Morpho-Lexicon for standard Moroccan Amazigh

Fadoua Ataa Allah[1,*], and *Siham* Boulaknadel[1]

[1]EDP CEISIC, Royal Institute of Amazigh Culture, Madinat El Irfane, Rabat, Morocco

Abstract. Standardized resources are key components for the development of applications related to human language technology. Therefore, it is important to adopt it for designing lexical resources, especially for less commonly resourced languages such Amazigh.

This language is spoken by many North African communities, including Morocco. Due to historical, geographical and sociolinguistic factors, the Amazigh language is characterized by the proliferation of many intervarieties, which has led to a complex morphology. This latter poses significant challenge to NLP tasks, especially that Amazigh language belongs to the Afro-Asiatic language (Hamito-Semitic) family, known by its non-concatenative morphology based on root and pattern.

Face to the scarcity of Amazigh language resources dealing with morphemes encoding, orthographic changes, and morphotactic variations, the elaboration of a standardized lexical resource will certainly ensure a large exchange and exploitation. In this context, this paper describes ongoing work for elaborating a morphological lexicon, based on inflected forms, for the standard Moroccan Amazigh language.

.

1 INTRODUCTION

Amazigh language is a prominent element of the Moroccan cultural heritage. However, it was not integrated on the education system until in 2003. For an effective use of this language in education and training, the development of NLP tools and resources is required, especially lexicons.

Various models of lexical resources have been designed and implemented during the last decade for specific purposes. These models vary between glossaries [1-4][a] and morphological lexicons of Nooj platform [5], Xerox FST tools [6] and UNL framework [7]. Nevertheless, each resource is structured in accordance with its project model.

In the purpose to capitalize on these resources, and make them useful for different steps of morphological tools' elaboration, including modelling, enrichment and evaluation, we proposed to build an inflected form lexicon within a standard lexical resource management.

Previous experiences in lexicon standardization have been undertaken by a series of projects like GENELEX [8], EAGLES [9], MULTEXT [10], ISLE [11] and LMF [12]. However, this later appears to be a synthesis and an abstraction over all the previous proposals. Thus, we have applied the LMF modelling framework for building an inflected form lexicon of the Moroccan standard Amazigh language.

Hence, this work describes the LMF modelling-based of an Amazigh lexicon, according to the morphological linguistic level with respect to the specificities of this language.

The remaining of this paper is structured as follow: Section 2 describes briefly the lexical markup framework. Section 3 introduces the Amazigh language and its morphology. Section 4 presents the Moroccan standard Amazigh morphological lexicon within LMF. Finally, Section 5 outlines conclusion and some perspectives.

2 Lexical Markup Framework

Lexical Markup Framework is the ISO International Organization for Standardization ISO/TC37 standard for natural language processing (NLP) and machine-readable dictionary (MRD) lexicons, emerged in 2008 as ISO 24613 [13].

The scope is "to provide a common model for the creation and use of lexical resources, to manage the exchange of data between and among these resources, and to enable the merging of large number of individual electronic resources to form extensive global electronic resources"[b].

[a] The electronic version of this dictionary 'DGLAI' is presented in http://tal.ircam.ma/dglai/.

[b] Cited from Wikipedia article on LMF (http://en.wikipedia.org/wiki/Lexical_Markup_Framework), consulted on Marsh 2018.

* Corresponding author: ataaallah@ircam.ma

2.1. LMF core components

The LMF core model has a structural skeleton that describes the basic hierarchy of information in a lexical entry. According to Romary et *al.* [14], the LMF core is composed of the following components (cf. Fig. 1).

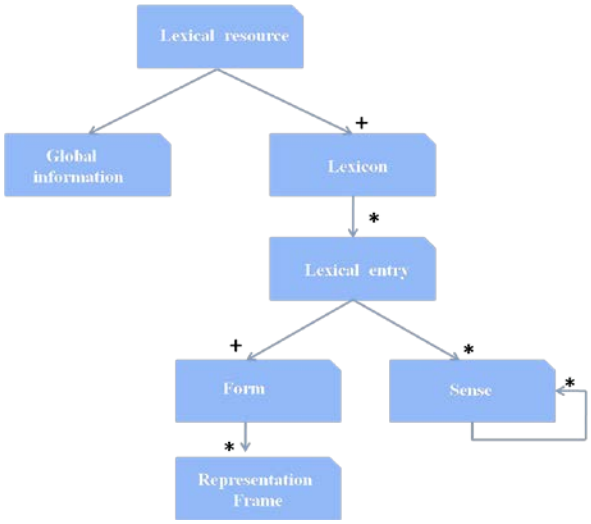

Fig. 1. LMF core model.

- The Lexical Database component, representing the entire resource and contains one or more lexicons.
- The Global Information component that gathers up administrative information and other general attributes.
- The Lexical Entry component, which includes the elementary lexical unit in a lexical database, and two subordinate components that are:
    - the Form component, representing the lexical entry and providing access to surface properties (phonological, graphical, and inflectional features); and
    - the Sense component, which specifies or identifies the meaning and context of the related form.

The Lexical Entry manages the relationship between sets of related forms and their senses. If there is more than one orthography for the word form (e.g. transliteration), the Form class may be associated with one to many Representation Frames.

- The Representation Frame contains a specific orthography and one to many data categories that describe the attributes of that orthography.

2.2 Data categories

LMF provides a mechanism for specifying the content of the core metamodel components by using three basic types of data categories [13]:

- Data categories that may be considered as rather specific to the domain of lexical description.
- Data categories that relate to a specific level of linguistic description such as morphology, syntax, etc. This type of data enforces coherence with other standardization activities.
- Data categories representing metadata descriptors used to document production and maintenance of lexical database, lexical entry and any component in lexical structure.

3 Moroccan Amazigh language features

3.1 Historical background

Amazigh language or Tamazight (ⵜⴰⵎⴰⵣⵉⵖⵜ [tamazight]), is belonging to the African branch of the Afro-Asiatic language family, also referred to Hamito-Semitic in the literature [15]. It is the native language of North Africa, from the Siwa Oasis to the Canary Isles, and from the Senegal river, in the Sahara, to the Mediterranean Sea. Since antiquity, it has its own writing system called "Libyco-Berber" (Tifinaghe in Amazigh). This system dates back more than 40 centuries [16, 17]. However, the appearance form of its signs has been undergoing many modifications: since its inception "the Libyan" to the neo-Tifinaghe in the late sixties and Tifinaghe-IRCAM in 2003 [18][c].

In Morocco, the Amazigh language was an oral tradition, spoken as dialects divided, due to historical, geographical and sociolinguistic factors, into three main varieties: Tarifit in the North, Tamazight in the Center and Tachelhit in the South. Nevertheless, since the creation of the Royal Institute of Amazigh Culture (IRCAM)[d] in 2001, this language is undergoing a progressive linguistic and technological standardization process [19, 20]. At present, the standard Amazigh language represents the model taught in schools, and it is widely used on Amazigh media and on/offline newspapers published in Morocco.

3.2 Amazigh inflection features

Amazigh is a language with a rich non-concatenative morphology. It has a highly inflection and complex derivation word system. The main morphosyntactic categories in Amazigh are: noun, adjective, verb, adverb, preposition, pronoun, particle, conjunction, interjection and numeral [21]. In this work, we focus on noun and verb categories.

Verb. The Amazigh verb has a great structural importance. It represents a wide morphological class and allows for other morphological class derivation. It occurs in two forms: basic and derived one.

[c] Tifinaghe-IRCAM is the official graphic system, proposed by the Royal Institute of Amazigh Culture, for writing the Amazigh language in Morocco. This system is written from left to right. It contains 33 graphemes corresponding to 27 consonants, 2 semi-consonant and 4 vowels [17].

[d] IRCAM is the abbreviation of the French name “Institut Royal de la Culture Amazighe”, which is the Moroccan academic institute charged with the development and the promotion of the Amazigh language and culture (www.ircam.ma).

The basic form is composed of a root and a pattern to compose a radical (root: ⵡⴹ [wḍ], pattern: ⴰCC [aCC], radical: ⴰⵡⴹ [awḍ] *to arrive*); while the derived one is obtained by the combination of a basic verb with one of the following prefixes morphemes: ⵙ/ⵙⵙ [s/ss], ⵜⵜ [tt] and ⵎ/ⵎⵎ [m/mm] (basic form: ⴰⵡⴹ [awḍ], derived form: ⵙⵙⴰⵡⴹ [ssawḍ] *to bring / to bring back*).

The verb, whether basic or derived, has three moods: indicative, imperative and participial. Each mood is characterized by its own personal markers.

In the indicative mood, the verb displays namely four aspectual forms: aorist, imperfective, positive perfect and negative one.

While, in the imperative and participial moods, it has two forms: simple and intensive one. These forms are respectively the combination of the mood personal markers with the aorist and the imperfective aspects. Furthermore, the verb inflects for gender (feminine, masculine), number (singular, plural), and person (first, second, third) [21, 22] (cf. Table 1).

Table 1. Inflection data categories for Amazigh verb.

Data Category	Identifier	Conceptual Range
Grammatical	Mood	{indicative, imperative, participial}
Grammatical	Aspect	{aorist, imperfective, positive perfect, negative perfect}
Grammatical	Person	{first, second, third}
Grammatical	Number	{singular, plural}
Grammatical	Gender	{masculine, feminine}

Noun. In Amazigh language, noun appears in several forms: simple form (ⴰⵔⴳⴰⵣ [argaz] the man), compound form (ⴱⵓⵀⵢⵢⵓⴼ [buhyyuf] the famine), or derived one (ⴰⵎⵢⴰⵡⴰⴹ [amyawaḍ] the communication). It belongs to one of the following types: proper, common or kinship. The two latter types vary in gender (masculine, feminine), number (singular, plural) and state (free, construct).

Similarly to Semitic languages, the state of Amazigh noun varies according to its grammatical syntax. The noun is in free state when it appears as a direct object, anteposed subject, or as an adjective. Also, when it comes after the prediction particle ⴷ [d], a vocative, an interrogative determinant or after one of the following morphemes: ⵢⴰⵙ/ⵢⵉⵔ [ɣas]/[ɣyr] only and ⵡⴰⵔ/ⵜⴰⵔ/ⵖⴰⵔ [war]/[tar]/[gar] without. While, the construct state covers the form in which noun appears as subject; or after a coordinating, attribution term, numeral noun or a preposition at the exception of ⴰⵍ/ⴰⵔ [al/ar] until ⴱⵍⴰ [bla] without.

Moreover, the noun varies in person (first, second, third) when it represents a kinship noun [22] (cf. Table 2).

Table 2. Inflection data categories for Amazigh noun.

Data Category	Identifier	Conceptual Range
Grammatical	Number	{singular, plural}
Grammatical	Gender	{masculine, feminine}
Grammatical	Person	{first, second, third}
Grammatical	State	{free, construct}

4 USING LMF FOR AMAZIGH INFLECTION LEXICON

With the aim to provide a list of all inflected form, for the Moroccan standard Amazigh language, useful for different steps of morphological tools' elaboration, including modelling, enrichment and evaluation, we proposed to adapt some LMF core model specification.

4.1 Formatting the title, authors and affiliations

Due to the separate management of the LMF core model and elementary linguistic descriptors, the LMF proposal appeared to be suitable to our purpose. However, we have made some minor accommodation to take into account the Amazigh morphological features. Thus, we have added a gloss and a set of data categories attached to the lexical entry component, informing about the part of speech. Furthermore, we have chosen the canonical form 'lemma' that corresponds to the masculine singular form for noun and the second person of the simple imperative for verb.

4.2 Data model

Lexical entries have been obtained by gathering a set of lexical entries from various sources [1, 2], [4-7], [23, 24]. Actually, we have achieved for the nominal and verbal categories 14 757 lexical entries and 267 266 inflectional forms. The statistic information of each category is shown in Table 3.

Table 3. The morphological lexicon statistics.

Lexical Category	Lexical entries	Inflectional forms
Noun	9 449	9 316
Verb	3 637	257 950

Our morphological lexicon contains lexical entries, their grammatical category and gloss structured as shown in Fig. 2 and Fig. 3. Each lexical entry includes a <lemmatizedForm>, describing the lemma spelling <orthography> and its grammatical category <grammaticalCategory>; a French translation <gloss>; and a <formset> that contains a set of inflected forms <inflectedForm>, specifying inflectional features according to the grammatical category.

5 Conclusion

In the purpose to take advantage of lexical resources, and make them useful for NLP tasks, we have proposed, in this paper, the first version of a large-coverage morphological lexicon for the Moroccan standard Amazigh language. The lexicon is built within LMF standard lexical resource management. Actually, it is restricted to the inflection forms of noun and verb categories. However, we plan, in the near future, to add the inflection forms of other categories, and the derivational forms; then, to follow our lexical development work by a validation step.

```
<lexicalEntry id="N_347">
    <lemmatizedForm>
        <orthography>ⴰⴹⴱⵉⴱ</orthography>
        <grammaticalCategory>Noun</grammaticalCategory>
    </lemmatizedForm>
    <gloss>médecin</gloss>
    <formSet>
        <inflectedForm>
            <orthography>ⴰⴹⴱⵉⴱ</orthography>
            <grammaticalNumber>singular</grammaticalNumber>
            <grammaticalGender>masculine</grammaticalGender>
            <grammaticalState>free</grammaticalState>
        </inflectedForm>
            ...
        <inflectedForm>
            <orthography>ⵜⴰⴹⴱⵉⴱⵜ</orthography>
            <grammaticalNumber>singular</grammaticalNumber>
            <grammaticalGender>feminine</grammaticalGender>
            <grammaticalState>free</grammaticalState>
        </inflectedForm>
            ...
        <inflectedForm>
            <orthography>ⵜⵉⴹⴱⵉⴱⵉⵏ</orthography>
            <grammaticalNumber>plural</grammaticalNumber>
            <grammaticalGender>feminine</grammaticalGender>
            <grammaticalState>construct</grammaticalState>
        </inflectedForm>
    </formSet>
</lexicalEntry>
```

Fig. 2. Moroccan Amazigh morphological lexicon sample for the noun ⴰⴹⴱⵉⴱ [aḍbib] *doctor*.

```
<lexicalEntry id="V_24">
    <lemmatizedForm>
        <orthography>ⴰⵣ</orthography>
        <grammaticalCategory>Verb</grammaticalCategory>
    </lemmatizedForm>
    <gloss>valoir mieux</gloss>
    <formSet>
            ...
        <inflectedForm>
            <orthography>ⵉⵣⵣ</orthography>
            <grammaticalMood>indicative</grammaticalMood>
            <grammaticalAspect>positive perfect</grammaticalAspect>
            <grammaticalPerson>first</grammaticalPerson>
            <grammaticalNumber>singular</grammaticalNumber>
            <grammaticalGender>masculine</grammaticalGender>
        </inflectedForm>
            ...
        <inflectedForm>
            <orthography>ⵜⵜⵜⴰⵣ</orthography>
            <grammaticalMood>indicative</grammaticalMood>
            <grammaticalAspect>imperfective</grammaticalAspect>
            <grammaticalPerson>second</grammaticalPerson>
            <grammaticalNumber>singular</grammaticalNumber>
            <grammaticalGender>feminine</grammaticalGender>
        </inflectedForm>
            ...
    </formSet>
</lexicalEntry>
```

Fig. 3. Moroccan Amazigh morphological lexicon sample for the verb ⴰⵣ [af] *to worth better*

References

1. M. Ameur, A. Bouhjar, A. Boumalk, N. El Azrak, R. Laabdelaoui, *Vocabulaire des médias* (IRCAM, Rabat, 2009)
2. M. Ameur, A. Bouhjar, A. Boumalk, N. El Azrak, R. Laabdelaoui, *Vocabulaire grammatical* (IRCAM, Rabat, 2009)
3. M. Ameur, K. Ansar, A. Bouhjar, N. El Azrak, *Terminologie amazighe de l'audiovisuel* (IRCAM, Rabat, 2012)
4. M. Ameur, K. Ansar, A. Boumalk, N. El Azrak, R. Laabdelaoui, *Dictionnaire général de la langue amazighe* (IRCAM, Rabat, 2017)
5. F. Z. Nejme, S. Boulaknadel, D. Aboutajdine, Finite State Morphology for Amazigh Language, CICLing **1**, 189-200 (2013)
6. F. Ataa Allah, Finite-State Transducer for Amazigh Verbal Morphology, LLC (2014)
7. I. Taghbalout, F. Ataa Allah, M. El Marraki, Amazigh Noun Inflection in the Universal Networking Language, IJEIT **9** 122-128 (2015)
8. Genelex, Projet EUREKA GENELEX. Rapport sur la couche sémantique, ASSTRIL, Gsi-Erli, IBM France, SEMA GROUP, 2.1 édition (1994)
9. Eagles, Expert Advisory Group on Language Engineering Standards. Reports of the Computational Lexicons Working Group (1996)
10. N. Calzolari, M. Monachini, Multext-Common Specifications and Notation for Lexicon Encoding (1996) www.lpl.univ-aix.fr/projects/multext/LEX/ LEX1.html
11. F. Bertagna, N. Calzolari, A. Lenci, A. Zampolli, ISLE – Computational Lexicons Working Group. The Multilingual ISLE Lexical Entry (MILE): a discussion paper (2000) www.tagmatica.fr/doc.htm
12. G. Francopoulo, M. George, ISO/TC 37/SC 4 N130 Rev.7. Language resource management – Lexical markup framework (LMF) (2005) www.tagmatica.fr/doc.htm

13. G. Francopoulo, LMF Lexical Markup Framework, Wiley-ISTE (2013)
14. L. Romary, S. Salmon-Alt G. Francopoulo, Standards going concrete: from LMF to Morphalou, Coling (2004)
15. J. Greenberg, The Languages of Africa, 2nd ed. (The Hague, Mouton, 1966)
16. M. Hachid, "Les premiers berbères", Entre Méditérranée, Tassili et Nili (Edisud-Ina-Yas, Alger, 2000)
17. A. Skounti, A. Lemjidi, E. Nami, Tirra aux origines de l'écriture au Maroc (IRCAM, Rabat, 2003)
18. M. Ameur, A. Bouhjar, F. Boukhris, A. Boukouss, A. Boumalk, M. Elmedlaoui, E. Iazzi H. Souifi, Initiation à la langue amazighe (IRCAM, Rabat, 2004)
19. F. Ataa Allah, S. Boulaknadel, Toward Computational Processing of less Resourced Languages: Primarily Experiments for Moroccan Amazigh Language, *Theory and Applications for Advanced Text Mining* (2012)
20. F. Ataa Allah, S. Boulaknadel, La Promotion de l'Amazighe à la Lumière des Technologies de l'Information et de Communication, Asinag, **9** (2014)
21. F. Ataa Allah, S. Boulaknadel, H. Souifi, H. Jeu d'Etiquettes Morphosyntaxiques de la Langue Amazighe, Asinag, **9** (2014)
22. F. Boukhris, A. Boumalk, E. El Moujahid, H. Souifi, La nouvelle grammaire de l'amazighe (IRCAM, Rabat, 2008)
23. R. Laabdelaoui, A. Boumalk, E. Iazzi, H. Souifi, K. Ansar, Manuel de Conjugaison de l'Amazighe (IRCAM, Rabat, 2012)
24. F. Ataa Allah, S. Boulaknadel, Amazigh Verb Conjugator, LREC (2014)