An XML model for an Arabic historical dictionary

Rim Laatar, Chafik Aloulou and Lamia Hadrich Belguith

University of Sfax, MIRACL Laboratory, Tunisia
rimlaatar@yahoo.fr, chafik.aloulou@fsegs.rnu.tn ,
l.belguith@fsegs.rnu.tn

Abstract. A historical dictionary is a dictionary which deals with the detailed history of words since their first appearance in language as well as the evolution of their meaning and use throughout history. The main objective of building a historical dictionary is to monitor the semantic evolution of each word of the language through its historical ages by determining the date of its first appearance and the date of its transformation. This work is a study which helps to trace the evolution of the meanings of a given word throughout time. We propose a structured Extensible Markup Language (XML) model that captures this transformation. This model will help linguists to create the Historical Dictionary for the Arabic Language.

Keywords: Historical dictionary, Arabic language, Extensible Markup Language (XML).

1 Introduction

Arabic is one of the oldest Semitic languages in the world. It is also one of the main languages spoken in the world by many people in various Arab countries using it as their mother tongue. However, despite its richness, the Arabic language does not have a historical dictionary yet.

In fact, the historical dictionary is a linguistic dictionary derived from the written human heritage, throughout history and in different places, in sciences, arts and literature. It includes the vocabulary of the language, the meanings and derivatives of words, their details and their history of use. It also studies the development of its structures and meanings throughout the successive stages of the language [1].

Therefore, building the historical dictionary for the Arabic language is very important, as it helps to understand the language, its richness, its origins and its evolution. It also gives an overview of the compilation of a historical lexicon of the Arabic language and its development over the last two millennia.

However, after several centuries, the Arab-Muslim culture does not yet have a historical dictionary which would reflect its speech, its thoughts and its rich heritage. It is so crucial to have such a dictionary for Arabic to historicize its past and get prepared for its future changes. Therefore, according to many unanimous researchers, creating a historical Arabic dictionary has become a necessity in the era of the computerization of knowledge of all kinds.

To begin with, dictionaries are known for their classical forms. In fact, the invention of the computer in the mid-1940s and the evolution of computing made it possible to digitize information and make it available in electronic forms. Soon after, the development of storage media facilitated the archiving of large amounts of information. As a result, the set of information stored in a paper dictionary gave way to an electronic version.

Therefore, the main objective of this work is to help the linguists to build a historical dictionary for Arabic by proposing a structured model describing the variation of the meaning of a word through time. Indeed, words in Arabic have undergone a historical process in which their growth is marked by significance and expectation. There are words that change their vocations over time and others have completely disappeared from literature.

In this paper, we present a model for capturing and storing the history of Arabic words in an XML format.

The rest of this paper is divided as follows. We will present in section two a state of the art reviewing some works on existing historical dictionaries in different languages, as well as the different attempts made to achieve these dictionaries in the Arabic language. In section three, we will present the main objectives of our work. In section four, we will talk about the proposed model, and finally, in section five, we will draw a conclusion and suggest some future work ideas.

2 Historical dictionaries: an overview

The idea of creating historical dictionaries appeared during the second half of the 19th century following the appearance of the method of historical analysis [2]. Several international projects have been launched in different countries whose purpose was to develop a historical dictionary.

In this Section, we are going to review some existing works related to the historical dictionary for other languages.

2.1 The German Historical Dictionary

The German historical dictionary "Deutsches WörterBuch" (DWB) is the most important German dictionary since the 16th century. It is also called the Grimm dictionary, referring to the names of its creators the Grimm brothers (Jacob and Wilhelm Grimm), who began working on it in 1838 with more than 80 collaborators who delt with more than 600,000 entries. It is a historical dictionary that traces the history of each word using many quotes. Indeed, the purpose of this dictionary is to analyze and explain exhaustively the origin and use of each German word. Each user of the dictionary could thus know the old and the modern use of each word.

The original work was to have six to seven volumes and was completed after ten years of work. Wilhelm Grimm who wrote the entries for the letter D died in 1859 and Jacob who managed to finish the letters A, B, C and E died in 1863.The following generations of linguists resumed the work. It was with the 380th delivery that appeared after 123 years of work, the 32nd and the last volume of the dictionary which represented about 320 000 entries [2].

The figure below shows an excerpt[1] from the dictionary for the word ' WÖRTERBUCH/Dictionary'.

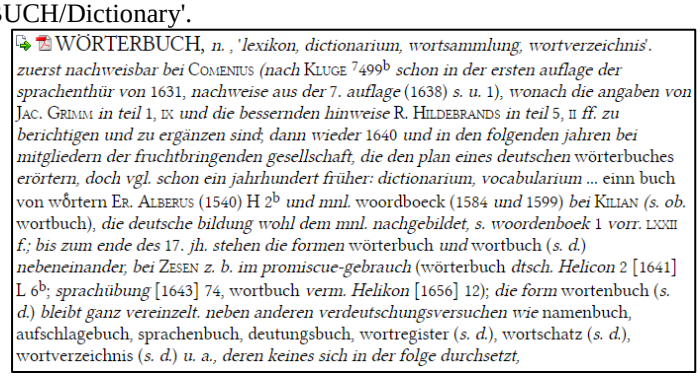

WÖRTERBUCH, *n.* , *'lexikon, dictionarium, wortsammlung, wortverzeichnis'. zuerst nachweisbar bei* COMENIUS *(nach* KLUGE 7499b *schon in der ersten auflage der sprachenthür von* 1631, *nachweise aus der* 7. *auflage* (1638) *s. u.* 1), *wonach die angaben von* JAC. GRIMM *in teil* 1, IX *und die bessernden hinweise* R. HILDEBRANDS *in teil* 5, II *ff. zu berichtigen und zu ergänzen sind; dann wieder* 1640 *und in den folgenden jahren bei mitgliedern der fruchtbringenden gesellschaft, die den plan eines deutschen* wörterbuches *erörtern, doch vgl. schon ein jahrhundert früher: dictionarium, vocabularium* ... einn buch von wôrtern ER. ALBERUS (1540) H 2b *und mnl.* woordboeck (1584 *und* 1599) *bei* KILIAN *(s. ob.* wortbuch), *die deutsche bildung wohl dem mnl. nachgebildet, s. woordenboek* 1 *vorr.* LXXII *f.; bis zum ende des* 17. *jh. stehen die formen* wörterbuch *und* wortbuch (*s. d.*) *nebeneinander, bei* ZESEN *z. b. im promiscue-gebrauch* (wörterbuch *dtsch. Helicon* 2 [1641] L 6b; *sprachübung* [1643] 74, wortbuch *verm. Helikon* [1656] 12); *die form* wortenbuch (*s. d.*) *bleibt ganz vereinzelt. neben anderen verdeutschungsversuchen wie* namenbuch, aufschlagebuch, sprachenbuch, deutungsbuch, wortregister (*s. d.*), wortschatz (*s. d.*), wortverzeichnis (*s. d.*) *u. a., deren keines sich in der folge durchsetzt,*

Fig. 1. An excerpt from the DWB dictionary

According to this excerpt, we note that the search for a given word in the German electronic history dictionary makes it possible to present the various synonyms of the searched word as well as its inflected forms, the articles which contain this word classified by date of appearance, links to its articles, and links to search in other dictionaries.

2.2 English Historical Dictionary OED

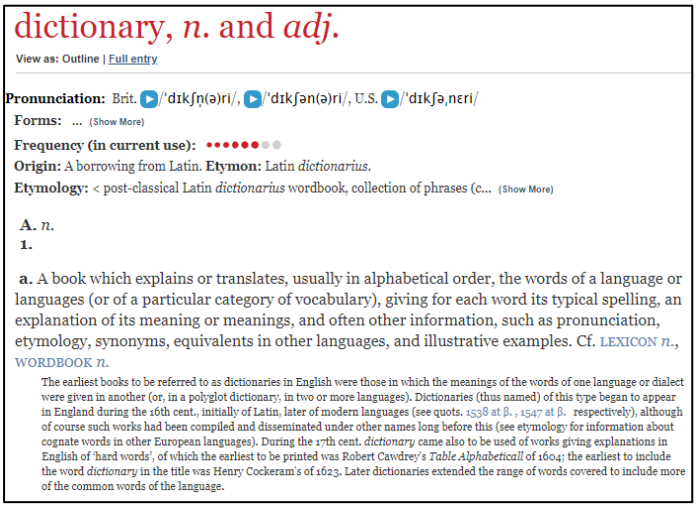
The Oxford English Dictionary (OED) is a reference dictionary for the English language. It is published by The Oxford University Press. The idea of this project was suggested in a lecture given in 1859 at the British Library by the poet Richard Chenevix Trench [3]. He showed the failings of English dictionaries and he proposed the development of an English historical dictionary. The essential task of a dictionary would be to trace the history and trajectory of each word, illustrating with quotations the nuances of meaning and uses that have emerged over time.

This dictionary contains words from the UK and various parts of the English-speaking world: North America, South Africa, Australia, New Zealand, Caribbean. As with other major dictionaries, the editorial work spanned several decades. The first complete edition, comprising twenty volumes, is published in 1928. It has been regularly updated since its publication. This edition has a total of 15,487 pages, with 414,825 articles illustrating 1,827,306 citations. A first supplement appeared in 1933 and four others were published between 1972 and 1986. In 1989, the whole is reissued. The book was published on CD-ROM in 1992 and became available online [2].

Figure 2 below shows an excerpt[2] from the dictionary for the word 'Dictionary'.

[1] http://woerterbuchnetz.de/cgi-bin/WBNetz/wbgui_py?sigle=DWB&mode=Vernetzung&lemid=GW27036#XGW27036

[2] http://www.oed.com/view/Entry/52325


dictionary, *n.* and *adj.*

View as: Outline | Full entry

Pronunciation: Brit. /ˈdɪkʃn(ə)ri/, /ˈdɪkʃən(ə)ri/, U.S. /ˈdɪkʃəˌnɛri/

Forms: ... (Show More)

Frequency (in current use):

Origin: A borrowing from Latin. **Etymon:** Latin *dictionarius*.

Etymology: < post-classical Latin *dictionarius* wordbook, collection of phrases (c... (Show More)

A. *n.*

1.

a. A book which explains or translates, usually in alphabetical order, the words of a language or languages (or of a particular category of vocabulary), giving for each word its typical spelling, an explanation of its meaning or meanings, and often other information, such as pronunciation, etymology, synonyms, equivalents in other languages, and illustrative examples. Cf. LEXICON *n.*, WORDBOOK *n.*

The earliest books to be referred to as dictionaries in English were those in which the meanings of the words of one language or dialect were given in another (or, in a polyglot dictionary, in two or more languages). Dictionaries (thus named) of this type began to appear in England during the 16th cent., initially of Latin, later of modern languages (see quots. 1538 at β. , 1547 at β. respectively), although of course such works had been compiled and disseminated under other names long before this (see etymology for information about cognate words in other European languages). During the 17th cent. *dictionary* came also to be used of works giving explanations in English of 'hard words', of which the earliest to be printed was Robert Cawdrey's *Table Alphabeticall* of 1604; the earliest to include the word *dictionary* in the title was Henry Cockeram's of 1623. Later dictionaries extended the range of words covered to include more of the common words of the language.

Fig. 2. An excerpt from the OED dictionary

The latter shows how the search for a word in the OED online dictionary gives information on the different pronunciation of this word, its etymology and its different meanings.

According to the study of the state of the art on the creation of historical dictionaries of other languages, we can see that the objectives of these dictionaries revolve around the study of the evolution of the meanings of the words since their first appearance. In the coming section of this work, we will give an overview of the different attempts to complete a historic dictionary for Arabic language.

2.3 Attempts to build a historical dictionary for Arabic: an overview

After several centuries, the Arab-Muslim cultures has not yet have its historical dictionary which would be the faithful reflection of its speech, its thought and its heritage, like any other great human civilization. Some unanimous researchers have said that the historical Arabic dictionary has become a necessity in the era of the computerization of knowledge of all kinds. For this purpose several projects have been launched in different countries (such as Germany, Egypt, Qatar, and America) for the realization of an Arabic electronic historical dictionary.

In 1935, a first attempt of the historical dictionary was launched and directed by August Fischer in Egypt. Fischer used texts from the Koran, ancient Arabic poems, etc. as material to construct this dictionary. After the death of Fischer in 1949, the German Oriental Company "Deutsche Morgenländische Gesellschaft" entrusted the project management to Jörg Kraemer, Helmut Gätje, Anton Spitaler, and Manfred Ullmann in Germany. The first edition of the dictionary was launched in 1970 concerning the letter "kaf / ك", followed by the second edition which includes the letter "lam / ل" and was divided into four volumes. The first one appeared in 1983, the second in 1991, the third one in 1999 and the last in 2009. The work is not yet complete [2].

In April 2004, the Historical Dictionary of the Arabic Language Committee was founded by a decision of the Arab Language and Science Counseling Union in Cairo (Egypt). This project aims to create a historical dictionary of the Arabic words and their uses in order to indicate the change of their meanings through time and space [4]. It is still under the studying stage of the corpus.

A third attempt of the project is the attempt at the Arab Center for Research in Doha in 2013, whose initial steps have been to prepare a reference bibliography of the sources of the linguistic corpus of the dictionary. The bibliography of the second phase, up to the year 500 AH according to Home Page of Doha Historical Dictionary [5].

As a best of our knowledge, the creation of historical dictionaries for other languages does not use the natural language processing (NLP) tools, whereas for the attempts that have been made to create the historical dictionary for Arabic, they were not successful.

As for the attempt at the Arab Center for Research in Doha, the team pointed out the need to rely on NLP tools to build such a dictionary [6].

So, we will focus on using NLP tools for the automation of a set of tasks that facilitate the work of linguists for the creation of a historical dictionary for Arabic.

3 Objectives

The idea of creating a historical dictionary for Arabic is important for the following two reasons. On the one hand, it offers the opportunity to understand our language and its evolution over the centuries in order to facilitate the understanding of our intellectual, scientific and cultural heritage by discovering the detailed history of each word. On the other hand, the realization of such a dictionary allows the Arabic language to have its own historical dictionary.

As mentioned previously, several efforts have been made to build a historical dictionary for the Arabic language. Unfortunately, these works were not at an advanced level and there is still no historical dictionary for Arabic.

Our objective is to help the linguists build a historical dictionary for Arabic by proposing a standardized structure using the XML language of a dictionary of meanings. This structure gives a historical account of the oldest use of a given Arabic word, as well as the variation of the meanings of the word through time.

Indeed, following the study of the structures of the historical dictionaries of the other languages, we noticed that the last ones study the variation of the meanings of the words through the time since the first dates of their appearance without giving information on the geographical aspect , neither the first places of appearance of words.

In this work, we aim not only to study the variation of the meaning of Arabic words over time but we also intend to give information on the first places of appearance of the word as well as their places of rebroadcasting.

It is worth pointing out that we made XML as our prior choice because it has a compatible hierarchical text format that is simple to implement, and easily interpreted by both humans and computers.

4 Proposed model

In this work, we propose a simple and structured model designed to capture and preserve historical information of an Arabic word and the variation of its meaning through time, as well as to help the linguists build a historical dictionary for Arabic language.

In fact, Doha site[3] submitted descriptions of their aims behind the draft of the historical lexicon of Doha for the Arabic language and the contents of this lexicon of historical information.

We reviewed these descriptions, and with the help of some members of the project team we developed a list of elements required to represent this information in an XML format.

The model contains a description of elements which enables us to monitor the Arabic word throughout history by capturing and storing the following information:

- The oldest use of the word, together with its meaning, users and sources,
- The meaning of the word, its historical appearance in the blog, with its indications, dates of use, its users and its places.

In order to extract historical information of a word, an indexing phase is necessary. This phase consists of representing the texts of our corpus in an XML form.

The texts of the corpus are thus represented in the form of two files under different extension (TXT and XML): the TXT extension contains the value of the text and the XML extension contains the title, the author, the period and the geography of appearance of the text as well as its value (see figure 3).

In fact, in our work we have used the Historical Arabic Dictionary Corpus (HADC) [7] which contains texts in Classical Arabic and Modern Standard Arabic with more than 116 millions of words. The list includes many text types such as poetry , the Quran, literary prose, Hadiths, history and genealogy, religions and doctrines, encyclopedias and dictionaries, journalistic texts, geography and travel literature. Indeed, the title of each document in the corpus is recorded under the headings of: Author's name - Date of death. The death of the author is so important as it sheds light on the history of the linguistic document.

[3] https://www.dohadictionary.org/AR/Lexical_Services/Pages/Bibliography.aspx

```
<text>
<titre>1349 - شعر جبران خليل جبران</titre>
<auteur>جبران خليل جبران</auteur>
<periode>1349 </periode>
<géogrpahie>الشام</géogrpahie>
<valeur>
قبس بدا من جانب الصحراء
هل عاد عهد الوحي في سيناء
أرنو إلى الطور الأشم فأجتلي
إيماض برق واضح الإيماء
حيث الغمامة والكليم مروع
أرست وقوراً أيما إرساء
دكناء مثقلة الجوانب رهبة
مكظومة النيران في الأحشاء
</valeur>
</text>
```

Fig. 3. Structure of the "xml" file of a text

After indexing the texts of the corpus, the objective of this work is to retrace the detailed history of an Arabic word in order to build the desired dictionary.

The following figure shows an excerpt from the dictionary described in XML.

```
<dictionary>
<word value="">
<first_date_of_appearance></first_date_of_appearance>
<sense></sense>
<first_places_of_appearnce>
<place></place>
</first_places_of_appearnce>
<meanings>
<meaning id="">
<value></value>
<beginning></beginning>
<authors>
<name_author></name_author>
</authors>
<first_places>
<place></place>
</first_places>
<places_of_spreading>
<place></place>
</places_of_spreading>
</meaning>
</meanings>
</word>
</dictionary>
```

Fig. 4. XML description of the dictionary of meaning

The element <first_date_of_appearance> contains the first date of appearance of a word and its first meaning <sense>.

However the element <first_places_of_appearance> contains the places where this word first appeared.

The element <meanings> contains the different meanings of the word, historically ranked according to its appearance in the corpus, specifying for each meaning the

history of its oldest use (<beginning>), its oldest place of appearance (<first_places>), its users (<authors>) and its places of spread (<places_of_spread>).

For example, in order to retrace the history of the word "القطار"(train), we chose documents containing this word from the HADC corpus. The following table presents the different texts chosen.

Table 1. Some contexts of uses extracted from the HADC corpus

Document	Title	Context of use
1	0060 - شعر مالك بن الريب (Poetry of Malek Ibn Raib)	نَقِيّ اللّونِ عَذبٍ كَما شيفَ الأقاحي بالقطار أتجزع أن عَرَفتَ بِبَطنِ (So pure the color is, pure like daisies made beautiful by rain, do you fret if you know that in Batin)
2	0458 – المخصص (Amukasses)	وقال بعير مقطور إلى آخر مشدود إلى القطار من الإبل (It is said that a camel is trailed to the end of a tugged train of camels)
3	0030 - شعر عروة بن حزام (Poetry of Iroua Ibn Hizam)	وتحتهما حقفان قدْ ضربتهما قطارٌ منَ الجوزاءِ ملتبدان (They were struck by a train of noble Gemini)

According to the first document named 0060 - شعر مالك بن الريب (Poetry of Malek Ibn Raieb), and referring to the XML extension of this document, the structure of the dictionary of meaning is as follows:

```
<dictionary>
<word value="القطار">
<first_date_of_appearance>0060</first_date_of_appearance>
<sense>جمع قطْر وهو المطر</sense>
<first_places_of_appearnce>
<place>الجزيرة</place>
</first_places_of_appearnce>
<meanings>
<meaning id="1">
<value>جمع قطْر وهو المطر</value>
<beginning>0060</beginning>
<authors>
<name_author>مالك بن الريب التميمي</name_author>
</authors>
<first_places>
<place>الجزيرة</place>
</first_places>
<places_of_spreading>
<place>الجزيرة</place>
</places_of_spreading>
</meaning>
</meanings>
</word>
</dictionary>
```

Fig. 5. Dictionary of meanings: first iteration

Finally, according to this method, the model for tracing the variation of the word's meaning through time is described by the following figure.

```
<dictionary>
<word value="القطار">
<first_date_of_appearance>0030</first_date_of_appearance>
<sense>جمع قطْر وهو المطر</sense>
<first_places_of_appearnce>
<place>الجزيرة</place>
</first_places_of_appearnce>
<meanings>
<meaning id="1">
<value>جمع قطْر وهو المطر</value>
<beginning>0030</beginning>
<authors>
<name_author>عروة بن حزام العذري</name_author>
</authors>
<first_places>
<place>الجزيرة</place>
</first_places>
<places_of_spreading>
<place>الجزيرة</place>
</places_of_spreading>
</meaning>
<meaning id="2">
<value>من الإبل عددٌ منها بعضُهُ خَلْفَ بعض على نَسَق واحد</value>
<beginning>458</beginning>
<authors>
<name_author>علي بن إسماعيل بن سيده الأندلسي</name_author>
</authors>
<first_places>
<place>الأندلس</place>
</first_places>
<places_of_spreading>
<place>الأندلس</place>
</places_of_spreading>
</meaning>
</meanings>
</word>
</dictionary>
```

Fig. 6. An XML model of a dictionary of meanings

5 Conclusion

In this article we have proposed an XML model for the representation of information needed to create the historical dictionary of the Arabic language. This model describes, for a given word, the variation of its meaning through time in order to trace its history.

For our future work, we are going to develop our method for the extraction of information described by the proposed model. The latter is based on the corpus designed specifically for the creation of the historical dictionary for Arabic.

Acknowledgment

We would like to thank Almoataz B. Al-Said for his incessant help and precious pointers in the preparation of this dictionary.

References

1. A.B. Al-said, Computerizing Historical Arabic Dictionary, al-Lisan al-arabi journal, al-Ribat, (2014).
2. A.B.Al-Said, A Corpus-based Historical Arabic Dictionary: Linguistic & Computational processing. PhD Dissertation, Cairo University, (2011).
3. J. A. H , Murray , A New English Dictionary on Historical Principles: Founded Mainly on the Materials Collected by the Philological Society. The Clarendon Press, (1888).
4. M. H. AbdelAziz, A Historical Dictionary for Arabic Language: Documents and examples, P.166, (2008)
5. Doha Historical Dictionary for Arabic Language. 2018. https://www.dohadictionary.org/AR/Lexical_Services/Pages/Bibliography.aspx
6. Al-Said, A. B., Computerizing Historical Arabic Dictionary, al-Lisan al-arabi journal, al-Ribat , vol. 74, (2014).
7. Al-Said, A. B., and L. Medea--García, The Historical Arabic Dictionary Corpus and its Suitability for a Grammaticalization Approach, 5th international conference in linguistics, Grammar and Corpora, Poland, (2014).