the boy takes the book |
| Syntactic behavior | O S V | |

Figure 6: Example of syntactic behavior "VSO"

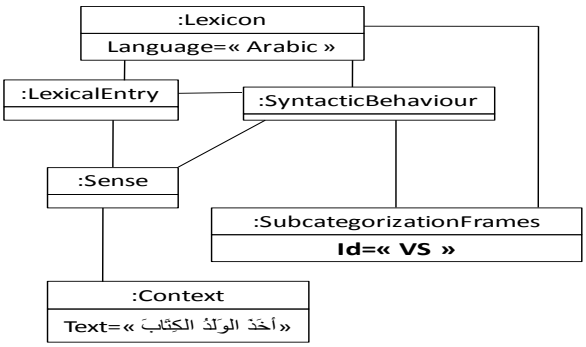


Figure 7: Illustration of the anomaly "incoherence between example and syntactic behavior"

The second anomaly related to the syntactic level is the incoherence between example and information in the LexemeProprety class. The example presented in figure 8 "أَخَذَ الوَلَدُ الكِتَابَ- the boy takes the book" is in the active voice. But, it can have an anomaly as it was mentioned in figure 9 such as the voice of example is passive voice.

| | | |
|---|---|---|
| Example | أَخَذَ الوَلَدُ الكِتَابَ | the boy takes the book |
| Voice of example | Active Voice | |

Figure 8: Example in active Voice

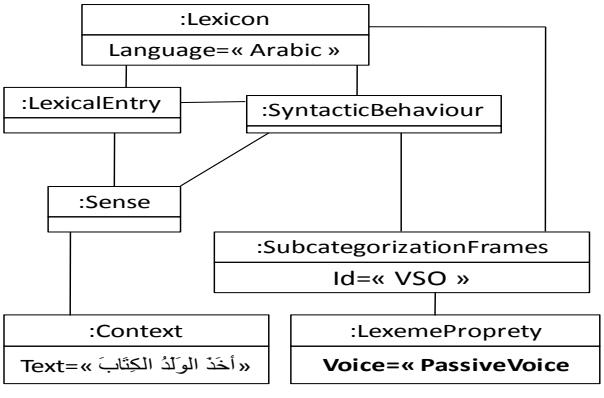


Figure 9: Illustration of anomalies of propriety values related to context and Lexeme Proprety class

### 3.4 Semantic anomalies

The senses of word may be general or specific to one field and may belong to a semantic class. In addition, The SenseRelation allows us to connect the senses belonging to different lexical entries with several types of relationships such as the synonym, the antonym. The SenseExample represents an instance of a given sense. SubjectField and Context are two classes from MRD extension. The first class is used when the meaning is specific to a particular area and the second one represents an example of using a LE in the frame of a given sense. Furthermore, the standard has represented the overlap between syntax and the semantics in the semantic extension.

For this model, we might find the following anomalies: incoherence sense (in Definition class), incoherence domain (in SubjectFieled class), redundancy of examples and senses, incoherence between example (in Context class) and sense, lack of explanation like definitions based on references (null pointer, synonymy or antonym), false semantic relations and incoherence between example and semantic class. Figure 10 shows semantic relations between three lexical entries. The sense 1 of word 1 is a synonym with the sense 3 of word 2 and the sense 2 of word 3 is a synonym with the sense 3 of word 2. Therefore, transitively speaking, the sense 2 of word 3 and the sense 1 of word 1 are synonyms. Nevertheless, in figure 11 presents the two senses (sense 2of word 3 and sense 1 of word 1) described previously as antonyms.

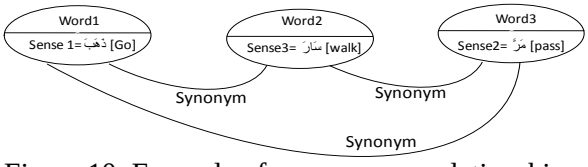


Figure 10: Example of synonymous relationships

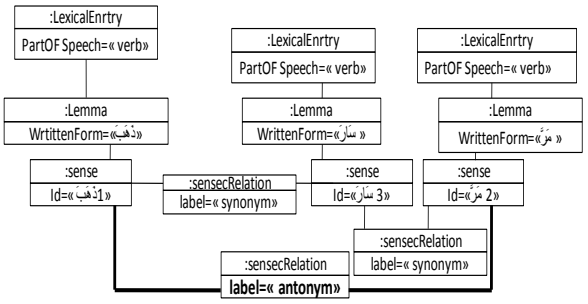


Figure 11: Illustration of semantic relations anomaly

The figure12 schematized below, presents an attribute value of semantic class" human" for the subject "the boy الوَلَدُ". But, it can cause an anomaly as shown in figure13 and presented the semantic class of subject like inanimate concrete.

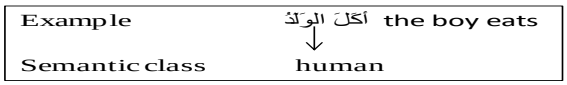


Figure 12: Example of semantic class

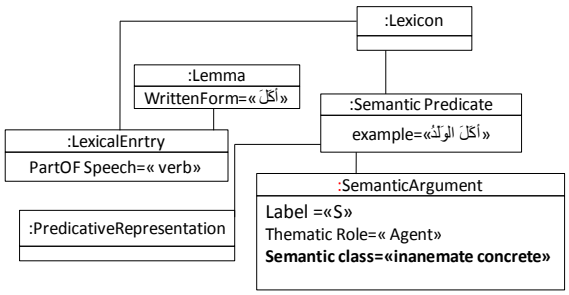


Figure 13: Illustration of anomalies attribute values for a semantic class

## 4 Overview of the approach

In this section, we give an overview of the approach that we propose for detecting anomalies in the content of LMF-standardized dictionaries. This approach consists mainly of three stages as shown in Figure 14. Firstly, we check the structure of dictionaries according to the DTD of LMF. Secondly, we proceed to verify the validity of the properties inside classes and finally we deal with coherence of properties that have connections outside classes. In the following figure, we detail the three stages of the proposed approach.

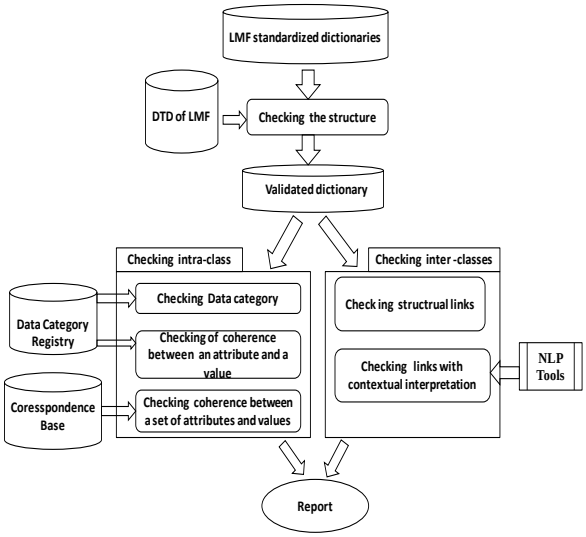


Figure 14: The approach overview for detecting anomalies in LMF-standardized dictionaries

### 4.1 Check of the structure

In this initial stage, we intend to check the structure of the dictionary dealt with. In the case of

encoding with XML (eXtensible Markup Language), this step is simple to perform. It consists of verifying the dictionary structure with respect to the DTD (Data Type Description) of the standard LMF. In the case of a relational encoding of the dictionary database, an appropriate reference schema should be used.

### 4.2 Check intra-class

The second stage consists in verifying the properties (Attributes and values) inside each class by checking at the beginning the used Data Category (DC) with respect to the Data Category Register (DCR). Then, we check the coherence between the used attributes and the associated values. Each selected attribute from the DCR has its appropriate values which are also specified in the DCR. Finally, we check the coherence between two DC, using a set of correspondence rules according to language specificities.

### 4.3 Check inter-classes

The purpose of this final stage is to verify the coherence between properties (attributes and values) situated in different classes. To achieve this, we inspect all existing links between the classes of the LMF-standardized dictionaries. For instance, in the morphological extension, we can have false structural links like LE1, which has a root LE2 and has a stem LE3, LE2 has a stem LE3. Also, in the semantic extension, we might have structural links anomaly such as LE1 is synonym with LE2, LE2 is synonym with LE3 and LE3 is antonym with LE1. Afterwards, for each extension of LMF-standardized dictionary, we verify the links with contextual interpretation by applying various NLP tools. For example, the verification of coherence between example and syntactic behavior requires primarily the use of a parser to obtain the syntactic tree of the example and then verify this structure with syntactic behavior described in the Syntactic Behavior class.

## 5 Case study: detection of morphological and syntactic anomalies in LMF-standardized Arabic dictionary

The proposed approach was applied to a case study and the experiment was carried out on the Arabic language. This choice is explained both by the great deficiency of work in evaluating electronic Arabic dictionaries and the availability within the research team of an LMF standardized Arabic dictionary containing about 37.000 entries.

To automatically perform the stages of the proposed approach, we developed a system using Java and NetBeans IDE7.2 environment (see Figure 16).

### 5.1 Fundamentals of the Arabic morphology

Arabic is a derivational and a flexional language. The base of the derivation process is a root composed of three out of four letters. Then, the obtained lemma can be a stem for another lemma. Each one is characterized by a schema that consists of presenting the model of its derivation. The base of the schema is composed of the three letters f [ف], E [ع], l [ل]. The schemas are classified according to the Parts Of Speech (POS).
Moreover, in the Arabic standard, the words contain vowels associated with their letters. The vowels are used to distinguish words that are composed of the same sequence of consonants but they are semantically different such as “kabar” [كَبَرَ], “kabur” [كَبُرَ] and “kabir” [كَبِرَ] [16]. Moreover, these vowels must be coherent to the indicated schema and can have an influence on the flexion process.
These characteristics are, among others, considered in the LMF normalized model of the used dictionary.

#### 5.1.1 Steps of the morphological detecting process

The proposed process is composed of the following four steps: (i) the verification of vowels, (ii) the verification of the coherence between POS and schemas, (iii) verification of the coherence between the stems and the lemmas (iv) the verification of the coherence between the roots, the schemas and the lemmas. The two first steps belong to the stage of validity intra-classes whereas the third and the fourth steps belong to the stage of inter-classes coherence. Figure 15 given below synthesizes the morphological detection process.

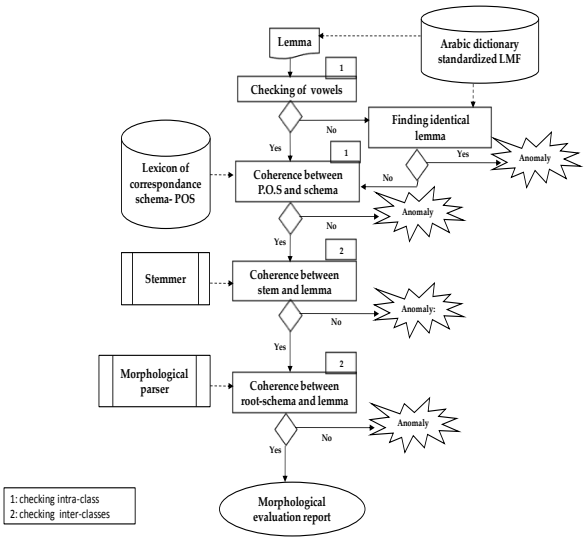


Figure 15: Morphological evaluation process

**Verification of vowels:** The aim of this step is to verify the used vowels of all the lemmas in the dictionary. In this step, we detect an anomaly if there are two lemmas like LE, that are using the same sequence of letters and one of them or both have no vowels.

**Verification of the POS-schema coherence:** The second step is to verify the coherence between POS and schema. To check this coherence, we need a lexicon of correspondence between Arabic schemas and its POS. At this phase, we used a lexicon which is enriched manually by an expert.

**Verification of the stem- lemma coherence:** This step consists of checking the coherence between stems and lemmas. According the standard LMF-ISO 24613, the stem is a sequence of morphs that is smaller than or equal to the form of a single lexeme and that may be affected by an inflectional, agglutinative, compositional or derivative process.

Moreover, the link between a lemma and its stem is presented through the RelatedForm class of the morphological extension. The stem does not need to be identical to the root of the word. In this stage, we used the "khoja Arabic stemmer" (S.Khoja, 2001) developed in Java. It removes the longest suffix and prefix. It then matches the remaining word with verbal and noun patterns to extract the stem.

**Verification of the root-schema-lemma coherence:** The last step consists of verifying the coherence between the root, the schema and the lemma that are based on the available information in the LexicalEntry (schema), Lemma (lemma) and RelatedForm (root) classes. For checking this coherence, we need a morphological parser. In our work, we used the MORPH parser (Chaabane et al., 2010).

### 5.1.2 The obtained results

Figure 16 illustrates the detection process and gives the obtained results at the end of this process. The percentage of incoherent entries can be due either to an inconsistency or absence of entry in the data base of the systems used (MORPH, khoja Arabic stemmer).
As shown in this Figure:

- The verifying of vowels: 96% of the entries contain vowels and 4% of them are without vowels.

- The coherence between schema and POS: the rate of coherent entries is 69% and the rate of incoherent entries (incorrect entries + unrecognized entries) is 31%.

- The coherence between stem and lemma: the rate of coherent entries is 75% and the rate of incoherent entries (incorrect entries + unrecognized entries) is 25%. This is explained by the absence, until now, of links between lemmas and their stems in the available dictionary.

- The coherence between root, schema and lemma: the rate of coherent entries is 57, 14% and the rate of incoherent entries (incorrect entries + unrecognized entries) is 42, 85%.

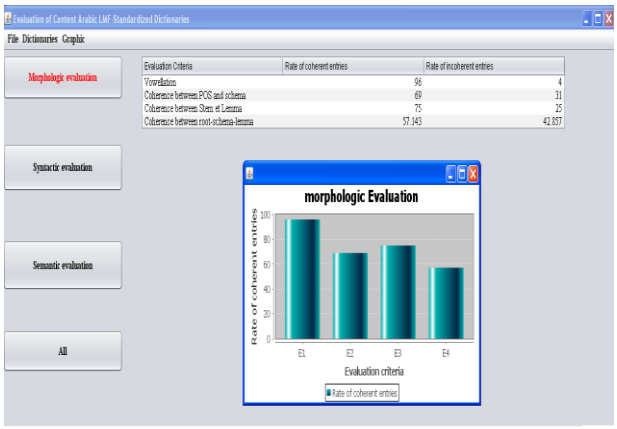


Figure 16: System outputs

## 5.2 The bases of the Arabic syntax

Parsing Arabic sentences is a difficult task due to the following reasons (Othman et al., 2003): first, the Arabic sentences are long and complex. Second, the Arabic sentence is syntactically ambiguous and complicated due to the frequent usage of grammatical relations, the order of words and phrases, conjunctions, etc. For the last two decades, concentration of the Arabic lan-

guage processing has focused on morphological analysis. In contrast, there were fewer works related to on syntactic analysis of Arabic.

To detect the anomalies of the syntactic level, we use the platform NOOJ[1].

NOOJ is a linguistic environment of development that can analyze a large corpus in real time. It includes tools to build, test and maintain formalized descriptions of natural languages (in the form of electronic dictionary or grammar) (M.Salbeztein, 2005).

NOOJ can build lemmatized concordances for a large text using finite state grammar, and can also perform transformations on texts hidden in order to annotate or produce paraphrases. The lexical module of NOOJ used in the detection of syntactic anomalies, is based on syntactic grammar.

This grammar is represented in the form a finite-state nodes. It represents sequences of grammatical categories corresponding to the production of a sentence. Although these grammatical categories are predefined by NOOJ (e.g. <V> verb, <S> subject, <PREP> preposition, <PRON> pronoun, <LOC> noun of place, etc.)

### 5.2.1 Steps of the syntactic detecting process

The proposed detection process is based primarily on the study of an example in order to compare the structure of the example with the syntactic behavior described in the Arabic standardized LMF dictionary and verifies the coherence between the voice of the example (passive voice or active voice) and the information presented in the Lexeme Proprety class.

Figure 17 given below synthesizes the syntactic detection process.

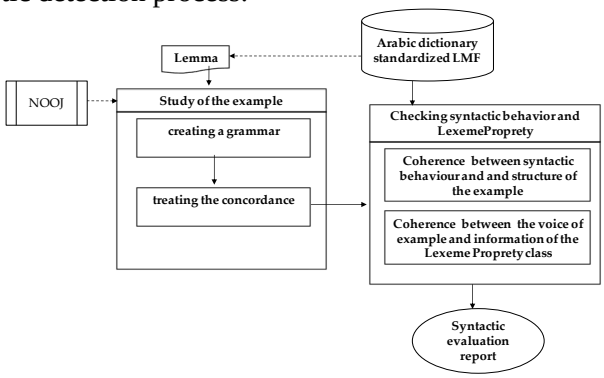


Figure 17: Syntactic evaluation process

**Study of the example**: in the platform NOOJ, we create a grammar corresponding to the examples presented in the Arabic standardized LMF dictionary so generate the concordance for verify the coherence with syntactic behavior and the information of the Lexeme Proprety class. This grammar is formed by seven nodes, besides to the two nodes: start and end. The nodes that are used: <V> verb, <N> noun, <PRON> pronoun, <PREP> preposition, <PREF> prefix, <ADJ> adjective, <LOC> noun of place.

**Verification of syntactic behavior and Lexeme Proprety:** in this step, we check the coherence between the syntactic behavior presented in the Arabic standardized LMF dictionary and the syntactic behavior described in the concordance table.

Also, NOOJ annotates for each verb the voice which is appropriate. This information is compared to information in *lexeme Proprety* class in the Arabic standardized LMF dictionary.

## 6 Conclusion

In this paper, we proposed an approach based on a typological study of the potential anomalies that can occur in LMF standardized dictionaries. The originality of this approach lies in the use of a unique, finely-structured source, rich in lexical and conceptual knowledge at the morphological, syntactic and semantic levels. Our method consists of three stages. It starts with verifying the structure of LMF dictionaries with respect to the DTD of LMF. Then, it performs the verification of properties in each class. Finally, it verifies the inter-classes links. In addition to, the experiment of the proposed approach carried out on an available LMF-standardized dictionary of Arabic language.

This experiment is related to the morphological and syntactic levels. For future works, we aim to deal with the automatic detection of semantic anomalies. In addition to that, we plan to extend the experiment to cover other languages.

## References

Akhatib A. 1967. "*Arabic dictionary between the past and the present*". Nachiroun library, Liban

Alchidyâq A. F.1899،*The spy on the dictionary*". Sâdir library, Beirut

Antoni-lay MH. Francopoulo G. and Zayssern L. 1994. *A generic model for reusable lexicons: The genelex project*, Literary and Linguistic Computing, 1994.

[1] The download free and the user manual of a linguistic platform NOOJ are available at: http://www.nooj4nlp.net

Asfour M. 2003 *Problems in modern English-Arabic lexicography*, journal of Zeitschrift für arabische Linguistik, 2003, N°42, pp 41-52.

Ben Mrad I. 1987 "*Studies in the Arabic dictionary",* dar Algharb Alislâmi, Beirut, Liban.

Calzolari N. Bertagna F. Lenci A. Monachini M. 2003. *Standards and best Practice for Multilingual Computational Lexicons. MILE (The Multilingual ISLE Lexical Entry)*. ISLE CLWG Deliverable D2.2 et 3.2 Pisa.

Calzolari N. McNaught J. Zampolli A. 1996. *Eagles, editors introduction.* http://www.ilc.cnr.it/EAGLES96/edintro/edintro.html

Chaâben N. Hadrich Belguith L. and Ben Hamadou A. 2010. *The MORPH2 new version: A robust morpholgical analyser for Arabic texts*. In the proceedings of the 10 international days on statistical analysis of data (JADT 2010), Rome,Italy, 9-11 June 2010.

Fersoe H. and Morachina M. 2004. *ELRA Validation Methodology and Standard Promotion for Linguistic Resources*, LREC 2004, Lisbon, Portugal.

Francopoulo G. and George M. 2008. *Language Resource Management.*2008. *Lexical Markup Framework (LMF). Technical report, ISO/TC 37/SC 4 N453 (N330 Rev.16)*, 2008.

Hamzaoui M. 1986. "*Arab dictionary issues*" Dar Algarb alislâmi, Beirut, Liban.

Kasimi A. 1998. "*Problematic in Arabic lexicographical significance*". Articles literary forum integraeted and literary library.

Khemakhem A. Gargouri B. and Ben Hamadou A. 2012 *LMF standardized dictionary for Arabic language*. In the proceedings of the 1st International Conference on Computing and Information Technology (ICCIT 2012), Al-Madina, Saudi Arabia, 12-14 March 2012.

Khoja S. 2001. *Stemming Arabic Text*. http://zeus.cs.pacificu.edu/shereen/research.htm, 2001

Khoury M. 1996. *Dictionnaires arabes bilingues, présentation historique et étude comparative*, thèse de maîtrise présentée à l'école des études supérieures et de la recherche de l'université de Ottawa, Ontario, 1996

Othman E. Shaalan K. and Refea A. 2003. *Achart parser for analysing modern standard Arabic sentence*. In proceedings of the MT summit IX workshop on machine translation for semetic languages : issues and approaches, USA , pp 33-39, 2003.

Rodrigues P. Zagic D. Buckwalter T. Maxwell M. and AntonRytting C. 2011. *Quality control for digitized dictionaries*. The 9th conference of the Association for Machine Translation in the Americas, workshop on developing up dating and coordination technologies, Dictionary and Lexicons for Terminological Consisentency, October 2011.

Salibeztein M. 2005. *NOOJ's dictionaries*. In actes LTC 2005, Poznan.

Veronis J. and Ide N. 1996. "*Encodage des dictionnaires électroniques: problèmes et propositions de la TEI*". In D. Piotrowsky (Ed.), Lexicograpbie et informatique - Autour de l'informatisation du Trésor de la Langue Française. Actes du Colloque International de Nancy (pp. 239-261). Paris: Didier Erudition, 1996.

Zagic D. Maxwell M. Doermann D. Rodrigues P. and Bloodgood M. 2011. *Correcting Errors in Digital Lexicograhic Resources Using a Dictionary Manipulation Language*. Proceedings of eLex 2011, pp. 297-301.