Structuring Arabic lexical and morphological resources using TEI: theory and practice

Ouafae Nahli, ILC-CNR, Italy and Angelo Mario Del Grosso, ILC-CNR, Italy
{name.surname}@ilc.cnr.it

***Abstract*—An Arabic word can be described according to its lexical and morphological information. The lexical information, conveyed by the root, consists of both semantic meaning and syntactic properties (e.g. parts of speech). The morphological information, encoded by patterns, is useful to group the words having similar syntactic, inflectional and semantic behaviour.**

Lexical analysis and morphological analysis have been separately described since the very first studies of the Arabic language. Although several scholarly works have illustrated Arabic lexicon models that encode semantic meanings, a systematic description of word patterns is still strongly lacking. In this work, we have implemented an exhaustive resource consisting of two levels: lexical and morphological. The lexical level collects information extracted from the dictionary *al=qāmūs al=muḥīṭ*. The morphological level describes pattern formalization, which allows to enrich word descriptions with additional semantic, morphosyntactic and inflectional information.

To build our digital resource, taking into account primary source, lexical requirements, and reusability, we followed the guidelines provided by the Text Encoding Initiative (abbreviated as TEI). In particular, we adopted the TEI module for the encoding of digital dictionaries and lexicons to formally represent the medieval *al=qāmūs al=muḥīṭ* dictionary. Given the complexity of describing the morphological information present in the patterns, we also used the TEI module devoted to encoding feature structures.

Consequently, we are building an exhaustive resource formed by the lexical and the morphological blocks. These two components are distinct but complementary resources where the lexical data are connected to morphological information. In addition, the morphological resource can be used as a stand-alone tool that allows the morphological analyzers to capture aspects of meaning that cannot be identified by current systems.

***Index Terms*—classical Arabic dictionary, digital lexicography, *al=qāmūs al=muḥīṭ*, word patterns, TEI, feature structures**

I. Introduction

THE Arabic language is a non-concatenative Semitic language described at two distinct but complementary levels:

- Lexicographic studies focus on describing words and their semantic characteristics. The Arabic lexicon is organized considering the roots which are the main vehicular axes of the semantic fields and to which the derived words are linked. For example, the words *katab-a*[1] ‘he wrote’[2], *maktab* ‘office; desk’ and *istaktab-a* ‘he asked to write; he dictated’, derive from - and are listed under - the root *ktb*, which transmits the semantic field of “writing”.
  The order of the consonants constituting the root is also significant. The words *bakat-a=hu* ’he hit him’ and *kabat-a=hu* ’he prostrated him, he humbled him’ derive from the roots *bkt* and *kbt*, respectively, which have different semantic fields.
- Morphological studies deal with morphological, syntactic and semantic aspects, by identifying word patterns. The patterns serve to generate words in combination with the roots. For example, the word patterns `R1aR2aR3-a`, `R1āR2iR3` and `maR1R2aR3`, convey morphosyntactic and semantic meanings about “dynamic, active and perfective verbs”, “agent name executing the action” and “noun of place where the action unfolds”, respectively. Furthermore, word generation also depends on the dominant meaning embodied in the root. For example, the “Noun of place where action unfolds” can derive from the root *ktb*, whose dominant semantic field is the active action of “writing”, but it cannot derive from the root *kbr* whose dominant meaning is stative, ‘to be great, to be large’.

Several studies have discussed the modelling of the Arabic lexicon. For example, [18] analyzes the Arabic inflection paradigms of verbs, while [21] describes the Arabic syntactic characteristics of verbs. The work of [25] concerns a customization of the lexical structure devoted to Arabic lexicons by adopting the TEI dictionary schema. The authors of this work have also created an Arabic lexicon editor, equipped with a check tool, based on both the ISO standard LMF (Lexical Markup Framework) and the TEI guidelines [5]. Unfortunately, the developed tool has not been published yet.

In [28] [29] [30], the authors state that the first goal of Arabic computational morphology is to formalize the inflectional morphology. Their goal is to create a procedure for a complete morpho-syntactic annotation of Arabic texts, leveraging an

This paper has been submitted for review on 2 May, 2021.

Ouafae Nahli works at the Institute for Computational Linguistics (ILC-CNR), Pisa, Italy (e-mail: ouafae.nahli@ ilc.cnr.it).

Angelo Mario Del Grosso works at the Institute for Computational Linguistics (ILC-CNR), Pisa, Italy (e-mail: angelo.delgrosso@ ilc.cnr.it).

[1] For the syntax and semantics of interlinear glosses, we used the Leipzig Glossing Rules [11]. Interlinear morpheme-by-morpheme glosses give information about the meanings and grammatical properties of individual words and parts of words. Segmentable morphemes are separated by hyphens and clitic boundaries are marked by an equals sign. For example, in *katab-a=hu* ’he wrote it’, the final vowel *–a* is the inflectional suffix of the third singular person and the suffix *=hu* is the accusative pronoun.

[2] Arabic words are transliterated and accompanied by the corresponding English meaning within single quotation marks.

advanced FST (finite-state transducer) implementation and using a root-and-pattern system. They created an inflectional resource for Arabic with a large coverage of inflected forms. Even for this work, the created resources are not available.

Finally, several efforts resulted in the Arabic WordNet, which is an Arabic lexical resource similar to EWN (see [1] [2] [4] [6]).

In contrast, there is no linguistic model of the morphological system providing a systematic description of patterns and their functions [35]. The word pattern description is well-established within the linguistic tradition but, to the best of our knowledge, it has never been faced in a computational perspective.

Consequently, our work aims to develop an Arabic resource able to consider two perspectives, i.e. the lexical level, which defines the formal representation of the lexicon - extracted from the primary source *al=qāmūs al=muḥīṭ (henceforth qāmūs)*; and the morphological level, which formalizes the inflectional, semantic and syntactic characteristics of the word patterns. As the quotation states:

> Medieval grammarians and lexicographers had designed Arabic morphology and lexicography for human minds tooled up with paper, whereas we should design Arabic computational morphology for humans equipped with processors and memory devices. [30] (page 7)

Hence, one of the challenges is to select suitable digital models and approaches. Many formal models and practical approaches have been proposed which describe and implement electronic lexical resources. Among these, alongside the well-known model provided by the Text Encoding Initiative, the digital lexicography community recently obtained significant results, including a new multi-modular release of the Lexical Markup Framework (LMF)[3] [20] as well as a promising proposal intended to provide a higher degree of compatibility among lexical resources encoded by using TEI, called TEI Lex-0[4] [19].

With respect to the aforementioned initiatives, the special feature of the TEI model takes into account, at the same time, a formal representation of the lexical elements and of the primary sources [32]. For these reasons, we decided to follow the guidelines provided by TEI to obtain a digital systematic resource. In particular, since TEI is designed to be modular, a lexical resource can be encoded by identifying suitable elements (XML tags) explicitly declared within specialized modules. For instance, the main encoding elements that we used for our work are defined within the “dictionary” module (9th module of the guidelines)[5] and the “feature structures” module (18th module of the guidelines).[6]

Module 9 defines elements for encoding dictionaries and lexical resources. Module 18 defines elements for encoding complex structures - known as feature structures - which group different properties in a bundle of nested entities.

In the light of these considerations, the elements of the TEI Module 9 have been adopted to encode the part of the resource concerned with lexical data. On the other hand, the elements of the TEI Module 18 have been adopted to encode the part of the resource concerned with the description of word patterns. As a matter of fact, the expressiveness of the feature structures allows us to represent word patterns by exploiting their general purpose “linguistic metalanguage”. Thanks to this choice, any descriptive morphological unit within a particular structure can make reference to elements encoded in any other feature structure.

In our previous work [27], we described a first attempt to model the morphological word patterns by using the TEI interpretation approach. This model makes it possible to separately encode the two levels (lexical and morphological), and at the same time to link them together. However, during the encoding work, we realized that the expressiveness of the `<interp>` TEI element was not sufficient to structure the complexity and richness of the morphological information conveyed by the patterns. Consequently, the “patterns structure” led us to select the encoding elements defined within module 18 of the guidelines, i.e. Feature Structures. This paper is an extension of our previous conference paper [27], where we illustrated some issues we addressed when constructing the aforementioned multi-perspective resource and some choices that we made to overcome the problems. We explain why we adopted the *feature structures* of the TEI encoding system to describe the morphological perspective.

The paper is organized as follows. Section II provides an overview of the morphological characteristics of the Arabic language. In section III, we discuss the primary source of our work *qāmūs* and explain the steps we followed to obtain the digital Arabic lexical resource. Section IV describes the encoding model we chose to design our resource according to the TEI guidelines. This section encompasses two topics: (i) representation of the lexical semantic level; (ii) representation of the morphological level. In the final part of the article (sections V and VI), we discuss and summarize our research purposes.

II. Arabic language Characteristics

A. Modern Standard Arabic vs Classical Arabic

In the introduction of his work, Wehr sought to define the Arabic language from both a diachronic and a diatopic point of view [33]. Throughout the Arabic world, the vocabulary and phraseology of modern written Arabic are found in the prose of books, newspapers, periodicals and letters and are used in formal public communication (radio and television).

[3] The current LMF standard is the ISO 24613-1:2019, for more details see https://www.iso.org/obp/ui/#iso:std:iso:24613:-1:ed-1:v1:en

[4] In order to delve into TEI lex-0 data format, see also the documentation at the following web address
https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html

[5] TEI-Dictionary, 9th module of the TEI guidelines is devoted to encoding lexical resources of all kinds. https://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html.

[6] TEI-FS, 18th module of the TEI guidelines is devoted to encoding analytical descriptions of all kinds. https://tei-c.org/release/doc/tei-p5-doc/en/html/FS.html

Consequently, the morphology and syntax of written Arabic are essentially the same in all Arabic countries.

Traditional adherence to the ancient linguistic norms and models of classical literature, especially the *Qur'an*, has allowed the language to remain intact over the centuries. As a result, Arabic phonology, morphology, syntax and much of the vocabulary have remained relatively unchanged since the earliest times. As pointed out by Wehr, quotations from the Qur'an and classical literature in modern literature show that it is not easy to distinguish between living and obsolete usage.

Thus, modern Arabic is a written language heavily influenced by traditional norms that must express a multitude of new foreign concepts, not for a single but for many countries covering large geographical areas. Therefore, vocabulary differences, both diachronic and diatopic, are mainly limited to the domain of specialized vocabulary that should be included in a separate dictionary. Likewise, colloquialisms and dialectal expressions that have spread in written form must be listed in appropriate dialect dictionaries or glossaries that vary from one country to the other.

In conclusion, Wehr distinguishes three types of resources. The dictionary of the Arabic written language has undergone few diachronic and diatopic variations, dialect glossaries and resources of specialized domain vocabularies.

In this work, we aim to construct a dictionary of the Arabic written language.

B. *Arabic morphological characteristics*

The Arabic language structure is based on a discontinuous morphological system. Words result from the combination of two abstract morphemes, namely the root and the word pattern, which intermingle and emerge in a discontinuous manner. The root is exclusively consonantal and carries a dominant meaning that is found in all derived words. Tables I and II illustrate some examples of words deriving from the root *ktb* and *kbr*, respectively.

TABLE I: Some examples derived from the root {*ktb*}

Word	English translation	Pattern
katab-a	he wrote	R_1aR_2aR_3-a
kattab-a	he did write	R_1aR_2R_2aR_3-a
istaktab-a	he asked to write; he dictated	istaR_1R_2aR_3-a
maktab	place for writing, office	maR_1R_2aR_3
maktabah	library; bookstore	maR_1R_2aR_3ah

TABLE II: Some examples derived from the root {*kbr*}

Word	English translation	Pattern
kabar-a	he was older than..	R_1aR_2aR_3-a
kabir-a	he was, became full-grown; he was old	R_1aR_2iR_3-a
kabur-a	he was great, big; he was important	R_1aR_2uR_3-a
kabbar-a	he raised	R_1aR_2R_2aR_3-a
istakbar-a	he considered great or important	istaR_1R_2aR_3-a
kabīr	great, big, large, voluminous, spacious	R_1aR_2īR_3
kibar	greatness; largeness	R_1iR_2aR_3

The root *ktb* transmits the semantic field of 'writing' and the root *kbr* transmits the semantic domain of 'greatness; largeness'. Therefore, words deriving from the same root constitute a derivational family and are morphologically, phonologically and, to some extent, semantically related [10].

To give a phonological structure to a word, vowels and sometimes specific consonants are added to the radical consonants. In order to indicate the nature and position of the potential added affixes, we use the word pattern. For example, we use the pattern $maR_1R_2aR_3$ to derive the 'name of place where action takes place', like *maktab* 'office (place of writing)'.

In addition, the word pattern also conveys morphosemantic and morphosyntactic information. For example, the patterns $R_1aR_2uR_3$-*a* and $R_1aR_2aR_3$-*a* indicate that the corresponding word is an active and perfective verb. However, according to $R_1aR_2uR_3$-*a*, a constructed verb belongs to the category of 'stative verbs' and is necessarily intransitive, like *kabur-a*. Instead, according to $R_1aR_2aR_3$-*a*, a constructed verb belongs to the category of 'dynamic verbs' and can be transitive or intransitive.

Summing up, root and word patterns are complementary in capturing the information conveyed by a word. Indeed, current psycholinguistic studies show that roots and word patterns are considered lexical units that govern the entire word recognition process [7]–[11]. These two perspectives are ultimately formalized in terms of autosegmental phonology, [22]–[24], as shown in Fig. 1.[7]

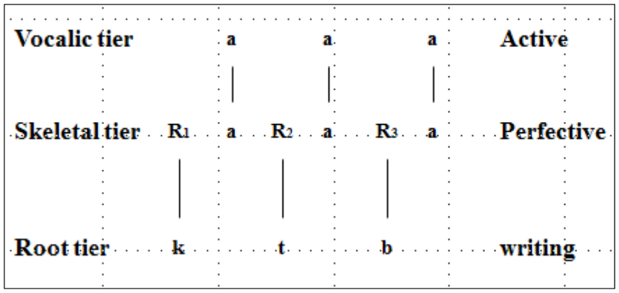


Fig. 1: Example of multi-linear representation: triliteral verb *kataba*

III. PRIMARY RESOURCE

A. *Characteristics of qāmūs*

We chose to work with the Arabic lexicon *qāmūs* for a number of different reasons, primarily, because of the authoritative status in the Arabic speaking world and the comprehensiveness of its entries. In his introduction, the author, 'al-fīrūz'ābādī (1329-1414), states that the lexicon was created by merging together several pre-existing dictionaries. As part of the process of compilation, 'al-fīrūz'ābādī greatly reduced the original contents from the source dictionaries he was using by eliminating examples, Quranic quotations, poetry and some grammatical information. The fact that AQAM is a well-structured lexicon containing short lexical items makes it an excellent candidate for conversion into a computational lexicon.

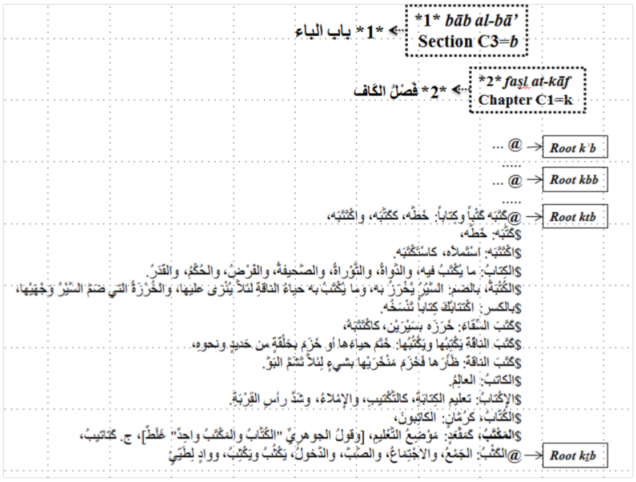
[7] Word pattern level is called skeletal tier by McCarthy.

B. Macrostructure of *qāmūs*

The entries of *qāmūs* were arranged by following the rhyme system. Therefore, lexical items were recorded as a unique group under the root from which they derive, and the order of the roots was obtained according to the following schema: firstly, the third radical consonant, then the first, and finally the second [3]. For example, the entry *maktab* is recorded under the root *ktb*, within the section of the consonant *b* (*bāb al=bāʾ*), then within the chapter of the consonant *k* (*faṣl al=kāf*), and finally according to the consonant *t*.

To obtain the digital version of *qāmūs*, our project went through various phases described in [26] [16] [17], which are summarized below. The multi-level implicit information was extracted and tagged using formal rules written from the information provided by the dictionary itself. To identify the parts that characterize the macrostructure and each lexical entry along with its definition, we manually marked the text with a set of symbols that helped us to segment the text. As shown in Fig. 2, the text is constituted by:

- sections corresponding to the third consonant of the roots and marked by the symbol *1*.
- Each section is divided into chapters corresponding to the first radical and marked by the symbol *2*.
- The chapters are also divided into root families. The symbol @ marks the beginning of the family as well as the first lexical entry. Finally, the other entries of the family are marked with the symbol $.



1 باب الباء
1 *bāb al-bāʾ*
Section C3=*b*

2 فَصْلُ الكَاف
2 *faṣl at-kāf*
Chapter C1=k

... @ → *Root kʾb*
.....
... @ → *Root kbb*
.....
@كَتَبَه كَتْباً وكِتاباً: خَطَّه، ككَتَّبَه، واكْتَتَبَه، → *Root ktb*
$كَتَّبَه: خَطَّه،
$اكْتَتَبَه: استَمْلاه، كاسْتَكْتَبَه.
$الكِتابُ: ما يُكْتَبُ فيه، والدَّواةُ، والتَّوْراةُ، والصَّحيفةُ، والفَرْضُ، والحُكْمُ، والقَدَرُ.
$الكُتْبَةُ، بالضم: السَّيْرُ يُخْرَزُ به، وما يُكْتَبُ به حياءُ الناقةِ لِئَلاَّ يُنْزَى عليها، والخُرْزَةُ التي ضَمَّ السَّيْرُ وَجْهَيْها،
$بالكسر: اكْتِتابُكَ كِتاباً تَنْسَخُه.
$كَتَبَ السِّقاءَ: خَرَزَه بسَيْرَيْن، كاكْتَتَبَهُ،
$كَتَبَ الناقةَ يَكْتِبُها ويَكْتُبُها: خَتَمَ حياءَها أو خَزَمَ بحَلْقَةٍ من حَديدٍ ونحوِه،
$كَتَّبَ الناقةَ: ظَأَرَها فَخَزَمَ مَنْخَرَيْها بشيءٍ لِئَلاَّ تَشَمَّ البَوَّ.
$الكاتِبُ: العالِمُ.
$الإِكْتابُ: تعليم الكِتابَةِ، كالتَّكْتيبِ، والإِمْلاءُ، وشَدُّ رأسِ القِرْبَةِ.
$الكُتَّابُ، كرُمَّانٍ: الكاتِبونَ،
$المَكْتَبُ، كمَقْعَدٍ: مَوْضِعُ التَّعْليمِ، [وقولُ الجوهريِّ "الكُتَّابُ والمَكْتَبُ واحِدٌ" غَلَطٌ]، ج. كَتاتيبُ،
@الكَثْبُ: الجَمْعُ، والاجْتِماعُ، والصَّبُّ، والدُّخولُ، يَكْثُبُ ويَكْثِبُ، وَوادٍ لِطَيِّئٍ → *Root kṯb*

Fig. 2: Macrostructure of *qāmūs*

A structural and linguistic rules-based approach was designed for automatic extraction of the information. The addition of superficial markers allows us not only to segment the text but also to extract information about the value of the radical consonants. Fig. 3 illustrates the text segmentation phases, while Fig. 4 illustrates the result of text encoding in a proprietary XML format.

C. Microstructure characteristics

Lexical information extraction depends on the knowledge of the deep structure of *qāmūs*, the style of its author, and the systematic control of punctuation inside the lexical entries. To highlight the complexity of the microstructure, Fig. 5 illustrates the verb *ṣaḥib-a* as an example. Here, the commas and colons separate the different parts of information.

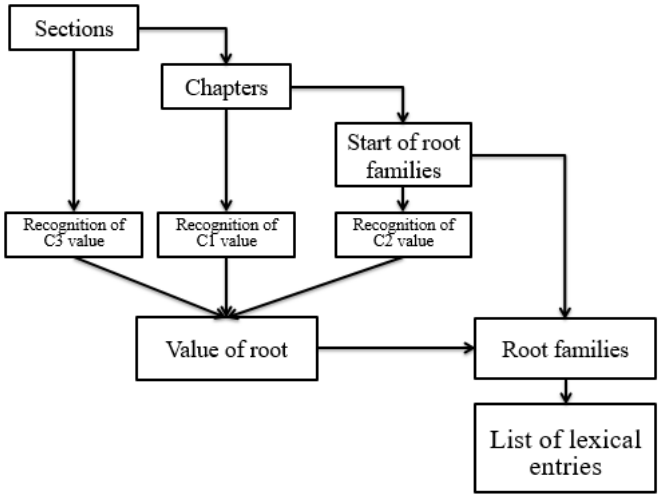
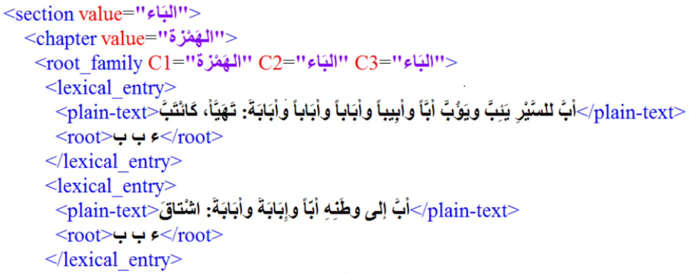


Fig. 3: Phases of text segmentation according to the macrostructure



```
<section value="البَاء">
  <chapter value="الهَمْزة">
   <root_family C1="الهَمْزة" C2="البَاء" C3="البَاء">
    <lexical_entry>
     <plain-text>أَبَّ للسَّيْرِ يَئِبُّ ويَؤُبُّ أَبّاً وأَبِيباً وأَبَاباً وأَبَابَةً: تَهَيَّأَ، كَائْتَبَّ</plain-text>
     <root>ء ب ب</root>
    </lexical_entry>
    <lexical_entry>
     <plain-text>أَبَّ إلى وطَنِهِ أَبّاً وإِبَابَةً وأَبَابَةً: اشْتَاقَ</plain-text>
     <root>ء ب ب</root>
    </lexical_entry>
```

Fig. 4: Text conversion to XML format

- The headword *ṣaḥib-a=hu* is formed by the word *ṣaḥib-a*. It is a verb, because it matches with one of the defined verbal patterns, $R_1aR_2iR_3$-*a*.[8] In addition, it is transitive because it is followed by the accusative pronoun =*hu*, which functions as a surface tag.
- Given the peculiarity of the Arabic script, the vowels in the manuscript may have been omitted or copied erroneously. The lexicographer introduced a better-known word after the headword to control its vocalization. In the example of Fig. 5, the headword *ṣaḥib-a=hu* is followed by the word *ka=samiʿ-a=hu*, formed by the conjunction *ka* 'as', the verb *samiʿ-a* 'he heard', and the accusative pronoun =*hu* 'it'. In addition, the vocalic alternation of the verb *samiʿ-a* is better known as **i/a**.[9] Therefore, the verb *samiʿ-a* 'he heard' is quoted to control vocalization of the headword *ṣaḥib-a=hu* and to indicate the inflection class of the lemma.

[8] With regard to verbs, we defined lists of inflectional prefixes, of inflectional suffixes, of accusative pronouns, and of 34 regular expressions that present possible inflected and derived verbal patterns.

[9] The perfect stem vowel is **i** (*samiʿ-a*) and the imperfect stem vowel is **a** (*ya-smaʿ-u*).

- After the control-word, the related *maṣdars* (verbal infinitive forms) are listed in the accusative indefinite forms: *ṣaḥābat-an* and *ṣuḥbat-an*.
- The lexical information is placed after the colon and it is composed by a single verb *ʿāšar-a=hu* ‘he was on intimate terms with him’. According to the style of the author, when the meaning consists of a single word, it is synonymous with the lemma.
- Other derivational information follows. The adjective *ṣāḥib-un* ‘companion’ is introduced by *wa=huwa* ‘and him/it (is)’. The adjective is accompanied by broken plurals introduced by *ǧ*, abbreviation of the plural (*ǧamʿ*). The words *wa=huwa* and *ǧ* act as surface patterns to capture adjectives and plurals, respectively.

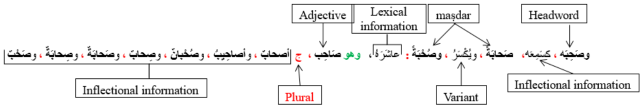

Adjective
Lexical information
maṣdar
Headword
وصَحِبَه ، كسَمِعَه، صَحابَةً ، ويُكْسَرُ ، وصُحْبَةً : عاشَرَهُ، وهو صَاحِب ، ج أصحابٌ ، وأصاحِيبُ ، وصُحْبانٌ ، وصِحابٌ ، وصَحابَةٌ ، وصِحابَةٌ ، وصَحْبٌ
Inflectional information
Plural
Variant
Inflectional information

Fig. 5: Microstructure of the lexical entry *ṣaḥiba*

This example justifies even more our decision to select *qāmūs* as dictionary, because it includes all inflectional information, (i.e. imperfective information or broken plurals) as well as derivational information (i.e. *maṣdars* and adjectives).

As summarized in Fig. 6, the set of symbols, the stylistic patterns, and punctuation make it possible to automatically perform text segmentation, information identification, and extraction, by structuring the data in electronic documents thanks to a descriptive formalism like the XML markup language.

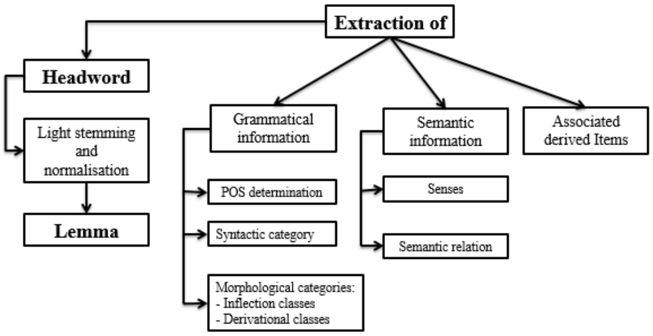
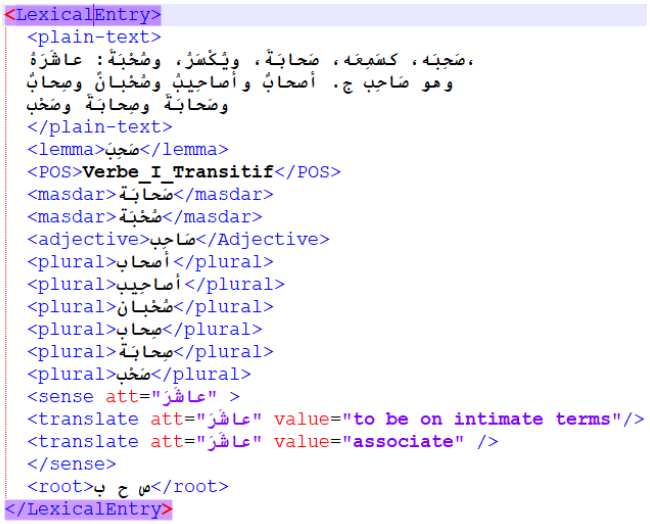


Fig. 6: Information extraction according to microstructure

All the sections of *qāmūs* have been encoded by means of an intermediate XML document.

Fig. 7 shows the intermediate XML of the extraction for the lexical entry *ṣaḥiba*. The XML tag `<plain_text>` contains the original part of *qāmūs*, from which we can extract information. Other XML tags make it possible to recognize the lemma, together with its morphosyntactic and semantic information (POS, transitivity, imperfectiveness, and sense) and corresponding derived words (*maṣdars*, adjectives and plurals of adjectives). In addition, we exploited the bilingual dictionary “An Advanced Learner’s Arabic-English Dictionary”[10] [31] to provide the English translations of the lemma.



```
<LexicalEntry>
  <plain-text>
  صَحِبَه ، كسَمِعَه ، صَحابَةً ، ويُكْسَرُ ، وصُحْبَةً : عاشَرَهُ
  وهو صَاحِب ج. أصحابٌ وأصاحِيبُ وصُحْبانٌ وصِحابٌ
  وصَحابَةٌ وصِحابَةٌ وصَحْب
  </plain-text>
  <lemma>صَحِبَ</lemma>
  <POS>Verbe_I_Transitif</POS>
  <masdar>صَحابَة</masdar>
  <masdar>صُحْبَة</masdar>
  <adjective>صَاحِب</Adjective>
  <plural>أصحاب</plural>
  <plural>أصاحِيب</plural>
  <plural>صُحْبان</plural>
  <plural>صِحاب</plural>
  <plural>صِحابَة</plural>
  <plural>صَحْب</plural>
  <sense att="عاشَرَ" >
  <translate att="عاشَرَ" value="to be on intimate terms"/>
  <translate att="عاشَرَ" value="associate" />
  </sense>
  <root>ص ح ب</root>
</LexicalEntry>
```

Fig. 7: Intermediate XML for the extraction *ṣaḥiba*

D. *Digital version of qāmūs*

The designed system was able to automatically convert all the sections of *qāmūs* to XML. The corresponding files are released as an open access bundle by means of the *CLARIN-IT* infrastructure.[11] CLARIN provides a descriptive file of the lexicon by linking the folders corresponding to each section. Each dossier corresponds to a section, and it contains:

- plain text files enriched with indicators that tag the start of section, chapters, root families and lexical entries;
- a folder containing XML files divided according to grammatical categories of lexical entries (Verb; Nouns; Adjectives). The proper names are presented in separate files according to their corresponding semantic classes (Plant, Food, Animal, Proper Name, Geographical name, Water, Group, and Others);
- a folder containing XML files of verbs, nouns and adjectives enriched with English translations.

IV. TEI REPRESENTATION OF THE RESOURCE

A. *Representation of the lexical semantic level*

In our previous works, we presented two formalization studies of *qāmūs* adopting standard formats like LMF (Lexical Markup Framework) [26] and LEMON (LExicon Model for ONtology) [16]. These studies showed that LMF and LEMON do not cover all Arabic language specificity.[12]

[10] We used the TEI version of “An Advanced Learner’s Arabic-English Dictionary”, published by Trustees of Tufts University, under the Perseus Project: http://www.perseus.tufts.edu.

[11] http://hdl.handle.net/20.500.11752/ILC-97. We can also find the registered dossiers for site search at: https://vlo.clarin.eu/, using *qāmūs* as word search.

[12] The LMF standard we adopted in the cited work is the ISO 24613:2008, which is under revision.

Consequently, we adopted the model, the encoding schema and the corresponding XML vocabulary designed and maintained by the Text Encoding Initiative (TEI).[13] This choice was motivated considering the expressiveness of the TEI to structure and share digital dictionaries as well as electronic lexicons and to describe complex structures. Indeed, such a schema takes into account the representation of the primary sources modelling lexical data and interpretative structures.

The TEI provides an excellent balance of the multiple dimensions of our resource. This framework was designed to represent complex textual material with complex structures, functions, and meanings. Moreover, the provided encoding schema has a modular design, so that it is possible to define suitable profiles able to appropriately represent the characteristics of the textual phenomena of interest like the dictionary structure. Finally, the framework provides two articulated and well-designed modules: the former devoted to the management of digital dictionaries and electronic lexicons (i.e. module n. 9 of the TEI guidelines), the latter devoted to the representation of complex feature structures (i.e. module n. 18 of the TEI guidelines).

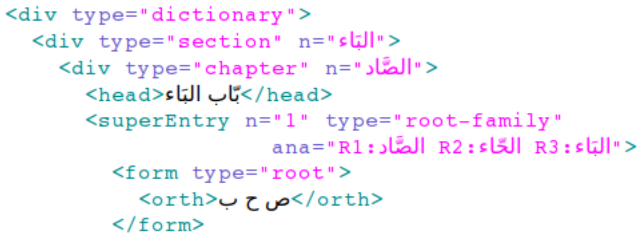
As concerns our resource, the main structure of the digital representation lies in the main structure of the primary source (*qāmūs*). In paper [27] we presented the TEI description extract from the lexical entry *ṣaḥib-a* that we also explain here, fragment by fragment.

The hierarchical segmentation and the divisions are described in Fig. 8, which outlines the main levels of *qāmūs*:

1) the top level corresponds to the whole dictionary;
2) the section level corresponds to the third radical;
3) the chapter level corresponds to the first radical;
4) the “super entry” corresponds to the root family.

Both the section and the chapter levels have been encoded by using the `<div>` element, adopting different values for the `@type` attribute.

The `<superEntry>` element denotes the root-family level and it is accompanied by the `@ana` attribute, which records the values of the radical consonantsR_1, R_2 and R_3.

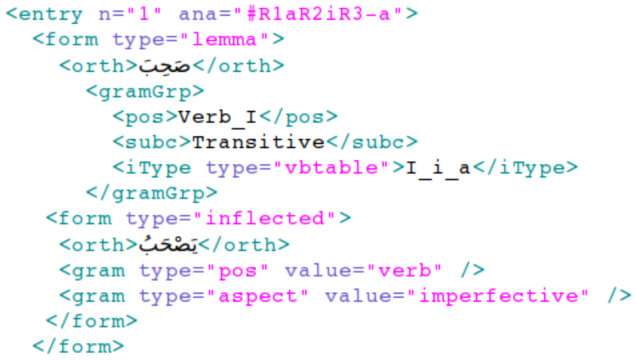


```
<div type="dictionary">
  <div type="section" n="البَاء">
    <div type="chapter" n="الصَّاد">
      <head>بّاب البَاء</head>
      <superEntry n="1" type="root-family"
                    ana="R1:الصَّاد R2:الحَاء R3:البَاء">
        <form type="root">
          <orth>ص ح ب</orth>
        </form>
```

Fig. 8: XML-TEI: Hierarchical structure

Fig. 9 illustrates the lemma *ṣaḥib-a*. The entry records the global attribute `ana=#R1aR2iR3-a`, which includes the lemma *ṣaḥib-a* within the group of verbs with the same word pattern `R1aR2iR3-a`.

We adopted the `@ana` attribute to associate the two coexisting and distinct levels, namely the lexical description, where we describe the lemma *ṣaḥib-a*; and the morphological description, where we describe the word pattern `R1aR2iR3-a`.



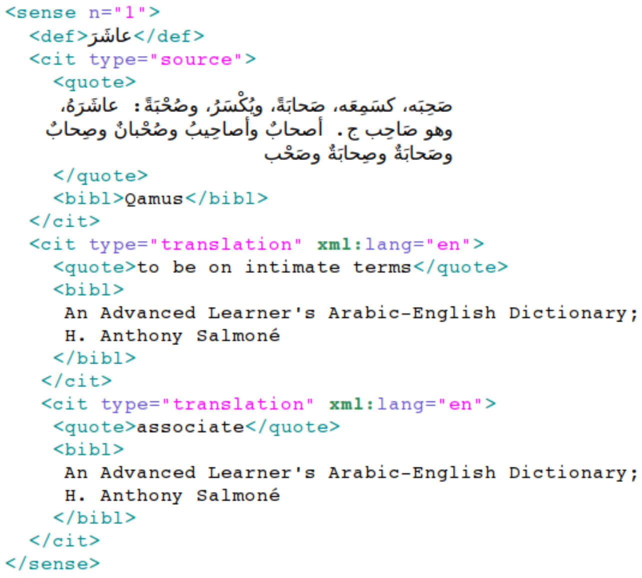
```
<entry n="1" ana="#R1aR2iR3-a">
  <form type="lemma">
    <orth>صَحِبَ</orth>
      <gramGrp>
        <pos>Verb_I</pos>
        <subc>Transitive</subc>
        <iType type="vbtable">I_i_a</iType>
      </gramGrp>
   <form type="inflected">
    <orth>يَصْحَبُ</orth>
    <gram type="pos" value="verb" />
    <gram type="aspect" value="imperfective" />
   </form>
  </form>
```

Fig. 9: XML-TEI: Example of the lemma *ṣaḥib-a*

The linguistic information about the lemma is also encoded in the `<GramGrp>` element:

- Part of speech: *Verb_I* (triliteral verb or form I).
- Syntactic information: Transitive.
- Inflectional information: the `<iType>` element indicates the inflectional class: *I_i_a*, i.e. it is a triliteral verb (form I) with the vocalic alternation i/a.
- An additional level of inflectional information encodes the imperfective form *ya-ṣḥab-u*.

Fig. 10 shows the XML fragment encoding the sense *ʿāšar-a* accompanied by two types of citations. The first type records the corresponding text extracted from the primary source *qāmūs* (`@type="source"`). The second type (`@type="translation"`) corresponds to the English translations (`@xml:lang="en"`) extracted from the bilingual dictionary.



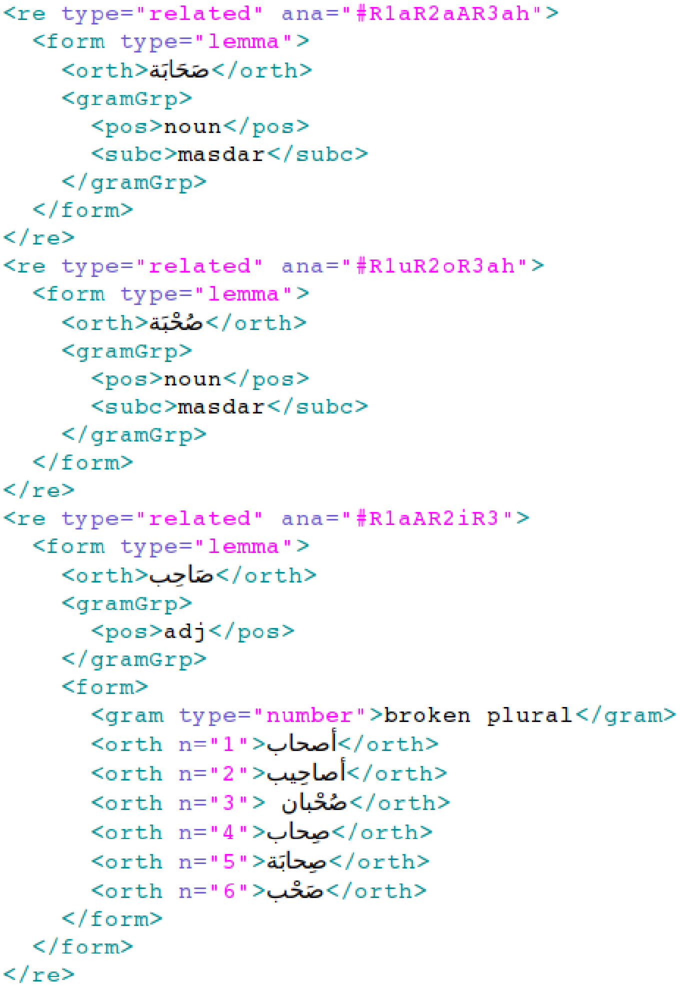
```
<sense n="1">
  <def>عاشَرَ</def>
  <cit type="source">
    <quote>
     صَحِبَه، كسَمِعَه، صَحابَةً، ويُكْسَرُ، وصُحْبَةً: عاشَرَهُ،
    وهو صَاحِب ج. أصحابٌ وأصاحِيبُ وصُحْبانٌ وصِحابٌ
                                وصَحابَةٌ وصِحابَةٌ وصَحْب
    </quote>
    <bibl>Qamus</bibl>
  </cit>
  <cit type="translation" xml:lang="en">
    <quote>to be on intimate terms</quote>
    <bibl>
     An Advanced Learner's Arabic-English Dictionary;
     H. Anthony Salmoné
    </bibl>
   </cit>
   <cit type="translation" xml:lang="en">
    <quote>associate</quote>
    <bibl>
     An Advanced Learner's Arabic-English Dictionary;
     H. Anthony Salmoné
    </bibl>
  </cit>
</sense>
```

Fig. 10: XML-TEI: Encoding of senses and citations

Adjectives and *maṣdars* are considered ‘related entries’ embodied in the main entry. Therefore, they have been encoded as related entries by using the `<re>` element, suitable to

[13] https://tei-c.org/ [Retrieved in May 2021].

our purposes because it contains the same elements as the principal entry [32]. Fig. 11 illustrates the three related entries that accompany the main entry *ṣaḥib-a*. Two related entries concern the respective *maṣdars*, *ṣaḥābah* and *ṣuḥbah*. The third one concerns the adjective *ṣāḥib*, accompanied by the broken plurals that characterize it.[14]



```
<re type="related" ana="#R1aR2aAR3ah">
  <form type="lemma">
    <orth>صَحَابَة</orth>
    <gramGrp>
      <pos>noun</pos>
      <subc>masdar</subc>
    </gramGrp>
  </form>
</re>
<re type="related" ana="#R1uR2oR3ah">
  <form type="lemma">
    <orth>صُحْبَة</orth>
    <gramGrp>
      <pos>noun</pos>
      <subc>masdar</subc>
    </gramGrp>
  </form>
</re>
<re type="related" ana="#R1aAR2iR3">
  <form type="lemma">
    <orth>صَاحِب</orth>
    <gramGrp>
      <pos>adj</pos>
    </gramGrp>
    <form>
      <gram type="number">broken plural</gram>
      <orth n="1">أصحاب</orth>
      <orth n="2">أصاحِيب</orth>
      <orth n="3"> صُحْبان</orth>
      <orth n="4">صِحاب</orth>
      <orth n="5">صِحابَة</orth>
      <orth n="6">صَحْب</orth>
    </form>
  </form>
</re>
```

Fig. 11: XML-TEI: Encoding of related entries

B. Representation of the morphological level

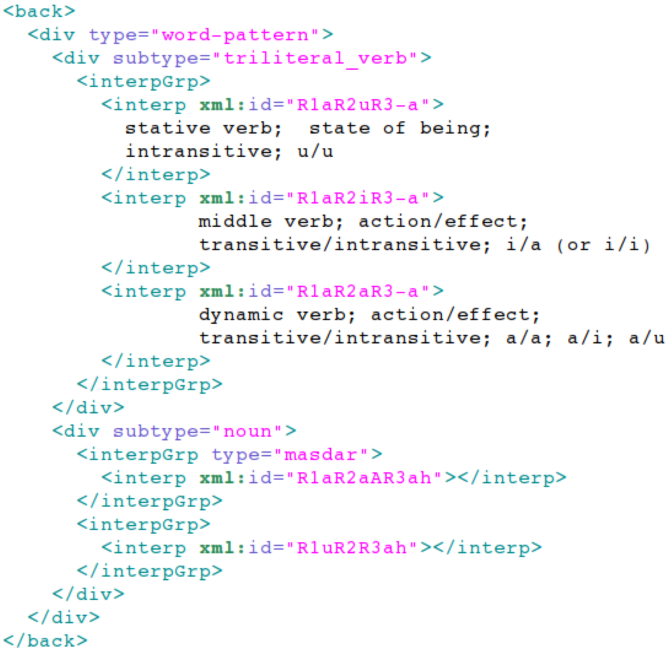
1) General characteristics of TEI: The morphological level concerns the description of word patterns typical of Arabic words. Consequently, the same TEI document, which represents the lexical resource, defines an additional XML block led by the `<back>` element encompassing nested `<div>` tags (see listing 1).

The `@type` attribute of the outermost `<div>` indicates that this part of the resource deals with the data describing word patterns. In turn, the further two nested `<div>` elements define morphosyntactic classes. These classes are declared by means of the `@subtype` attribute, grouping the verb class for the first `div` and the noun class for the second `div`.

```
<back>
   <div type="word-pattern">
   <div subtype="verb_pattern">
     [...]
   </div>
   <div subtype="noun_pattern">
          [...]
   </div>
   </div>
   </back>
```

Listing 1: The main TEI-XML block for encoding word patterns

As far as the representation of word patterns is concerned, during the first phase of our work we adopted the `<interp>` element to encode their characterizing semantic, morphosyntactic and inflectional information (see figure 12) [27].



```
<back>
  <div type="word-pattern">
    <div subtype="triliteral_verb">
      <interpGrp>
        <interp xml:id="R1aR2uR3-a">
          stative verb;  state of being;
          intransitive; u/u
        </interp>
        <interp xml:id="R1aR2iR3-a">
                middle verb; action/effect;
                transitive/intransitive; i/a (or i/i)
        </interp>
        <interp xml:id="R1aR2aR3-a">
                dynamic verb; action/effect;
                transitive/intransitive; a/a; a/i; a/u
        </interp>
      </interpGrp>
    </div>
    <div subtype="noun">
      <interpGrp type="masdar">
        <interp xml:id="R1aR2aAR3ah"></interp>
      </interpGrp>
      <interpGrp>
        <interp xml:id="R1uR2R3ah"></interp>
      </interpGrp>
    </div>
  </div>
</back>
```

Fig. 12: XML-TEI: First attempt of word patterns encoding by using the `<interp> element`

Since the inherent structure of word patterns describes different properties that can also be organized into complex hierarchical substructures, we soon realized that the `<interp>` tagset was not sufficiently capable of expressing all the necessary features. Therefore, we decided to adopt the TEI Feature Structures tagset (module 18 of the TEI guidelines)[15], which allowed us to describe recursively both simple-structured properties and complex rich-structured values with multiple hierarchical levels.

The TEI guidelines provide encoders with a general but rigorous method to represent analytical and interpretative information. This flexible method - called "feature structures"

[14] During the encoding word patterns, we used the Buckwalter transliteration (a character-by-character transliteration), to represent all the word characters, for example: *sukūn*: (o); *ā*: (aA). For more details, see: http://www.qamus.org/transliteration.htm [Retrieved in October 2020].

[15] https://tei-c.org/release/doc/tei-p5-doc/en/html/FS.html

- has been defined by a specific module within the TEI infrastructure (Module 18). Basically, each piece of information can be represented by a set of properties, each one characterized by a name and a value. This special pair is the building block of the descriptive strategy behind the feature structures method. The fundamental representation of analytical data is constituted by the `<fs>` element that contains zero or more `<f>` elements. The value can be as simple as illustrated in listing 2.

```
<fs type="example-simple">
 <f name="property">
  <string>
    value of the property
  </string>
 </f>
</fs>
```

Listing 2: Example of a simple TEI-XML fragment in the feature structure

Otherwise, a feature value can be as complex as a new nested feature structure, as in listings 3 and 4 which show some descriptive options in representing nested elements.

```
<fs type="example-complex">
 <f name="property">
   <fs type="nested">
    <f name="nested-property-alt">
     <vAlt>
      <symbol value="a"/>
      <symbol value="b"/>
     </vAlt>
   </f>
  </fs></f></fs>
```

Listing 3: Example of a TEI-XML fragment in the feature structure using the `vAlt` element

```
<fs type="example-complex">
 <f name="property">
   <fs type="nested">
    <f name="nested-property-coll">
     <vColl org="set">
      <symbol value="a"/>
      <symbol value="b"/>
     </vColl>
   </f>
  </fs></f></fs>
```

Listing 4: Example of a TEI-XML fragment in the feature structure using the `vColl` element

In listing 3, the XML-TEI fragment shows the use of the `vAlt` element that stands for *value alternation* and can be used as a nested element of a feature description. This type of encoding strategy is useful to define exclusive values for a single data representation.

Listing 4 illustrates the use of the `vColl` element which represents an inclusive group of values structured as a list of values, or as a bag of values, or finally, as a set of values.

In addition, the `vColl` and `vAlt` elements can operate jointly to represent an alternation of collections or a collection of alternations as demonstrated in the following example (in listing 5):

```
<fs type="example-complex">
 <f name="property">
   <fs type="nested">
    <f name="nested-property-alt-coll">
     <vAlt>
      <vColl org="list">
       <vAlt>
        <symbol value="a"/>
        <symbol value="b"/>
       </vAlt>
       <symbol value="k"/>
      </vColl>
      <vColl org="set">
       <symbol value="c"/>
       <symbol value="d"/>
      </vColl>
     </vAlt>
   </f>
  </fs></f></fs>
```

Listing 5: Example of combination of the vAlt and vColl elements

2) Encoding of verb patterns: At the current stage of the work, we have encoded the patterns related to the verbal words.

In Western grammars of Arabic, the verbal patterns are identified by a Roman numeral. The number I indicates that the verbal pattern is composed of only short vowels together with radical consonants. The other numbers indicate that the verb is derived.

When the verb is basic, it can be triliteral or quadriliteral when it derives from a triconsonantal or a quadriconsonantal root, respectively. In addition, the basic triliteral verbs consist of three variants, distinguished in the perfective by the quality of the second vowel, also called the theme vowel [15]. In this context, the word patterns are described by means of a hierarchical structure. The first level of the hierarchy groups the five main characteristics of the patterns:

- **vocalized form**: it represents the fully vocalized form;
- **morphological form**: it identifies the Roman number of the word pattern and the nature of the root, namely triconsonantal (T) or quadriconsonantal (Q);
- **syntax**: it expresses transitivity/intransitivity properties;
- **inflectional**: it indicates the imperfective verbal form;
- **semantic**: it determines the semantic traits presented in word patterns.

As an example, in listing 6 we illustrate the TEI-XML description of the encoding of the morphological, syntactical and inflectional characteristics of the R_1aR_2**u**R_3-a pattern:

- it is characterised by the vowel /u/ as theme vowel in the perfective;

- words with the R_1aR_2**u**R_3-a pattern are basic verbs (form I) that derive from a triconsonantal root (T);
- they are intransitive because the subject is just an entity located in a state;
- they also have the vowel /u/ in the imperfective: $pref_1$-R_1oR_2**u**R_3-$suff$.[16]

```
<f name="Vocalized_form">
 <symbol value="R1aR2uR3-a"/></f>
<f name="morphological_form">
 <symbol value="I_T"/></f>
<f name="syntax">
 <symbol value="intransitive"/></f>
<f name="flexion">
 <symbol value="pref1-R1oR2uR3-suff"/>
 </f>
```

Listing 6: morphological, syntactic and inflectional description of R_1aR_2**u**R_3-a

In listing 7, we present the semantic description of the R_1aR_2**u**R_3-a pattern. The value of the semantic property has a complex structure. It is described by a nested structure, `<fs type="sem-definition">`, which is composed by two properties:

- `<f name="type">`, which provides a semantic explanation of the pattern.
  As an example, the R_1aR_2**u**R_3-a pattern defines verbs which are pure-stative, i.e. they are used to describe entities in a simple property state.
- `<f name="description">`, made of a nested feature structure, provides a conceptual description together with an Arabic example and its English translation.
  In our example, the R_1aR_2**u**R_3-a pattern expresses an 'adjectival concept' [14], which has the form 'to be + adjective', as in the verb *karuma* "to be generous".

```
<f name="semantic">
 <fs type="sem-definition">
  <f name="type">
   <string xml:id="R1aR2uR3a_1">
   pure stative</string></f>
  <f name="description">
   <fs type="sem-description">
    <f name="conceptual_meaning">
      <string corresp="#R1aR2uR3a_1">
      to be + adjective
      </string></f>
    <f name="example">
      <fs type="R1aR2uR3a_1"
        corresp="#R1aR2uR3a_1">
       <f name="ar">
           <string
         xml:lang="ar">كَرُمَ</string></f>
          <f name="en">
              <string xml:lang="eng">
             to be generous</string></f>
           </fs>
          </f></fs></f></fs></f>
```

Listing 7: Semantic description of the R_1aR_2**u**R_3-a pattern

There are patterns that have more complex characteristics. By means of the "Feature Structure" we can definitely formulate very complex properties.

For example, the R_1aR_2aR_3-a pattern represents a structure where the subject of the verb has an initiator role. The R_1aR_2aR_3-a verbs are described within the `vColl` element as dynamic, active and action verbs (listing 8) [13].

```
<fs type="sem-definition">
<f name="type">
  <vColl>
     <symbol value="dynamic"/>
     <symbol value="active"/>
     <symbol value="action"/>
  </vColl>
 </f>
</fs>
```

Listing 8: Example of the `vColl` element in the feature structure for semantic description

The R_1aAR_2aR_3-a pattern encodes two relations, and the subject of the verb represents the Initiator of one of these relations, and the Endpoint of the other.

Different types of dual verbs that are created using R_1aAR_2aR_3-a can be shared events, interaction verbs, transaction verbs, competition verbs, verbs of opposition, verbs of cooperation, stimulus-response verbs, and then verbs formed from roots which lexicalize symmetrical concepts (terms in [13], pages 143-163).

These semantic properties are formalized within the `<f name="type">` element by using the `vAlt` element (value alternation). The `<vAlt>` element represents the value part of a feature-value specification that contains a set of values, only one of which can be valid.[17]

Each property has an identifier `xml:id` that links it to corresponding conceptual meanings within the `<f name="conceptual_meaning">` element, and to the Arabic examples and English translation in the `<f name="example">` element.

In listing 9, we present some semantic properties of the R_1aAR_2aR_3-a pattern:

- Verbs that express *"shared event"* with `id=III_T_1` and have the conceptual meaning of *"to do an action with someone"*, for example, the verb *sākana* 'to live with'.
- Verbs that express *"contact verbs"* with `id=III_T_2` and have the conceptual meaning of *"to bring about contact between the subject and the object"*, for example, the verb *ṣāfaḥa* 'to shake the hand of'.

[16] The imperfective prefixes consist of one of the particles, ʾ, *t*, *y* or *n*, and a vowel that is /u/ ($pref_2$) in case the verb consists of four consonants, otherwise it is /a/ ($pref_1$). On the other hand, suffixes ($suff$) are the same for all verbs.

[17] www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-vAlt.html.

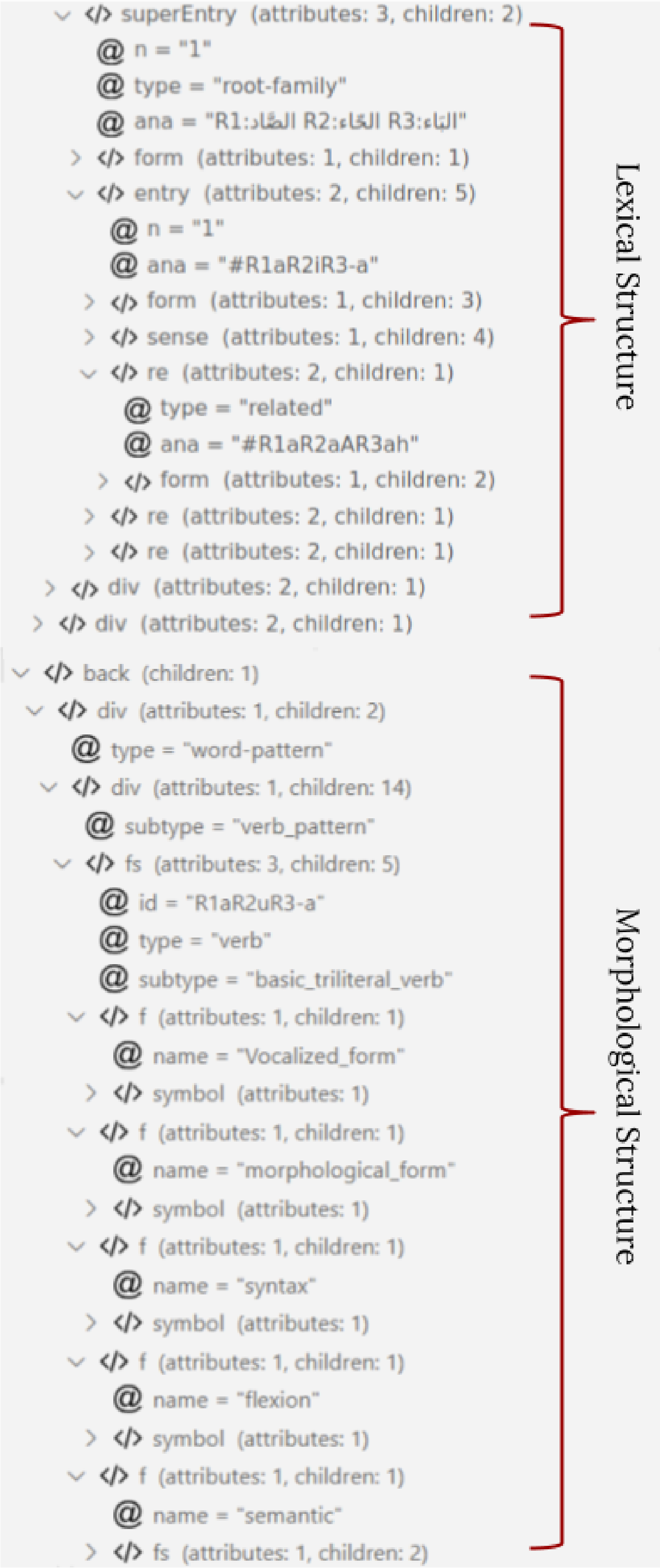
```
  <f name="semantic">
   <fs type="sem-definition">
    <f name="type">
     <vAlt>
      <string xml:id="III_T_1">
      shared event verbs</string>
      <string xml:id="III_T_2">
      contact verbs</string>
    ----
  </vAlt>
  <f name="description">
   <fs type="sem-description">
    <f name="conceptual_meaning">
     <vAlt>
      <string corresp="#III_T_1">
      to do an action with someone
      </string>
      <string corresp="#III_T_2">
      to bring about contact between
      the subject and the object
      </string>
     ---
    </vAlt>
 </f>
 <f name="example">
  <fs type="III_T_1" corresp="#III_T_1">
   <f name="ar">
    <string>سَاكَنَ</string></f>
   <f name="en">
    <string>to live with</string></f>
 </fs>
 <fs type="III_T_2" corresp="#III_T_2">
  <f name="ar">
    <string>صَافَحَ<string></f>
  <f name="en">
    <string>to shake the hand of</string>
  </f>
    ---
 </fs>
 </f></fs></f>
```

Listing 9: Example of some semantic properties of the R_1aAR_2aR_3-a pattern

V. DISCUSSION

Figure 13 presents the two main structural hierarchies of our Arabic resource. The upper part of the model shows the lexical structure, while the lower part shows the morphological structure of the resource.

On the one hand, the lexical resource, extracted from the *qāmūs* dictionary, is structured in super-entries. Each super-entry groups the lemmas deriving from the same root. Each lemma is accompanied by different (syntactic and morphological) types of information, POS and Arabic senses extracted from *qāmūs*. The description of the lemma is also enriched by English translations extracted from the bilingual dictionary. The entry can have nested <re> elements encoding the relative lemmas, such as adjectives and *maṣdars*. These are considered 'related entries' embodied in the main verbal entry.



```
    ⌄ </> superEntry  (attributes: 3, children: 2)
        @ n = "1"
        @ type = "root-family"
        @ ana = "R1:الضاد R2:الكاء R3:الياء"
      › </> form  (attributes: 1, children: 1)
      ⌄ </> entry  (attributes: 2, children: 5)
          @ n = "1"
          @ ana = "#R1aR2iR3-a"
        › </> form  (attributes: 1, children: 3)
        › </> sense  (attributes: 1, children: 4)
        ⌄ </> re  (attributes: 2, children: 1)
            @ type = "related"
            @ ana = "#R1aR2aAR3ah"
          › </> form  (attributes: 1, children: 2)
        › </> re  (attributes: 2, children: 1)
        › </> re  (attributes: 2, children: 1)
   › </> div  (attributes: 2, children: 1)
  › </> div  (attributes: 2, children: 1)

⌄ </> back  (children: 1)
  ⌄ </> div  (attributes: 1, children: 2)
      @ type = "word-pattern"
    ⌄ </> div  (attributes: 1, children: 14)
        @ subtype = "verb_pattern"
      ⌄ </> fs  (attributes: 3, children: 5)
          @ id = "R1aR2uR3-a"
          @ type = "verb"
          @ subtype = "basic_triliteral_verb"
        ⌄ </> f  (attributes: 1, children: 1)
            @ name = "Vocalized_form"
          › </> symbol  (attributes: 1)
        ⌄ </> f  (attributes: 1, children: 1)
            @ name = "morphological_form"
          › </> symbol  (attributes: 1)
        ⌄ </> f  (attributes: 1, children: 1)
            @ name = "syntax"
          › </> symbol  (attributes: 1)
        ⌄ </> f  (attributes: 1, children: 1)
            @ name = "flexion"
          › </> symbol  (attributes: 1)
        ⌄ </> f  (attributes: 1, children: 1)
            @ name = "semantic"
          › </> fs  (attributes: 1, children: 2)
```

Fig. 13: The XML-TEI hierarchy of the encoded resource

On the other hand, the morphological block describes the semantic, morphological and syntactic traits characterizing the word patterns. The structure of this block is composed of different TEI-XML elements taken from the feature structures module. The representation of the word patterns is located within the `<back>` element of the TEI document. Therefore, two `<div>` elements group the verb and the noun patterns. The entry point element of the morphological description is the `<fs>` element, which is accompanied by three attributes:

- `@xml:id`, used to link the lexical entry with the corresponding word pattern;
- `@type`, used to classify the pattern category;
- `@subtype`, used to further characterize the pattern subgroup.

The hierarchy deepens into the description of the properties by using five feature elements (each of which expressed by the `<f>` tag and the `@name` attribute):

- `<f name="vocalized_form">`
- `<f name="morphological_form">`
- `<f name="syntax">`
- `<f name="flexion">`
- `<f name="semantic">`

The value of the first four features is expressed by the `<symbol>` element, while the last feature is a complex structure, expressed by nesting `<fs>` elements.

Word pattern formalization will be exploited to build a morphological resource providing meta-linguistic information that allows the classification of lemmas into classes with the same systematic distribution of inflectional properties. The most exploitable inflectional regularities are found in word patterns that provide enough information to infer the entire inflectional paradigm of a verbal entry (see [28], [29] and [30]). Consequently, the addition of the formalized pattern plays a distinctive role in the lexical entries, since all other inflected forms of the same verb can be derived on-the-fly from the abstract pattern.

The most important part consists in specifying which semantic aspects of the pattern can characterize a lemma. In this case, the lexicographer's work is manual and quite complex. However, the final and enriched framework is intended to provide a more theoretically-grounded lexical architecture for the representation of discontinuous morphologies. The present study represents a theoretically-sound but also practical approach to a rigorous, formal representation of a complex root-and-pattern system, where the obtained resource will be of great value for systematic studies since it allows to correlate lexicological, derivational and semantic data.

Furthermore, the TEI-based architecture is intended to spell out the wealth of information provided by the dictionary and provided by the patterns. It is important for us to guarantee the maximum degree of reusability and compatibility between different lexical resources and processing tools. In this scenario, the current representation of our lexico-morphological resource is considered just one of the possible serialization modalities. Indeed, the abstract model we defined for the lexical entries and the morphological descriptions can be instantiated by means of different standards as long as these standards are mutually isomorphic and sufficiently expressive. For example, the current TEI representation can be profitably mapped into the TEI Lex-0 schema or into the LMF ISO format.

VI. CONCLUSIONS

In this paper we present the work we are conducting for the creation of a digital Arabic resource representing both lexical data and morphological description of word patterns. The main objective of our research is to define the theory and practice we adopted to formalize lexical information of Arabic lemmas and to encode an exhaustive description of word patterns.

The construction of the morphological resource constitutes the originality of this work, which has the aim of filling the gap in the field of linguistic resources, due to the current lack of such tools. Our morphological resource is built from scratch and the information is taken from classical manuals. It is an innovative, autonomous and self-contained resource that can be exploited for many other linguistic tasks. For example, it can be adopted to automatically recognize the pattern of an unknown word as well as to reduce the linguistic fields (lexical, morphosyntactic and even semantic) to which a word may belong.

In order to build our richly-structured artifact which takes into account primary source, lexical requirements, and reusability, we chose to adopt two main modules of the TEI guidelines: Module 9 and Module 18. However, we intend to consider also other data representation formats and serialization methods based on both LMF and TEI Lex-0 models. Consequently, as further future works, we intend to deepen the morphological representation of the Arabic words by adopting lexicographical standards other than TEI.

Finally, the next steps will be devoted to the actual encoding of the lexicon discussed in this article. Our plan is to release an improved version of the *qāmūs* lexical resource in TEI format and of the morphological resource in which all the verbal patterns are encoded.

Linking each lexical entry with the corresponding pattern will take more time and will imply a considerable amount of manual work.

REFERENCES

[1] La. Abouenour, K. Bouzoubaa, and P. Rosso. "On the evaluation and improvement of Arabic WordNet coverage and usability". Lang. Resour. Eval. 47, 3 (2013), 891–917. 2013.

[2] M. Alkhalifa and H. Rodríguez. "Automatically extending NE coverage of Arabic WordNet using Wikipedia". In Proceedings of the 3rd International Conference on Arabic Language Processing (CITALA'09). 2009.

[3] R. Baalbaki. "The Arabic Lexicographical Tradition : From the2nd/8th to the 12th/18th Century". Brill, Leiden, Boston, 2014.

[4] G. Badaro, H.Hajj, and N. HABASH. "A Link Prediction Approach for Accurately Mapping a Large-scale Arabic Lexical Resource to English WordNet". ACM Trans. Asian Low-Resour. Lang. Inf. Process, Vol. 19, No. 6, Article 80. Publication date: October 2020.

[5] S. Ben Ismail, H. Maraoui, K. Haddar, and L. Romary. "ALIFeditor for generating arabic normalized lexicons". In proceed-ings of the 8th International Conference on Information andCommunication Systems (ICICS), pp. 70–75, Irbid, 2017.

[6] W. Black, S. Elkateb, H. Rodriguez, M. Alkhalifa, P. Vossen, A. Pease, and C. Fellbaum. "Introducing the Arabic WordNet project". In Proceedings of the 3rd International WordNet Conference. Citeseer, 295–300. 2006.

[7] S. Boudelaa and W. D. Marslen-Wilson. “Discontinuous mor-phology in time: Incremental masked priming in arabic”. Language, Cognition and Neuroscience, vol. 20(1–2), pp. 207–260,2005.

[8] S. Boudelaa and W. D. Marslen-Wilson. “Productivity andpriming: Morphemic decomposition in arabic”. Language, Cog-nition and Neuroscience, vol. 26(4–6), pp. 624–652, 2011.

[9] S. Boudelaa. “Is the Arabic Mental Lexicon Morpheme-Based or Stem-Based? Implications for Spoken and Written Word Recognition”. Dordrecht: Springer, 2014. Pp. 31–54

[10] S. Boudelaa and W. D. Marslen-Wilson. “Structure, form, andmeaning in the mental lexicon: evidence from arabic”. Language,Cognition and Neuroscience, vol. 30(8), pp. 955–992, 2015.

[11] B. Comrie, M. Haspelmath, and B. Bickel. “The leipzig glossing rules: Conventions for interlinear morpheme-by-morphemeglosses”. Max Planck Institute for Evolutionary Anthropology, 2008.

[12] A. M. Del Grosso, A. Bellandi, E. Giovannetti, S. Marchi, and O. Nahli, “Scanning is Just the Beginning: Exploiting Text and Language Technologies to Enhance the Value of Historical Manuscripts”. In 2018 IEEE 5th International Congress on Information Science and Technology (CiSt), pp. 214–219, IEEE, 2018.

[13] P. J. Glanville. “The Arabic verb root and stem and their contribution to verb meaning”. Doctoral dissertation, The University of TexasatAustin, 2011.

[14] P. J. Granville. “The Lexical Semantics of the Arabic Verb”. Oxford University Press, Oxford, 2018.

[15] C. Holes. “Modern Arabic: Structures, Functions and Varieties”. Washington, DC:Georgetown University Press. 2004.

[16] M. Khalfi, O. Nahli, and A. Zarghili. “Classical dictionary al-qamus in lemon” in Proceeding of the 4th IEEE International Colloquium on Information Science and Technology (CiSt), pp. 325–330, IEEE, 2016.

[17] M. Khalfi, A. Zarghili, and O. Nahli. “A new rich lexical resource for classical arabic”. COMPUSOFT: An International Journal of Advanced Computer Technology, 9(10), pp. 3863-3885, 2020.

[18] A. Khemakhem, B. Gargouri, A. Abdelwahed, and G. Fran-copoulo. “Modélisation des paradigmes de flexion des verbes arabes selon la norme lmf-iso,” (Toulouse), pp. 24613, 2007.

[19] R. Laurent, e T. Tasovac. “TEI Lex-0: A Target Format for TEI-Encoded Dictionaries and Lexical Resources.” In TEI Conference. Tokyo: ADHO, 2018.

[20] R. Laurent, M. Khemakhem, A. F. Khan, J. Bowers, N. Calzolari, M. George, M. Pet, and P. Banski. ”LMF Reloaded”. The 13th International Conference of the Asian Association for Lexicography, Istanbul, Turkey. ASIALEX 2019. https://doi.org/10.5281/zenodo.3606434.

[21] N. Loukil, K. Haddar, and A. Benhamadou. “A syntactic lexiconfor arabic verbs”. In proceedings of Seventh International Conference on Language Resources and Evaluation LREC, (Malta), pp. 269–272, 2010.

[22] J. J. McCarthy. “Formal problems in Semitic phonology and morphology”. Unpublished phd. dissertation, MIT, 1979.

[23] J. J. McCarthy, “A prosodic theory of non-concatenative morphology”. Linguistic Inquiry, vol. 12, pp. 373–418, 1981.

[24] J. J. McCarthy, “Prosodic templates, morphemic templates, and morphemic tiers”. In The structure of phonological representations, pp. 191–223. 1982.

[25] H. Maraoui and K. Haddar. “Automatisation de l’encodage des lexiques arabes en tei”. In proceedings of 2nd conference onCEC-TAL, 2015.

[26] O. Nahli, F. Frontini, M. Monachini, F. Khan, A. Zarghili, andM. khalfi. “Al qamus al muhit, a medieval arabic lexicon in lmf”. In proceedings of LREC 2016, ELRA, 2016.

[27] O. Nahli and A. M. del Grosso. “Creating Arabic Lexical Resources in TEI: A Schema for Discontinuous Morphology Encoding”. The 6th IEEE Congress on Information Science and Technology (CiSt), pp. 178-187, 2020.

[28] A. A. Neme. “A lexicon of Arabic verbs constructed on the basis of Semitic taxonomy and using finite-state transducers”. In Proceedings of the International Workshop on Lexical Resources WoLeR, ESSLLI International Workshop on Lexical Resources, Ljubliana, Slovenia. 2011.

[29] A. A. Neme, É. Laporte. “Pattern-and-root inflectional morphology: the Arabic broken plural”. Language Sciences, Vol 40, November 2013, Pages 221-251. 2013.

[30] A. A. Neme. “An arabic language resource for computational morphology based on the semitic model”. Computation and Language [cs.CL]. Université Paris-Est, 2020. English. ⟨NNT : 2020PESC2013⟩.

[31] Salmoné, H. A. “An Advanced Learner’s Arabic-English Dictionary”. Librairie du Liban, Beirut. 1889.

[32] TEI Consortium,TEI P5: Guidelines for Electronic Text En-coding and Interchange. [version 3.6.0], 2019.

[33] H. Wehr. “A Dictionary of Modern Written Arabic”. ISBN 0-87950-001-8. 3rd edition ed., Spoken Language Service, Inc. 1976.

[34] Wright, W. “A Grammar of the Arabic Language”. Cambridge University Press, London. 1896

[35] A. Yahya. “Encoding Semantic Relations Between the Predicate and Participants Through Arabic Verbal Morphology: A Systems Interaction Approach”. Phd Thesis. University of Colorado at Boulder, ProQuest Dissertations Publishing, 2019. 13898002.