Available online at: https://ijact.in

Date of Submission	14/07/2020
Date of Acceptance	10/10/2020
Date of Publication	31/10/2020
Page numbers	3863-3885 (23 Pages)

ISSN:2320-0790

COMPUSOFT
An International Journal of Advanced Computer Technology

A NEW RICH LEXICAL RESOURCE FOR CLASSICAL ARABIC

Mustapha Khalfi[1], ArsalaneZarghili[2], Ouafaenahli[3]

[1,2]Intelligent Systems and Applications Laboratory, Faculty of Science and Technology, Sidi Mohamed Ben Abdellah University, Fez, 30000, Morocco.

[3]Istituto di Linguistica Computazionale ``A. Zampolli'', CNR, Pisa, 56124, Italy

[1]mustapha.khalfi@usmba.ac.ma, [2]arsalane.zarghili@usmba.ac.ma, [3]ouafae.nahli@ilc.cnr.it

Abstract: Currently, large lexical resources are getting a high potential relevance for information systems and need of Lexical resources in Natural Language Processing (NLP) fields is paramount. To contribute meet these needs, we build a lexical resource from the famous dictionary al=qāmūs al=muḥīṭ (AQAM). Using a rule based approach, we have designed a system that allows extracting morpho-syntactical, semantics and lexical information from the famous dictionary. So, we obtained a digitized and structured version of AQAM, enriched by morpho-syntactical and lexical explicit information. In addition, the obtained resource is enriched by English translations of lemma and accompanying senses using a bilingual English-Arabic dictionary. Then we present an overview of an experiment alignment of the section of the letter bā' on Princeton's WordNet (PWN) and Suggested Upper Merged Ontology (SUMO). This experience turned out to be interesting because it revealed that mapping an Arabic lexical resource on an English resource shows commonality between the two languages, but it allows especially to emphasize the non-equivalences between them. All obtained resources are represented in XML format and distributed under free license.

Keywords:Information Extraction; Arabic Lexicon; Al Qamus Al Muhit; Machine-readable dictionary; Arabic Lexical Resource;

I. INTRODUCTION

Currently, large lexical resources are getting a high potential relevance for information systems because they are essential in most NLP fields, such as information extraction, information retrieval, text summarization, classification and machine translation. However, Arabic NLP suffers from lack of digital linguistic resources [2]. In addition, Arabic language is highly inflectional and derivational, with a richmorphology and complex syntax as well [3] which puts Arabic NLP in front of big challenge and issue.

Fortunately, different efforts are made for face up to these task challenging and recently some works were developed in Arabic resources field.Arabic WordNet (AWN) is the Arabic lexical database constructed based on the English WordNet. AWN is constructed on lists of suggested Arabic translations for different words contained in English synsets[4]; [5]. Several attempts to enrich and extend AWN content has been made [6]; [7]; [8]; [9]. However, gap between AWN and Arabic language remained one of limitations for its use, because it doesn't take some particularities of Arabic language and it doesn't cover some specific concepts of Arabic culture[9].Alkhatib et al. propose, in [13], anapproach to develop a WordNet based on *al=hadīt*[1] by building semantic connections between words in order to provide a better understanding of meanings of *al=hadīt* words. Boudelaa and Marslen-Wilson presented in [10] a lexical database for Modern Standard Arabic named Aralex and based on a contemporary text corpus of 40 million words. The authors in [11] propose a lexical database based on aModern Standard Arabic corpus constituted of 1,089,111,204 words.Namly and Bouzoubaa presented in [12] the LMF conversion of the Contemporary Arabic language dictionary *mu'ğam al=luḡah al=mu'āsirah*[2].

[1] It refers a record of the words, the actions, and the silent approval of the prophet Mohammed

[2]http://arabic.emi.ac.ma/alelm/?q=Resources (the last accessed date: 14/09/2020).

However, we note that computational lexical resources concern Modern Standard Arabic and there is shortage for resources dedicated to Classical Arabic.So to made for face up to these lack of machine-readable resources for Classical Arabic, our work is aimed to make a new rich lexical resource based on a medieval Classical Arabic dictionary, القَامُوس المُحِيط*al=qāmūs al=muḥīṭ* (henceforth AQAM). The obtained resource could be used in NLP projects, more precisely in Classical Arabic heritage treatment and it can be considered as a starting point for diachronic and historical study of Arabic language. So, the digital and enriched version which is released open source will help Classical Arabic studies to gain on several fronts (lexicography, semantics, philology, etc.).

After this introductory section, the next section deals with digital conversion phases of AQAM. Particular and detailed preparation of the AQAM plain text allows us to identify its macrostructure and microstructure and stylistic particularity. Consequently, we are able to identify numerous regular expressions and morphological and stylistic rules that allow extraction of lexical, morphological and syntactical information and permit that AQAM digital conversion is carried out automatically. Afterwards we discuss some examples of lexical entries and some obtained statistical results.To enrich our resource, we used a bilingual dictionary to get the English translation necessary to map the digital AQAM to existing ontological resources, i.e. PWN (the Princeton WordNet) and SUMO (The Suggested Upper Merged Ontology). The thirdsection describes the followed methodology and explain the algorithm used the mapping process. This phase has been applied to the chapter of the letter *bā’*. The dictionary is full of proper names, so we had to face the task of identifying an automatic methodology which makes it possible to identify and extract them. Proper names extraction task is carried out based on a defined tags list consisting of keywords used by AQAM’s author.Finally, a conclusion ends the article and discusses future perspectives for continuing our research which has already produced.

II. AQAM’s Digital Conversion

A. Data preparation and segmentation phases

In order that recognition of lexical entries and components occurs automatically and correctly, it is necessary to study macrostructure and microstructure and to know of abbreviated conventions and dictionary layout.As Ide and Véronis elucidate in [15], dictionaries in general exhibit a strong duality between their surface structure (the text) and their deep structure (the information content), and much of the deep structure information is not explicit in the surface structure. In this precise study, we also had to face the fact that medieval Arabic dictionary structure is very variable and different from that modern.

1) Choice of the lexicon al=qāmūs al=muḥīṭ:We chose AQAMfor its authoritative status and its comprehensiveness in terms of number of entries within it. In fact, AQAM is one of the most used Arabic dictionaries to date, especially for study of Classical Arabic manuscripts, poetry and literature. In addition, it is composed of the collection of different dictionaries previous, but the author,*al=fīrūz’ābādī*(1329-1414), reduced original content from source dictionaries, by eliminating examples, Quranic quotations, poetry etc. Consequently, AQAM is well structured and contains short lexical items and may be considered an excellent candidate for conversion into a computational lexical resource and the succinct style used by the author facilitates information extraction[16].

Arabic lexicography schools are especially distinct according to order of presenting lexical entries (alphabetical or phonetical order - order by roots or by lexemes). But, generally, Arabic language structure does not allow classifying lexical entries in alphabetical order and words are brought back to root from which they are derived.AQAM belongs to lexicographical tradition ordering lexical entries according to root from which they derive and following the rhyme system. In this system, lexical elements are classified in alphabetical order on the basis of the radicals of their roots, starting with the last followed by the first, then intermediate radicals, i.e. the last radical is considered in first, followed by the first radical and the other radicals are arranged in front. Hence, the lexical orderis(3 + 1 + 2) in triliterals and (4 + 1 + 2 + 3) in quadriliterals.[3]So, AQAM is divided in sections (أَبْوَاب). Each section (بَاب) is devoted to a consonant constituting the last radical and it is divided into chapters (فُصُول) ordered according to the first radical consonant. Each chapter(فَصْل) is also divided into various parts gathering root family, i.e. all lexical entries that have same root. Within chapter, roots are listed alphabetically according to the second (and the third) radical consonant. Finally, lexical entries are grouped together under the root from which are derived.

2) Data Preparation and phases of text segmentation:In order to allow text[4] segmentation, process begins with preparation phase by identifying and marking essential parts that characterize macrostructure. As shown in the AQAM fragment illustrated in Figure.1,we prepared the plain text and added several indicators and signs[5].

- The symbols ***1*** and ***2*** indicate respectively section and chapter. Therefore, they explicit the values of the third (C3) and the first radical consonant (C1), respectively. For example, in Fig. 1, the symbol *1* indicates the section of the letter *bā’*, consequently the third radical is *bā’*. By cons, the symbol *2* indicating the chapter *dāl* gives also the first radical value that is *dāl*.
- The symbol **@** indicates change of the second radical. Therefore, root family is beginning by the symbol **@**. So, between two successive symbols **@**, we find lexical items derived from the same root. In Figure. 1, it is highlighted the root family characterized by C1= *dāl,* C2=*‘ayn* and

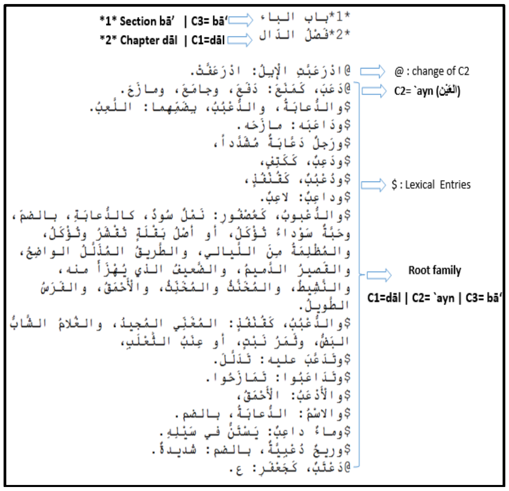
[3] see [30], pages 363 and 394.

[4] We used a free access copy of AQAM available at: https://ar.wikisource.org (the last accessed date: 16/09/2020). However, as reference, we used an edited copy [18].

[5]for more details, see [16] and [29]

C3=*bā’*.Note that the symbol **@**not allows the identification of C2 which is done in alphabetical order.

• The symbol **@** indicates also the first lexical entry and successive lexical entries begin by the symbol **$**. So lexical entry is identified by **@** or **$** and finish by the character "carriage return".


1 Section bā' | C3= bā'
2 Chapter dāl | C1=dāl
@ : change of C2
C2= `ayn (الغَيْن)
$: Lexical Entries
Root family
C1=dāl | C2= `ayn | C3= bā'

*1*باب الباء،
*2*فصْل الدّال

@اذرَعَبَّتِ الإبِلُ: اذرَعَفَّتْ.
@دَعَبَ، كَمَنَعَ: دَفَعَ، وجامَعَ، ومازَحَ.
$والدُّعابَةُ، والدُّعْبُبُ، بِضَمِّهِما: اللَّعِبُ.
$وداعَبَه: مازَحَه.
$ورَجلٌ دَعّابَةٌ مُشَدَّداً،
$ودَعِبٌ، كَكَتِفٍ،
$ودُعْبُبٌ، كَقُنْفُذٍ،
$وداعِبٌ: لاعِبٌ.
$والدُّعْبوبُ، كَعُصْفورٍ: نَمْلٌ سُودٌ، كالدُّعابَةِ، بالضمِّ،
وحَبَّةٌ سَوْداءُ تُؤْكَلُ، أو أصْلُ بَقْلَةٍ تُقْشَرُ وتُؤْكَلُ،
والمُظْلِمَةُ مِنَ اللَّيالي، والطَّريقُ المُذَلَّلُ الواضِحُ،
والقَصيرُ الدَّميمُ، والضَّعيفُ الذي يُهْزَأُ منه،
والنَّشيطُ، والمُخَنَّثُ والمُخَنِّثُ، والأَحْمَقُ، والفَرَسُ
الطَّويلُ.
$والدُّعْبُبُ، كَقُنْفُذٍ: المُغَنِّي المُجيدُ، والغُلامُ الشّابُّ
البَضُّ، وثَمَرُ نَبْتٍ، أو عِنَبُ الثَّعْلَبِ،
$وتَدَعَّبَ عليه: تَدَلَّلَ.
$وتَداعَبُوا: تَمازَحُوا.
$والأَدْعَبُ: الأَحْمَقُ،
$والاسْمُ: الدُّعابَةُ، بالضم.
$وماءٌ داعِبٌ: يَسْتَنُّ في سَيْلِهِ.
$وريحٌ دُعْبِيَّةٌ، بالضم: شَديدَةٌ.
@دَعْثَبٌ، كَجَعْفَرٍ: ع.

Figure 1: Macro-structure of AQAM

3)Phases and results of the AQAM text segmentation:Text segmentation is done following steps summarized in Figure 2. AQAM is divided into sections whose title allows us to recognize the last radical consonant value. Each section is divided into chapters whose title allows us to recognize the value of the first radical consonant.Inside each chapter, beginning of each root family allows us to identify the second (and the third eventually) radical consonant.Finally, we can identify roots value and lexical items lists that derive from them.Table 1 illustrates the content of *al=qāmūs al=muḥīṭ*.

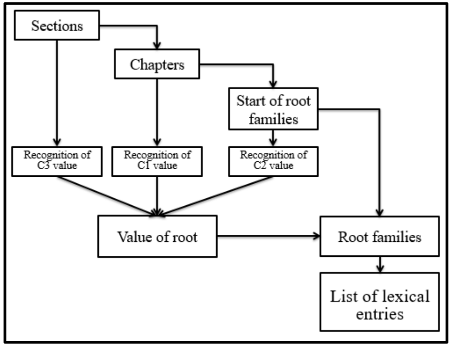


Figure 2: Text segmentation according to macrostructure.

Section (bāb)	chapter (faṣl)	Roots	Lexical entries
hamza	25	272	1499
bā’	28	650	5065
tā’	28	239	987
ṯā’	25	203	894
ǧim	28	461	1921
ḥā’	23	302	2191
ḫā’	26	209	883
dāl	27	458	3544
ḏāl	24	126	490
rā’	28	858	7942
zāy	24	284	1450
sīn	24	522	2564
šīn	25	284	1276
ṣād	21	224	1289
ḍād	21	114	1056
ṭā’	27	345	1902
ẓā’	18	88	337
`ayn	26	426	4025
ġayn	23	137	668
fā’	27	436	3746
qāf	26	468	3617
kāf	26	252	1501
lām	28	833	6400
mīm	28	740	5586
nūn	28	551	3619
hā’	25	154	784
wāw&yā’ *wāw*	28	374	2877
yā’	28	331	2778
’alif	-	32	-
Total			
29	687	10373	70992

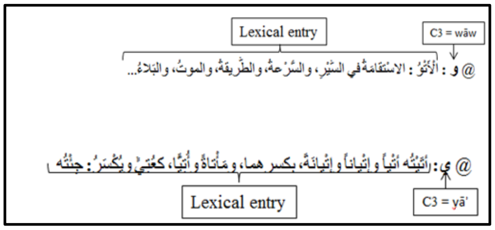
Table I: Some AQAM statistics

We have obtained 29 files in TXTformat that are released as open sources by means of the *CLARIN-IT infrastructure*[6], including:

• 26 sections corresponding to consonants from *hamza* to *hā’*;

• a section for *al=’alif al=layyinah*, i.e. the *’alif* that marks the lengthening the vowel /*a*/ to /*ā*/. This section contains only 32 entries of particles, for example, *’ilā* that can be preposition (to, toward; up to, as far as; till) or conjunction (until);

• a section is dedicated to the two consonants *wāw* and *yā’*. The section title being “*section of wāw and yā’*”, it is not possible to identify the C3 value. But, the author specified C3 at root family beginning,as shown in Figure. 3. So, we have divided the section into two sub-sections, one for *wāw* and the other for *yā’*.


Lexical entry
C3 = wāw
Lexical entry
C3 = yā'

@ و : الأَتْوُ : الاسْتِقامَةُ في السَّيْرِ، والسُّرْعَةُ، والطَّريقَةُ، والمَوْتُ، والبَلاءُ...

@ ي : أَتَيْتُه أَتْياً وإتْياناً وإتْيانَةً، بكسرِهما، ومَأْتاةً وأُتِيّاً، كعُتِيٍّ ويُكْسَرُ: جِئْتُه

Figure 3: Root families beginning with C3=wāw or C3=yā’

B. *Phases of Information extraction*

1) Characteristics of lexical entries microstructure:Example 1 illustrates a verbal lexical entry starting with the orthographic form وَآبَهُ *wa=’āb-a=hu* which

[6] http://hdl.handle.net/20.500.11752/ILC-97 (the last accessed date: 16/09/2020).

is composed by: i) the conjunction *wa*= "and" as proclitic; ii) the verb *’āb-a* composed by the verbal stem *’āb* separated by hyphens from the inflectional suffix *-a* indicating the third person singular and iii) the accusative pronoun *=hu* as enclitic[7].According to Classical Arabic lexical conventions, it is important to take into consideration the accusative pronoun presence because it indicates that the verb is transitive. After the colon, the orthographic form أَبْعَدَهُ*’ab`ad-a=hu*(composed by the verb *’ab`ad-a* and the accusative pronoun *=hu*) corresponds to the sense of the verb *’āb-a=hu*: "He sent it away".

Example 1

وآبَهُ اللَّهُ: أَبعَدَهُ

wa=’āb-a=hu allah: ’ab`ad-a=hu

Example 2 illustrates a nominal lexical entry and begins with the orthographic form وَالأَبَابُ*wa=al=’abāb-u*, composed by the conjunction *wa*= "*and*", the defined article *al*=; the noun *’abāb* and the nominative definite suffix *-u*. In this case, it is important to take into consideration the presence of the definite article *al*= because it denotes that the word to be defined is a noun. After the colon, we find the means *al=mā’-u "Water"* and *wa=s=sarāb-u "mirage"* separated by commas.

Example 2.

وَالأَبَابُ: المَاءُ، وَالسَّرَابُ

wa=al=’abāb-u: al=mā’-u, wa=s=srāb-u

As illustrated in examples 1 and 2, lexical entry generally consists of two parts. The part located before the colon begins with the word to define and composed by an inflected verbal form or a composed nominal form. In this article, orthographic form that begins the entry and from which it is possible to extract the lemma is called Headword. The part located after the colon is constituted by definitions separated by commas.

In addition,we note that almost other lexical entries start with the conjunctionوَ*wa "and"*. At sentence beginning, the conjunction *wa* is considered a filler word with no lexical content and therefore can be deleted without losing any information. Consequently, another phase of preparation consists in removing the conjunction وَ*wa "and"* which abounds at lexical entries beginning. The operation occurs after checking that the first radical consonant is not *wāw*. In this latter case, the removal took place after manual controls.

In many cases, lexical entries can be more complex. Figure4gives idea about richness and complexity of lexical entry in AQAM. The verbal headword, أَبَّ*’abb-a*, is followed by inflectional information concerning related imperfect forms of the lemma (*ya-’ibb-u* and *ya-’ubb-u*) and information concerning related derived words as masdars (verbal nouns) in accusative indefinite form (*’abb-an*, *’abīb-an*, *’abāb-an* and *’abābat-an*). Finally, located after the colon, the sense تَهَيَّأَ*tahayya’-a* is equivalent in meaning to the English predicate *"to prepare"*.So we had to recognize information different types that may exist in a lexical entry and find methods to extract it.

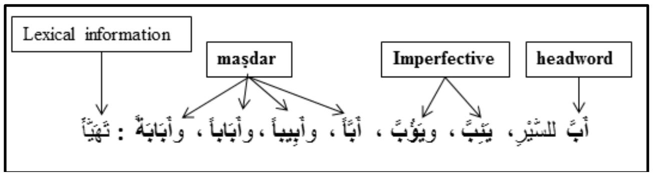

Lexical information
maṣdar
Imperfective
headword
أَبَّ للسَّيْرِ، يَئِبُّ ، ويَؤُبُّ ، أَبّاً ، وأَبِيباً ، وأَبَاباً ، وأَبَابَةً : تَهَيَّأَ

Figure 4: Example of microstructure complexity

2) Phases of Information extraction

As shown in Figure. 5, according to microstructure characteristics, we have identified different steps of information extraction:

- Extraction of the headword that must undergo some treatment to get the corresponding lemma in according to current lexical conventions.[8]
- Grammatical information groups morphosyntacticand semantic information about lexical unit: POS (Part Of Speech); syntactic information (for example transitivity/intransitivity of a verbal lemma); morphological information that indicate for example flexion category associated with a lexical unit, and derivational category (triliteral or quadriliteral or derived verb).
- Semantic information groups different senses and when it is possible to identify the relationship between lemma and sense, such as synonymy, antonymy.
- Extraction of associated derived words, like *maṣdar* when is cited with verbal lemma.

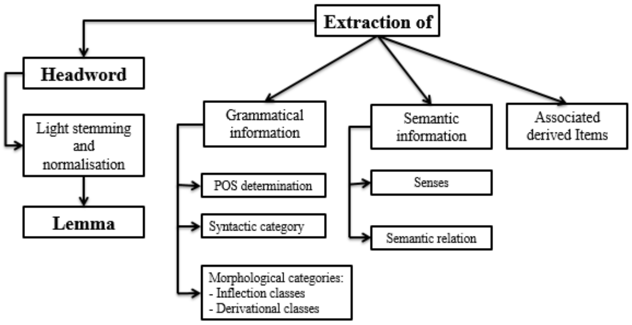


Figure 5:Information extraction according to AQAM microstructure

[7] For syntax and semantics of interlinear glosses, we use “The Leipzig Glossing Rules: Conventions for Interlinear Morpheme-by-morpheme Glosses” (Max Planck Institute for Evolutionary Anthropology, 2015. Available in: http://www.eva.mpg.de/lingua/resources/glossing-rules.php (the last accessed date: 21/09/2020)). Interlinear morpheme-by-morpheme glosses give information about the meanings and grammatical properties of individual words and parts of words. So, segmentable morphemes are separated by hyphens and clitic boundaries are marked by an equals sign. In fact, in Example 1, conjunction, verb and accusative pronoun are separated by an equals sign.

[8] Canonical verbal lemma in Arabic dictionary is the third person singular masculine of the past tense. By cons, canonical lemma of nouns is the singular and of adjectives is the singular masculine.

Results depend much on knowledge of AQAM deep structure, the style of its author and systematic control of punctuation inside lexical entries. In following sections, we will explain followed methodologies in each phase to obtain a complete digital version of AQAM.

C. *Extraction and normalization of lemma*

1) Cases of lexical entries with headword: Headword candidate selection is done based on root radicals, i.e. it must contain all radicals. To obtain the lemma, it is necessary to specify grammatical nature of the headword in order to determine which treatment it must undergo.The Algorithm 1 illustrates the adopted procedure. In a first phase, it is easier to define morphological patterns of verbs because their number is limited. Firstly, we defined Verbal Inflectional Suffixes list and Transitivity Suffixes list (see Table XV in the Appendix A). Secondly we defined a list regrouped 34 regular expressions which present possible inflected and derived verbal patterns (some examples are presented in the algorithm 1 core). The headword is a verbal form if it matches one of defined verbal rules. In case it doesn't correspond to any verbal pattern, we test if it starts with the definite article *al*=, either it ends with the *tanwīn-un*, or its final letter equals to ة which denotes that it is a nominal entry case. Finally, we have to undergo the headword to a normalization process in order to obtain the lemma according to the modern lexical convention.

Algorithm 1: Headword Surface Form Classification

input : Headword W, List NominativeSuffixes, List TransitivitySuffix, List Verbal_derived_Form, List Nominal_Form, List Adjectival_Form

output : R:Headword Surface Form

```
/* Examples of Verbal_derived_Form          */
/* fa`alaفَعـَل */
verbI_a  =C+a+"{0,1}?"+C+a+C+NominativeSuffixe
+ TransitivitySuffixe +"{0,1}?";
/* qaāla                          قَالَ*/
verbI_cAc=C + a +"{0,1}?"+ "ا" C +
NominativeSuffixe+TransitivitySuffixe+"{0,1}?";
/* danāدناا*/
verbI_A   =C + a +"{0,1}?"+ C + a + "{0,1}?"+
"ا";
/* mašaY                              مَشَى*/
verbI_Y   =C + a +"{0,1}?"+ C + a + "{0,1}?"+
"ى";
/* ʾabba                              أَبَّ*/
verbI_CC =C + a +"{0,1}?"+ C + C + a;
/* daḥraǧa                            دَحْرَجَ*/
verbe_Q  =C + a + C + o +C + a + C +
NominativeSuffixe+TransitivitySuffixe+"{0,1}?";
/*********************************************/
/*Examples of Nominal_Surface_Form*/
/* wardat وَرْدَة */
noun_un = Arabic_Word + "ة" + vowel + "{0,1}?";
/* al=kitāb-uالكِتَاب*/
noun_al  ="ال"  +  Arabic_Word  +  vowel  +
"{0,1\}?";
/* kitāb-u ar=rajul-i كِتَابُ الرَّجُلِ*/
noun_add  = Arabic_Word + u + Arabic_Word + i ;
/*********************************************/
/*Examples of Adjectival_Surface_For*/
/* hallāb-un                 هَلَّابٌ*/
adj_un  = Arabic_Word + un + Arabic_Word + un;
/* huwamu'tašib-un هُوَ مُؤْتَشِبٌ*/
adj_Pronoun=pronoun+Arabic_Word + un;
/*********************************************/
for v inVerbal_derived_Formdo
   IfW.Matches(v)  then
        R ← Verb;
end
for n inNominal_Surface_Formdo
   ifW.Matches(n)  then
        R ← Noun;
end
for ainAdjectival_Surface_Formdo
   ifW.Matches(a)then
        R ← ADJ;
end
return R;
```

Lemma normalization

In Example 3 the headword *taṣāḫab=ū* is constituted by the stem *taṣāḫab* and the nominative pronoun *ū* which indicates that it is verb. After the colon,we find the two senses *taṣāyaḥ=ū* and *wa=taḍārab=ū* with the meanings *"they shout at each other"* and *"they brawl or they strike one another"* respectively. By cons, Example 4 illustrates a case of nominal lexical entry.The headword *at=tawb-u* is accompanied by the sense*al=libās-u* with the meaning "*clothes*".

Example 3

تَصَاخَبُوا: تَصَايَحُوا، وَتَضَارَبُوا

taṣāḫab=ū : taṣāyaḥ=ū, wa=taḍārab=ū

Example 4.

الثَّوْبُ: اللِّبَاسُ

at=tawb-u: al=libās-u

To obtain canonical forms, the lemma and the senses must undergo some procedures of stemming.As shown in Table II, complete stemming process reduces words to their roots, by removing clitics, derivational and inflectional affixes. By cons, light stemming process removes only clitics and inflectional prefixes and suffixes.Therefore, the light stemming respects the semantic characteristics of words[17]. So for these reasons we have adopted the light stemming in order to process headwords and senses.

Word	Stemming	Light stemming
taṣāḫab=ū	*ṣḫb*	*taṣāḫab*
taṣāyaḥ=ū	*ṣyḥ*	*taṣāyaḥ*
wa=taḍārab=ū	*ḍrb*	*taḍārab*
aṯ=ṯawb-u	*ṯwb*	*ṯawb*
al=libās-u	*lbs*	*libās*
Result	Root	Stem

Table II: Stemming vs light stemming

The applied light stemming procedure depends on results of headword grammatical nature, returned by the algorithm 1. In fact, we use special treatment for each category and the process is done automatically following instructions of the algorithm 2 described below.

Algorithm 2: Headword Light-Stemming

```
input    :  Lemma X, POS P, List NominativeSuffixes,
            List TransitivitySuffixes
output   :  R:Light stemmed word
if X.getRoot().getC1() <> "و"then
X←eliminatePrefixe("و");
ifX.getPOS() == "Verb" then
fortsinTransitivitySuffixesdo
  IfX.endsWith(ts)then
       X←eliminateSuffixes(X,ts);
    end
fornsinNominativeSuffixesdo
  IfX.endsWith(ns)then
       X←eliminateSuffixes(X,ns);
    end
R ←X;
else ifX.getPOS() == "Noun"then
  ifX.endsWithVowel()then
  X←X-(Last_Vowel);
  ifX.startsWith("ال") then
  X←X-("ال");
  ifX.charAt(1)== šadda then
  X←eliminateGemination(X);
  R ←X;
else ifX.getPOS() == "ADJ"then
  ifX.endsWithVowel()then
   X←X-(Last_Vowel);
R ←X;
return R;
```

In case of verbs, we delete suffixes, using the two predefined lists containing nominative suffixes and possible accusative suffixes. As result, the presented algorithm 2 returns the light-stemmed headword. For instance, from *taṣāẖab=ū* we obtain *taṣāẖab* that requires one more step adding the vowel *-a* in order to get the verbal canonical form *taṣāẖab-a*.

In case of nouns, we remove the definite article and the final vowel. In addition, if the first radical is one of the sun letters, the light stemming algorithm permits to remove the gemination mark.[9] For example, from the headword التَّوْبُ*alt~awobu*[10]we obtain the light stemmed word ثَوْب*tawob*which already matches the canonical form without other procedures.

Note that when the headwords do not contain the radical consonants, entries cannot be selected by following this procedure. They are excluded and grouped in a separate file which allowed usto study them and identify other cases.

2) Case of word-mold

Sometimes, the author does not mention the headword, but he uses a well-known word that act as patterns to build the lemma. The lexical entry, in Example 5, belongs to the root family *ḥbb*, but the initial word isكَكِتَابٍ*ka=kitāb-in*composed by the conjunctionكَ*ka* "*like*" followed by the word *kitāb "book"* which can be used as a mold to produce the lemma. After identifying the root family *ḥbb*, we produce automatically the lemma by replacing each consonant by the corresponding radical. Consequently, the lemma constructed isحِبَاب*ḥibāb*, having the pattern *C1iC2āC3 "like kitāb*". In addition, the obtained lemma (*ḥibāb*) is classified as noun based on the final vowel of the word-mold, *tanwīn al=kasra (in),* that is one of case vowels of indefinite nouns.

Example 5.

كَكِتَابٍ: المُحَابَبَةُ

ka=kitāb-in: al=muḥābabat-u

By cons, in Example 6, the entry belongs to the root family *`šb*. But, the initial word starts with the conjunctionكَ*ka*followed by the verb *fariḥa"he was happy"*. In this case, the lemma constructed عَشِبَ*`ašiba*recognised as triliteral verb because has the pattern *C1aC2iC3a*.

Example 6.

كَفَرِحَ: يَبِسَ

ka=fariḥa: yabisa

3*) Cases of vocalization pattern*

In other cases, without mentioning the lemma, the author mentions vocalization patternto indicate its vocalization relying on the previous lemma.In Example 7, the first lexical entryhas the headword *al=ğabāb-u*.By cons, the second lexical entry starts with *bi=al=kasr-i "with the vowel \i\"*.Following the author's instructions in the introduction, to obtain the headword of the second lexical entry, we must change the vowel of the first consonant of the previous lemma (*al=ğabāb-u*) with the vowel \i\":*al=ğibāb-u*.

Example 7.

$الجَبَابُ: القَحْطُ الشّدِيدُ
$بِالكَسْرِ: المُغَالَبَةُ

The list of vocalization patterns is explained in the introduction of AQAM ([18], page 20) and summarized in Table IV. The change of vowel respect to the previous lemma only concerns the first radical consonant C1. In this case, the author begins the lexical entry with the preposition بِـ "*with*" attached to the name of the new vowel which may be *fatḥah* /*a*/, *ḍammah* /*u*/ or *kasrah* /*i*/. Only one case, signalled by the pattern *bi=at=taḥrīk-i*, where *taḥrīk-i* also signifies the vowel /*a*/ but, in this case, it will take the place of vowel of the first and second radical (C1 and C2).

Keyword	**change vowel of**
بِالفَتْح *withthefatḥah*	C1 to /*a*/
بِالضَّمّ *withtheḍammah*	C1 to /*u*/
بِالكَسْر *withthekasrah*	C1 to /*i*/
بِالتَّحْرِيك *withthetaḥrīk*	C1 and C2 to /*a*/

Table III: Vowel change markers

4) Cases of Adjectives

Examples 8 and 9 explain microstructure of some adjective lexical entries. According to ClassicalArabic lexical

[9] If the first radical is one of the sun letters listed in Table **Error! Main Document Only.** (Appendix B), it assimilates the letter *lām* of the definite article *al=* resulting in a doubled consonant expressed by the mark of gemination*šaddah* on the first radical.

[10] We use the Buckwalter transliteration because it's a character-by-character transliteration, in order to represent all the word characters including: *sukūn* (o); *šaddah* (~) *hamza-over-alif* (>). For more details, see: http://www.qamus.org/transliteration.htm (the last accessed date: 21/09/2020).

conventions, adjective can be introduced as predicate in a nominal sentence. In Example 8, we see that there is an **indefinite** subject رَجُلٌ*rağul-un*"*a man*"followed by the adjective *’alūb-un*" to definite.

Example 8.

رَجُلٌ **أَلُوبٌ**: سَرِيعُ إِخْرَاجِ الدَّلْوِ، أَوْ نَشِيطٌ

A man [is] ***’alūb-un****: [he is] quick to get the bucket out of the well, or [he is] active*

In this case, we verify if the sequence before colon matches the modeled rule (1), where we check if the first word W (supposed an indefinite subject) and the second words (supposed an ADJ candidate) ends with the indefinite nominative suffix *un*.

- Modeled rule (1): W-*un* + ADJ-*un*.

By cons, in Example 9, subject of nominal sentence isthe nominative pronoun هُوَ*huwa* "*he/it]*" followed by the adjective **مُؤْتَشِبٌ***mu’tašib-un*.

Example 9.

وَهُوَ**مُؤْتَشِبٌ**: غَيْرُصَرِيحٍفِينَسَبِهِ

He [is] ***mu’tašib-un****: he isn’t clear to declare his dynasty*

The nominative pronoun *huwa* is considered a surface pattern to capture adjective. In this case, we verify that entry sequence match the modeled rule (2). The first word is one of nominative pronouns[11]and the second word finishes by indefinite nominative suffix *un*.

- Modeled rule (2): Nominative pronoun + ADJ-*un*.

Then make certain that the second word contains all radicals before deciding that it is the candidate headword. After removing the nominative suffix *un*, the obtained lemma is classified as adjective.

5) Part Of Speech

Cases of adjectives

To define lemma’s Part Of Speech (POS), we use some specific techniques to each grammatical category. As already explained, the adjective is automatically tagged when introduced by an undefined subject or nominative pronoun.

Cases of verbs

About verbs, we are looking for whether the obtained lemma corresponds to one of morphological patterns listed in TableV. As we know, Arabic text may not be entirely vocalized. In our work, we checked during preparation phase that text was well vocalized. However, to recognize that a lemma is verb, it is required that all vowels are present, otherwise manual control is required.

Verb	Transliterated pattern	Arabic pattern
I	*C1aC2(a|u|i)C3-a*	فَعَلَ – فَعُلَ – فَعِلَ
II	*C1aC2~aC3-a*	فَعَّلَ
III	*C1AC2aC3-a*	فَاعَلَ
IV	*>aC1oC2aC3-a*	أَفْعَلَ
V	*taC1aC2~aC3-a*	تَفَعَّلَ
VI	*taC1AC2aC3-a*	تَفَاعَلَ
VII	*AinoC1aC2aC3-a*	انْفَعَلَ
VIII	*AiC1taC2aC3-a*	افْتَعَلَ
IX	*AiC1oC2aC3~-a*	اِفْعَلَّ
X	*isotaC1oC2aC3-a*	اسْتَفْعَلَ
Q_I	*C1aC2oC3aC4-a*	فَعْلَلَ
Q_II	*taC1aC2oC3aC4-a*	تَفَعْلَلَ
Q_III	*AiC1oC2anoC3aC4-a*	اِفْعَنْلَلَ
Q_IV	*AiC1oC2aC3aC4~-a*	اِفْعَلَلَّ

Table IV: Examples of verbal morphological patterns

Cases of nouns

Selected headword X is tagged as nouns if it matches one of modeled rules:

- X ends with the letter T *bound tā’* which is specific orthographic form of nouns.
- (***al*=X-*u***) : X starts with the definite article "الـ" *al* and ends with the nominative definite vowel *ḍamma* /*u*/.
- (**X-*u***) (***al*=Y-*i***): X constitutes the first word of genitive construction where X doesn’t start with the definite article but it ends with the nominative definite vowel /*u*/ providing that Y starts with the definite article.

However, work quality requires manual control especially of Nouns file because there are cases, like in Example 10, where the lemma is recognized as noun but in fact it’s an adjective.

Example10.

المُغَلَّبُ: المَغْلُوبُ مِرَارًا، وَالمَحْكُومُ لَهُ بِالغَلَبَةِ

*al=****muḡallab****-u: defeated several times, and doomed to fail*

D. *Extraction of morphosyntactic and derivational information*

1) Cases of verbs

Although there are some phonological and semantic rules, imperfect forms of triliteral verb remain ambiguous and varying according to the median vowel. Likewise *maṣdars* or verbal nouns of triliteral verbs are many varied and not predictable. For this reason, imperfect forms and verbal nouns are indicated in dictionaries and it is fruitful to extract them. Example 11 illustrates in which way those information are reported in AQAM. We note that there are two imperfect forms *ya’ibb-u* and *ya’ubb-u* that follow the verbal headword *’abb-a*. After imperfect forms, we find four infinitive verbs which finish with the indefinite accusative suffix -*an*: *’abb-an*, *’abīb-an*, *’abāb-an* and *’abābat-an*.

Example 11.

أَبَّ لِلسَّيْرِ، يَئِبُّ، وَيَؤُبُّ، أَبًّا، وَأَبِيبًا، وَأَبَابًا، وَأَبَابَةً: تَهَيَّأَ

’abb-a *for walking,* ***ya’ibb-u****, and* ***ya’ubb-u****,* ***’abb-an****, and* ***’abīb-an****, and* ***’abāb-an****, and* ***’abābat-an****: to prepare*

To extract imperfect forms, we follow instructions presented in the algorithm 3. We go through the sequence after the headword, looking for a word starting by the

[11] Nominative pronouns are: هُوَ*huwa* - هِيَ*hiya* - هُمَا*humā* - هُمْ*hum* and هُنَّ*hunna*.

imperfect prefix يـَ-*ya*, finishing with -*u* and containing all radical consonants.

Algorithm 3: Imperfect Form Extraction

input : String PlainText, char[] root

output : R:Vector of imperfect Form

```
seq ←getSeqBeforeColon(PlainText);
tokens ←seq.tokenizer();
fortokenintokensdo
  iftoken.startsWith("يــ")&&
  token.endsWith(u) &&
  token.contains(root)then
  R add(token);
end
return R;
```

*Maṣdars*extraction is also done automatically by checking that the candidate token ends with the indefinite accusative suffix *-an* and obviously contains all root radicals. Before affecting extracted *maṣdars*to the corresponding verbal entry, we apply them the automatic light stemming to remove possible particles and the ending vowel.

We also note that derived verbs, especially cause verbs, are frequently indicated in the same lexical entry of the corresponding triliteral verb, as shown in Example 12:

Example 12.

تَوَهَّمَ: ظَنَّ، **وَأَوْهَمَهُ وَوَهَّمَهُ** غَيْرُهُ

tawahham-a*: to believe, and someone* ***’awham-a=hu****, and* ***wahham-a=hu***

The verbal lemma***tawahhama*** is intransitive and accompanied by the sens *ẓanna* with the meaning "*to think, believe, suppose*". In addition we note the nested verbs, *’awhama=hu* and *wahhama=hu* with the mean "*someone else makes him believe*". Presence of accusative pronoun *=hu "him"* indicates that they are transitive. We note also that are derived from the same root as the verbal lemma and match a causality patterns *>aC1oC2aC3a* (أَفْعَلَ) and *C1aC2C2aC3a* (فَعَّلَ). After eliminating the conjunction *wa "and"* and accusative pronoun, and checking the existenceof all root radicals, the verbs, *’awham-a* and *wahham-a*, are tagged as *cause*.

2) Cases of nouns and adjectives

Broken or internal plural is characteristic of nouns and adjectives and it is impossible to know how many plurals a name can have except by resorting to dictionaries like AQAM. Detection of broken plurals is very simple thanks to specific surface patterns which precede them, and they are validated and extracted based on root radicals. Let’s study Example 13:

Example 13.

جَذَبَتِ النَّاقَةُ: قَلَّ لَبَنُهَا، **فَهِيَ** جَاذِبٌ وَجَاذِبَةٌ وَجَذُوبٌ، **ج.** جَوَاذِبُ وَجِذَابٌ

ǧa<u>d</u>aba-ti the camel: his milk has diminished, and she [is]ǧā<u>d</u>ib-un, and ǧā<u>d</u>ibat-un and ǧa<u>d</u>ūb-un

The headword *ǧa<u>d</u>aba-ti* is an inflected form of the lemma *ǧa<u>d</u>aba* and listed under family root *ǧ<u>d</u>b*. Before the meaning *qalla laban-u=hā* "*her milk has decreased*", we find three adjectives derived by the same root *ǧ<u>d</u>b* and introduced by the adjective surface form triggers *fa=hiya "then she [is]"*:*ǧā<u>d</u>ib-un*, *ǧā<u>d</u>ibat-un* and *ǧa<u>d</u>ūb-un* andtwo broken plurals of adjectives,*ǧawā<u>d</u>ib-u* and *wa=ǧi<u>d</u>āb-un*, preceded by the surface pattern ج which is the acronym of جَمْع*"plural"*.

3) Cases of nested entries

Consider Example 14 the verb triliteral *batt-a*is nested into the lexical entry of the adjective *al=bātt-u*.

After the light stemming process, the extracted lemma *bātt* has three senses: 1) *mahzūl"gaunt; weak"*; 2) *’aḥmaq* "*stupid*" and 3) *sakrān* "*drunk*".

Example 14.

البَاتُّ: المَهْزُولُ، **وَقَدْ** بَتَّ يَبِتُّ بُتُوتًا، وَالأَحْمَقُ، وَالسَّكْرَانُ

*al=****bātt****-u*: *al=mahzūl-u*, ***wa=qad****batt-aya-bitt-ubutūt-an*, *wa=l=’aḥmaq-u*, *wa=s=sakrān-u*.

After the first sense, the perfect verb *batt* is preceded by the specific particle *wa=qad*[12] that precedes only verbs. In AQAM, the author style is very precise and the particle *wa=qad* is used only for introduce nestedverb into adjectival entry. So, the particle *wa=qad* is considered the surface pattern that triggers automatic controls. As a result, the system checks whether the next word (which is surely a verb) is formed from the same radicals as the main headword. Then it automatically creates the specific lexical entry for the verb.In addition, the verb concerns only the sense that precedes it. Therefore the verb *batt-a* concerns only the first sense and it means *he became gaunt*. We also note that it is followed by the imperfect form *ya-bitt-u* and the *maṣda*r *butūt-an*.

E. *Extraction of senses and lexical information*

After recognizing and extracting lemma and different morphosyntactic characteristics, we go to lexical level to define senses, as well as to settle different relations between lemma and senses.

1) ***Synonymy relationship***

The author’s style is very concise, i.e. sense which is synonym consists of a single word (that is one of reasons we choose AQAM). For instance, in Example 14, the senses *mahzūl* "*gaunt; weak*"; *’aḥmaq* "*stupid*" and *sakrān* "*drunk*" are automatically tagged as synonyms of the lemma *bātt*.

2) Antonymy relationship:

In Example 15, the definition is composed of a sentence *ḍidduǧamada* in which the first word means "*opposite*". So the second word *ǧamada* "*to freeze*" is tagged as antonym of the lemma *dāba* "*to melt, to liquefy*". There are two surface patterns ضِدُّ*ḍidd-u* and عَكْس*’aks-a* whose meaning is "*opposite to*" and allowed to detect the antonymy relation between lemma and senses.

Example 15.

ذَابَ ذَوْبًا وَذَوَبَانًا: **ضِدُّ جَمَدَ**

<u>d</u>āba*, <u>d</u>awb-an, wa=<u>d</u>awabā-an : opposite [to] freeze*

[12] Following perfect verb, the particle *wa=qad* indicates the termination of an action; sometimes corresponding to English *"already"*.

3) Explicative sentence

Otherwise definition can be an explicative sentence, like in Example 16 when the lemma *sabbābah* "*the index finger*" is defined by the sentence "*follow the thumb*"!

Example 16.

السَّبَّابَةُ: تَلِي الابْهَامَ

*as=**sabbābat**-u: "follow the thumb"*

F. *Some XML Conversion Examples*

After extraction of various morpho-syntactic and semantic information, we present in this part an overview of obtained results, which is outlined in XML conversion of some rich examples.XML Example 1 shows XML conversion of the entry discussed in Example 1 and we find the following XML tags:

- root_family to which the entry belongs. In this example, the root family is C1=*"hamza"*; C2=*"wāw"*; C3=*"bā'"*
- lexical_entry: delimits each entry.
- plain_text: contains the original part of AQAM text from which we extract information.
- lemma: contains the lemma extracted from the plain text. In this example, the lemma is *'āba*.
- pos: to declare the lemma grammatical category.
- morphological_information: to designate the morphological pattern of the lemma. The lemma 'āba is triliteral verb of form I.
- morphosyntactic_information: he lemma *'āba* is transitive.
- sense: holds in the of the lemma. The verb *'ab`ad-a*is synonym of the lemma *'āba*.
- root: from which the lemma is derive, presented as a letter script form of the root_family.

```
<root_familyc1="الهَمْزة"c2="الواو"c3="البَاء">
<lexical_entry>
<plain_text>آبَهُ اللّٰهُ : أبْعَدَهُ</plain_text>
<lemma>آبَ</lemma>
<pos>Verb</pos>
<morphological_information>I
</morphological_information>
<morphosyntactic_information>Transitif
</morphosyntactic_information>
<senseatt="أبْعَدَ">
<synonymatt="أبْعَدَ"/>
</sense>
<root>ء و ب</root>
</lexical_entry>
</root_family>
```

XML Example 1. XML conversion of lexical entry of 'āba

XML Example 2 presents the conversion results of the verbal entry exemplified in Example 12. The system extracts the lemma *tawahhama* "*to believe*" and recognizes it as verb inside the tag <pos> having form V as morphological pattern. Verbs *'awhama* and *wahhama* "*to make believe*" are extracted and classified as causality action of the lemma. Finally, the sens *ẓanna* is marked as synonym of the lemma *tawahhama*.

```
<root_familyc1="الواو"c2="الهَاء"c3="المِيم">
<lexical_entry>
<plain_text> تَوَهَّمَ: ظَنَّ، وَأَوْهَمَهُ وَوَهَّمَهُ
        غَيْرَهُ</plain_text>
<lemma>تَوَهَّمَ</lemma>
<pos>Verb</pos>
<morphological_information>V
</morphological_information>
<cause>أَوْهَمَ</cause>
<cause>وَهَّمَ</cause>
<senseatt="ظَنَّ">
<synonymatt="ظَنَّ"/>
</sense>
<root>و هم</root>
</lexical_entry>
```

XML Example 2. XML conversion of lexical entry of tawahhama

XML Extract 3 shows the XML conversion of the entry in Example 13. The lemme *ğadab-a* is marked as verb of form I. The system recognises three adjectives derived from the verb marked by the <adjective> tag and extracts two broken plurals of adjectives marked by <plural> tag. Finally, the <sense> tag contains the sentence *"his milk has diminished"* that explains the lemme *ğadab-a*.

```
<root_familyc1="الجَيم"c2="الذَال"c3="البَاء">
<lexical_entry>
<plain_text>جَذَبَتِ النَّاقَةُ: قَلَّ لَبَنُهَا، فَهِيَ جَاذِبٌ وَجَاذِبَةٌ
        وَجَذُوبٌ، ج. جَوَاذِبُ وَجِذَابٌ، كَنِيَام</plain_text>
<lemma>جَذَبَ</lemma>
<pos>Verb</pos>
<morphological_information>I
</morphological_information>
<adjective>جَاذِبٌ</adjective>
<adjective>جَاذِبَةٌ</adjective>
<adjective>جَذُوبٌ</adjective>
<plural>جَوَاذِبُ</plural>
<plural>جِذَابٌ</plural>
<senseatt="قَلَّ لَبَنُهَا" />
<root>ج ذ ب</root>
</lexical_entry>
</root_family>
```

XML Example 3. XML conversion of ğadab-a

The XML Example 4 illustrates information extracted from the lexical entry the lemma *dāba*ricognised as verbof form I. The verbal nouns *dawb and dawabān* are extracted and tagged as *maṣdars*. Then the verbs *'adāba* and *dāwwaba* "*to dissolve some thing*" are marked by the <cause> tag. Finally the verb *ğamada* is marked as antonym of the lemme *dāba*. XML Example 4. XML conversion of dāba

From the entry mentioned in Example 14, two separate lexical entries are extracted. The first one, as shown in XML Example 5, concerns the lemma *bātt* accompanied by tree definitions tagged as synonyms: *mahzūl "emaciated", 'aḥmaq "foolish" and sakrān "drunk"*.

An other lexical entry is created separately for the nested verb *batt-a* and shown in the XML Example 6. Having the same plain text as the brooder entry, the lemma *batta* is followed by its imperfect form *ya-bitt-u*. The <pos> and <morphological_information> indicate that it is Form I Verb. Finally we find the masdar form *butūt* and the root(*btt*).

```
<root_familyc1="البَاء"c2="التَّاء"c3="التَّاء">
<lexical_entry>
<plain_text> البَاتُّ: المَهْزُولُ، وَقَدْ بَتَّ يَبِتُّ بُتُوتًا،
        وَالأَحْمَقُ، وَالسَّكْرَانُ</plain_text>
<lemma>بَاتّ</lemma>
<pos>Noun</pos>
<senseatt="مَهْزُول">
<synonymatt="مَهْزُول"/>
</sense>
```

```
<senseatt="أَحْمَق ">
<synonymatt="أَحْمَق "/>
</sense>
<senseatt="سَكْرَان">
<synonymatt="سَكْرَان"/>
</sense>
<root>تتب</root>
</lexical_entry>
</root_family>
```

XML Example 5. XML conversion of bātt

```
<root_familyc1="البَاء"c2="التَّاء"c3="التَّاء">
<lexical_entry>
<plain_text> البَاتُّ: المَهْزُولُ، وَقَدْ بَتَّ يَبِتُّ بُتُوتًا،
        وَالأَحْمَقُ، وَالسَّكْرَانُ</plain_text>
<lemma>بَتَّ</lemma>
<imperfective>يَبِتُّ</imperfective>
<pos>Verb</pos>
<morphological_information>I
</morphological_information>
<masdar>بُتُوتًا</masdar>
<root>تتب</root>
</lexical_entry>
</root_family>
```

XML Example 6. XML conversion of batta

G. Cases of Proper Names

AQAM contains many proper nouns which are introduced by the author in specific manner. Considering the lexical entry in Example 17 composed by the headword *al=ǧanāb-u* and senses that accompany it consisting of isolated words separated from each other by commas.

Example 17.

الجَنَابُ: اَلفِنَاءُ، وَالرَّحْلُ، وَالنَّاحِيَةُ، وَجَبَلٌ، وَعَلَمٌ

*al=ǧanāb-u: **the** patio, and **the** luggage, and the direction, and a mountain, and a proper name of a person.*

The first three definitions consist of definite words, and we can interpret the lexical entry:

the*ǧanāb-u* : [is] **the** patio, and [is] **the** luggage, and [is] **the** direction.[13]

In this case, the senses correspond to common meanings. After light stemming procedures, each sense is considered synonym for the lemma, as shown in XML Example 7.

```
<root_familyc1="الواو"c2="الهَاء"c3="البَاء">
<lexical_entry>
<plain_text> الجَنَابُ: الفِنَاءُ، والرَّحْلُ، والناحِيَةُ،
        وجَبَلٌ، وعَلَمٌ</plain_text>
<lemma>جَنَاب</lemma>
<pos>Noun</pos>
<senseatt="فِناء">
<synonymatt="فِناء"/>
</sense>
<senseatt="رَحْل">
<synonymatt="رَحْل"/>
</sense>
<senseatt="ناحِية">
<synonymatt="ناحِية"/>
</sense>
<root>ج ن ب</root>
</lexical_entry>
</root_family>
```

XML Example 7. XML conversion ofǧanāb

By cons, the last two senses are undefined and end with the nominative suffix (*un*). The definition, *wa=ǧabal-un*, does not mean "*and [it is] a mountain*", but it implies "*and [it's the proper name of] a mountain*". The last definition, *wa=ʼalam-un*, is more explicit and it implies "*and [it is] a proper name of a person*". In summary, a sense indicates a proper name when it is undefined and ends with the *tanwīn* nominative suffix, *un*. In this case, "light stemming" procedures consist in eliminating only possible particles (conjunctions) and the nominative suffix and the lemma must keep the definite article: *al=ǧanāb*. In addition, as illustrated in the XML Example 8 below, the lemma is accompanied by the POS as Proper Noun.

```
<root_familyc1="الواو"c2="الهَاء"c3="البَاء">
<lexical_entry>
<plain_text> الجَنَابُ: الفِناءُ، والرَّحْلُ، والناحِيةُ،
        وجَبَلٌ، وعَلَمٌ</plain_text>
<lemma>الجَنَابُ</lemma>
<pos>Proper_Noun</pos>
<senseatt="جَبَلٌ"/>
<senseatt="عَلَمٌ"/>
<root>ج ن ب</root>
</lexical_entry>
</root_family>
```

XML Example 8. XML conversion of al=ǧanāb

H. Digital version of AQAM

The designed system is able to automatically convert all sections of *al=qāmūs al=muhīt* to XML. The corresponding files are released as open sources by means of the *CLARIN-IT infrastructure*,[14] where there is a descriptive file of the lexicon AQAM and to which the dossiers corresponding to sections are linked. Each dossier corresponds to a section and contains:

- TXT text enriched with indicators which tags the chapters start; the root families start 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 separated files according to their corresponding semantic classes (Plant, Food, Animal, Proper Name, Geographic, water, Group and Others).
- a folder containing XML files of verbs, nouns and adjectives enriched with English translations.

Processes and methodologie that allow lemmas and senses translation and permit to identify semantic classes of proper names will be discussed in the next section.

Table VI illustrates briefly number of verbs, nouns, adjectives and proper names contained in each section. On the other hand, Table VII illustrates number of triliteral, quadriliteral and derived verbs contained in each section. By analyzing the obtained results presented in the tables VI and VII, our system is able to recognize a very large number of AQAM entries.

Table V: Some AQAM XML conversion statistics

Section (bāb)	Verbs	Nouns	Adjs	Proper Names

[13] The words in square brackets do not exist in the Arabic text. They are used to make the translated text more understandable.

[14] http://hdl.handle.net/20.500.11752/ILC-97 (the last accessed date: 21/09/2020).

hamza	637	569	19	93
bā’	1242	2961	385	1054
tā’	285	487	48	182
t̲ā’	227	483	31	164
ğim	507	1024	98	336
ḥā’	661	1078	138	346
hā’	322	435	63	116
dāl	855	1716	163	763
d̲āl	92	268	23	108
rā’	2052	4231	353	1735
zāy	453	738	68	264
sīn	575	1356	104	508
šīn	447	612	73	189
ṣād	448	623	79	128
dād	376	388	53	118
tā’	649	866	77	255
zā’	121	148	15	16
`ayn	1306	1846	227	494
ğayn	278	289	32	70
fā’	1135	1767	209	469
qāf	1071	1728	201	589
kāf	468	732	70	243
lām	1834	3191	319	1096
mīm	1450	2685	232	1068
nūn	835	1756	142	874
hā’	271	307	43	59
wāw	638	1182	47	373
yā’	440	1156	60	383
Total				
28	19675	34233	3204	12095

Due to the huge amount of data, we have controlled obtained results for the *d̲āl* section, which contains 95 verbs, 270 Nouns, and 27 Adjectives.

Table VIII presents statistical results concerning the *d̲āl* section. Our system has an accuracy of 100%, 95% for recall and 97% for F-measure for Verbs. For nouns we have 99% for precision, 97% for recall and 98% for F-measure. Finally for Adjectives we got 100% for precision, 85% for recall and 92% for F-measure.

Table VIII: Statistics of AQAM entries extraction

Class	Precision	Recall	F-Measure
Verbs	100%	95%	97%
Nouns	99%	97%	98%
Adjectives	100%	85%	92%

Some case should be resolved manually, because the system hasn’t been able to recognize. For example, In Example 18 the lemma عَلْبَى *`albā* derived from the triliteral root (ع ل ب) has an unusual verbal pattern annexed to the quadrilateral verbal pattern. An other example is illustrated in Example 19 which illustrates an Idiomatic expression. Examples 20 and 21 illustrates two lemmas which have unusual pattern ending with kasra and they have different POS.The first one*hā’i*is the imperative mood of the verb*hā’a*and means "*Give! Bring!"*.By cons, the second one is the noun *ğabādi*.

Example 18.

عَلْبَى الرَّجُلُ: ظَهَرَتْ عَلَابِيُّهُ كِبَراً

***`albā*** *the man: zaharat `alābiyyuh-u kibar-an*

Table VII: STATISTICS OF EXTRACTED AQAM VERBAL ENTRIES

Section	I	II	III	IV	V	VI	VII	VIII	IX	X	Q_I	Q_II	Q_III	Q_IV	Total
hamza	397	25	25	66	24	6	8	5	-	11	56	10	4	-	637
bā'	624	109	36	192	77	17	27	19	1	36	71	17	-	16	1242
tā'	180	17	9	31	11	1	11	9	1	5	9	-	-	1	285
t̲ā'	126	18	1	15	19	3	9	6	1	11	13	4	-	1	227
ğīm	234	36	7	85	44	4	14	18	3	13	38	6	3	2	507
ḥā'	324	54	23	106	45	14	12	14	2	17	37	8	5	-	661
ḫā'	166	37	6	36	23	5	10	9	6	3	18	3	-	-	322
dāl	358	84	25	165	74	15	27	12	4	42	27	5	1	16	855
d̲āl	36	6	5	18	1	2	5	-	1	5	7	6	-	-	92
rā'	744	220	61	321	174	43	91	41	12	82	173	40	6	44	2052
zāy	216	37	14	55	35	11	23	5	-	14	25	9	3	6	453
sīn	255	32	21	68	45	25	21	12	5	5	63	20	3	-	575
šīn	196	37	8	52	37	9	34	9	1	10	32	15	2	5	447
ṣād	212	47	9	56	25	7	23	16	4	4	37	8	-	-	448
ḍād	171	40	10	72	25	3	19	5	5	13	9	4	-	-	376
ṭā'	312	42	20	75	50	10	35	21	2	14	45	12	3	8	649
ẓā'	72	7	4	15	6	-	5	1	1	4	6	-	-	-	121
`ayn	637	103	28	203	109	14	39	51	2	36	57	20	7	-	1306

ğayn	133	20	6	45	29	2	3	13	-	5	18	4	-	-	278
fā'	487	100	32	242	69	16	41	28	3	49	48	9	1	10	1135
qāf	487	100	35	183	69	7	33	53	9	19	56	15	5	-	1071
kāf	297	11	4	27	25	13	37	5	9	10	19	8	3	-	468
lām	715	159	57	286	153	31	79	29	7	70	181	30	1	36	1834
mīm	666	129	40	241	120	13	57	17	6	29	82	29	8	13	1450
nūn	394	77	31	141	53	15	31	3	4	29	28	4	-	25	835
hā'	132	35	14	31	24	6	1	2	-	3	18	5	-	-	271
wāw	588	3	10	21	2	3	7	-	1	-	3	-	-	-	638
yā'	375	19	12	16	3	2	7	-	-	5	1	-	-	-	440
Total															
28	**9534**	**1604**	**553**	**2864**	**1371**	**297**	**709**	**403**	**90**	**544**	**1177**	**291**	**55**	**183**	**19675**

Example 19.

جَازِئُكَ مِنْ رَجُلٍ: نَاهِيكَ

ğāzi'uk-a min rağul-in: "this man will suffice thee"

Example 20.

هَاءِ، بِالكَسْرِ:هَاتِ

hā'i, bi=al=kasri: "Give! Bring!"

Example 21.

جَبَاذِ: المَنِيَّةُ، أَوْ النِّيَّةُ الجَابِذَةُ

ğabādi: "Fate, destiny; death", or "attractive idea"

III. AUTOMATIC SEMANTIC TAGGING OF AQAM USING PWN AND SUMO

Currently, semantic lexicons and lexical ontologies are important resources in semantic processing of natural language which is becoming increasingly important as technology advances. For these reasons, we thought of enriching the new obtained digital version of AQAM with semantic and ontological information. Our motivation is to increase the ability of Arabic language processing by linking the digital version of AQAM with preexisting English resources as the Princeton WordNet (PWN) and the Suggested Upper Merged Ontology (SUMO). The next section presents the methodology and results of a pilot stage of mapping between AQAM and PWN and SUMO.

A. Resources

1) The bilingual dictionary

We have chosen the bilingual dictionary "*An Advanced Learner's Arabic-English Dictionary*"[26], because the author specified that he used, among others, "*An Arabic English Lexicon*" of Edward William Lane, itself based on *tāğ al='arūs min ğawāhir al=qāmūs* for [19] which is the extended dictionary of AQAM.[15]

2) Princeton WordNet (PWN)

WordNet (PWN3.0) is a large lexical database for English [20].[16] It groups words together based on their meanings into synonyms sets. Rightly, synonym set is called synset. Each synset represents one underlying *concept*, i.e., a set of synonyms share the same meaning in a given context [21]. In addition, synsets are interlinked by means of lexical and conceptual-semantic relations: synonymy, antonymy, hyponymy, hypernymy, meronymy, troponymy, etc.

3) Suggested Upper Merged Ontology (SUMO)

In information science, an Upper Ontology, also known as Top level or Foundation or Universal Ontology, is defined as "an Ontology which describes very general concepts that are the same across all knowledge domains" [22]. The Suggested Upper Merged Ontology (SUMO)[17] was created by merging a number of existing upper-level Ontologies. Moreover, it's the only formal ontology that has been mapped to all of the WordNet lexicon[23]; [24]. The used database contains mappings from the WordNet lexical database to SUMO.[18]

Consider the following Example when the WordNet entry enriched by SUMO concept.The first part of the record states that the number 01828405 is the unique identifier of the verb synset {*hanker, long, yearn*}. The synset is connected with other synsets by lexical and hierarchical relations thanks to labeled pointers.[19] After the sign marker "|" we find the gloss defining the synset meaning (desire strongly or persistently).[20] Finally in last position the corresponding

[15] Fortunately, we have a machine-readable version in TEI format available online (the last accessed date: 21/09/2020): http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3atext%3a2002.02.0005.

[16] http://wordnet.princeton.edu/ (the last accessed date: 21/09/2020).

[17] http://www.adampease.org/OP/ (the last accessed date: 21/09/2020).

[18] The SUMO and PWN3.0 resources used in this paper are available for download at the link (the last accessed date: 21/09/2020): https://github.com/ontologyportal/sumo/blob/master/WordNetMappings/

[19] The "@" symbol indicates that this synset is subsumed by the synset {desire, want} whose identifier is 01825237. Each synset after the "+" symbol groups derivationally related nouns, for example, the synset 10270878 {longer, thirster, yearner}. Finally the symbol ~ indicates the troponymy relation with the synset 01805684 {ache, yearn, yen, pine, languish}.

[20] For more details of all the components constituting the line you can consult the following link:

SUMO concept (IntentionalPsychologicalProcess+) is designated with the "&%" symbol and ends with the "+" symbol that indicatesthat the synset 01828405 {hanker, long, yearn} is subsumed by the SUMO concept.[21]

01828405 37 v 03 hanker 0 long 2 yearn 0 007 @ 01825237 v 0000 + 10270878 n 0303 + 07486628 n 0302 + 10270878 n 0201 + 07486628 n 0201 + 07486922 n 0101 ~ 01805684 v 0000 01 + 22 00 | desire strongly or persistently
&%IntentionalPsychologicalProcess+

B. *Methodology*

The first step consist to translate lemma using the Arabic-English dictionary. However polysemy is one of major problems with machine translation. Although, polysemous word has only one meaning in a specific context.

We noted (generally) that the author of AQAM defined the lemma meaning by presenting it in a specific example/context and the sens is a single word considered synonym of the lemma. For instance, the lemma *'abba* is more polysemous in isolated context and can be translated as: "***prepare***, *get ready to*, ***yearn***, ***long***".However in the context "*'abba his homeland*" in Example 22, the lemma *'abba* has the meaning *ištāqa* translated as "*desire*, ***yearn, long***".By cons in the context "*'abba for the walk*" of the Example 23, it means *tahayya'a* which is translated: "***prepare for***".So to deal with polysemy problem, in addition to lemma translation, we also had recourse to translation of senses which accompany it.To disambiguate lemma meaning, we consider an algorithm which takes into account English translated words that result of the intersection between lemma translations and senses translations. In Example 22, the lemma *'abba* is translated as "***yearn, long***". By cons in Example 23, it is translated as "***prepare for***".

Example 22.

أَبَّ إِلَى وَطَنِهِ أَبًّا وَإِ بَابَةً وَأَبَابَةٌ: اشْتَاقَ

'abba *his homeland ..... :* ***iŝtāqa***

Example 23.

أَبَّ لِلسَّيْرِ يَئِبُّ وَيَؤُبُّ أَبًا وَأَبِيبًا وَأَبَابًا وَأَبَابَةً: تَهَيَّأَ

'abba *for the walk ..... :* ***tahayya'a***

In the specific case of proper name, we don't justly translate the lemma however we can to classify it thanks to the sens that accompanies it. In Example 24, the lemma *ǧal'ib-un* is a proper name of "***a mountain*** *in the Medina*" and therefore it is a toponym i.e. a place-name.

Example 24.

جَلْعِبٌ: جَبَلٌ بِالمَدِينَةِ

ǧal`ib-un:"a mountain in the Medina"

Figure. 6 presents a summary and stages of the adopted procedures. After translation process, following steps consist to search eventual translation equivalents in Princeton WordNet and SUMO in order to find synsets and concepts that can correspond to the Arabic lemma.

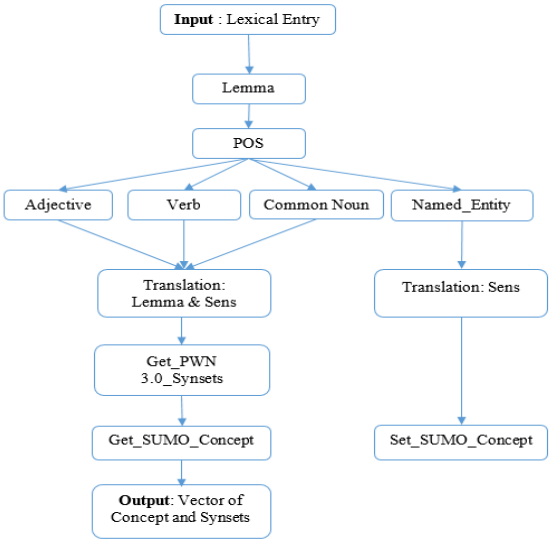


Figure 6: Mapping Process.

C. *Process of Translation*

The algorithm 4 illustrates followed steps. The available resources differ in many details. The Arabic lemmas are presented in the TEI digital version of bilingual dictionary using a non-standard transliteration. So the first step is to transliterate words candidates using the same transliteration system.Afterwards we search transliterated form in the Arabic-English Lexicon to get English translations. However, the AQAM text which we have is not fully vocalized, by cons the bilingual lexicon is fully vocalized. So we define that each vowel found in AQAM candidate word must be obligatory in the matching process. In fact, as presented in the algorithm 4, we affect to searched vowels of the transliterated form a priority equals to 1. After having the corresponding regular expression, therefore we recover all corresponding translations.In addition the used bilingual dictionary author present verbs translations in simple past mode. In order to link AQAM to English WordNet (PWN), we must have the verbal infinitive form. So we use the JAWS[22](Java API for WordNet Searching), which allowing to get the basic form, based on searching of the obtained result in the PWN verbs. For instance, in Examples 22 and 23 the lemma 'abba is translated as "*prepared, got ready to, longed, yearned*", after converting them to their basic form we got "*prepare, get ready to, yearn, long*".

Algorithm 4: Translation of AQAM's Lemma and simple sens

```
input    :  String qAmusLemma, List ArabicVowels,
            Lexicon ALAED
output   :  Translation: Vector of matched words
```

https://wordnet.princeton.edu/documentation/wndb5wn

[21] There are three symbols, "=", "+" or "@", which indicates the precise relationship between the SUMO concept and the WordNet synset. They mean, respectively, that the WordNet synset is equivalent in meaning, subsumed or an instance of the SUMO concept.

[22] https://github.com/jaytaylor/jaws

Translation

```
qAmusLemmaTransliterated←Transliteration(qAmusLemma);
/* Transliteration of أَبَّ 'abba= ″A^ab~a″*/
Char[] splitedLemma←
splitToChar(qAmusLemmaTransliterated);
forcinsplitedLemmado
ifc.matches(ArabicVowels)then
    s←s+c+"{1}?";
else
        s←s+c+"{1}?.{0,1}?";
end
end
s←clean(s) $ //delete additional .{0,1}?
/*s= A{1}?^{1}?a{1}?b{1}?~{1}?a{1}?*/
forALAEDLemmainALAEDdo
ifs.matches(ALAEDLemma)then
addToTranslation(
ALAEDLemma.getTranslation());
end
/*Translation={Prepared, got ready to, for; Yearned for, longed to see}*/
ifqAmusLemma.getPOS()==″Verb″then
fort inTranslationdo
get_basic_form(t);
end
/*Translation={Prepare, get ready to, for; Yearn, long}*/
returnTranslation;
```

D. *Process of mapping*

Algorithm 5 illustrates the automatic mapping process of AQAM with PWN and SUMO that is summarized in the following steps:

1) Look for each element E of the lemma English translation list : L(X) in PWN.
2) Retrieve the Synsets list $L_S(E)$ for each element E.
3) For each element p of a given synset of $L_S(E)$, retrieve the corresponding SUMO concept.
4) Classify the result by concept in List Lemma Concepts $L_{LC}(X)$.
5) For each sense s, repeat from 1 to 4 and get the senses concepts of the $L_{SC}(s)$ list.
6) Compare the elements of $L_{LC}(X)$ and $L_{SC}(s)$.
   - If $L_{LC}(X) = L_{SC}(s)$: the sense concepts list and the lemma concepts list are equals, it means that the sense and the lemma are "*Perfect Synonym*".
   - $R = L_{LC}(X) \cap L_{SC}(s)$, if $R \neq \varnothing$: the sense concepts list and the lemma concepts list share some concepts, it means that the sense and the lemma are "*synonym*".

Algorithm 5: Mapping an AQAM's LE to PWN and SUMO

input : Lemma X, Sens S, Lemma_Translation L1, Sens_Translation L2
List Nominal_Form, List Adjectival_Form

output : R: Vector of corresponding SUMO concepts and PWN synsets

```
fore1inL1do
List_Lemma_Syn←getSynsets(e1);
forp1 inList_Lemma_Syndo
List_Lemma_Concepts←
getSUMOConcept(p1);
end
LLC←sortByConcept(List_Lemma_Concepts);
end
for e2inL2 do
List_Sens_Syn←getSynsets(e2);
forp2inList_Sens_Syndo
List_Sens_Concepts←
getSUMOConcept(p1);
end
LSC←sortByConcept(List_Sens_Concepts);
end
// Compare LLC && LSC
R←LLC ∩ LSC;
ifR == LSC && R == LLCthen
X and S are perfect Synonyms;
elseifR <>∅then
X and S are Synonym;
else
R←LLC;
end
return R;
```

E. *Discussion and analyse of some examples*

1) Example of the verbal lemma 'abba

Figure. 7 presents a diagram which summarizes workflow taking as example the original AQAM entry. Final result contains morphosyntactic and lexical level which concern lemma extraction accompanied with its grammatical category, enriched with different associated derived forms, and concludes by semantic component enrichment regarding senses extraction, translations and mapping to PWN and SUMO. In addition the system recognizes automatically:

1) *Translation of the lemma 'abba: "Prepare, get ready to, long, yearn" and translation of the sense ištāqa: "desire, yearn, long"*.
2) We keep only English equivalents who are common between lemma and sense translations: "*long, yearn*".
3) We search all corresponding PWN3.0 synsets for each common English equivalent.

   The verb *long* is part of one synset {hanker, long, yearn}: "desire strongly or persistently".

   The verb *yearn* is part of three synsets: {yearn} "have affection for; feel tenderness for"; {ache, languish, pine, yearn, yen}: " have a desire for something or someone who is not present" and {hanker, long, yearn}: "desire strongly or persistently".
4) We keep only those who are common between different sets of synsets: {hanker, long, yearn}: "desire strongly or persistently" that have ID=01828405.
5) We retrieve the corresponding SUMO Concept based on the gotten Synset-ID : **IntetionalPsychologicalProcess+**. The suffix '+' indicates that the synset ID = 01828405 {hanker, long, yearn} is subsumed by the SUMO concept IntetionalPsychologicalProcess.
6) There are some common concepts between the lemma and its sense, so the Mapping value "Synonym" is added automatically to the ExternalLinks Tag.

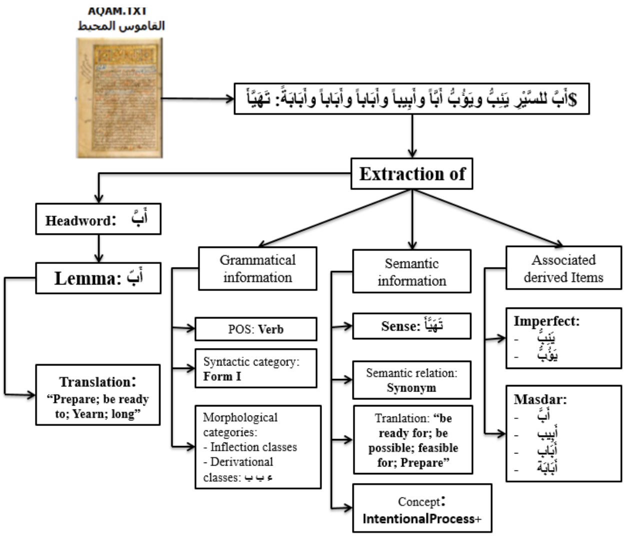

AQAM.TXT
القاموس المحيط
$أبَّ للسَّيْرِ يَئِبُّ ويَؤُبُّ أبّاً وأبِيباً وأبَاباً وأبَابَةً: تَهَيَّأَ
Extraction of
Headword: أبَّ
Lemma: أبّ
Translation: "Prepare; be ready to; Yearn; long"
Grammatical information
POS: Verb
Syntactic category: Form I
Morphological categories:
- Inflection classes
- Derivational classes: ء ب ب
Semantic information
Sense: تَهَيَّأَ
Semantic relation: Synonym
Tranlation: "be ready for; be possible; feasible for; Prepare"
Concept: IntentionalProcess+
Associated derived Items
Imperfect:
- يَئِبُّ
- يَؤُبُّ
Masdar:
- أَبٌّ
- أبِيب
- أَبَاب
- أَبَابَة

Figure 7: Workflow diagram

XML Example 11 illustrates the verbal entry concerning the lemma *’abba* conversion.

```
<sectionvalue="البَاء">
<chaptervalue="الهَمْزَة">
<root_familyc1="الهَمْزَة"c2="البَاء"c3="البَاء">
<lexical_entry>
<plain_text>أَبَّ إلى وطَنِهِ أَبّاً وإبَابَةً وأَبَابَةً: اشْتَاقَ</plain_text>
<lemma>أَبَّ</lemma>
<pos>Verb</pos>
<morphosyntactic_information>I </morphosyntactic_information>
<masdar>أَبّ</masdar>
<masdar>إبَابَة</masdar>
<masdar>أَبَابَة</masdar>
<lemma_translateatt="أَبَّ"value="prepare, getready to; yearn; long"/>
<senseatt="اشْتَاقَ">
<synonymatt="اشْتَاقَ">
<translatelang="eng"value="Desire, yearn, long"/>
<external_linksatt="mapping"value="Synonym"/>
<conceptatt="SUMO"value="IntentionalPsychologicalProcess+">
<synsetref="PWN"ver="3.0"id="01828405"syn="hanker, long,
             yearn"Gloss="desire strongly or persistently"/>
</concept>
</external_links>
</synonym>
</sense>
<root>ء ب ب</root>
</lexical_entry>
</root_family>
</chapter>
</section>
```

XML Example 11. Example of the verbal lemma ’abba

2) Example of the lemma kitāb

Example of the lemma كِتَابkitāb shows complicated lexical entry that requires more further study. The English translation of the lemma *kitāb* proves that it is very polysemous:

```
<lemma_translateatt="كِتَاب"value="writing; book, volume; letter; record,
   register; contract; Sacred Writings, Scriptures: the Pentateuch,
   Old Testament; the Gospel; the Kur’ān; Decree, ordinance, statute,
   prescript; prescription; sentence; Receptacle for ink"/>
```

IN AQAM, we tell seven senses which prove that the lemma is very polysemous:

الكِتَابُ: مَا يُكْتَبُ فِيهِ، وَالدَّوَاةُ، وَالتَّوْرَاةُ، والصَّحِيفَةُ، وَالفَرْضُ، وَالحُكْمُ، وَالقَدَرُ

Sense 1: mā yuktabu fīh

It consists of a whole phrase "*mā yuktabu fīh*" that means "[*material*] *on which one writes*". In this case, the manual intervention of the expert is necessary. After disambiguation, the explanatory sentence is mapped to the PWN3.0 at the synset {*book, volume*}, ID=02870092, defined by the Gloss "*physical objects consisting of a number of pages bound together*" and subsumed by the SUMO concept ***Book***.

```
<senseatt="مَا يُكْتَبُ فِيهِ">
<translatelang="eng"value="null"/>
      <external_linksatt="mapping"value="null">
<conceptref="SUMO"value="Book+">
           <synsetref="PWN"ver="3.0"id="02870092"syn="book,
           volume"Gloss="physical objects consisting of a number of
           pages bound together"/>
</external_links>
</sense>
```

Sense 2: dawāt

Initially, the system finds no common English equivalents between the lemma and the second sense "*dawāt*" translated as "*Inkbottle, inkstand*". However, after a detailed study, we note that the lemma has been translated, in the bilingual dictionary, as "*Receptacle for ink*", sentence who can explain "*Inkbottle, inkstand*". So, after manual revision, the second sense is tagged as "synonym", linked at the synset {*inkstand, inkwell*}, ID=03572321 with the Gloss="*a small well holding writing ink into which a pen can be dipped*" and subsumed by the SUMO concept ***Device***.

```
<senseatt="دَوَاة">
<synonymatt="">
<translatelang="eng"value="Inkbottle, inkstand"/>
<external_linksatt="mapping"value="Synonym">
<conceptref="SUMO"value="Device+">
<synsetref="PWN"ver="3.0"id="03572321"syn="inkstand,
            inkwell"Gloss="a small well holding writing ink into which a
            pen can be dipped"/>
</concept>
</external_links>
</synonym>
</sense>
```

Sense 3: tawrāt

The third sens "*tawrāt*" is translated as "*Torah, Pentateuch, Old Testament*". In PWN, the English word "*Torah*" is mapped to three synsets linked to the same SUMO concept ***Text***. This demonstrates that the PWN is very detailed and has a very fine grid, which does not facilitate mapping procedures.

```
<senseatt="تَوْرَاة">
<synonymatt="تَوْرَاة">
<translatelang="eng"value="Torah, Pentateuch; Old Testament"/>
<external_linksatt="mapping"value="Synonym">
<conceptref="SUMO"value="Text@">
<synsetref="PWN"ver="3.0"id="06451891"syn="Laws, Pentateuch,
            Torah"Gloss="the firs of three divisions of the Hebrew,
            Scriptures comprising the first five books of the Hebrew
            Bible considered as a unit"/>
```

```
<synsetref="PWN"ver="3.0"id="06452223"syn="Torah"Gloss="the
        whole body of the Jewish sacred writings and tradition
        including the oral tradition"/>
</concept>
<conceptref="SUMO"value="Text+">
<synsetref="PWN"ver="3.0"id="06408442"syn="Torah"Gloss="(Juda
        ism) the scroll of parchment on which the first five books of
        the Hebrew Scripture is written;is used in a synagogue
        during services"/>
</concept>
</external_links>
</synonym>
</sense>
```

Sense 4: ṣaḥīfah

The sense "*ṣaḥīfah*" is translated as "*leaf, paper, page*" and each English equivalent is more polysemous:

-*leaf* is part of three synsets of which the synset {*leaf, fogliage*} with gloss "*a sheet of any written or printed material (especially in a manuscript or book)*" and subsumed to the SUMO concept ***Text*** can correspond to the lemma meanings.

-*paper* is part of eight synsets of which only the synset {*paper*} defined by the gloss "*a material made of cellulose pulp derived mainly from wood or rags or certain grasses*" can correspond to the lemma meanings, and is equivalent to the SUMO concept ***Paper*** through the symbol "=".

-*page* is part of five synsets of which only the synset {*page*} with the gloss "*one side of one leaf*" corresponds to the lemma, and it is equivalent to the SUMO concept ***Page*** through the symbol "=".

```
<senseatt="صَحِيفَة">
<synonymatt="صَحِيفَة">
<translatelang="eng"value="leaf, page, paper"/>
<external_linksatt="mapping"value="Synonym">
<conceptref="SUMO"value="Text+">
<synsetref="PWN"ver="3.0"id="06256229"syn="folio, leaf"Gloss="a
        sheet of any written or printed material (especially in a
        manuscript ord book)"/>
</concept>
<conceptref="SUMO"value="Page=">
<synsetref="PWN"ver="3.0"id="06256697"syn="page"Gloss="one
        side of one leaf (of a book or magazine or newspaper or letter
        etc.) or the written or pictorial matter it contains"/>
</concept>

<conceptref="SUMO"value="Paper=">
<synsetref="PWN"ver="3.0"id="14974264"syn="paper"Gloss="a
        material made of cellulose pulp derived mainly from wood or
        rags or certain grasses"/>
</concept>
</external_links>
</synonym>
</sense>
```

Senses 5 and 6: farḍ and *ḥukm*

The senses *farḍ* and *ḥukm* are considered synonymous with the lemma because they have common translations"*statute, law, ordinance*". However, the relation is not really exact, because the word *kitāb* is used in a figurative sense in these cases, and means those senses by extension.

To explain these two meanings, *az=zabīdī*, in its lexicon *tāǧ al=`arūs* ([19], Volume 4, page 101) gives the examples that refer only to Allah.

In addition, Wehr in its lexicon [25], marks *farḍ* and *ḥukm* by the abbreviation "Isl. Law" (Islamic Law: *fiqh*) i.e. they are part of the traditional Islamic terminology:

farḍ: religious duty (Isl. Law).
hukm: legal consequence of the facts of a case (Isl. Law).

In PWN and SUMO, no concept considers terminology of Islamic Law that represents example of ontological non-equivalence. Consequently, the external links remain as gaps.

```
<senseatt="فَرْض">
<synonymatt="فَرْض">
<translatelang="eng"value="religious duty"/>
<external_linksatt="mapping"value="null">
<conceptref="SUMO"value="null"/>
<synsetref="PWN"ver="3.0"id="null"syn="null"Gloss="null"/>
</concept>
</external_links>
</synonym>
</sense>
<senseatt="حُكْم">
<synonymatt="حُكْم">
<translatelang="eng"value="legal consequence of the facts of a case"/>
<external_linksatt="mapping"value="null">
<conceptref="SUMO"value="null"/>
<synsetref="PWN"ver="3.0"id="null"syn="null"Gloss="null"/>
</concept>
</external_links>
</synonym>
</sense>
```

Sense 7: qadar

The sense*qadar* is translated as *fate, destiny; divine decree; providence*. However, *az=zabīdī*, in its lexicon [19], specifies that the sens *qadar* concerns the Qur'ān:

وَكِتَابُ اللَّهِ: قَدَرُهُ

the Book of Allah: [is] "its ***qadar***".

In this precis context, we link the lemma *kitāb* in PWN 3.0 to the synset ID=06461609 that identify the synset {*Book, Koran, Quran, al-Qur'an*} subsumed to SUMO concept ***Book***.

```
<senseatt="قَدَر">
<synonymatt="قَدَر">
<translatelang="eng"value="statute, law, ordinance"/>
<external_linksatt="mapping"value="null">
<conceptref="SUMO"value="Book+"/>
<synsetref="PWN"ver="3.0"id="06461609"syn="Book, Koran, Quran,
        al-Qur'an"Gloss="the sacred writings of Islam revealed by
        God to the prophet Muhammad during his life at Mecca and
        Medina"/>
</concept>
</external_links>
</synonym>
</sense>
```

3) Example of the lemma ṯawāb

الثَّوَابُ: العَسَلُ، وَالنَّحْلُ، وَالجَزَاءُ، كَالمَثُوبَةِ، وَالمَثْوَبَةِ

In "An Advanced Learner's Arabic-English Dictionary", the lemma *ṯawāb* is translated as:

(a) reward, recompense. (b) Price. (c) Honey.[23]

We note that meanings are classified according to their frequency and the English equivalents are accompanied by the symbol (a), (b) or (c) (etc.) that represent the first, the second, the third, (etc.) meanings of the lemma. For example, the sens *`asal* "*honey*" that is no longer used in Modern Arabic is classified as the third sens (c).

*Senses 1 and 2: `asal*and *naḥl*

In-depth studies of AQAM allow us to note that there are some recurrent definitions, for example, the meanings *`asal* "*honey*" and *naḥl* "*bee*" respectively recur 42 and 31 times. For reasons of optimization, we identify those "*Tag key words*" that can speed up the mapping and linking automatically the lemma with the corresponding concept. The "Tag key words" identified in the section *bā'* are listed in appendix C. So, the two first senses are linked automatically to their corresponding PWN and SUMO concepts, respectively to ***Honey=*** and ***Bee=.*** In addition, the system separates them into different files, respectively, File_Food and File_Animal.

```
<senseatt="عَسَل">
<synonymatt="عَسَل">
<translatelang="eng"value="honey"/>
<external_linksatt="mapping"value="synonym">
<conceptatt="SUMO"value="Honey=">
<synsetref="PWN"ver="3.0"id="15305917"syn="honey"Gloss="An
            edible, sweet substance produced by some species of Bees"/>
</concept>
</external_links>
</synonym>
</sense>
<senseatt="نَحْل">
<synonymatt="نَحْل">
<translatelang="eng"value="bee"/>
<external_linksatt="mapping"value="synonym">
<conceptatt="SUMO"value="Bee=">
<synsetref="PWN"ver="3.0"id="15330350"syn="bee"Gloss="A hairy
            insect, some species of which produce honey and/or sting"/>
</concept>
</external_links>
</synonym>
</sense>
```

Sense 3: ǧazā'

The third meaning *ǧazā'* is translated as: "*repayment; remuneration, recompense, requital, compensation*". Having common translations with lemma, it is automatically considered by the system as synonymous of the lemma ***tawāb***.

However, *ṯawāb* is specifically related to the Islamic Law Domain[24] and senses "*repayment; remuneration, recompense, requital, compensation*" are related to "*good or bad deeds*" which are not planned in PWN and SUMO.

So, as cited by [27]), linking those concepts that have a religious connotation to PWN and SUMO would miss their Muslim identity.

```
<senseatt="جَزَاء">
<synonymatt="جَزَاء">
<translatelang="eng"value="repayment; remuneration, recompense,
            requital, compensation"/>
<external_linksatt="mapping"value="synonym">
<conceptref="SUMO"value="null"/>
<synsetref="PWN"ver="3.0"id="null"syn="null"Gloss="null"/>
</concept>
</external_links>
</synonym>
</sense>
```

Senses maṯūbah and *maṯwabah*

The last two senses meanings are a specific cases, cited according to the author's style. Introduced by the conjunction *ka* "*as*", they are derived from the same root of the lemma and they inherits the last meaning that precedes. After checking that *maṯūbah* and *maṯwabah* contains the lemma radical consonants, they inherits the translation of *ǧazā'*.

```
<senseatt="مَثُوبَة">
<synonymatt="مَثُوبَة">
<translatelang="eng"value="repayment; remuneration, recompense,
            requital, compensation"/>
<external_linksatt="mapping"value="synonym">
<conceptref="SUMO"value="null"/>
<synsetref="PWN"ver="3.0"id="null"syn="null"Gloss="null"/>
</concept>
</external_links>
</synonym>
</sense>
<senseatt="مَثْوَبَة">
<synonymatt="مَثْوَبَة">
<translatelang="eng"value="repayment; remuneration, recompense,
            requital, compensation"/>
<external_linksatt="mapping"value="synonym">
<conceptref="SUMO"value="null"/>
<synsetref="PWN"ver="3.0"id="null"syn="null"Gloss="null"/>
</concept>
</external_links>
</synonym>
</sense>
```

In conclusion, obtained results demonstrate that lemmas enriched with concepts extracted from PWN and SUMO denote common and general meanings to Arabic and English cultures and languages. They also make it possible to emphasize differences between them, because lemmas which denote specific concepts of Arabic language and culture don't have equivalents and present gaps in mapping.

F. Statistical results analyse

Review process is complicated and time-consuming, as shown in the previous section there are complicated cases that require in-depth study and manual review. Consequently, this part of our work was experimented on the Section of bā '. The following tables and the resulting statistical analyzes give a general idea of the organization of AQAM and the complexity of the work.

According to Table IX, the *bā'* chapter consists of 1244 verbal lemmas, 2455 nominal lemmas and 216 adjectives. Lexical entries can be without definition generally extracted from nested entries. Other lexical entries can have monosemous lemma with a single definition; or polysemous lemma having multiple definitions. Note that polysemous lemmas (especially Nouns lemmas) outnumber those monosemous.

Table VIII: Statistics of AQAM's Lemma types

Class	Lexical entries			
	Total	without definition	monosemous	polysemous

[23] [26], page 81.

[24] *'requital, recompense, reward (for good deeds);* ***(Isl. Law)*** *merit, credit (arising from a pious deed)* '[25], page 108.

Verbs	1244	122	386	276 (518 senses)
Nouns	2455	-	515	774 (1651 senses)
Adjs	216	8	71	21 (102 senses)

Table X illustrates definition types that can constitute by an explanatory sentence or simple word and we note that:

- more than half of lemmas are defined by explanatory sentence which cannot be automatically dealt and requires manual review and study.
- other items are defined by a simple word that is considered as lemma synonyms and can be automatically translated and mapped.

Table IX: Statistics of definition types

Class	Definition	
	Simple word	Explanatory sentence
Verbs	904	748
Nouns	2166	2918
Adjs	102	152

Table XI illustrates translation rate of verbal, nominal and adjectival lemmas and simple definitions. Note that bilingual dictionary coverage is not total and generally more than half of candidate words (lemmas and simple definitions) are not translated.

Table X: Statistics of AQAM's Lemmas and Senses Translation

Class	Lemma		Senses		Syn
	Lemma	Translated Lemma	Simple word	Translated sense	
Verbs	1244	941	904	477	109
Nouns	2455	1171	2166	1100	200
Adjs	216	59	102	32	3

Table XII illustrates manual control results and percentage of items without semantic correspondence in SUMO and PWN. Many cases of non-correspondence may be explained because meaning or translation is an explanatory sentence which prevents mapping. Other cases are done to different organization of Arabic and English lexicons [27].

Table XI: Manual control result

POS	Controlled Items	non correspondence	
		Total	Rate
Verbs	511	204	40%
Nouns	479	41	8.50%
Adjs	200	56	29.50%

For example, verb mapping statistics illustrated in Table XIII shows that an important number of verbal entries are unmapped.

Concerning triliteral Verbs, non-correspondence may be due to lexical organization between Arabic and English language. We got an important number of unmapped entries, due to the significant amount of translation which followed the form "*to be + participle*" or "*be + adjective*". Whereas they are expressed with a state verb in Arabic, they aren't present in English lexicon. For example, the verb "*karuma*" which means "*to be noble, magnanimous, generous*" isn't mapped in PWN and SUMO.

Derived verbs are also poorly mapped, because they express Arabic derivational concepts that are expressed in English by an entire sentence. For example, "*ta`assaba*" which means "*to wind the turban round one's head*".

As for nouns and adjectives, mapping is more important than for verbs. However, some cases of mismatch are due to conceptual and cultural differences, as already explained in the previous examples (for more details see [27]).

Table XII: Statistics of Mapped AQAM's Verbal Lexical Entries to SUMO

Verb form	Lemma	Translated Lemma	Senses	Translated Senses	Mapped (%)
I	624	551	540	281	33%
II	109	92	46	25	8%
III	36	29	26	10	5%
IV	192	138	101	44	8%
V	77	61	60	43	5%
VI	17	10	10	5	-
VII	27	17	12	8	4%
VIII	19	13	15	8	1%
IX	1	-	-	-	-
X	36	20	26	7	1%
Q_I	71	37	68	52	2%
Q_II	17	6	13	7	2%
Q_III	-	-	-	-	-
Q_IV	16	-	12	7	-

G. *Particular cases of proper names*

The class of proper names is defined by Van Langendonck as follows:

"*A proper name is a noun that denotes a unique entity at the level of 'established linguistic convention' to make it psychosocially salient within a given basic level category [pragmatic]. The meaning of the name, if any, does not (or not any longer) determine its denotation [semantics]. [...] Proper names do not have asserted lexical meaning but do display presuppositional meanings of several kinds: categorical (basic level), associative senses (introduced either via the name bearer or via the name form), emotive senses and grammatical meanings*". ([28], page 06)

In summary, categorical meaning of proper names that pertains to the basic level concepts is the only lexical meaning that proper names seem to have at the level of established linguistic convention ([28], page 86).

So, in this first mapping phase, we defined a tags/triggers List used by the AQAM author allowing to categorize Proper Names. Afterwards, we established for each trigger word the corresponding basic level concept which exists in SUMO. The table XVII in Appendix C contains the list of Proper Names trigger found in the section *bā'*.

Each categorized proper name corresponds to *an Instance Of* the SUMO basic level concept, e.g. man, woman, country, city, river, etc. Then, they are classified in different files in relation to semantic class to which they belong, for example: Astrology, Building, Group, Geographic Location, Human Proper Name, Water.

For example, the key word *ğabal-un* - which means *a mountain*- is a Proper Name trigger used by the AQAM author to designate Proper Noun of Mountain. So the lemma *ğal`ib*of the next lexical entryis mapped by the symbol '@' as an instance of SUMO Concept ***Mountain@***.

```
<root_familyc1="الجِيم"c2="اللَّام"c3="العَيْن"c4="البَاء">
<lexical_entry>
<plain_text>جَلْعِبٌ: جَبَلٌ بِالمَدِينَةِ</plain_text>
<lemma>جَلْعِب</lemma>
<pos>Proper_Noun</pos>
<conceptatt="SUMO"value="Mountain@"/>
</lexical_entry>
</root_family>
```

As shown table XIV, most of extracted Proper Names has been mapped correctly. By analyzing obtained results, our system is able to recognize all Proper Names of Geographic Location, Group, Building and Astrology classes. By cons, concerning Horse and Water classes, we obtained 83% for Precision, because surface patterns are ambiguous and manual revision is required.

Table XIII: Statistics of Proper Names extracted

Class	Precision	Recall	F-Measure
Geographic Location	100%	100%	100%
Group	100%	100%	100%
Building	100%	100%	100%
Astrology	100%	100%	100%
Proper Name	99%	99%	99%
Water	83%	100%	90%
Horse	83%	100%	90%

In XML Example 12, the headword *ḥallāb* is "*[a Proper Noun of]* ***a horse*** *[of the tribe] banī ṯa`lab*." So, the trigger word *faras-un* "*a horse*" indicates that the headword is Proper Noun automatically linked as instance of the SUMO basic level concept ***Horse***.

```
<root_familyc1="الحَاء"c2="اللَّام"c3="البَاء">
<lexical_entry>
<plain_text>حَلَّابٌ: فَرَسٌ لِبَنِي ثَعْلَبَ</plain_text>
<lemma>حَلَّابٌ</lemma>
<pos>Proper_Noun</pos>
<senseatt="فَرَسٌ لِبَنِي ثَعْلَبَ">
<concept>Horse@</concept>
</sense>
<root>ح ل ب</root>
</lexical_entry>
</root_family>
```

XML Example 12. Example of ***Horse*** *as Proper Noun.*

In other cases, Such as in XML Example 13, the system classify the headword *al=ğānib* as *instance of* the ***Horse*** Concept but the surface pattern *faras-un* "*horse*" is part of common sense explanatory sentence, "*[is] a horse that is wide between the legs*".

```
<root_familyc1="الجِيم"c2="النُّون"c3="البَاء">
<lexical_entry>
<plain_text>الجَانِبُ: المُجْتَنَبُ، المَحْقُورُ، وَفَرَسٌ بَعِيدُ مَا بَيْنَ الرِّجْلَيْنِ</plain_text>
<lemma>جَانِب</lemma>
<pos>ADJ</pos>
<senseatt="مُجْتَنَب">
<synonymatt="مُجْتَنَب"/>
</sense>
<senseatt="مَحْقُور">
<synonymatt="مَحْقُور"/>
</sense>
<senseatt="فَرَسٌ بَعِيدُ مَا بَيْنَ الرِّجْلَيْنِ"/>
<root> ج ن ب</root>
</lexical_entry>
</root_family>
```

XML Example 13. Example of ***Horse*** *part of explanatory sentence.*

Manual control is also necessary when confronting the XML Examples 14 where the surface pattern *mā'-un* "*a water*" indicates a Proper Name of *"[a] water (i.e. well) of banī al=`anbari*" and 15 where the same surface pattern is part of explanatory sentence "*what you see [at] noon like [a] water*".

```
<root_familyc1="الهَمْزَة"c2="الزَّاي"c3="البَّاء">
<lexical_entry>
<plain_text>إِزَابٌ، بِالكَسْرِ: مَاء لِبَنِي العَنْبَرِ</plain_text>
<lemma>إِزَاب</lemma>
<pos>Proper_Noun</pos>
<senseatt="مَاء لِبَنِي العَنْبَرِ">
<concept>Water@</concept>
</sense>
<root>ء ز ب</root>
</lexical_entry>
</root_family>
```

XML Example 14. Example of ***Water*** *as Proper Noun*

```
<root_familyc1="السِّين"c2="الرَّاء"c3="البَاء">
<lexical_entry>
<plain_text>اَلسَّرَابُ: مَا تَرَاهُ نِصْفَ النَّهَارِ كَأَنَّهُ مَاءٌ</plain_text>
<lemma>سَرَاب</lemma>
<pos>Noun</pos>
<senseatt="مَا تَرَاهُ نِصْفَ النَّهَارِ كَأَنَّهُ مَاءٌ"/>
<root>س ر ب</root>
</lexical_entry>
</root_family>
```

XML Example 15. Example of Water part of explanatory sentence

CONCLUSION

In this article, we described steps and methodology followed to build a new lexical resource for Classical Arabic based on convert and enrich the Classical Arabic dictionary AQAM into a machine-readable format. The conversion process is based on encoding, segmentation, extraction and represent as possible explicitly lexical information and classifying AQAM entries.

We also described our approach to link the mentioned Arabic Lexicon with external references, such as PWN and SUMO, through an Arabic-English Lexicon. This part of our work has been experimented on the Section of بَاء*bā'*. Cross-Lingual Ontology mapping obtained results between AQAM, PWN and SUMO prove that each language has its specific linguistic environment and cultural context. So this makes it necessary to add concepts that takes into consideration historical and cultural aspects of Arabic language. Then, results are deposited in *CLARIN-IT*infrastructure[25].

[25] http://hdl.handle.net/20.500.11752/ILC-97

As perspective, for the AQAM conversion level we plan to control and optimize used based-rules and regular expressions in information extraction, then check and validate more automatically generated results.

Recall that the designed resource is coded in a generic XML format, and in[16] and [29]we presented two conversion studies of AQAM in Standard format as LMF (Lexical Markup Framework) and LEMON (LExicon Model for ONtology), revealing that these Standards do not cover all Arabic language specificity, such as for their Morphological module which needs to be extended in order to represent some Arabic-specific morphological features (maṣdar, imperfect, scheme, root, etc). Thus, this pushes us to think about using TEI-encoding which is more convivial for historical data.

Furthermore, concerning semantic level, we intend to continue enrichment of the lexical resource, and using other bilingual resources based on classical Arabic to get more coverage and rise mapping rate. Rightly, disambiguation is complicated and time-consuming; we need to find ways to help solve control process, by using some knowledge-based semantic similarity measurement techniques.

Finally, we are considering keeping enriching our lexical resource, for purpose of using it in other research projects.

APPENDIX A

USED LIST OF NOMINATIVE AND TRANSITIVE SUFFIXES

Table XIV: Nominative and Transitive Suffixes

Nominative Suffixes		**Transitive Suffixes**	
Arabic Form	**BuckwalterTransliteration**	**Arabic Form**	**BuckwalterTransliteration**
◌َ	*a*	هُ	*hu*
◌َتْ	*ato*	هَا	*haA*
◌َتِ	*ati*	هُمْ	*humo*
◌ْتُ	*otu*	هُمَا	*humaA*
◌ْتَ	*ota*	هُنَّ	*hun~a*
◌ْتِ	*oti*	كَ	*ka*
◌ْتُنَّ	*otun~a*	كِ	*ki*
◌ْتُمْ	*otumo*	كُمَا	*kumaA*
◌ْنَا	*onaA*	كُنَّ	*kun~a*
◌ُوا	*uwA*	كُمْ	*kumo*
◌َا	*aA*	نَا	*naA*
-	-	نِي	*niy*

APPENDIX B

USED LIST OF SOLAR LETTERS

Table XV: SOLAR LETTERS

Script	**Transliteration**	**Letter**	**Transliteration**
ت	*t*	التَّاء	*tā’*
ث	*ṯ*	الثَّاء	*ṯā'*
د	*d*	الدَّال	*dāl*
ذ	*ḏ*	الذَّال	*ḏāl*
ر	*r*	الرَّاء	*rā'*
ز	*z*	الزَّاي	*zāy*
س	*s*	السِّين	*sīn*
ش	*š*	الشِّين	*šīn*
ص	*ṣ*	الصَّاد	*ṣād*
ض	*ḍ*	الضَّاد	*ḍād*
ط	*ṭ*	الطَّاء	*ṭā'*
ظ	*ẓ*	الظَّاء	*ẓā'*
ل	*l*	اللَّام	*lām*
ن	*n*	النُّون	*nūn*

APPENDIX C

Example of some keywords for Proper Names extraction

Tags	SUMO basic level concept	Semantic Class
اسْمٌ - صحابيٌّ - صحابِيَّاتٌ - صحابِيَّة - شاعرٌ - محدثٌ - (مَلِكٌ ل) رجلٌ - أبو قَبيلة - أسماءٌ - مِن أسْمائِهنَّ - لَقَبُ - لُقِّبُأبو قَبيلَةٍ - امْرَأَةٌ - [اسمُ رَجُلٍ] - [والِدُ + مضاف إليه] - شيخٌ - قارِئٌ - (بنتُ + مضاف إليه) - (أبو بَطْنٍ) - مُتَكَلِّمٌ - كُنْيَةُ - جَدٌّ - زاهِدٌ - عَلَمٌ - مَلِكٌ - فَرْدٌ - مَوْلىً - مُؤَرِّخٌ - تابِعِيٌّ - أَخْبارِيٌّ	Human	Proper Name
قَبِيلَةٌ - جَمَاعَةٌ - بَطْنٌ - قَبائِلُ - بُطُونٌ - قَوَمَةٌ	Group	Group
د – بَلَدٌ ة - قَرْيَةٌ - قَرْيَتَانِ - [كُورَةٌ] - مَكَّةُ - اسْمُ كُورَةٍ + مَدِينَةُ ع - مَوْضِعٌ - مَوَاضِعُ - مَوْضِعَانِ - المَكَانُ - نَاحِيةٌ (بَيْنَ – قُرْبَ - وَرَاءَ - مِنْ) - أَرْضٌ - مَحَلَّةٌ - جَبَلٌ - جَبَلانِ - أجْبُلٌ - وجِبالٌ حَجَرٌ - صَخْرَةٌ هَضَبَة - هِضَابٌ	Nation City Region Neighborhood Mountain Rock Hill	Geographic
مَاءٌ - [مَاءٌ لِ] - مَاءَةٌ - مُوَيْهَةٌ - مِيَاهٌ - مَسِيلٌ بِئْرٌ - الآبَارُ نَهْرٌ - وادٍ - وادِيانِ سَبَخَةٌ	Creek MineOrWell River Swamp	Water
نَباتٌ - نَبْتٌ - عُشْبَةٌ - بَقْلَةٌ - مايَنْبُتُ شَجَرٌ - [شُجَيْرَةٌ] - كُلُّ شَجَرٍ - نَخْلٌ	Plant Botanical Tree	Plant
طَعَامٌ - [مِن طَعَامٍ] شَرَابٌ - [شَرَابٌ مِن] تَمْرٌ - مِن تَمْرٍ	Food Beverage Date Fruit	Food
النَّحْلُ سَمَكٌ الذِّئْبُ - [الذِّئْبُ + نعتُ] - اسْمُ الذِّئْبِ فَرَسٌ م - فَرَسٌلِ-أفْراسٌلِ- [فَرَسُ + مضاف إليه] - اسْمُ فَرسٍ... قُنفُذٌ - دُوَيْبَةٌ ...	Bee Fish Canine Horse Animal	Animal
داءٌ - بَثْرٌ – مَرَضٌ دَواءٌ	Disease Or Syndrome Biologically Active Substance	Disease
لُعْبَةٌ شَهْرٌ اسمُ جِنِّيٍّ - قَبيلَةٌ مِنَ الجِنّ - اسْمُ الشَّيْطانِ كَوْكَبٌ ثَوْبٌ	Game Time Interval Cognitive Agent Planet Clothing	Others

References

[1] A. M. Al-Zoghby, A. Elshiwi, and A. Atwan, “Semantic Relations Extraction and Ontology Learning from Arabic Texts---A Survey,” in *Intelligent Natural Language Processing: Trends and Applications*, K. Shaalan, A. E. Hassanien, and F. Tolba, Eds. Cham: Springer International Publishing, 2018, pp. 199–225.

[2] K. Shaalan, “A Survey of Arabic Named Entity Recognition and Classification,” *J. Comput. Linguist.*, vol. 40, no. March 2013, pp. 469–510, 2014, doi: 10.1162/COLI

[3] S. Alqrainy, H. Muaidi, and M. S. Alkoffash, “Article: Context-Free Grammar Analysis for Arabic Sentences,” *Int. J. Comput. Appl.*, vol. 53, no. 3, pp. 7–11, 2012.

[4] S. Elkateb, W. Black, H. Rodríguez, M. Alkhalifa, P. Vossen, A. Pease,and C. Fellbaum, “Building a WordNet for Arabic,” in *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06)*, 2006

[5] H. Rodríguez, D. Farwell, J. Farreres, M. Bertran, M. Alkhalifa, M. A.Martí, W. Black, S. Elkateb, J. Kirk, A. Pease, P. Vossen, and C. Fellbaum, “Arabic WordNet: Current state and future extensions,” in *The fourth global WordNet conference*, 2008, pp. 387–405.

[6] M. AlKhalifa and H. Rodríguez, “Automatically extending NE coverage of Arabic WordNet using Wikipedia,” in *the 3rd international conference on Arabic language processing CITALA’09*, 2009.

[7] H. Rodríguez, D. Farwell, J. Ferreres, M. Bertrán, M. Alkhalifa, and M. A. Martí, “Arabic WordNet: Semi-automatic Extensions using Bayesian Inference,” in *the the 6th Conference on Language Resources and Evaluation LREC2008*, 2008.

[8] R. Del Gratta and O. Nahli, “Enhancing Arabic WordNet with the use of Princeton WordNet and a bilingual dictionary,” in *2014 Third IEEE International Colloquium in Information Science and Technology (CIST)*, 2014, pp. 278–284, doi: 10.1109/CIST.2014.7016632.

[9] Y. Regragui, L. Abouenour, F. Krieche, K. Bouzoubaa, and P. Rosso, “Arabic WordNet: New Content and New Applications,” in *Proceeding of the 8th Global Wordnet Conference (GWN 2016)*, 2016.

[10] S. Boudelaa,W.D Marslen-Wilson, "Aralex: A lexical database for Modern Standard Arabic". *Behavior Research Methods. vol.* 42, pp. 481–487. 2010. https://doi.org/10.3758/BRM.42.2.481.

[11] M. Attia, P. Pecina, A. Toral, L. Tounsi, and J. van Genabith, “A Lexical Database for Modern Standard Arabic Interoperable with a Finite State Morphological Transducer,” in *Systems and Frameworks for Computational Morphology*, 2011, pp. 98–118.

[12] D. Namly and K. Bouzoubaa, “LMF conversion of an editorial dictionary: the case of the Contemporary Arabic dictionary,” in *Journée d’étude Ressources langagières de l’arabe pour le TAL : construction, standardisation, gestion et exploitation*, 2015.

[13] M. Alkhatib, A. A. Monem, and K. Shaalan, “The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications,” *Procedia ComputerScience.*, vol. 117, no. 2, pp. 759–776, 2017, doi: https://doi.org/10.1007/s13369-017-2737-2.

[14] T. Zerrouki and A. Balla, “Tashkeela: Novel corpus of Arabic vocalized texts, data for auto-diacritization systems,” *Data inBrief.*, vol. 11, pp. 147–151, 2017, doi: 10.1016/j.dib.2017.01.011.

[15] N. Ide and J. Véronis, “Encoding Dictionaries,” in *Text Encoding Initiative: Background and Context*, N. Ide and J. Véronis, Eds. Dordrecht: Springer Netherlands, 1995, pp. 167–179.

[16] O. Nahli, F. Frontini, M. Monachini, F. Khan, A. Zarghili, and M. Khalfi, “Al Qamus al Muhit, a Medieval Arabic lexicon in LMF,” *inProceedingsof the Tenth International Conference on Language Resources andEvaluation (LREC 2016).* Paris, France: European Language ResourcesAssociation (ELRA), may 2016, pp. 943–950.

[17] R. Duwairi, M. Al-Refai, and N. Khasawneh, “Stemming Versus Light Stemming as Feature Selection Techniques for Arabic Text Categorization,” in *2007 Innovations in Information Technologies (IIT)*, 2007, pp. 446–450.

[18] M. D. al-fīrūz ‘ābādī, "*al=qāmūs al=muḥīṭ*", 8th edition. Beirut: mu’assasat ar-risālah, 1998.

[19] M. az=zabīdī, "*tāğ al=’arūs min ğawāhir al=qāmūs*". Maṭba`at al-Kuwayt, 2001.

[20] C. Fellbaum, “WordNet: An Electronic Lexical Database,” *MIT Press*, 1998.

[21] G. A. Miller, “WordNet: A Lexical Database for English,” *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995, doi: 10.1145/219717.219748.

[22] B. Ladislav, D. Petr, and O. Vojtech, “Knowledge Base Modeling and Design Procedure,” in *Information Modelling and Knowledge Bases XXIII*, H. Jaak, K. Yasushi, T. Takehiro, J. Hannu, and Y. Naofumi, Eds. IOS Press, 2012, pp. 331–343.

[23] A. Pease, I. Niles, and J. Li, “The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications,” *AAAI Tech. Rep. WS-02-11*, 2002.

[24] I. Niles and A. Pease, “Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology,” in *PROCEEDINGS OF THE 2003 INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE ENGINEERING (IKE 03), LAS VEGAS*, 2003, pp. 412–416.

[25] H. Wehr, “A Dictionary of Modern Written Arabic,” 3rd Editio., J. M. COWAN, Ed. Spoken Language Service, Inc, 1976, p. XI.

[26] H. A. Salmoné, "*An advanced learner’s Arabic-English dictionary": incl. an Engl. index*. Libr. du Liban, 1978.

[27] O. Nahli, “Arabic Language Alignment with English Ontologies: Some Ontological Reflections,” in *5th {IEEE} International Congress on Information Science and Technology, CiSt 2018, Marrakech, Morocco, October 21-27, 2018*, 2018, pp. 254–260, doi: 10.1109/CIST.2018.8596580.

[28] W. van Langendonck, "*Theory and Typology of Proper Names"*. ser.Studies and monographs. Bod Third Party Titles, 2007.

[29] M. Khalfi, O. Nahli, and A. Zarghili, “Classical dictionary Al-Qamus in lemon,” *Colloq. Inf. Sci. Technol. Cist*, vol. 0, pp. 325–330, 2016, doi: 10.1109/CIST.2016.7805065.

[30] R. Ba`labakkī, "*The Arabic Lexicographical Tradition: From the 2nd/8th to the 12th/18th Century"*. ser. Handbook of Oriental Studies. Brill, 2014.