LMF Standardized Model for the Editorial Electronic Dictionaries of Arabic

Feten Baccar Ben Amar[1], Aïda Khemakhem[1], Bilel Gargouri[1]
Kais Haddar[2] and Abdelmajid Ben Hamadou[3]

MIRACL Laboratory
[1] FSEGS, B.P. 1088, 3018 Sfax, Tunisia

[2] FSS, B.P. 802, 3018 Sfax, Tunisia

[3] ISIMS, B.P. 242, 3021 Sakiet-Ezzit Sfax, Tunisia

Abstract. This paper is interested in the development of the Arabic electronic dictionaries of human use (editorial use). It proposes a unified and standardized model for these dictionaries according to the future standard LMF (Lexical Markup Framework) ISO 24613. Thanks to its subtle and standardized structure, this model allows the development of extendable dictionaries on which generic interrogation functions adapted to the user's needs can be implemented. This model has already been carried out on some existing Arabic dictionaries using the ADIQTQ (Arabic DIctionary Query Tool) system, which we developed for the generic interrogation of standardized dictionaries of Arabic.

1 Introduction

With the advent of the data-processing tools and the proliferation of means of communication, the storage methods and information access have changed. In fact, the editorial electronic dictionaries are released of the constraint of their paper versions. Besides, they act in response to a certain number of practical needs of update and interrogation for heterogeneous users (e.g. the learners, the professional journalists or writers and researchers in linguistics) [1], [2] and [3]. That is why, several electronic dictionaries are today available on the Web or commercialized on CD-ROM, especially, for the Western languages.
The computerization of Western dictionaries has begun since the last half century. The objective was to make the users benefit from an extremely invaluable aid, provided that these electronic dictionaries allow a selective and fast access to information. These services consist, particularly, in finding a word by its approximate orthography, extracting a list of specialized terms referring to a field, seeking a list of words on criteria other than alphabetical, etc. Among the Western electronic dictionaries, some examples of the French and English languages can be mentioned: the

Baccar Ben Amar F., Khemakhem A., Gargouri B., Haddar K. and Ben Hamadou A. (2008).
LMF Standardized Model for the Editorial Electronic Dictionaries of Arabic.
In *Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science*, pages 64-73
DOI: 10.5220/0001737800640073

Trésor de la Langue Française informalisé[1] (TLFi), the Dictionnaire de l'Académie Française[2], Le Nouveau Petit Robert électronique, the Bibliorom Larousse, Oxford English Dictionay[3] (Second Edition) and Collins Electronic English Dictionary & Thesaurus.

In addition, the structural diversity of the dictionary resources available made the exchange and the fusion of their data obviously difficult and complex. Actually, a unified and standardized lexical representation is necessary and preliminary to any exploitation of these resources out of their own design context [4]. Thus, two major proposals for the standardization of electronic dictionaries have emerged, namely the TEI[4] (Text Encoding Initiative) [5] and LMF (Lexical Markup Framework) [6].

However, the Arabic language could not fully profit from these studies and realizations. The development of Arabic electronic dictionaries, in particular, those of human use, is in a phase that can be described as primal. Indeed, it seems that the quasi-total of the Arabic electronic dictionaries available are nothing but a digitalization of a non-structured and less beneficial dictionary content (under format: ".doc", ".pdf", or ".html") [7]. In addition, the associated interrogation tools of these dictionaries offer only one type of search, which can be qualified as primitive (search for the lexical entries in one or more dictionaries) as proved by the realizations of the well-known projects Ajeeb[5] (of Sakhr[6] System), Alburaq[7], Kalimet[8] and LisanElArab/Al-qamous-Al-Mouhit[9]. These interrogation limits are, mainly, due to a structuring weakness of the dictionary entries used.

Starting from these observations and in order to give to the Arabic language its appropriate position as a widespread language, we propose an LMF modeling for the development of extendable and incremental Arabic editorial dictionaries (possibility of adding entries and descriptions within the entries). This modeling will make it possible to benefit from the richness of the Arabic language unifying as subtle as possible what exists with an evolutionary structure, from which, it is possible to fulfill generic interrogation functions adapted to the various needs of the users.

This paper starts with a presentation of the main features of the Arabic dictionaries with a specific focus on their structural specificities. Then, evidence is provided for the choice of modelling LMF for the development of the Arabic electronic dictionaries of human use. Next, the standardized model proposed is detailed. Finally, a computerized implementation and an experimentation of the model are described, starting with the standardization of the editorial dictionary (El-Ghani) « الغنيّ » and of a set of extracts relative to other dictionaries. The interrogation is realized by the system ADIQTO (Arabic DIctionary Query TOols) which is developed in the present paper for the generic interrogation of the standardized Arabic dictionaries.

[1] www.atilf.fr/tlfi

[2] www.atilf.fr/academie8 and www.atilf.fr/academie9

[3] http://www.oed.com/

[4] http://www.tei-c.org

[5] http://lexicons.ajeeb.com/Results.asp

[6] http://sail-technology.com/press/pdf/sakhr2003.pdf

[7] http://www.alburaq.net/mukhtar/root.cfm

[8] http://www.kl28.com/lesanalarab.php

[9] http://www.content.com.sa/Languages/LisanElArab/default.aspx

2 Main Characteristics of the Arabic Dictionaries

The Arabic lexicography is a very ancient discipline. All along its history, it has known different schools each of which having its own specificities. Some of these schools propose a classification of the lexical entries according to the phonetic order (e.g. العين « Al-Ain », تهذيب اللغة « Tahdhib Al-Loghah »). Others use the alphabetical order either direct (e.g. لسان العرب «LisanElArab», القاموس المحيط « Al-qamous-Al-Mouhit », الصحاح « Assihah ») or reversed (e.g. الغنيّ « El-Ghani », لاروس: المعجم العربي الحديث « Larousse : Al-moojam Alarabi Al-Hadith »). What is also noticed is the classification according to the alphabetical increasing order of the root of the word which represents one of the specificities of the Arabic dictionaries (e.g. الوسيط «Al-Wassit», المصباح المنير « Al-Mosbah Almounir »). This organization makes it possible to include all the verbal and nominal forms derived from a root under the same entry.

This diversity can be summarized in the following points:

- Content: the description of the contents of the entries varies from a dictionary to another. Some are satisfied with the examples to define the word meaning, others give the definition along with examples. The information of a morphological and syntactic nature is more or less rich, etc.
- Macrostructure: the organization of the lexical entries differs from a dictionary to another (according to the lexicographical school). For example, the word *MaKTaBaTun* « مَكْتَبَةٌ » (library) represents a lexical entry by itself in the dictionary (El-Ghani) « الغنيّ » whereas it appears as a sub-entry of the lexical entry of *KaTaBa* « كَتَبَ » (to write) in the dictionary (Al-Wassit) « الوسيط » since it is a word derived from *KaTaBa.*
- Microstructure: 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. For example, the plural and/or the female of a lemma of the type noun «اسم» or adjective «صفة» are at the beginning of the article in the dictionary (El-Ghani) « الغنيّ » and at the end of the article, if they exist in (Al-wassit) « الوسيط ».
- Consultation: There is not a single method to consult these dictionaries. In most cases, the user is compelled to know the internal structure of a dictionary to be able to find the information s/he seeks.

The question raised is how to benefit from this structural and informational richness in the development of new Arabic electronic dictionaries?

3 Choice and General Presentation of LMF

3.1 Choice of LMF

In order to propose a unified model of the editorial dictionaries of the Arabic language, our choice was settled on LMF for the following reasons. First of all, this future standard, which will be published very soon by the ISO under number 24613, allows the specification of monolingual and multilingual lexicons intended at the same time for the editorial use and NLP one. Moreover, it ensures an extendable and modular modelling covering all the levels of linguistic description (e.g. morphology, syntax, semantics). Moreover, its modelling flexibility promotes the representation of the characteristics of the Arabic language, especially, its derivational and inflectional aspects (e.g. possibility of the representation of the roots and the designs) [8]. Eventually, modelling LMF lends itself to a use online with the Web services [9].

3.2 General Presentation of LMF

From a generic perspective, LMF proposes a meta-model made up of a mandatory core part and optional extension models covering, in particular, the morphological, syntactic, semantic aspects and MRD (Machine Readable Dictionary) [6].

The LMF modularity is ensured by the accumulation mechanisms of the data categories having meta-model components. In fact, elementary linguistic descriptors make it possible to decorate the meta-model classes with the specificities of the language and the lexicon (e.g. /Part of speech/, /Grammatical number/, etc.). They are organized in a register of data categories, which is consultable and editable online. Thus, their standardization follows the principles defined by the standard ISO 12620 [10].

4 Development of the Standardized Model of the Arabic

This model constitutes a unified representation of the existing Arabic dictionaries. This project adopted the latest revision of LMF (rev.15) [6].

4.1 Steps of the Creation of the Standard Model

This paragraph represents an outline of the normative and representative data model of the Arabic dictionaries. The steps followed for the development of this model respect the process recommended by LMF [6]. Therefore, we started with the selection of the classes necessary to represent dictionary information of Arabic. The classes, which we kept, constitute sub-models of the target one and belong to the morphological, MRD, syntactic and semantic extensions. These classes are embedded on the LMF core which essentially specifies the concepts of lexicon, word, form and sense. The presence of the core is obligatory according to LMF. After that, the data categories adequate to Arabic are selected from the standardized register. Finally, the meta-model part selected is decorated by the various data categories selected.

4.2 Morphological Extension Model

The model of this extension makes it possible to represent morphological information of the lexicons since the Arabic language has derivational and inflectional aspects. The classes *Lemma*, *RelatedForm* and *WordForm*, kept from the LMF morphological extension, are used to describe, respectively, the scheme and the lemma, the root, and the properties of the inflected terms of this lemma. Fig.1a describes the bonds of these morphological classes with those of the core. Moreover, it specifies the categories of data used for the classes of this extension.

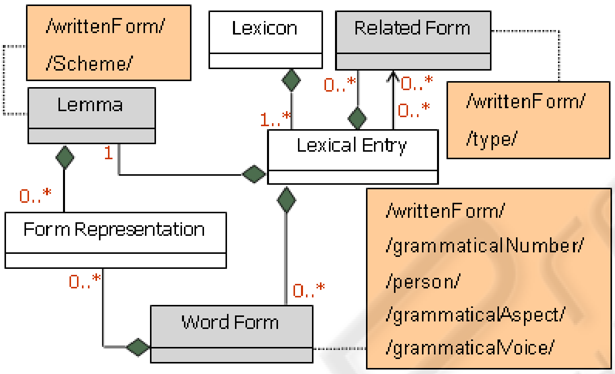
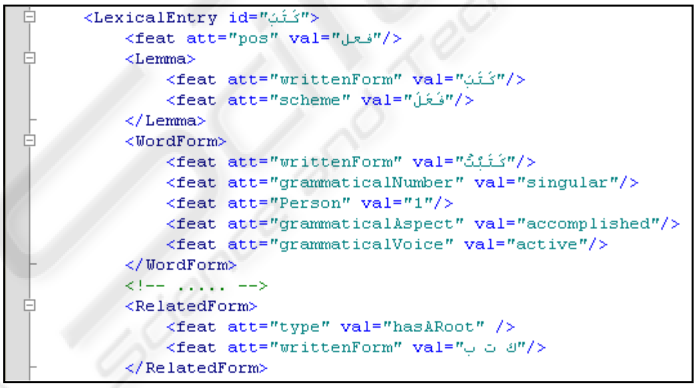


Fig. 1a. Morphology class model.

Fig. 1b presents an XML extract which illustrates the instantiation of this model for the instance of the lexical entry *kataba* « كَتَبَ » *(to write)*.



```
<LexicalEntry id="كَتَبَ">
    <feat att="pos" val="فعل"/>
    <Lemma>
        <feat att="writtenForm" val="كَتَبَ"/>
        <feat att="scheme" val="فَعَلَ"/>
    </Lemma>
    <WordForm>
        <feat att="writtenForm" val="كَتَبْتُ"/>
        <feat att="grammaticalNumber" val="singular"/>
        <feat att="Person" val="1"/>
        <feat att="grammaticalAspect" val="accomplished"/>
        <feat att="grammaticalVoice" val="active"/>
    </WordForm>
    <!-- ..... -->
    <RelatedForm>
        <feat att="type" val="hasARoot" />
        <feat att="writtenForm" val="ك ت ب"/>
    </RelatedForm>
```

Fig. 1b. XML extract of an instance of the morphological model.

4.3 MRD Extension Model

This extension is particularly used in the description of the electronic dictionaries of human use. Fig. 2 presents the association of the MRD classes which are kept; Context and SubjectField with the Sense class of the core. The first class represents an example of the lemma use in a phase. As far as the second class is concerned, it defines the field of use corresponding to a given sense. The third MRD class which is suggested by LMF, was excluded, namely Equivalent, since it relates to a multilingual context.

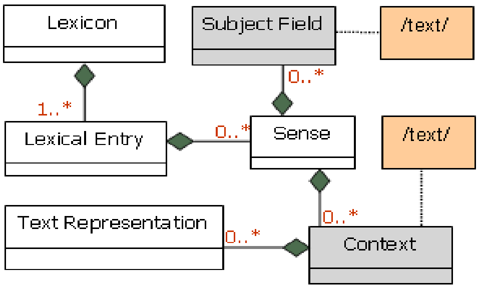


Fig. 2. MRD class model.

4.4 Syntactic Extension Model

The objective of this model is to describe the syntactic properties of the most important word in case of a several-word construction. The classes deemed essential to describe the Arabic syntactic information are *SyntacticBehaviour*, *Subcategorization-Frame* and *SubcategorizationFrameSet*, which are attached to the core as shown in Fig.3. The role of these classes, indicated in LMF, corresponds well to descriptions of a syntactic nature used in the majority of Arabic dictionaries where the focus on a verb suggests the description of its valence.

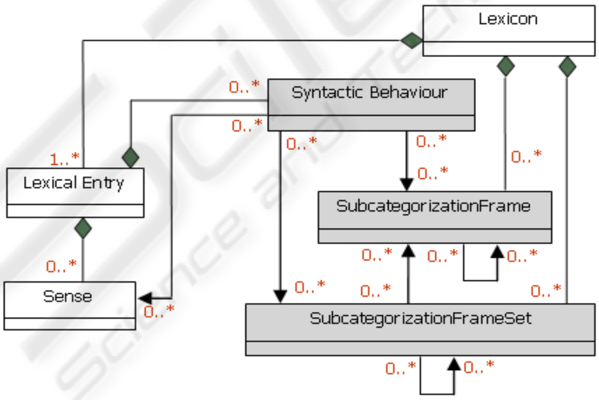


Fig. 3. Syntactic class model.

4.5 Semantic Extension Model

Concerning the semantic extension of the present model, the classes SenseRelation and SenseExample are kept. The choice of the SenseRelation class, instead of Synset, also proposed by LMF, is justified by the fact that we do not have Arabic dictionaries of WordNet type which is based on the concept of Synset (sets of synonyms). The SenseExample class proves very useful for the Arabic language. Indeed, all dictionaries of this language insist on the examples, mainly to define the chief words. The two classes kept are connected to the Sense class of the core as shown in Fig. 4.

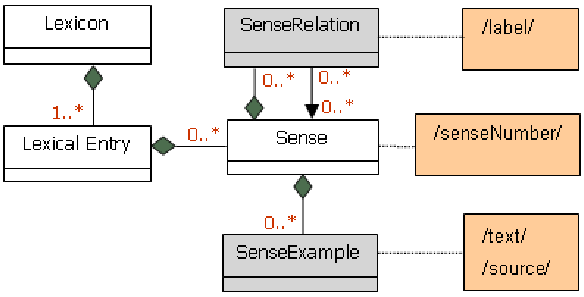


Fig. 4. Semantic class model.

5 Implementation and Experimentation of the Model

Aiming at the experimentation of the elaborated model, the standardization of certain dictionaries available, especially (*El-Ghani*) « الغنيّ » were first carried out. Then, a set of interrogation functions were defined according to the level of the users. Eventually, these functions were developed and tested on the standardized versions of dictionaries within the framework of the implementation of the system of interrogation ADIQTO.

5.1 Standardization of the Existing Dictionaries

The standardization started with the dictionary (El-Ghani) « الغنيّ » whose HTML version is available. Indeed, this dictionary is characterized by its reduced and invariable microstructure, as well as by a set of markers facilitating the transformation of its plain content into an XML structured version. Afterwards, this conversion was done in a quasi-automatic method using the NLP tools available in our laboratory MIRACL (e.g. segmentor, morphological analysor).

Furthermore, some extracts of other dictionaries such as (Al-Wassit) «الوسيط», (Lisan-El-Arab) « لسان العرب » and (Al-Mouhit) « المحيط», whose structure is more complex, were standardized. Their structure varies from one entry to another and it is characterized by a quasi-absence of markers. Besides, the standardization was carried out manually bearing in mind the difficulty of the installation of an automatic system of analysis and conversion.

5.2 Development of the Interrogation Functions

To define the interrogation functionalities of the dictionaries, two criteria were considered: the user intervention and the query complexity. Hence, the interrogation services can be classified in three types: (i) simple interrogation (ii) guided interrogation and (iii) advanced interrogation (complex). The research can be carried out in one or more dictionaries chosen by the user.

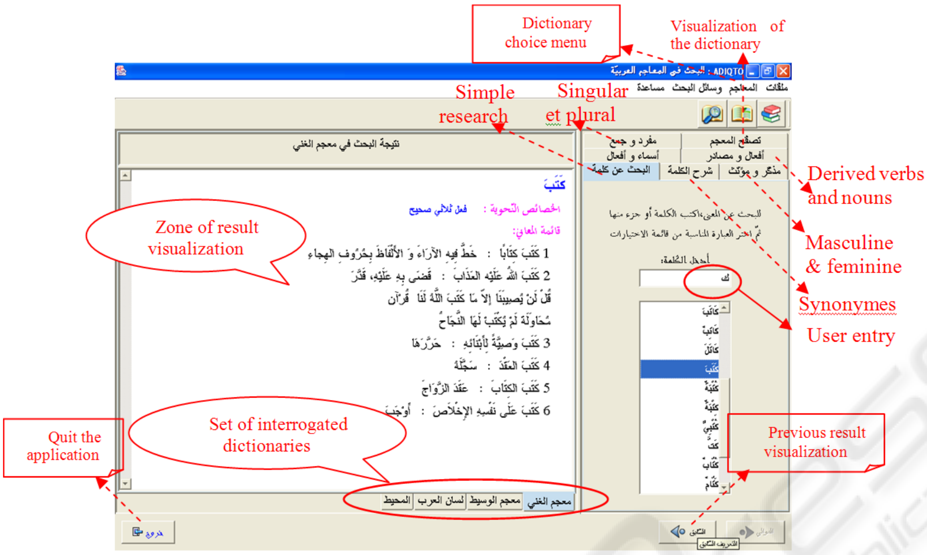
- Simple interrogation
  It allows simple access to word found as a lexical entry in the dictionary. All information relating to the required word will be displayed: morphological characteristics, definitions, examples, etc.

- Assisted interrogation
  The user of this type of research enjoys assistance through a correction of the introduced word. Moreover, in case of ignoring the lexical entry, the user can introduce some words used in its definition to attend the desired lexical entry. In addition, the preset queries will be within the users' reach, and consistent with the research of the synonyms, masculine or feminine, singular or plural and finally derived verbs and nouns.

- Advanced interrogation
  The third type relates to complex needs expressed by experts who seek to calculate the statistics on the morphology of the words, the link between words, etc or to calculate specific results not stored in the dictionary as inflection. Within this framework, the graphic language LMF-QL [11] will be re-used. LMF-QL makes it possible to generate an XQuery starting with a dynamic selection of the entries which will be combined with logical operators (*not*, *and* and *or*) and a dynamic selection of the entries. Thanks to this type of consultation, the user is able to reach any information and to make crossings according to the considered needs (e.g. to seek the words which have the same root).

5.3 Implementation of the ADIQTO System

The implementation of ADIQTO system is carried out in Java along with the use of API Saxon9 to ensure the execution of the XQuery queries. This implementation is based on predetermined queries ensuring, namely, simple and assisted researches (macros). As for the advanced research, its queries will be generated automatically thanks to the use of LMF-QL language. Fig.5 illustrates the simple research of the word *KaTaBa* « كَتَبَ » *(to write).*



نتيجة البحث في معجم الغني

كَتَبَ

الخصائص النّحوية : فعل ثلاثي صحيح

قائمة المعاني:

1 كَتَبَ كِتَابًا : خَطَّ فِيهِ الآرَاءَ وَ الأَلْفَاظَ بِحُرُوفِ الهِجاءِ

2 كَتَبَ اللهُ عَلَيْهِ العَذَابَ : قَضَى بِهِ عَلَيْهِ، قَدَّرَ

قُلْ لَنْ يُصِيبَنَا إلاّ مَا كَتَبَ اللّهُ لَنَا قُرآن

مُحَاوَلَة لَمْ يُكْتَبْ لَهَا النَّجَاحُ

3 كَتَبَ وَصِيَّةً لِأَبْنَائِهِ : حَرَّرَهَا

4 كَتَبَ العَقْدَ : سَجَّلَهُ

5 كَتَبَ الكِتَابَ : عَقَدَ الزَّوَاجَ

6 كَتَبَ عَلَى نَفْسِهِ الإِخْلاَصَ : أَوْجَبَ

للبحث عن المعنى،اكتب الكلمة أو جزء منها

ثمّ اختر العبارة المناسبة من قائمة الاختيارات

كَاتَبَ
كَاتِبٌ
كَاتَلَ
كَتَبَ
كُتْبَةٌ
كِتْبَةٌ

Fig. 5. Graphic interface of simple research in ADIQTO system.

6 Conclusions

The model proposed in this paper came in conformity with LMF-ISO 24613 standard for the representation of the editorial electronic dictionaries of Arabic language. This model has the merit to unify the various models suggested in the lexicography of Arabic and offering the possibilities of extension.

The implementation of ADIQTO system, enabled us to show the value of this model which was tested on the dictionary « الغنيّ » (*El-Ghani*) and on fragments of other dictionaries. The evolutionary and subtle structure of this model has allowed a selective access to the various fields related to a lexical entry by means of a set of research (simple, aided and advanced) satisfying the needs of a diversity of users (e.g., language learner, writer, journalist, linguist, etc.)

In future work, we intend to develop Web services allowing to put ADIQTO system online, provided with the standardized dictionaries available, so as to widen the spectrum of its use.

Finally, we plan to approach the standardization of the existing editorial dictionaries of Arabic, mainly, out of the numerical versions we dispose of, (we have ten of these dictionaries in: ".doc", ".html" and ".pdf" format).

References

1. Dendien, J. & Pierrel, J.-M.: Le Trésor de la Langue Française informatisé. Un exemple d'informatisation d'un dictionnaire de langue de référence. TAL 44, Numéro 2 (2003) 11-39

2. Wooldridge R. : L'informatisation du Dictionnaire de l'Académie française (DAF). DictA 1998, Table ronde sur l'informatisation des dictionnaires anciens. Limoges, 19-20 novembre (1998)
3. ARREGI X. & AL.: Semiautomatic of the Euskal Hiztegia Basque Dictionary to a queryabale electronic form. L'objet 8/2002, LMO' 2002, 45-57
4. Francopoulo G. : Proposition de norme des lexiques pour le traitement automatique du langage. INRIA/LORIA-ACTION SYNTAXE, Version-1.10, (13 mai 2004).
5. Sperberg-Mcqueen C.M., Burnard L.: TEI P5 − Guidelines for Electronic Text Encoding and Interchange, TEI Consortium. (January 2005)
6. Francopoulo G., George M. ISO/TC 37/SC 4 Rev.15. Language resource management − Lexical markup framework (LMF) (2008)
7. Ait Taleb S. : Dictionnaires électroniques arabes : le modèle des dictionnaires de Sakhr, *revue de l'Association Marocaine des Etudes Lexicographiques*, Numéro 3-4, 15-31 (2005)
8. Khemakhem A., Gargouri B., Abdelwahed A., Francopoulo G. : Modélisation des paradigmes de flexion des verbes arabes selon la norme LMF - ISO 24613. *Traitement Automatique des Langues Naturelles : du 5 au 8 juin 2007 à Toulouse* (2007)
9. Ben Abderrahmen M., Chaari. F, Gargouri B., Jmaiel M. : Des services orientés besoin pour l'exploitation des bases lexicales normalisées. *10th Maghrebian Conference on Software Engineering and Artificial Intelligence, 07-09 Décembre 2006, Agadir, Maroc* (2006)
10. Ide N., Romary L.: International standard for a linguistic annotation framework. *International Journal of Natural Language Engineering*, 10 Numéro 3-4 (2004) 211-225
11. Ben Abderrahmen M., Gargouri B., Jmaiel M.: LMF-QL: A graphical Tool to Query LMF databases. *Third Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics October 5-7, Poznań, Poland* (2007)