An approach for generating personalized views from normalized electronic dictionaries : A practical experiment on Arabic language

Aida Khemakhem
MIRACL Laboratory
FSEGS, B.P. 1088,
3018 Sfax, Tunisia
khemakhem.aida@gmail.com

Bilel Gargouri
MIRACL Laboratory
FSEGS, B.P. 1088,
3018 Sfax, Tunisia
bilel.gargouri@fsegs.rnu.tn

Abdelmajid Ben Hamadou
MIRACL Laboratory
ISIMS, Cité Ons,
3021 Sfax, Tunisia
abdelmajid.benhamadou@isimsf.rnu.tn

Abstract

Electronic dictionaries covering all natural language levels are very relevant for the human use as well as for the automatic processing use, namely those constructed with respect to international standards. Such dictionaries are characterized by a complex structure and an important access time when using a querying system. However, the need of a user is generally limited to a part of such a dictionary according to his domain and expertise level which corresponds to a specialized dictionary. Given the importance of managing a unified dictionary and considering the personalized needs of users, we propose an approach for generating personalized views starting from a normalized dictionary with respect to Lexical Markup Framework LMF-ISO 24613 norm. This approach provides the re-use of already defined views for a community of users by managing their profiles information and promoting the materialization of the generated views. It is composed of four main steps: (i) the projection of data categories controlled by a set of constraints (related to the user's profiles), (ii) the selection of values with consistency checking, (iii) the automatic generation of the query's model and finally, (iv) the refinement of the view. The proposed approach was consolidated by carrying out an experiment on an LMF normalized Arabic dictionary.

1 Introduction

Electronic dictionaries are very useful in nowadays society, with the globalization and the increase of world communication and exchanges. There are clearly identified needs of dictionaries for human use as well as for automatic processing use.

Given the importance of having recourse to standards when constructing such lexical resources in order to promote the reuse and the fusion, the standardization committee ISO TC37/SC4 has recently validated the Lexical Markup Framework norm (LMF) project under the standard ISO 24 613 (Francopoulo and George 2008). LMF provides a common and shared representation of lexical objects that allows for the encoding of rich linguistic information, including among others morphological, syntactic, and semantic aspects. The LMF proposal is distinguished by the separate management of the hierarchical data structure (meta-model) and elementary linguistic descriptors (data categories) which promotes to cover several languages.

A normalized dictionary covers wide areas that include all lexical information of a given language and which are useful both for human use and for Natural Language Processing use in accordance with the kind of the user (linguist, lexicographer, developer, etc.), the level of the user (learner, expert, etc.) and the domain of the use (linguistic, medicine, biology, etc.). These dictionaries are characterized by a complex structure that supports the richness of natural languages. Therefore, dealing with a unique and complete dictionary is well for the manage task. However, such dictionaries are large and can be time consuming when querying their contents especially on the web. Moreover, displaying all details when some of which are not useful for the field of the query research is a

Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 953–960,
MIT, Massachusetts, USA, 9-11 October 2010. ©2010 Association for Computational Linguistics

nuisance for the user. So, it will be interesting to reduce the displayed details according to the domain or to the expertise level of the user by generating personalized views (virtual or materialized) in order to appropriate the use of such dictionaries to the user needs.

The idea of creating document views is not a new concept but applying it on LMF normalized dictionaries is a new one. Indeed, it has been some attempts for dictionary creating in accordance to the TEI consortium (Véronis and Ide 1996) but the problem was the fact that created textual views (corresponding to the surface structure) or data base views (corresponding to the deep structure) were not customized. Others propositions are very interesting but they concern the ontology domain when dealing with the concept of point of view (Corby and al 2005).

In this paper, we propose an approach that favors the use of normalized dictionaries by generating virtual/materialized personalized views. This approach is based on the profiles information associated to user's community which helps to retrieve already defined views stored in a library of views. For the illustration, we use an Arabic LMF normalized dictionary (Baccar and al. 2008, Khemakhem and al., 2009) developed in the framework of an Arabic project supervised by the ALECSO (The Arab League Educational, Cultural and Scientific Organization) and founded by the University of King Abdul-Aziz in the Kingdom of Saudi Arabia[1]. An environment supporting the proposed approach was implemented. At present, it concerns the Arabic language.

The present paper is outlined as follows. We will start by giving an overview of projects that use LMF notably for the construction and the exploitation of electronic dictionaries. Then, we will present the foundation of the proposed approach related to the profile and the view concepts. After that, we will explain the different steps of the view's generating approach. Finally, we will bring back the experimentation that we carried out on a normalized Arabic dictionary using the proposed approach.

2 State of the art of projects using LMF

After the emergence of LMF, some projects were launched in order to construct or exploit electronic dictionaries in accordance with this norm. Among others we note LEXUS (Kirsch 2005) (Kemps-Snijders and al 2006), LIRICS (LIRICS 2005) and LMF-QL (Ben Abderrahmen and al 2007). All these labors have been recourse to Web service technology that favors to invoke, locally or afar, appropriate services for the management or the exploitation of a normalized dictionary.

LEXUS offers an interface permitting to the user to define formats of lexical bases with a perspective to enable the construction of lexical bases according to the LMF model. However, it does not allow the verification of the compliance between the model and the norm.

LIRICS proposes some APIs that focus especially on the management of lexical data base. These APIs offer the possibility to work on the structure of the LMF base by adding, modifying or deleting components of LMF model. However, there is no interface which facilitates the use of these APIs.

LMF-QL provides Web services to be used while developing lingware systems. These services offer the exploitation of a normalized lexical base without having any piece of information about its content and its structure. The results of these services may be personalized dictionaries given in an XML format. However, it covers only the morphological level.

Concerning the construction of personalized dictionaries using the works mentioned above, we can notice that the user must have an idea about the content of the desired dictionary and its structure. He must also have acquaintances with queries generation to satisfy his requirements.

Finally, we note an absence of works dealing with the generation of views starting from LMF standardized dictionary.

3 Foundation of the approach

An electronic dictionary can be used by many users who have different requirements. Indeed, by being a language learner, a researcher in linguistics or a teacher's, needs and uses are not the same. Therefore, it will be better to have a tool (editor) allowing generating a suitable view. The making of a view of the dictionary might be difficult for some kinds of users, so the recourse to user profiles may facilitate this task. One can note that the user profile is very important to guide the user through the retrieval and the reuse of existent views corresponding to his profile.

[1] www.almuajam.org

3.1 A user profile definition

Generally, all features characterizing a user or a group of users can be grouped under the term of a user profile. For electronic dictionaries, a user profile is a dataset that concerns a community of users of the dictionary.

Every profile is characterized by a category, a level of expertise and a field. Indeed, we classify the views of the dictionary according to a profile that is based on a selection of these three criteria. The formal representation of a profile is the following:

P : < K, L, F>

K: the kind of user: lexicographer, linguist, lingware system developer, etc.

L: the level of the expertise: beginner, student, expert, etc.

F: the field of user: medicine, sport, biology, general, etc.

3.2 A View definition

A view of a dictionary is a synthesis of a dictionary interrogation query. We can consider it as a specialized or lexical dictionary, supported by a query.

A dictionary view allows to filter some lexicographic information and to hide others that are useless for some users.

The formal representation of a view:

View : < D, P, C >

D: dictionary: each view is specific to a normalized dictionary.

P: profile of the view (see previous section).

C: it is a set of properties which characterizes the model of the view. Each property has the following representation:

C : <A, V, W>

A: attribute is a simple representation of a characteristic model. This characteristic may be a class (Lemma, Sense…), a feature (definition, pos, genre,…) or a relationship (RelatedForm, SenseRelation…) as indicated in Figure 3.

V: value: each attribute can have a set of values. For example, the values verb and noun for the attribute POS (Part Of Speech).

W: weight of a property. It may take the values 0 or 1.

- If the weight equals to 0, then this property is mandatory only for part of the lexical entry.
- If the weight equals to 1, then this property is mandatory for each lexical entry of this view.

3.3 Different types of views

There are two types of views:

- Virtual view: the results of this view are calculated upon request. In this case, interrogating queries of this view might generate a composed query that interrogates directly the principal dictionary (underlying).
- Materialized view (physical): a physical copy faithful to the view definition which is stored and maintained.

4 Proposed approach

In this section we describe the proposed approach through the detail that we will give for each one of its four steps. These steps are illustrated in Figure 1.

- Projection of the view model: includes the specification of data categories (linguistic information), controlled by constraints in order to build a suitable normalized and valid model. It can be started by already existing profiles.
- Selection and checkout of the coherence: concerns the specification of some values for data categories (DC) already specified. We use coherence rules to check the constancy knowing that there are strong dependencies between some DCs and values of other DCs (see section 4.2).
- Automatic generation of the model and the query: includes the model refinement and the checking of constraints by priority.
- Refinement of the view: involves the validation of a new view of dictionary, by adding the elements that are related to the lexical entries of this view.

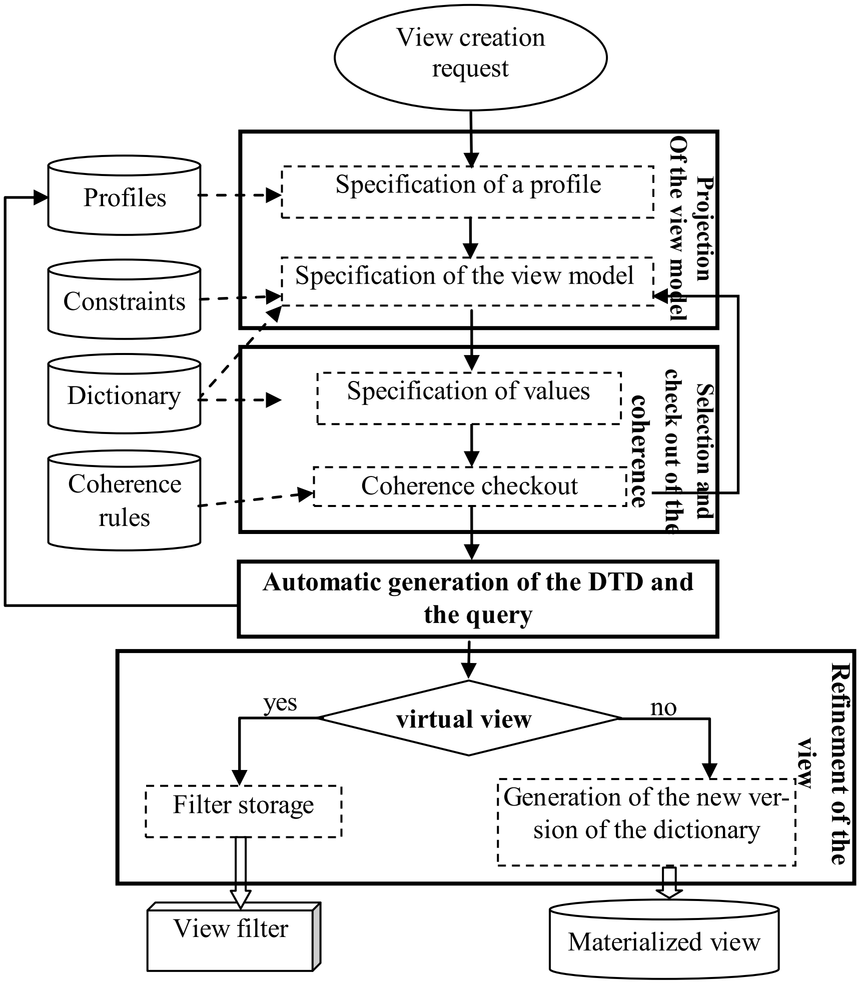

View creation request
Profiles
Constraints
Dictionary
Coherence rules
Specification of a profile
Specification of the view model
Projection Of the view model
Specification of values
Coherence checkout
Selection and check out of the coherence
Automatic generation of the DTD and the query
yes
virtual view
no
Filter storage
Generation of the new version of the dictionary
Refinement of the view
View filter
Materialized view

Figure 1.The approach for generating personalized dictionary views.

4.1 Projection of the view model

The UML model of the dictionary is difficult to understand. So, we suggest a simpler representation which is more abstract. The choice can be started by already existing profiles in order to avoid views redundancy by the reuse of the previous ones and help users.

a. The specification of a profile

A user profile is a description that corresponds to a user community. We use the features of profile to filter the DCs. Indeed, we offer to the user only DCs that correspond to its field and its level based on the weight assigned to each DC. We assigned these weights according to a study on the needs of each user level. This study is based on the specific documents and dictionaries for each user level. By example for beginner level, we studied the school books to extract the information (root, schema,…) needed at this level.

It also facilitates and accelerates the task of needs specification and permits to avoid views redundancy. The projection phase is started by the specification of the user profile that requires the choice of its category, its level and its field. Then, if the user wants to consult the previous views of his profile, we display all the views associated to this profile. Otherwise, we offer the DCs specific to its field and its level.

b. Constraints

The abstraction of the model can hides relations between DCs. Indeed, during their specification there is a risk of having views with a non valid model. So, the DC specification must be controlled by constraints rules such as:

- If we select the field, the semantic class or the nature then the definition must be selected.
- If the relation between syntactic behaviors and senses is selected then the definition must be selected.

4.2 Selection and checkout of the coherence

Most of data categories use a list of values such as part-of-speech (pos), scheme, field, etc. Values specification of some DCs might influence the presence or the absence of the other DCs. For example, if the user has chosen DCs: pos, root, scheme, gender and number; then, he has fixed the value du pos =”particle” (it means that he needs only particles) and the value of the number =”singular”. In this case, we note an incoherence problem since the DC “number” is among particles characteristics but it concerns nouns. In this case, we must request the user to rectify the specifications. The selection must contain a checkout phase of coherence of DCs specified with the already existing data in the dictionary. This phase is based on coherence rules which ensure consistency between DCs and the specified values.

4.3 Automatic generation of the DTD and the query

We use the Document Type Definition (DTD) of the LMF norm and DC specifications of the model to automatically generate the DTD of the view. We use algorithms to generate DTD elements, respecting the order of the participating classes and ensuring classes relations.

The automatic generation of the query (i.e. using XQuery) involves two steps: the first one concern the specification of conditions for the selection of a lexical entry and its related information (i.e., semantic, syntactic). The second step permits the definition of an XML representation of a lexical entry. This step is based on the projection specified by the user and the priority order of DCs. There are DCs that influence the presence of the lexical entry and others that only influence the presence of the sense.

4.4 Refinement of the view

The steps of this phase depend on the type of the view. If it’s a virtual one, then it’s necessary to save the query already generated. This query will be used during the operation of this view. If it’s a materialized one, then query results are a part of the dictionary and must be saved for the operation of the view. The second case may give us a non valid XML base, especially when there are two lexical entries in relation and our query will select one of them that have an identifier of the other entry that is not selected. We recall that the lexical entries may have morphological links with other lexical entries and semantic links with other senses.

Indeed, after recording query‘s results, we move to the step of refinement. This step consists to valid the new personalized dictionary, adding lexical entries, senses and syntactic frame in relation with these results.

5 Experimentation

5.1 The normalized Arabic dictionary

In order to experiment our approach, we are going to use the normalized Arabic interactive dictionary containing more than 38000 lexical entries and developed in the framework of an Arabic project[2] supervised by the ALECSO. This dictionary is modeled according to the meta-model proposed by LMF[3] (ISO 24613) and uses data categories generated by the DCR[4] norm (ISO 12620). The dictionary pattern is composed of classes selected from the kernel or from one of its extensions (morphological, semantic, syntactic, MRD) in order to see a dictionary covering most of new dictionary’s needs (Baccar and al 2008). Since there are many information that can be classified in multi extensions in the same time, the norm’s editor have chosen to put them in one of these extensions. For this reason, we did not use only Machine Readable Dictionary (MRD) extension. This pattern valorizes derivation phenomenon in Arabic language and neutralizes the differences between lexicographical schools, ensuring language evolution. In fact, we have considered

[2] www.almuajam.org/

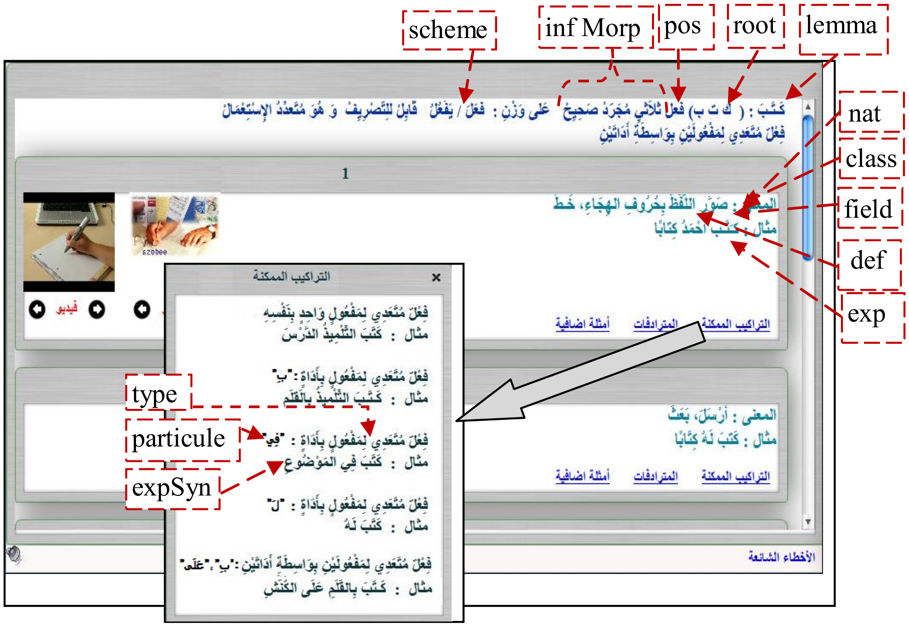
[3] www.lexicalmarkupframework.org/

[4] www.isocat.org/

roots (ك ت ب “k t b”), derived forms (كَتَبَ “kataba” (write), كَاتِب ”kâtib” (writer)), invariable words (إنَّ ”inna” (indeed), حَتَّى “hattâ” (in order)) and non-Arab origin words (كَمْبْيُوتَر “computer”, أَنْتَرْنَات “internet”) as lexical entries that can have morphological relations (i.e., RelatedForm relation).

In addition, this dictionary is rich with semantic information (i.e., definitions, examples, subject field, semantic class) and syntactic information (i.e., subcategorisation frame). It ensures the link between senses and their possible syntactic behaviors.

The Figure 2, given below, shows a part of the lexical entry “كَتَبَ” ”kataba” (write) which gives an idea about the structure of this dictionary.



كَتَبَ : (كَ ت ب) فعل ثلاثي مُجَرَّدٌ صَحِيحٌ عَلى وَزْنِ : فَعَلَ / يَفْعُلُ قَابِلٌ لِلتَّصْرِيفْ وَ هُوَ مُتَعَدِّدُ الإِسْتِعْمَالْ
فِعْلٌ مُتَعَدِي لِمَفْعُولَيْنِ بِوَاسِطَةِ أَدَاتَيْنِ

المعنى : صَوَّرَ اللَّفْظَ بِحُرُوفِ الهِجَاءِ، خَطَّ
مثال : كَتَبَ أحْمَدُ كِتَابًا

المعنى : أَرْسَلَ، بَعَثَ
مثال : كَتَبَ لَهُ كِتَابًا

التراكيب الممكنة

فِعْلٌ مُتَعَدِي لِمَفْعُولٍ وَاحِدٍ بِنَفْسِهِ
مثال : كَتَبَ التِّلْمِيذُ الدَّرْسَ

فِعْلٌ مُتَعَدِي لِمَفْعُولٍ بِأَدَاةٍ :"بِ"
مثال : كَتَبَ التِّلْمِيذُ بِالْقَلَمِ

فِعْلٌ مُتَعَدِي لِمَفْعُولٍ بِأَدَاةٍ : "فِي"
مثال : كَتَبَ فِي المَوْضُوعِ

فِعْلٌ مُتَعَدِي لِمَفْعُولٍ بِأَدَاةٍ : "لَ"
مثال : كَتَبَ لَهُ

فِعْلٌ مُتَعَدِي لِمَفْعُولَيْنِ بِوَاسِطَةِ أَدَاتَيْنِ :"بِ"،"عَلَى"
مثال : كَتَبَ بِالقَلَمِ عَلَى الكُنَّشِ

pos : part of speech
nat : nature
class : class
field : field
exp : example for a sense
inf Morp : Morphlogical piece of information
def : definition
expSyn : example of using a type
type : type of a syntactic frame

Figure 2. Example of a lexical entry of the normalized Arabic dictionary.

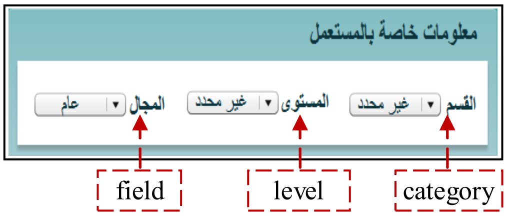
In this Figure, we highlight some properties of the used dictionary such as:

- The diversity of information: morphology, semantics, syntax, image, video, etc.
- The sense relations (i.e., synonym) link two senses and not two lexical entries
- The precision of the syntactic behavior. Indeed, each syntactic behavior has a type; the particles needed an example and its definition.
- The structure of a lexical entry varies according to its part of speech

5.2 Experiment of the approach

We illustrate in this section the generation process of a personalized dictionary view. We have setup a computing system online, that allows the user to make his view of the dictionary in the format of an interactive Web page (so independent of all material or software owner) in which he will be able to define, create and enhance his personal view.

Before starting the generation of the view, the user must specify his profile. If the views are associated with this profile, we will display their description to reuse the existing views and to avoid redundancy.



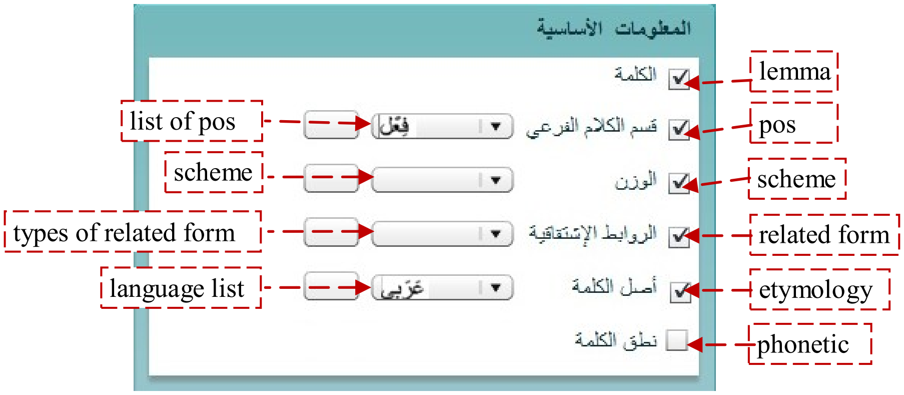
معلومات خاصة بالمستعمل

القسم [غير محدد ▾] المستوى [غير محدد ▾] المجال [عام ▾]

Figure 3. Specification of the user profile.

The user must set its category, its level and its field.

He can select an existing view corresponding to its needs or he can specify its purpose without using the existing. Then, he chooses the categories of data needed. Next, he can set their values.



المعلومات الأساسية

☑ الكلمة

☑ قسم الكلام الفرعي [فِعْل ▾]

☑ الوزن [▾]

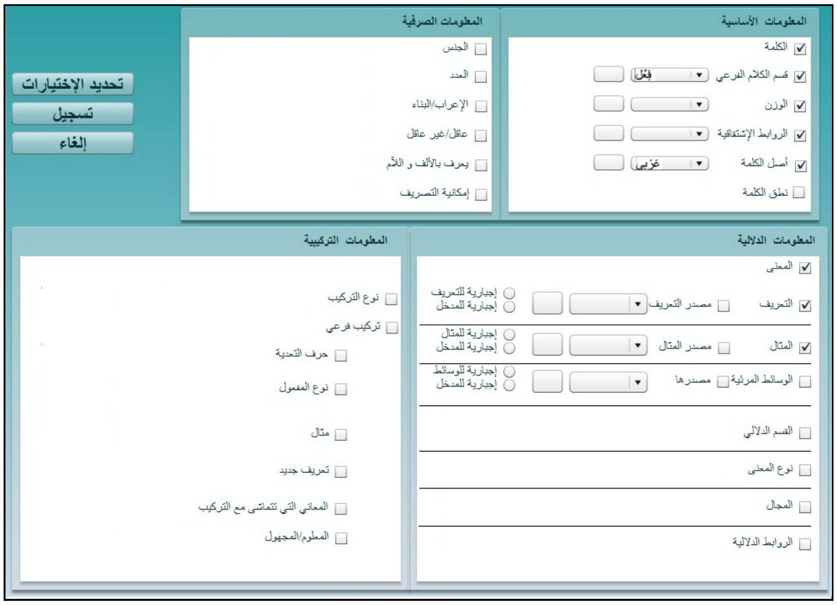
☑ الروابط الإشتقاقية [▾]

☑ أصل الكلمة [عَرَبِي ▾]

☐ نطق الكلمة

Figure 4. Representation of some categories of data and the list of values for selected DCs

In the following Figure 5, we present the key information in the dictionary. We give an example of view that includes only the schema, the derivational relations, sense, examples of all the Arabic verbs (pos = verb). Then from 38000 lexical entries, our view has 7000 verbs and 3000 roots i.e. only 10,000 entries.



المعلومات الأساسية

☑ الكلمة

☑ قسم الكلام الفرعي [فِعْل ▾]

☑ الوزن [▾]

☑ الروابط الإشتقاقية [▾]

☑ أصل الكلمة [عَرَبِي ▾]

☐ نطق الكلمة

المعلومات الصرفية

☐ الجنس

☐ العدد

☐ الإعراب/البناء

☐ عاقل/غير عاقل

☐ يعرف بالألف و اللآم

☐ إمكانية التصريف

المعلومات الدلالية

☑ المعنى

☑ التعريف ☐ مصدر التعريف [▾] ○ إجبارية للتعريف ○ إجبارية للمدخل

☑ المثال ☐ مصدر المثال [▾] ○ إجبارية للمثال ○ إجبارية للمدخل

☐ الوسائط المرئية ☐ مصدرها [▾] ○ إجبارية للوسائط ○ إجبارية للمدخل

☐ القسم الدلالي

☐ نوع المعنى

☐ المجال

☐ الروابط الدلالية

المعلومات التركيبية

☐ نوع التركيب

☐ تركيب فرعي

☐ حرف التعدية

☐ نوع المفعول

☐ مثال

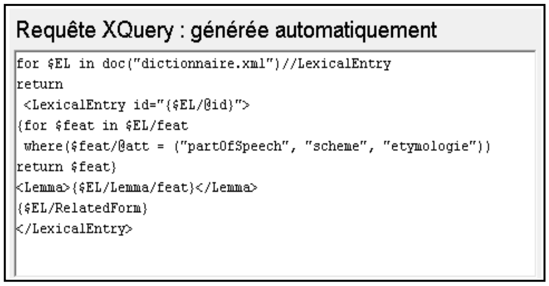
☐ تعريف جديد

☐ المعاني التي تتماشى مع التركيب

☐ المعلوم/المجهول

Figure 5. Interface for creating a view.

In the Figure 5, the user has selected the lemma, pos, schema, the derivational relations, etymology, sense, definition and example. For pos, he fixed the value of "فِعْل" (verb). According to the specification of requirements, the user clicks the save button. The system checks the consistency of the chosen data categories and values, then, it generates a query in the XQuery language (see Figure 6).



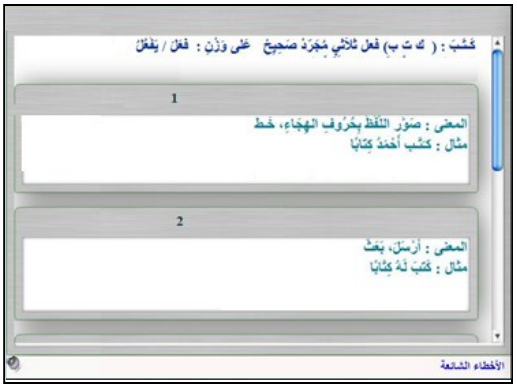
Requête XQuery : générée automatiquement

```
for $EL in doc("dictionnaire.xml")//LexicalEntry
return
 <LexicalEntry id="{$EL/@id}">
{for $feat in $EL/feat
 where($feat/@att = ("partOfSpeech", "scheme", "etymologie"))
return $feat}
<Lemma>{$EL/Lemma/feat}</Lemma>
{$EL/RelatedForm}
</LexicalEntry>
```

Figure 6. Example of a generated query

If the user chooses a materialized view, the system must save the query result in the user's computer after refinement (add missing information to validate the XML document).

For the verification of results, the user must choose the view before starting the search in the dictionary. In the following Figure 7, we present the results of research in the view already specified in Figure 5.



كَتَبَ : (ك ت ب) فعل ثلاثي مجرّد صحيح   على وزن : فَعَلَ / يَفْعُلُ

1

المعنى : صَوَّرَ اللفظ بحروف الهجاء، خطّ
مثال : كَتَبَ أحمدُ كِتابًا

2

المعنى : أرسلَ، بَعَثَ
مثال : كَتَبَ لَهُ كِتابًا

Figure 7. Displayed results of query applied on a generated view.

6 Conclusion

The construction of specialized dictionaries is an old concept. However, it has not been used after the publication of LMF standard in spite of the complexity and the richness of normalized dictionaries. In this paper, we proposed an approach allowing the generation of specialized and personalized views of dictionaries according to users' profiles in order to benefit from the management of a unique dictionary and give appropriate services.

A Practical experiment was carried out on a normalized Arabic dictionary using an appropriate tool that permits to manage users' profiles and views' generation. We successfully performed some empirical illustrations starting from the normalized dictionary.

In the future, we will consider the experimentation of the developed tool in the generation of various personalized views both in virtual and materialized versions. Also, we plan to put up our system on the Web. Also, we plan to experiment our approach on others languages.

References

Baccar, F., Khemakhem, A., Gargouri, B., Haddar, K. and Hamadou, A.B.. 2008. *LMF Standardized Model for the Editorial Electronic Dictionaries of Arabic*. In Proceedings of NLPCS'08, pp. 64–73. Barcelona, Spain.

Ben Abderrahmen M., Gargouri B. and Jmaiel M. 2007. *LMF-QL: A Graphical Tool to Query LMF Databases for NLP and Editorial Use*. Book Chapter In: Human Language Technology. Challenges of the Information Society, Third Language and Technology Conference, LTC 2007, Poznan, Poland.

Corby O., Dieng-Kuntz R., Faron-Zucker C., Gandon F., Giboin A. 2005. *Le moteur de recherche sémantique Corese*. In Proc. of the Workshop Raisonner le web sémantique avec des graphes, AFIA platform. Nice.

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

Kemps-Snijders M., Nederhof M.-J. and Wittenburg P. 2006. *LEXUS, "A web-based tool for manipulating lexical resources"*. In LREC 2006.

Kirsch K. 2005. *LEXUS Manual*. LEXUS version 1.0.

LIRICS (Linguistic Infrastructure for Interoperable Resource and Systems). 2005. "*Guidelines and tools for producing standards, test-suites and API(s)*".

Véronis, 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 (1996) (pp. 239-261). Paris: Didier Erudition.