Building a standardized Wordnet in the ISO LMF for aeb language

Nadia B.M. Karmani(1) nadia.karmani.tn@ieee.org
Hsan Soussou(2) hsan.soussou@gmail.com
Adel M. Alimi(1) adel.alimi@ieee.org
(1) REGIM: REsearch Groups on Intelligent Machines
University of Sfax, National Engineering School of Sfax (ENIS)
BP 1173, Sfax, 3038, Tunisia
(2) MD Soft, Tunisia

Abstract

Internet communication plays a considerable part in economic, financial and even politic domains. It is greatly influencing the politic revolution of many Arabic countries. That allows Internet communication to take more and more scale especially in an Arabic context. In this case, we notice that Internet communication is based on textual interchange using Arabic dialects more than Arabic language. However, few efforts were made for Arabic dialect processing particularly for aeb[1] language. In this case, we suggest building a standardized aeb Wordnet, which is a basic tool for Natural Language Processing (NLP) of aeb language. In this article, we present an extended Wordnet-LMF model acquired to aeb language specificities used to represent aeb Wordnet and we describe building steps.

1 Introduction

Wordnet, firstly developed for English language, cover newer days many others languages and even dialects. In an Arabic case, many efforts were make to build a Wordnet for Modern standard Arabic but no real attempt has been made for Arabic dialects.

Arabic dialects represent Arabic language variations often spoken. However, they are written in some press articles, theater pieces, poetic books and Internet based communication such as email, instant messaging, forums, blogs, social networks, etc.

With the politic revolution of several Arabic countries like Tunisia (i.e. also Egypt, Syria, etc.), Arabic dialect processing takes more and more scale in Arabic countries and particularly in Tunisia.

In this case, we suggest building a powerful semantic lexicon for Tunisian dialect (aeb language): aeb Wordnet using the expand approach which is formulated according to an adapted format of Wordnet-LMF. The use of standardized Lexical Markup Framework (LMF) ISO 24613 format allows interchange between aeb Wordnet and other standardized lexicons.

2 Challenges

Developing an aeb wordnet faces many constraints associated to resources, language characteristics and use.

Generally, building a Wordnet needs a lot of resources. But, for aeb language few written resources are found: an electronic bilingual dictionary eng[2]-aeb, some press articles, some theater pieces, some poesy books, etc. Indeed, aeb like other Arabic dialects is sometimes written and it's not educated.

In addition to the lack of resources, we notice many language specificities: absence of standard transcription, use of six variations (i.e. Tunis, Sahel, Sfax, occidental north, occidental south and oriental south) and estrangement from English language and even from Arabic language. Indeed, aeb language is characterized by the absence of standard transcription: the same word can be represented by different transcriptions e.g. the word [إِتْوَقَّعْ /ʔitwaqqaʔˤ/[3]] "anticipate" can be transcribed as [إِتْوَقَّعْ /ʔitwaqqaʔˤ/] or [تْوَقَّعْ /twaqqaʔˤ/]. Also, it uses six variations e.g. the variations of the personal pronoun "I" are illustrated by the Table 1.

[1] aeb is the ISO 639-3 language code for Tunisian Arabic.

[2] eng is the ISO 639-3 language code for English.

[3] phonetic transcription according to International Phonetic Alphabet (IPA).

aeb variations					
Tunis	Sahel	Sfax	Occidental north	Occidental south	oriental south
آنَا	آنِي	آنَا	نَا	أنَا	أنِي

Table 1. Variation of the personal pronoun [آنَا/ʔa:na:/] "I"

In addition, aeb is a Semitic language like Arabic very different from English at morphological, lexical and syntactical levels and also different from Arabic seeing that its alphabet counts three consonants which aren't used in Arabic (i.e. [ڤ /v/],[ڤ /q'/] and [پ /P/]), its lexicon is full of foreign words (e.g. the word [دَاكُورْدُو /da:ku:rdu:/] "all right" is borrowed from Italian language) and it uses Arabic roots to express other meanings (e.g. the Arabic root [خدم/xdm/] meaning "to serve" is used in aeb language as [خْدِمْ/xdim/] to express "to work").

Also, the use of aeb language raises other constraints. Indeed, aeb use covers spoken and written forms. The last form can be diacritized, not diacritized or partially diacritized e.g. the word [إِتْوَقَّعْ] "anticipate" can be transcribed as [إِتْوَقَّعْ], [إتوقع] [إِتْوقَّعْ]. It can be also scripted with Arabic, Latin or a mixed script e.g. the word [إِتْوَقَّعْ] "anticipate" can be transcribed as [إِتْوَقَّعْ], [etwa99a3] or [etو993].

3 Wordnet-LMF

Towards generation of a standard model representing lexicons, many works are made around LMF. Wordnets, seen that they are considered as semantic lexicons, can use LMF. They precisely can use Wordnet-LMF formed by the components described below. These components are not sufficient to express correctly aeb particularities such as the use of many transcriptions for the same lexical entry, the variation, the phonetic sight or the inflected and derived forms. So, we add others LMF components as extension.

3.1 Components

Wordnets, all over the world, share the same basic concepts (i.e. word, verb, noun, adjective, adverb, synset, etc.) and organization (i.e. sets of synonyms, each representing a lexicalized concept (Miller, 1995)) but they have different representations. Some efforts were made, thought the project Knowledge-Yielding Ontologies for Transition-Based Organization (KYOTO[4]), to propose a standardized model for Wordnets: KYOTO-LMF or Wordnet-LMF. This model is an LMF dialect. It is a Wordnet adapted version of the common standardized framework for representing natural language processing (NLP) lexicons: ISO 24613 LMF (Soria and Monachini, 2008).

This model is composed of twenty one elements (Soria et al., 2009). LexicalResource is the root element with three children representing general information (i.e GlobalInformation), the lexicon associate to a defined language (i.e. Lexicon) and a bracketing element grouping together SenseAxis (i.e. SenseAxe). The root element describes the resource that can be a monolingual or a multilingual Wordnet. The other elements can be distributed over three different packages, i.e. the morphological, the NLP semantic and the NLP multilingual notations package.

The morphological package contains five elements describing a lexeme in a given language (i.e. LexicalEntry), a word that can be a root, a stem, an inflected form or a multiword expression (i.e. Lemma), one meaning of a LexicalEntry (i.e. Sense), a link between a Sense and another resource (i.e. Monolingual-ExternalRef) and a bracketing element grouping together MonolingualExternalRef (i.e. Mono-lingualExternalRefs).

The NLP semantic package is formed by seven elements representing a set of shared meanings within the same language (i.e. Synset), the gloss associate with one synset (i.e. Definition), an example of use associate to one synset (i.e. Statement), a relation between synsets (i.e. SynsetRelation), a bracketing element grouping together RelationSynset (i.e. RelationSynsets), a link between a Synset and another resource (i.e MonolingualExternalRef) and a bracketing element grouping together MonolingualExternalRef (i.e. Monolingual-ExternalRefs).

And finally, the NLP Multilingual notations package containing four elements used only to describe multilingual Wordnets.

This model contains also an element describing administrative information (i.e. Meta) used with LexicalEntry, MonolingualExternal-Ref, Synset and SynsetRelation.

3.2 Wordnet-LMF vs aeb language

Wordnet-LMF is a model adopted, in the project

[4] KYOTO (project nr. 211423) FP7-ICT-2007-1

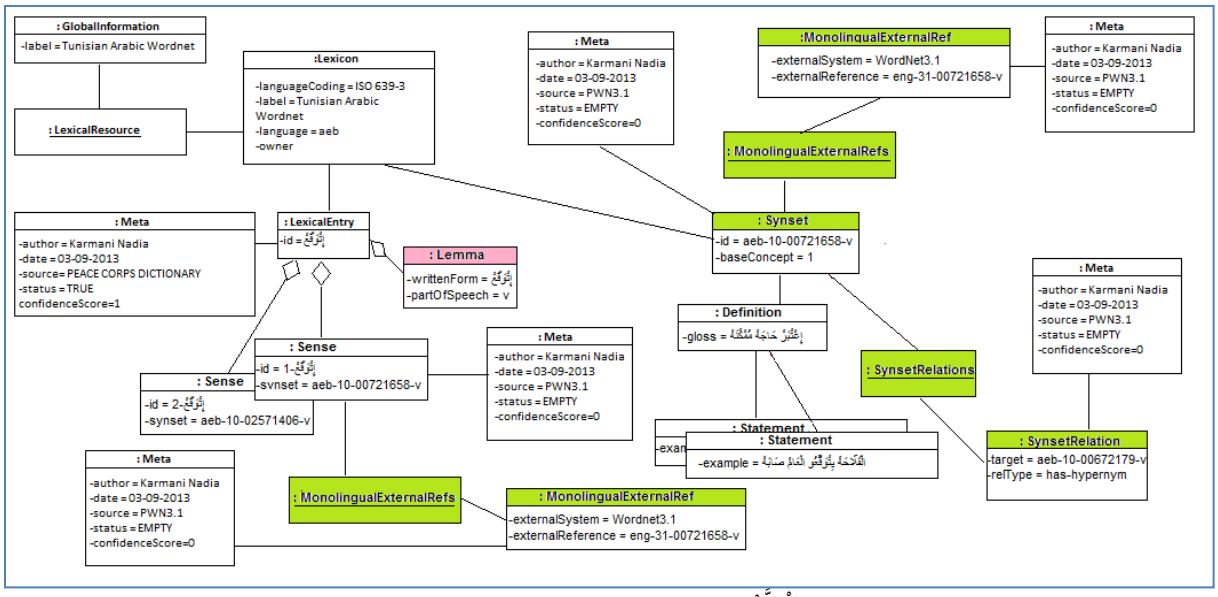


Figure 1. Wordnet-LMF object diagram of the word [إِتْوَقَّعْ /ʔitwaqqaʔˤ/] "anticipate"

KYOTO, to represent Wordnets of English, Dutch, Italian, Basque, Spanish, Chinese and Japanese (Soria et al., 2009). These languages are different from aeb language. So, the use of Wordnet-LMF to represent aeb language can't preserve the language specificities.

The Figure 1 represents the Unified Modeling Language (UML[5]) object diagram of Wordnet-LMF associated to the word [إِتْوَقَّعْ /ʔitwaqqaʔˤ/] "anticipate". It illustrates the limits of Wordnet-LMF for aeb language description. Indeed, Wordnet-LMF model doesn't express the use of many transcriptions for the same lexical entry, the variation, the phonetic and the inflected forms of a lexical entry and the structure of Semitic languages (i.e. derivation phenomena).

3.3 Extension

To express properly the aeb language specificities, we suggest extending Wordnet-LMF using ISO LMF.

In this case, we firstly propose to replace the cardinality "1..1" of the association between LexicalEntry and Lemma by the cardinality "1..*".That allows the affection of more than one Lemma to the same LexicalEntry as it is shown in Figure 3, e.g. the lexicalEntry [إِتْوَقَّعْ /ʔitwaqqaʔˤ/] "anticipate" has two lemmas illustrated by the figure below.

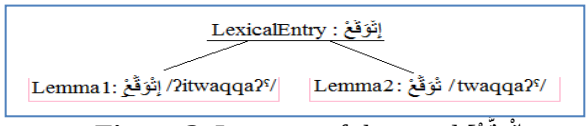


Figure 2. Lemmas of the word [إِتْوَقَّعْ /ʔitwaqqaʔˤ/] "anticipate"

Secondly, we propose to add two attributes for the entity Lemma: script and orthographyName, seen that aeb transcription uses Arabic or Latin script.

Finally, we suggest adding three ISO LMF elements: FormRepresentation, WordForm and RelatedForm to represent respectively the phonetic and the variation, the inflected and the derived forms of a lexical entry (ISO 24613, 2008). E.g. these elements are integrated for the word [إِتْوَقَّعْ /ʔitwaqqaʔˤ/] "anticipate" like it is shown in the Figure 3.

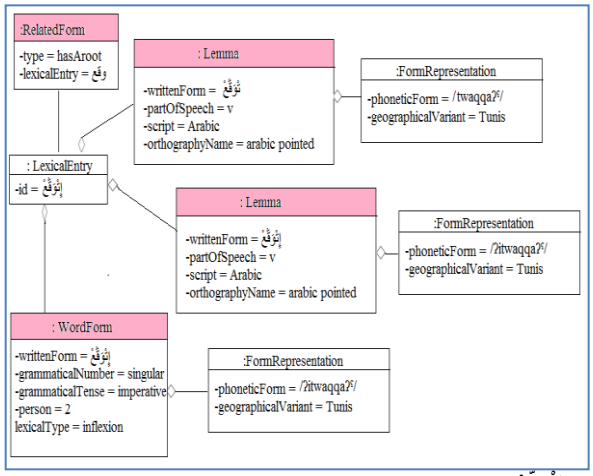


Figure 3. Object diagram of the word [إِتْوَقَّعْ /ʔitwaqqaʔˤ/] "anticipate"

[5] UML is a standardized (ISO/IEC 19501:2005), general-purpose modeling language in the field of software engineering (Wikipedia).

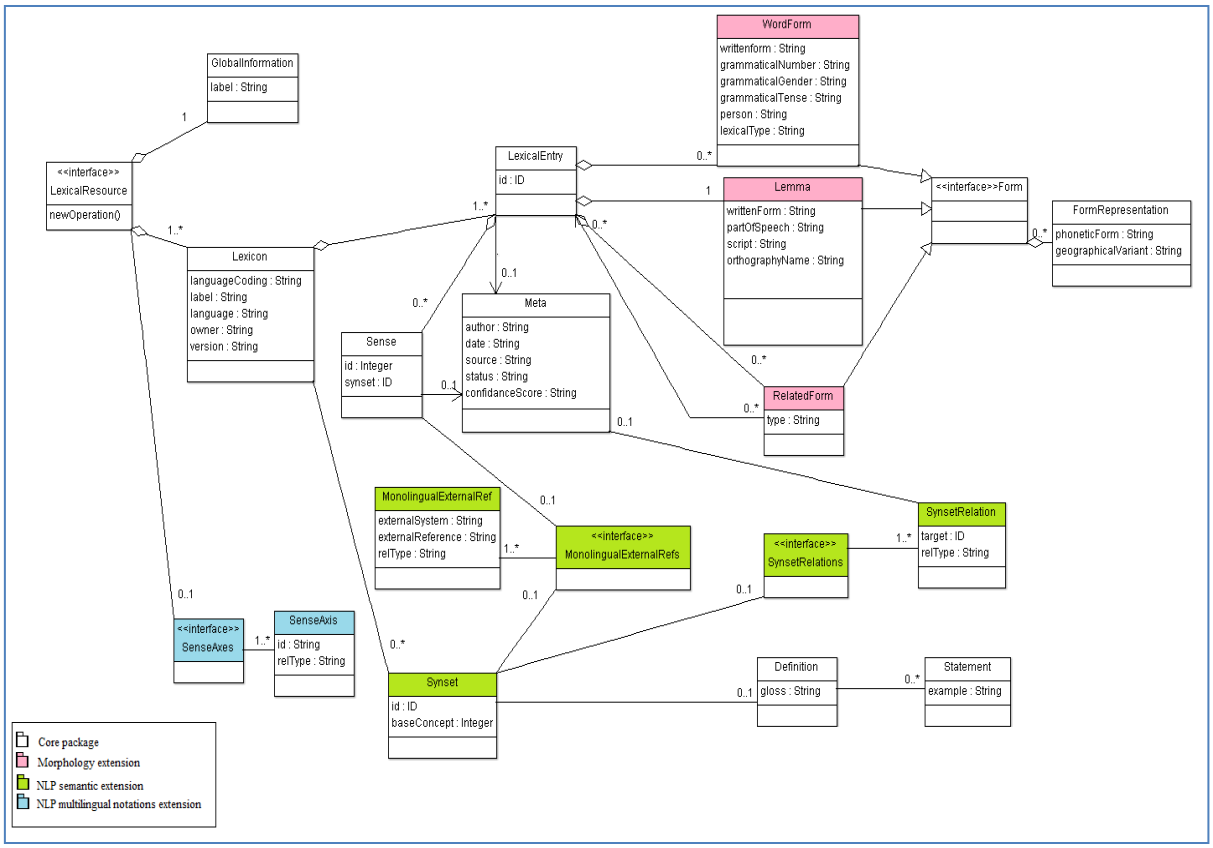


Figure 4. Extended Wordnet-LMF model for aeb language

Consequently, we get the Wordnet-LMF extended model accurate to aeb language shown by UML class diagram in the Figure 4.

4 aeb Wordnet construction

In general, building Wordnet can be done by merged or expanded approach (Vossen, 1998). The merged approach consists on the creation of synsets and synset relations using language resources. But, the expand approach generates synsets and synset relations from the widely used WordNet: Princeton WordNet (PWN) by translation.

The first approach save the language specificities but it is complex and need a lot of language resources. The second one is easy and needs only PWN and bilingual dictionaries but it generates a biased Wordnet to PWN.

In the case of aeb Wordnet development, the first approach can't be used because of the lack of aeb resources. So, we adopt the expanded approach. We use PWN 3.1 released in 2011 and the only bilingual dictionary found for English and Arabic Tunisian: Peace Corps dictionary of Rached Ben Abdelkader, Abdeljelil Ayed and Aziza Naouar edited in July 1977 listing about 6000 aeb words.

Aeb Wordnet is developed manually, with the format XML[6], throw three steps: creation, validation and extension.

4.1 Creation

Wordnet is a set of lexical entries L= {l} and a set of synsets S={s}. A lexical entry l is composed of one word w∈W at least, which can be a Lemma or a WordForm. A synset s is composed of a subset of lexical entries L' and a set of synset relations R={r}.

To generate aeb Wordnet (L_{aeb}, S_{aeb}) we use both PWN and Peace Corps dictionary D. Indeed, for every translation t∈D, we generate a subset of lexical entries L'_{aeb} and a subset of synsets S'_{aeb}.

A translation t can be monosemous (t = (w_{eng}, w_{aeb})), divergent polysemous (t= {(w_{eng}, $w1_{aeb}$), (w_{eng}, $w2_{aeb}$), {(w_{eng}, $w3_{aeb}$) …}) or represent a lexical lacuna (t= (w_{eng}, ø)).

In the first case, the translation t = (w_{eng}, w_{aeb}) generates one lexical entry l_{aeb} for w_{aeb}

[6] XML is a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable (Wikipedia).

considered as a lemma and a subset of synset S'_{aeb} like it is presented in the Figure 5. $S'_{aeb}=\{s\ (L'_{aeb}, R)\}$ is equivalent to $S'_{pwn}=\{s\ (L'_{pwn}, R)\}$ i.e. synsets of w_{eng} in PWN and L'_{aeb} is obtained from the translation of words in L'_{pwn} using D.

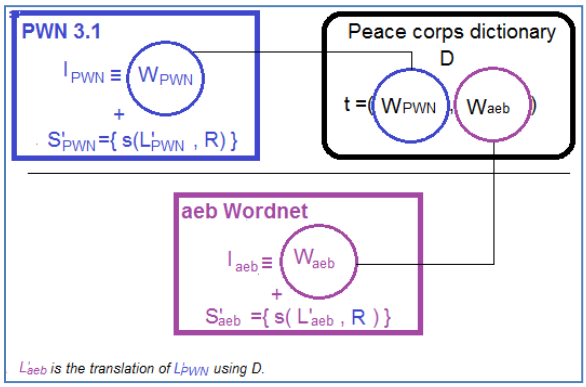


Figure 5. Monosemous translation

E.g. the translation t1= {("anticipate", "إِتْوَقَّعْ")} generates the lexical entry l_{aeb} described by the Figure 3 with six senses (i.e. equivalent to the english word "anticipate" senses) as it is shown below.

```
<LexicalEntry id="إِتْوَقَّعْ">
    …
    <Sense id="1_إِتْوَقَّعْ" synset="aeb-10-00721658-v">
    <Meta author="Karmani Nadia" date="2013-09-03"
source="PWN3.1" status="EMPTY"/>
    <MonolingualExternalRefs>
    <monolingualExternalRef externalSystem="WordNet 3.1"
externalReference="eng-31-00721658-v" />
    </MonolingualExternalRefs>
    </Sense>
    <Sense id="2_إِتْوَقَّعْ" synset="aeb-10-02571406-v">
    <Meta author="Karmani Nadia" date="2013-09-03"
source="PWN3.1" status="EMPTY"/>
    <MonolingualExternalRefs>
    <monolingualExternalRef externalSystem="WordNet 3.1"
externalReference="eng-31-02571406-v" />
    </MonolingualExternalRefs>
    </Sense>
    <Sense id="3_إِتْوَقَّعْ" synset="aeb-10-00722732-v">
    <Meta author="Karmani Nadia" date="2013-09-03"
source="PWN3.1" status="EMPTY"/>
    <MonolingualExternalRefs>
    <monolingualExternalRef externalSystem="WordNet 3.1"
externalReference="eng-31-00722732-v" />
    </MonolingualExternalRefs>
    </Sense>
    <Sense id="4_إِتْوَقَّعْ" synset="aeb-10-00919743-v">
    <Meta author="Karmani Nadia" date="2013-09-03"
source="PWN3.1" status="EMPTY"/>
    <MonolingualExternalRefs>
    <monolingualExternalRef externalSystem="WordNet 3.1"
externalReference="eng-31-00919743-v" />
    </MonolingualExternalRefs>
    </Sense>
    <Sense id="5_إِتْوَقَّعْ" synset="aeb-10-01808928-v">
    <Meta author="Karmani Nadia" date="2013-09-03"
source="PWN3.1" status="EMPTY"/>
    <MonolingualExternalRefs>
    <monolingualExternalRef externalSystem="WordNet 3.1"
externalReference="eng-31-01808928-v" />
    </MonolingualExternalRefs>
    </Sense>
    <Sense id="6_إِتْوَقَّعْ" synset="aeb-10-00343295-v">
    <Meta author="Karmani Nadia" date="2013-09-03"
source="PWN3.1" status="EMPTY"/>
    <MonolingualExternalRefs>
    <monolingualExternalRef externalSystem="WordNet 3.1"
externalReference="eng-31-00343295-v" />
    </MonolingualExternalRefs>
    </Sense>
</LexicalEntry>
```

Also, it creates six synsets in S'_{aeb}={aeb-10-00721658-v, aeb-10-02571406-v, aeb-10-00722732-v, aeb-10-00919743-v, aeb-10-01808928-v, aeb-10-00343295-v} .The synset $s_{aeb\text{-}10\text{-}00721658\text{-}v}$ is detailed below.

```
<Synset id="aeb-10-00721658-v" baseConcept="1">
<Meta author="Karmani Nadia" date="2013-09-03"
source="PWN3.1" status="EMPTY"/>
<Definition gloss="إِعْتَبَرْ حَاجَهْ مُمْكَنَهْ" ><Statement example=" الْفَلَاحَهْ
يِتْوَقْعُو الْعَامْ صَابَهْ"/></Definition>
<SynsetRelations>
    <synsetRelation target="aeb-10-00672179-v" relType="has-
hypernym"><Meta author="Karmani Nadia" date="2013-09-03"
source="PWN3.1" status="EMPTY"/>
    </synsetRelation>
    …
    </SynsetRelations>
    <MonolingualExternalRefs>
    <monolingualExternalRef externalSystem="WordNet3.1"
externalReference="eng-31-00721658-v" />
    </MonolingualExternalRefs>
    </Synset>
```

In the second case, the translation t= {(w_{eng}, $w1_{aeb}$), …, (w_{eng}, wn_{aeb})} shown in Figure 6 generates a subset of lexical entries L'_{aeb} (i.e. L'_{aeb} = {$l1_{aeb}$, …, ln_{aeb} } /$l1_{aeb}$ and ln_{aeb} have respectively $w1_{aeb}$ and wn_{aeb} as Lemmas) and a subset of synset S'_{aeb}= {s (L'_{aeb}, R)} equivalent to S'_{pwn}={s (L'_{pwn}, R)}.

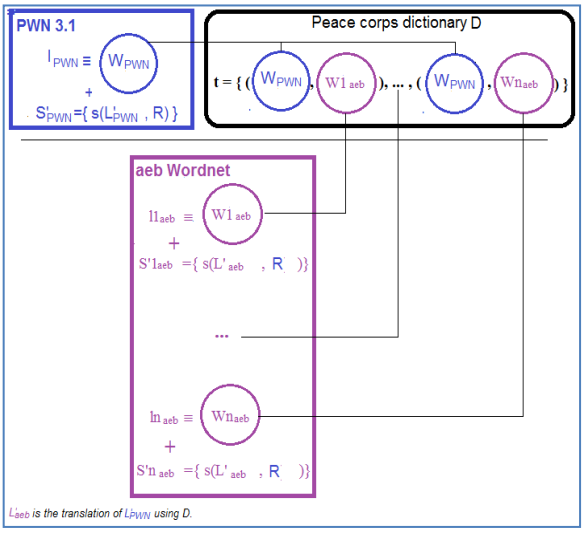


Figure 6. Polysemous translation

The subset of synsets S'_{aeb} includes synsets of the lexical entries in $L' = \{ l1_{aeb}, \dots, ln_{aeb}\}$/ $S'_{aeb} = S'l1_{aeb} \cup (S'l2_{aeb} - S'l1_{aeb} \cap S'l2_{aeb}) \cup \dots \cup (S'ln_{aeb} - S'ln\text{-}1_{aeb} \cap S'ln_{aeb})$ e.g. the translation t2={("work", "خْدِمْ"), ("work"," خَدِّمْ ")} generates L'_{aeb}={"خْدِمْ","خَدِّمْ"} composed of two lexical entries and S'={aeb-10-02418610-v, aeb-10-02415985-v, aeb-10-02531113-v, aeb-10-01528454-v , aeb-10-02449024-v , aeb-10-00100305-v, aeb-10-02413117-v , aeb-10-02441810-v, aeb-10-02121463-v, etc} illustrated by the Figure 7.

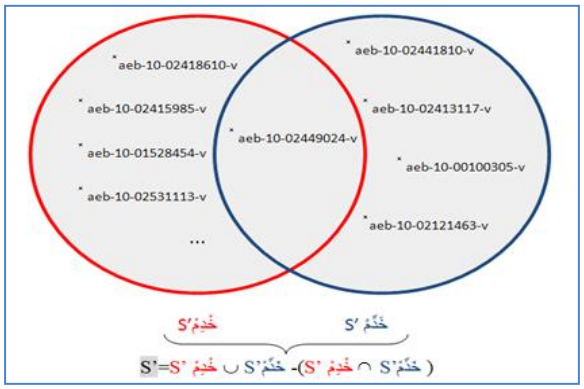


Figure 7. Synsets distribution between lexical entries generated from the translation t2

Finally, the third case presented in Figure 8 doesn't affect aeb Wordnet e.g. the translation t3= {("fir", ø)}.

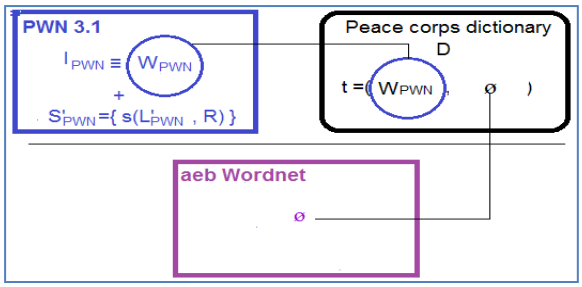
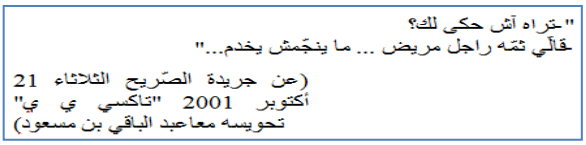


Figure 8. Lexical lacuna

4.2 Validation

Some aeb Wordnet elements need validation through or after creation. LexicalEntry is validated through creation but Sense, Synset and SynsetRelation are validated after creation. The validation consists on the affection of "True" value to the attribute status of Meta element. We validate an element when we find it in a confident resource such as dictionary, press article, theater piece, poetic book, etc.


"تراه آش حكى لك؟
قالّي تمّه راجل مريض ... ما ينجّمش يخدم..."
(عن جريدة الصّريح الثلاثاء 21 أكتوبر 2001 "تاكسي ي ي"
تحويسه معاعبد الباقي بن مسعود)

E.g. from the press article تحويسه معا عبد الباقي بن مسعود from الصّريح October 2001 at the top, we validate the Synset s aeb-10-00590283-v and the Sense 2_خْدِمْ of the lexicalEntry خْدِمْ as it is shown below.

```
<Synset id=" aeb-10-02415985-v" baseConcept="1">
<Meta author="Karmani Nadia" date="2013-09-03" source="
تحويسه معا عبد الباقي بن مسعود from الصّريح October 2001"
status="TRUE"/>
…
</Synset>
```

```
<LexicalEntry  id="خْدِمْ ">
    …
    <Sense id="2_خْدِمْ" synset="aeb-10-02415985-v">
    <Meta author="Karmani Nadia" date="2013-09-03"
source="PWN3.1" status="TRUE"/>
    <MonolingualExternalRefs>
    <monolingualExternalRef externalSystem="WordNet3.1"
externalReference="eng-31-02415985-v" />
    </MonolingualExternalRefs>
    </Sense>
    …
    <LexicalEntry>
```

4.3 Extension

To create aeb Wordnet, we use Peace Corps dictionary containing about 6000 aeb words used in Tunis. This potential cannot be compared to PWN 3.1 potential counting about 147278 eng words[7]. It represents 24.54% of PWN 3.1 potential.

In this case, we suggest enriching aeb Wordnet lexicon by derivation, by variation and by corpus.

The first method consists on the generation of derived forms when it is possible (i.e. when the word is derivative, not fixed or borrowed). Indeed, aeb language is a Semitic language like Arabic. So, from a root we can build many words according to defined patterns e.g. from the root [شرب /ʃrab/] "to drink" we can generate five direct derived nouns like it is shown in the Table2 (Mejri et al, 2009).

Root [شرب / ʃrab/]				
Patient	**Predicative**	**Superlative**	**Locative**	**Agent**
مَشْرُوبْ /maʃru:b/	شُرْبْ /ʃurb/	شِرِّيبْ /ʃirri:b/	مَشْرَبْ /maʃrab/	شَارِبْ /ʃa:rib/

Table 2. Direct derivation of the root [شرب /ʃrab/]"to drink"

[7] WordNet homepage: `wordnet.princeton.edu`

The second method is based on searching existing varied forms for the elements Lemma or wordForm created in aeb Wordnet e.g. the Lemma [قَالْ /qaːl/] "says" has a varied form [قَالْ /qʾaːl/] used in the occidental north, the occidental south and the oriental south; the wordForm [شُوفْ /ʃuːf /] "see" of the LexicalEntry [شَافْ/ʃaːf /] "to see" has a varied form [أَرَى /ʔaraː/] used only in Sfax.

The third method consists on the use of an aeb corpus composed by aeb texts collected from press articles, theater pieces, poetic books, etc to search words absent in aeb Wordnet and to add them.

With the methods presented at the top, we widely support aeb Wordnet potential.

5 Conclusion

In this article, we presented aeb Wordnet building using the standard ISO LMF. We adapted Wordnet-LMF to aeb specificities based on ISO-LMF and we presented aeb building steps with the expand approach.

Building aeb Wordnet consists on processing PWN by translation to instantiate wordnet-LMF extended model for aeb language. The translation is based on a bilingual dictionary seen the lack of resources. It is supported by validation and extension steps. In this way, we create easily aeb wordnet from PWN and we save aeb language specificities.

This Wordnet is basic, standard and efficient NLP tool. It is an elementary tool with the lack of aeb NLP tools. Its standard structure allows easy interchange with other Wordnets and lexicons. And its current potential i.e. 6000 aeb words is acceptable with the absence of aeb lexicons. Moreover, the extension of aeb wordnet allows its potential to attempt potential of other Wordnets even PWN potential.

Aeb Wordnet is a necessary tool. It will greatly enhance NLP of aeb and so Internet communication monitoring witch become a real challenge with the unsteadiness of economic, finance, politic, etc in Tunisia. Also, it will be very useful to wrestle against terrorism witch disrupt the democratic transition.

Acknowledgments

This work is supported by the General Direction of Scientific Research (DGRS T), Tunisia, under the ARUB program.

References

Sabri Elkateb, William Black, Horacio Rodriguez, Musa Alkhalifa, Piek Vossen, Adam Pease and Christiane Fellbaum. 2006. Building a WordNet for Arabic. *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, Genoa, Italy.

Gil Francopoulo, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet, Claudia Soria. 2007. Lexical Markup Framework: an ISO Standard for Semantic Information in NLP Lexicons. *Proceedings of the Workshop on Lexical-Semantic and Ontological Resources of the GLDV Working Group on Lexicography*, Tübingen, 13-14 April 2007.

Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, Claudia Soria. 2006. Lexical Markup Framework (LMF). *Proceedings of LREC*, Genoa, Italy, 22-28 May 2006.

ISO 24613. 2008. *Language Resource Management – Lexical Markup Framework. ISO.* Geneva, 2008.

Salah Mejri, Mosbah Said, Ines Sfar. 2009. Pluringuisme et diglossie en Tunisie. *Synergies Tunisie n° 1*. pp 53–74.

George A. Miller.1995. WORDNET: a lexical database for English . *COMMUNICATIONS OF THE ACM*. November 1995/Vol. 38, No. 11.

Claudia Soria, Monica Monachini , Piek Vossen. 2009. KYOTO-LMF: fleshing out a standardized format for wordnet interoperability. *Accepted for publication at IWIC2009*. Stanford, Palo Alto, CA.

Claudia Soria and Monica Monachini. 2008. Kyoto-LMF wordnet representation format. Kyoto-LMF wordnet representation format. *KYOTO Working Paper WP02_TR002_V4_Kyoto_LMF*.

Piek Vossen (ed). 1998. EUROWORDNET a database with lexical semantic networks. (Reprinted from *Computers and the Humanities*, 32(2-3), 1998). Dordrecht: Kluwer Academic Publishers, Kluer Academic Publishers.