Towards an automatic conversion approach of editorial Arabic dictionaries into LMF-ISO 24613 standardized model

Aïda Khemakhem*, Imen Elleuch*, Bilel Gargouri*, Abdelmajid Ben Hamadou**
MIRACL Laboratory
* FSEGS, B.P. 1088, 3018 Sfax, Tunisia
khemakhem.aida@gnet.tn elleuch_imen@yahoo.fr bilel.gargouri@fsegs.rnu.tn
** ISIMS, B.P. 242, 3021 Sakiet-Ezzit Sfax, Tunisia
abdelmajid.benhamadou@isimsf.rnu.tn

Abstract

In this paper, we propose an automatic approach to convert electronic Arabic dictionaries of human (editorial) use into an LMF-ISO 24613 standardized model in order to benefit from the precious resources of Arabic language available as digitalization dictionaries on the Internet in different formats (i.e., word, txt, html). Our approach is composed of five steps (Identification of the markers, Definition of the kinds and the order of linguistic blocks, Macro segmentation, Micro segmentation, Mapping) that consider the noted diversities between Arabic dictionaries of human use such as the content, the macro structure and the micro structure. The proposed approach was validated by carrying out an experimentation on the Dictionary Al-Ghany. The content of the obtained standardized version of this dictionary was successfully queried using the ADIQTO system.

I- Introduction

Dictionaries are linguistic resources which are quite important for the learning and the maintaining of natural languages. Indeed, it is ages since editorial dictionaries (for human use) have been developed in paper versions for many natural languages. Furthermore, with the advent of Computer Sciences, several editorial electronic dictionaries are proposed to release from the constraint of their paper versions and benefit from the high capacity of storage and facilities of access methods.

In this context, several works attempted to convert a paper version of a dictionary to an electronic one using consortium guidelines for the modelling such as TEI (Text Encoding Initiative http://www.tei-c.org) (Arragi et al., 2002; Dendien et al., 2003).

For the Arabic language, the development of electronic dictionaries is in a phase that can be described as primal. Indeed, several of the available Arabic electronic dictionaries are nothing but a digitalization of a non-structured and less beneficial dictionary content (in different formats : ".doc", ".txt", or ".html") (Ait Taleb, 2005).

Recently, an ISO standard was proposed for modelling lexical resources of the majority of languages among them is the Arabic language (Khemakhem et al., 2006). This standard is called LMF (Lexical Markup Framework) and has the code ISO 24613 (Francopoulo & George, 2008).

Moreover, a unified and standardized model was proposed (Baccar et al., 2008) for the editorial Arabic dictionaries according to the LMF-ISO 24613. Thanks to its subtle and standardized structure, it was possible to develop the ADIQTO system (Arabic DIctionaries Query TOols) that implement generic interrogation of all LMF standardized Arabic dictionaries for editorial use (Baccar et al., 2008).

Starting from the standardized model, and in order to benefit from the precious resources of Arabic language available as digitalization dictionaries on the Internet, we propose an automatic conversion approach of these dictionaries into the LMF standardized model. Our approach takes into account the main diversities that we noted between the majorities of these dictionaries. These diversities are related to the variety of the Arabic lexicography schools and concern the content (the description of the contents of the entries varies from a dictionary to another), the macro structure (the organization of the lexical entries differs from a dictionary to another) and the micro structure (the organization of the linguistic information, on the level of the lexical entries, varies from a dictionary to another and even within the same dictionary).

The paper is organized as follows. First, we give an idea about the state of the art of the electronic Arabic dictionaries of human (editorial) use. Second, we present the normalized model that we developed for Arabic dictionaries according to the LMF-ISO 24613 standard. After that, we give the details of the proposed conversion approach. Finally, we describe the experimentation carried out on the Al-Ghany dictionary and present some of its results.

II- The state of the art of the electronic Arabic dictionaries of human use

The Arabic lexicography is a very ancient discipline. All along its history, it has known different schools each of which having its own specificities. Indeed, many Arabic dictionaries of human use are proposed and are very richness in content in spite of their various organizations. However, the development of the electronic versions is in a phase that can be described as primal.

Recently, several activities and projects (Ajeeb[1], Kalimet[2], Alburaq[3]…) have been reported to realize interrogation tools of Arabic dictionaries; these include a digitalization of old paper dictionaries such as the *lessen-elarab, Al-Ghany, muhit-elmuhit, elmuhit, qamus-elmuhit, elwasi* and *taj-elarous*. These electronic dictionaries are numerous and very rich in linguistic information, but they are poorly structured. So despite the richness and diversity of Arabic electronic dictionaries, search tools are not able to give good results because of the structuring weakness of the dictionary entries.

[1] http://lexicons.ajeeb.com/Results.asp
[2] http://www.kl28.com/lesanalarab.php
[3] http://www.alburaq.net/mukhtar/root.cfm

III- Normalized model for the Arabic dictionaries

The committee TC37/SC4 of ISO has published a new standard called LMF (Lexical Markup Framework) ISO 24613 for the lexical resources, notably for the construction of dictionaries (Francopoulo & George, 2008). However, we have carried out some experimentation on LMF since it was a project (Francopoulo, 2004) to make sure that it is suitable for Arabic language (Khemakhem et al., 2006).

Recently, we developed a normalized model for the Arabic dictionaries according to LMF- ISO 24613 (Baccar et al., 2008).

In this section, we start by giving a general idea of the LMF standard. After that, we describe the normalized model that we propose for Arabic dictionaries.

III-1 LMF – Lexical Markup Framework

LMF-ISO 24613 proposes a high-level conceptual model for lexical resources. It is a structural data model composed of a core package and a set of extension packages. These packages contain a set of lexical classes. These classes are described by a meta model of UML specification and can be decorated by a set of attribute-value pairs taken from a Data Category Registry (DCR) which is an ISO standard 12620 (http://www.isocat.org). Lexical classes and data categories provide the main building blocks for a common shared representation of lexical objects that allows the encoding of rich linguistic information. Data categories are a formalized representations of the most relevant linguistic concepts, such as ″part of speech″, ″grammatical gender″, etc.

III-2 The proposed model for the Arabic dictionaries

The structure of our target dictionary is constructed in two stages. In the first one, we select the necessary classes for our model. In the second, we decorate our model by the data categories that are specific to the Arabic language.

Indeed, we selected the following classes:

- ***Lexicon***: is the class that includes all lexical entries and subcategorisation frame for a given language.
- ***LexicalEntry***: is the main class to connect the lemma, inflected forms, meanings and syntactic behavior of a lexical unit.
- ***Lemma***: is a compulsory class. It contains the no conjugator lexical entry.
- ***WordForm***: contains an inflection form and morphological features.
- ***RelatedForm***: this class is very important to connect two lexical entries that have morphological link (eg, the relationship between a bare verb and its Masdar).
- ***Sense***: it includes the following semantic information: definitions, contexts (Example), domain.
- ***Definition***: is a class representing a narrative description of a sense.
- ***Context***: It represents an example of lexical entry in a sentence.
- ***SubjectField***: this class is required for the specific meaning to a particular field.
- ***SenseRelation***: this class is very important to connect two lexical entries that have semantic link (synonym, hyponym).
- ***SyntacticBehaviour***: it provides the link between meaning and syntactic behavior
- ***SubcategorizationFrame***: This class represents a syntactic construction.

In the following Figure, we present the relationship between these classes.

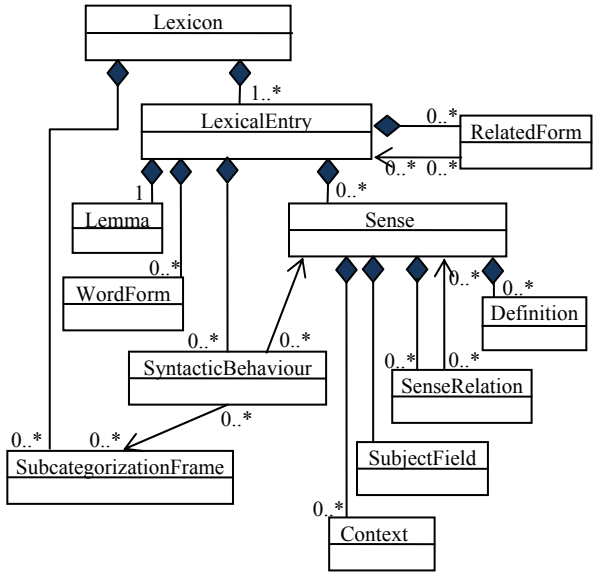

Figure 1: The model for the Arabic dictionaries

The specified attributes for each class are given in the following Table; the attributes and the different values are selected from the DCR-ISO 12620.

Class	Attributes
Lexicon	language
LexicalEntry	id partOfSpeech scheme type
Lemma	writtenForm
WordForm	writtenForm grammaticalNumber grammaticalGender verbFormAspect voice person
RelatedForm	targets type
Sense	id
Definition	text
Context	text source
SubjectField	label
SenseRelation	label
SyntacticBehaviour	id senses subcategorizationFrames
SubcategorizationFrame	id type

Table 1: List of attributes for each class

IV- The conversion approach

In this section we will present the basic idea of our approach of normalization of Arabic editorial dictionaries.

After that, we will detail the different steps of the proposed approach.

IV-1 Basic idea of the conversion approach

In order to benefit from the richness of editorial Arabic dictionaries, available in different formats (i.e., word, txt, html) on the Internet, we propose to develop a conversion system to make them in a queryable form.

It is well known that a lot of diversities are noted between these dictionaries and are related to their contents (the description of the content of the entries varies from a dictionary to another), their macro structures (the organization of the lexical entries differs from a dictionary to another) and their micro structures (the organization of the linguistic information, on the level of the lexical entries, varies from a dictionary to another and even within the same dictionary). Indeed, we use the normalized model as a target model that covers all these diversities.

IV-2 Steps of the conversion approach

The proposed approach for the normalization of Arabic editorials dictionaries is composed of five steps. They are: (i) Identification of the markers, (ii) Definition of the kinds and the order of linguistic blocks, (iii) Macro segmentation, (iv) Micro segmentation and (v) Mapping.

IV-2.1 Identification of the markers

This step consists of analyzing the dictionary in order to identify all the markers (indicators) used at the beginning of a particular kind of information. These markers can be signs not used in the Arab language, words of Arab language and/or abbreviations of Arab words. We can classify these markers into two categories: explicit markers and implicit ones.

- Explicit markers: they are clear and easy to detect.
- Implicit markers: they are specific and need a special analysis to be detected.

At this stage, the identification of the dictionary's markers is to be carried out manually by an expert person.

IV-2.2 Definition of the kinds and the order of linguistic blocks

Having detected a list of markers, now it is necessary to see which markers are used to define each block and what is the order of these blocks. This order varies from one dictionary to another and even from one unit to another in the same dictionary seeing the diversity of dictionaries structures. These blocks can be morphological, syntactic or semantic blocks or an overlapping between them.

IV-2.3 Macro segmentation

This step consists, first of identifying each lexical entry in the dictionary in a large block. Then, this block will be segmented according to the markers in specialized blocks awarding the type of each one clearly.

The result of this step is the lexical entry and blocks presented as a black box that may be of different types.

IV-2.4 Micro segmentation

The aim of this step is to undergo more segmentation of the identified blocks. As a result, if it is necessary, each block will be dealt separately and segmented again, in order to arrive to the elementary and particular information required.

IV-2.5 Mapping

Having detected specific and particular information, the last remaining step is to identify and recognize the XML tag of the target structure that will receive this information. In this step many controls and treatments must be realized to get a correct file XML valid and conform to the LMF structure of the Arabic editorial dictionaries.

All those steps can be summarized in the following Figure 2.

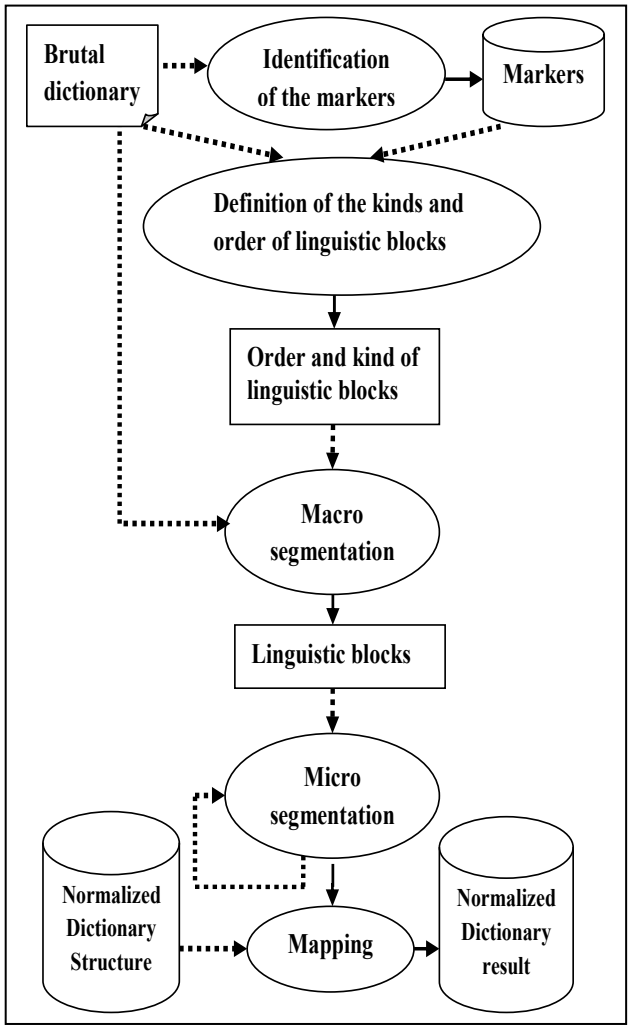

Figure 2: Steps of the conversion approach

V- Experimentation and results

V-1 Experimentation

In order to test our approach, we carried out an experimentation on the Al-Ghany «الغني» Arabic dictionary available in html format on Internet (http://www.alfaseeh.com accessed on 01/12/2006) while using JAVA and Eclipse for the implementation. The version of the electronic dictionary that we accessed is composed of about 30.000 html files at the rate of one file for each entry.

The first analyze of the dictionary have detected the list of markers. By way of examples, the Table 1 presents some of these markers.

متعد	(سيا)	(حو)	النَّوعِ مِن	المُخْتَصُّ بال	ج:	-
لازم	(فيز)	(نب)	المَرَّةُ مِن	أَفْعَلُ التَّفْضيل	ف:	.
مُثَنَّى	(طبخ)	(جغ)	اسم مِن	بِصيغَةِ الجَمْع	مؤ:	:
في ال	(كيما)	(تش)	نِسْبَةٌ إلى	مَبْنِيٌّ لِلمَجْهُول	جج:	،
خما	(حش)	(قا)	م. بحرف	مص.صِناعيٌّ	مذ	"
ثلا	(ريا)	(كما)	لازمتع	المَنْسُوبُ إلى	مذ:	[ ]
سدا	(عسك)	(طخ)	اسم أداة	صفة مشبّهة	مفع.	( )
ربا	(عس)	(نح)	صيغَةُ	مَبْنِيٌّ عَلى	فا.	مؤ
وَاحِدَةُ	(طب)	(هن)	مُثَنَّاهَا	مص.ميمي	ات	أ.
حديث	(فلك)	(فك)	بِمَعْنَى	م. بظرف	ون	1.
مف	(حسا)	(مو)	ظرف	تَصْغيرُ	مص	ن.
مفرد	مؤنث	(فل)	مُثَنَّاهُ:	فِي عِلْمِ	صف	وَقَدْ
وَاحِد	مُذَكَّرٌ	(فز)	قرآن	لِلْمُبالَغَة	جمع	ةٌ

Table 1: Some markers used in the Al-Ghany dictionary

Other markers are already used in the dictionary. Indeed, some markers exist in the dictionary in an inflected form. There are also some markers that are a combination of other ones. For example:

- The marker (مص.) : this abbreviation indicate that words after are a derived forms and the type of those derived forms is "masdar"
- The combination of markers مص. and ( ) gives a new marker :(مص.). The new marker indicates that the type of lexical entry is a "masdar".

In the following, we will use the fragment given in the Figure 3 to experiment our approach.

بَكَى - [ب ك ي]. (ف: ثلا. لازمتع. م. بحرف). بَكَيْتُ، أَبْكِي، اِبْكِ، مص. بُكاءٌ. 1."بَكَى الوَلَدُ" : سالَ دَمْعُهُ. "بَكَى صاحِبي لَمَّا رَأى الدَّرْبَ دونَهُ". (امرؤ القيس). 2."بَكَى الفَقيدَ" : رَثاهُ. 3."بَكَى صاحِبَهُ بُكاءً" : حَزِنَ، تَأَلَّمَ. "بَكَى عَلَيْهِ" "بَكَى لَهُ"

Figure 3: Fragment of the lexical entry"بَكَى"

Due to the markers already detected, we can segment this fragment using the following marker **.1** in two blocks: the first one is morpho-syntactic block and the second one is a semantic block. The block1 and the block 2 are the result of macro segmentation. We can view the result of the first segmentation in the Figure 4.

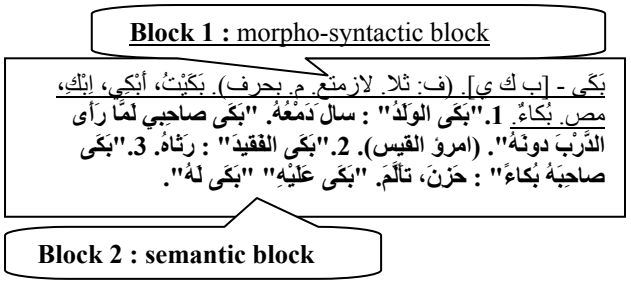

Figure 4: Marco segmentation of the lexical entry "بَكَى"

After detection of the block 1 and the block 2, it is necessary to segment each one of them separately.

✓ ***The segmentation of the block 1:***

In this segmentation we can detect the lexical entry, the root, a new syntactic block (block1.1), 3 words forms and a derived form.

The segmentation of block1 is presented in the Figure 5.

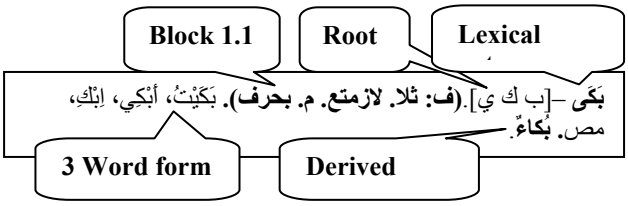

Figure 5: Segmentation of block 1

The details of this segmentation are presented in the following Table 2.

Micro segmentation of block 1		
	Marker	Information
Lexical entry	-	Lemma "بَكَى"
Root	[ ]	Root = "ب ك ي"
Type	ف:	Type="verb"
Block 1.1	(…….)	(ف: ثلا. لازمتع. م. بحرف). New Segmentation
Word form	(ف:.....). ...، ...، مص.	Word form1 =" بَكَيْتُ" Word form2 =" أَبْكِي" Word form3 =" اِبْكِ"
Derived form	مص.	Derived form =" بُكاءٌ" Type of derived form= "masdar"

Table 2: Results of the segmentation of the block 1

All the detected information are elementary aside a new block detected. So, the new block 1.1 must be segmented and the other elementary information will be replied in the XML structure.

❖ The segmentation of the block 1.1:

The new syntactic block1.1 must be segmented to obtain the syntactic behavior of the verb.

The Table 3 presents the results of the segmentation of the block1.1.

Micro segmentation of block 1.1		
	Marker	Information
Syntactic behavior	لازمتع. م. بحرف	Syntactic behavior="لازم/ متعدبواسطة"

Table 3: Results of the segmentation of the block 1.1

Now, the segmentation of the block 1 has been accomplished, and we have to segment the block 2.

✓ ***The segmentation of the block 2:***

The block 2 is a semantic block, and it necessitates to be segmented to obtain the different senses of the lexical entry. Here, we have 3 different senses of the word presented in the Figure 6.

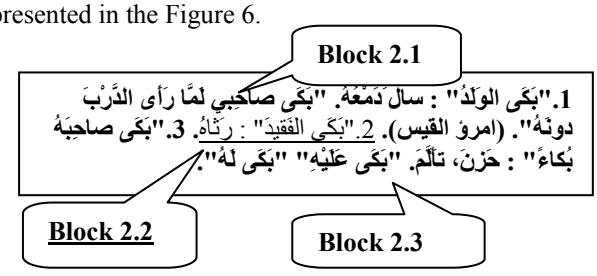

Figure 6: Segmentation of the block 2

The details of the segmentation of the semantic block 2 are described in the Table 4.

Micro segmentation of block 2		
	Marker	Information
Block 2.1	.1	Sense1 ➔ **سال : "بَكَى الوَلَدُ" دَمْعُهُ. "بَكَى صاحِبي لَمَّا رَأى الدَّرْبَ دونَهُ". (امرؤ القيس).** New Segmentation
Block 2.2	.2	Sense2 ➔ **: "بَكَى الفَقيدَ" رَثاهُ.** New Segmentation
Block 2.3	.3	Sense3 ➔ **."بَكَى صاحِبَهُ بُكاءً" : حَزِنَ، تَألَّمَ. "بَكَى عَلَيْهِ" "بَكَى لَهُ".** New Segmentation

Table 4: Results of the segmentation of the block 2

To refine the information into the identified blocks, the block 2.1, block 2.2 and block 2.3 must be segmented.

❖ The segmentation of the block 2.1:

The result of the first segmentation of the semantic block2 has given a new 3 blocks (block2.1, block 2.2 and the block 2.3). Because this result is a new blocks, we must segment each of them in order to obtain the using of the lexical entry in different phrases (example) and the different senses in many contexts or domains. Those examples of using the lexical entry are justified by other examples which refer to different sources like the Koran "قرآن", the Hadith "حديث", etc.

The segmentation of the first block 2.2 can recognize a context of using the lexical entry, the definition of this context, an example (context) that is putted from the writer Imroelkays "**امرؤ القيس** ".

This micro segmentation is described in the following Figure 7.

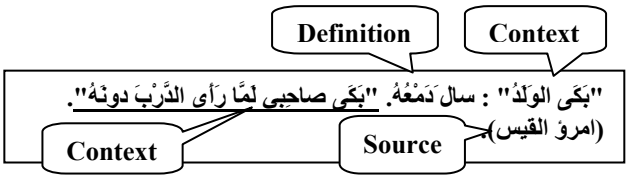

Figure 7: Segmentation of the block 2.1

The details of the micro segmentation of the block 2.1 are given in the following Table 5.

Micro segmentation of block 2.1		
	Marker	Information
Context and definition of using the lexical entry	" : " ( )	Context 1 of sense1= **"بَكَى الوَلَدُ"** Definition of context1= **" سال دَمْعُهُ."** Context 2 of sense1= **"بَكَى صاحِبي لَمَّا رَأى الدَّرْبَ دونَهُ"** Source of context2= **" امرؤ القيس"**

Table 5: Results of the segmentation of the block 2.1

❖ The segmentation of the block 2.2 and block 2.3 is analogous to the segmentation of block 2.1.

Now, after having detected the basic information we have to put them in the target XML structure.

All information detected after different segmentations are affected to the specific tags in the XML structure as shown in the following Figure 7.

```
<?xml version="1.0" encoding="UTF-8"?>
<Lexicon>
  <feat att="language" val="arab" />
  <LexicalEntry id="1بكي">
    <feat att="partOfSpeech" val="verb" />
    <Lemma>
      <feat att="writtenForm" val="بَكَى" />
      <feat att="Scheme" val="فَعَلَ" />
    </Lemma>
<WordForm>
    <feat att="WittenForm" val="بَكَيْتُ" />
    <feat att="GramaticalNumber" val="singular"/>
    <feat att="GramaticalGender" val="masculine"/>
    <feat att="Person" val="1" />
    <feat att="GrammaticalAspect" val=
"accomplished" />
    <feat att="GrammaticalVoice" val="active" />
</WordForm>
<WordForm>
    <feat att="WittenForm" val="أبْكِي"/>...
</WordForm>
    <WordForm>
    <feat att="WittenForm" val="اِبْكِ"/>...
</WordForm>
    <RelatedForm targets="بكي">
    <feat att="type" val="root" />
    </RelatedForm>
    <RelatedForm targets="2بكي ">
    <feat att="type" val="derivedForm" />
    </RelatedForm>
    <Sens id="1بَكَىPD1">
    <Context>
    <feat att="text" val="بَكَى الوَلَدُ" />
    </Context>
    <Definition>
   <feat att="text" val=" سالَ دَمْعُهُ  " />
    </Definition>
    <Context>
    <feat att="text" val="بَكَى
صاحِبي لَمَّا رَأى الدَّرْبَ دونَهُ " />
    <feat att="source" val="امرؤ القيس" />
    </Context>
    </Sens>
    <Sens id="1بَكَىPD2">… </Sens>
    <Sens id="1بَكَىPD3">… </Sens>
    <SyntacticBehaviour subcategorizationFrames=
"متعدبواسطة لازم" />
  </LexicalEntry>
  <LexicalEntry id="بكي">
    <feat att="type" val="root" />
    <Lemma>
      <feat att="writtenForm" val="بكي" />
    </Lemma>
    <RelatedForm targets="1بكي">
```

```
    <feat att="type" val="derivedForm" />
   </RelatedForm>
  </LexicalEntry>
  <LexicalEntry id="2بكي">
<feat att="partOfSpeech" val="masdar" />
   <Lemma>
    <feat att="writtenForm" val="بُكاءٌ " />
    <feat att="Scheme" val="فُعَالٌ" />
   </Lemma>
   <RelatedForm targets="1بكي">
    <feat att="type" val="tronc" />
   </RelatedForm>
  </LexicalEntry>
</Lexicon>
```

Figure 7: Extract of the XML result of the mapping step

V-2 Experimentation results

As a result of the experiment carried out on Al-Ghany dictionary, we have been able to recognize 25.000 lexical entries from the 30.000 ones. The recognized entries are classified as follow:

- 10.400 verbs
- 14.400 nouns
  - 2210 active participles
  - 1100 passive participles
  - 6300 masdars
  - 200 relation nouns
  - 400 once nouns
  - 4490 other type of noun
- 3850 roots (root is not a lexical entry in the original dictionary but in the final result normalized dictionary is a lexical entry).

All information that we have been able to recognize related to these entries are mapped into our normalized model aside some ones that necessitate a special treatment.

The obtained version of the dictionary Al-Ghany (in XML format) has been successfully queried using the ADIQTO system.

VI- Conclusion

In this paper, we presented an automatic conversion approach of Arabic electronic dictionaries of human use into a normalized model. This approach considers the main difference between the Arabic dictionaries such as their content, their macro structure and their micro structure. In order to experiment this approach, we tested it on the dictionary Al-Ghany.

We believe that the obtained results are very advantageous namely that we have been able to use the ADIQTO system to query the content of the obtained standardized version of the dictionary Al-Ghany. However, we plan to integrate a morphological parser to improve the efficiency of our system. In addition, we intend to carry out the proposed approach on others Arabic dictionaries and to attend a generic system that deals with different dictionaries.

VII- References

Ait Taleb S. (2005). Dictionnaires électroniques arabes : le modèle des dictionnaires de Sakhr. Revue de l'Association Marocaine des Etudes Lexicographiques. Numéro 3-4. 15-31.

Arregi X., & al. (2002). Semiautomatic of the Euskal Hiztegia Basque Dictionary to a queryable electronic form. L'objet 8/2002. LMO' 2002. 45-57.

Baccar F., Khemakhem A., Gargouri B., Haddar K., Ben Hamadou A.. (2008). Modélisation normalisée LMF des dictionnaires électroniques éditoriaux de l'arabe. TALN'08.Avignon, 9-13 juin.

Dendien J., Pascal M. & Pierrel, J.-M. (2003). Le Trésor de la Langue Française informatisé : Un exemple d'informatisation d'un dictionnaire de langue de référence. TAL 44. Numéro2. 11-39.

Francopoulo, G. (2004). Proposition de norme des lexiques pour le traitement automatique du langage. NRIA/LORIA-ACTION SYNTAXE, Version-1.10, 13 mai 2004.

Francopoulo G. & George M. (2008). ISO/TC 37/SC 4 N453 (N330 Rev.16). Language resource management — Lexical markup framework (LMF).

Khemakhem A., Gargouri B. & Abdelwahed A. (2006). LMF est-il convenable pour la langue arabe ? Journées sur le Traitement Automatique de la Langue Arabe JTALA'06. 05-06 Juin. Rabat. Maroc.