Toward an amazigh language processing

Fatima Zahra NEJME[1] *Siham BOULAKNADEL*[2] *Driss ABOUTAJDINE*[1]
(1) GSCM-LRIT, Université Mohamed V, BP 1014 Agdal-Rabat, Maroc
(2) IRCAM, Avenue Allal El Fassi, Madinat Al Irfane, Rabat-Instituts, Maroc
Fatimazahra.nejme@gmail.com, Boulaknadel@ircam.ma,
aboutaj@fsr.ac.ma

ABSTRACT

Since antiquity, the Amazigh heritage is expanding from generation to generation. In the aim of safeguarding it from being threatened of disappearance, it seems opportune to equip this language of necessary means to confront the stakes of access to the domain of New Information and Communication Technologies (ICT). In this context, and in the perspective to build tools and linguistic resources for the automatic processing of Amazigh language, we develop a lexicon and morphological rules using finite state technology within the linguistic developmental environment Nooj to parse amazigh texts.

Vers un traitement automatique de la langue Amazighe

Depuis l'antiquité, le patrimoine Amazighe est en expansion de génération en génération. Dans l'objectif de sauvegarder, exploiter ce patrimoine et éviter qu'il soit menacé de disparition, il semble opportun d'équiper cette langue de moyens nécessaires pour affronter les enjeux d'accès au domaine des nouvelles technologies de l'information et de la communication (NTIC) qui s'avère primordial pour promouvoir et informatiser cette langue. Dans ce contexte, et dans les perspectives de développer des outils et des ressources linguistiques pour le traitement automatique de cette langue, nous avons entrepris d'utiliser la plateforme d'ingénierie linguistique NooJ afin de créer un module pour la langue Amazighe standard (Ameur et al., 2004a). Notre premier objectif est l'analyse des textes Amazighe. A cet effet, nous commençons par la formalisation du vocabulaire Amazighe (Nom, Verbe et Particules). Dans cet article nous nous intéresserons à la formalisation de deux catégories, nom et de particules, permettant de générer à partir d'une entrée lexicale son genre (masculin, féminin), son nombre (singulier, pluriel) et son état (libre, annexion). Enfin, nous développons un dictionnaire électronique afin de l'utiliser, d'une part, pour tester nos règles de flexions et d'autre part pour l'analyse lexicale des textes Amazighe.

KEYWORDS: Amazigh language, NooJ, Natural language processing, Less-resourced language, lexical analysis, inflectional morphology, flexional grammar, dictionary.

KEYWORDS IN L_2 : La langue Amazighe, NooJ, Traitement automatique des langues naturelles, langue peu dotée, analyse lexicale, morphologie flexionnelle, grammaire flexionnelle, dictionnaire.

Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP), pages 173–180,
COLING 2012, Mumbai, December 2012.

Introduction

The Amazigh language in Morocco is considered as a prominent constituent of the Moroccan culture and this by its richness and originality. However it has been long discarded otherwise neglected as a source of enrichment cultural. Nevertheless, due to the creation of the Royal Institute of Amazigh Culture (IRCAM)[1], this language has been introduced in the public domain including administration, media also in the educational system in collaboration with ministries. It has enjoyed its proper coding in the Unicode Standard (Andries, 2008; Zenkouar, 2008), an official spelling (Ameur et al., 2006a), appropriate standards for keyboard realization and linguistic structures that are being developed with a phased approach (Ameur et al., 2006b; Boukhris et al., 2008). This process was initiated by the standardization, vocabularies construction (Kamel, 2006; Ameur et al., 2009a; Ameur et al., 2006a; Ameur et al., 2009b), Alphabetical Arrangement (Outahajala, 2007), spelling standardization (Ameur et al., 2006a) and development of rules grammar (Boukhris et al., 2008).

However, this not sufficient for a less-resourced language (Berment, 2004) as the Amazigh to join the well-resourced language in information and Communication Technologies, mainly due to the lack of already available language processing resources and tools. Therefore, a set of scientific and linguistic research are undertaken to remedy to the current situation. These researches are divided, on the one hand, on researches that are concentrated on optical character recognition (OCR) (Amrouch et al., 2010; Es Saady et al., 2010; Fakir et al., 2009), and in the other hand, on those that are focused on natural language processing (Iazzi and Outahajala, 2008; Ataa Allah and Jaa, 2009; Boulaknadel, 2009; Ataa Allah and Boulaknadel, 2010; Outahajala et al., 2010; Boulaknadel and Ataa Allah, 2011), which constitute the priority components of researches.

In this context, the present work deals with ongoing research efforts to build tools and linguistic resources for the Amazigh language. Our first main objective is to develop a morphological analyzer to parse amazigh texts. For this purpose, we begin by building a morphological analyzer for Amazigh nouns, implemented using the Finite State Technology within the linguistic developmental environment Nooj.

This paper is structured around five main sections: the first present a description of the Amazigh language particularities. The second expose the automatic Amazigh language processing, which includes an overview of NooJ environment, and the formalization of a set of rules. While the last section is dedicated to the conclusion and perspectives.

Amazigh language particularities

The Amazigh language also known as Berber or Tamazight (ⵜⴰⵎⴰⵣⵉⵖⵜ [tamazight]), is belonged to the African branch of the Afro-Asiatic language family, also referred to as Hamito-Semitic in the literature (Greenberg, 1966; Ouakrim, 1995). It is currently presented in a dozen countries ranging from Morocco, with 50% of the overall population[2] (Boukous, 1995), to Egypt, passing through Algeria with 25%, the Tunisia, Mauritania, Libya, Niger and the Mali (Chaker, 2003).

[1] Institution responsible for the preservation of heritage and the promotion of the Moroccan Amazigh culture and its development (see http://www.ircam.ma/).
[2] It present the Amazigh population largest in number.

In Morocco, we distinguish between three major Amazigh dialects. Tarifit is spoken in northern Morocco, Tamazight in the Middle Atlas and south-eastern Morocco, and Tashelhit in south-western Morocco and the High Atlas.

Today, the current situation of the Amazigh language is at a pivotal point. It holds co-official status in Morocco. Its morphology as lexical standardization process is still underway. At present, it represents the model taught in must schools and used on media and official papers published in Morocco.

Amazigh morphology

The Amazigh language presents a rich and complex morphology whose words can be classified into three morphosyntactic categories which we cite: the noun, the verb and particles (Boukhris et al., 2008; Ameur et al., 2004b). Practically, nouns and verbs are the base of the Amazigh morphology and the more important categories to focus on, as others can be derived from them. In this paper, we are interested in noun morphology.

1. Noun characteristics

The noun in the Amazigh language is always composed of one word between two spaces and formed from a root and a pattern. It is characterized by gender (masculine or feminine), number (singular or plural), and state (free or construct) (Boukhris et al., 2008).

Gender: the Amazigh noun is characterized by one of grammatical gender: masculine or feminine.

- The masculine noun: begins with one of the initial vowels: ⴰ [a], ⵉ [i] or ⵓ [u]. However, there are some exceptions as: ⵉⵎⵎⴰ [imma] “(my) mother”.
- The feminine noun: is marked with the circumfix ⵜ .... ⵜ [t....t]. However, there are some exceptions such as nouns which have only the initial ⵜ [t] or the final ⵜ [t] of morpheme of the feminine: ⵜⴰⴷⵍⴰ [tadla] “the sheaf”, ⵜⵔⵔⵓⵢⵜ [ṛṛuyt] “the fatigue”.

Number: the noun, masculine or feminine, has a singular and plural. This latter has four forms: the external plural, broken plural, mixed plural and plural in ⵉⵏ [id].

- The external plural: is formed by an alternation of the first vowel ⴰ /ⵉ [a/i] accompanied by a suffixation of ⵏ [n] or one of its variants.
- The broken plural: involves a change in the vowels of the noun.
- The mixed plural: is formed by vowels’ change accompanied, sometimes by the use of the suffixation by ⵏ [n].
- The plural in ⵉⵏ [id]: this kind of plural is obtained by ⵉⵏ [id] prefixing. It is applied to a set of nouns including: nouns with an initial consonant, proper nouns, parent nouns, compound nouns, numerals, as well as borrowed nouns.

State: we distinguish between two states: the free state and the construct one.

- The free state: is unmarked. The noun is in free state if it is: a single word isolated from any syntactic context, a direct object, or a complement of the predictive particle ⵏ [d].

- The construct state: involves a variation of the initial vowel. In case of masculine nouns, it takes one of the following forms: initial vowel alternation ⴰ [a] /ⵓ [u] or adding of ⵡ [w]; adding of ⵢ [y] to the nouns of vowel ⵉ [i]. For the feminine nouns, it consists to drop the initial vowel or maintaining of this vowel.

Automatic Amazigh language processing: development and evaluation

Nooj platform

NooJ[3], released in 2002 by Max Silberztein (Silberztein, 2007), is a linguistic development platform that provides a set of tools and methodologies for formalizing and developing a set of Natural Language Processing (NLP) applications. It presents a package of finite state tools that integrates a broad spectrum of computational technology from finite state automata to augmented/recursive transition networks. Thus, it presents a complete platform for formalizing various types of textual phenomena (orthography, lexical and productive morphology, local, structural and transformational syntax). For each of these formalization levels, NooJ propose a methodology, one or more formalisms, tools, software development and a corresponding parser that can be used to test each piece of the linguistic formalization over large corpora.

Given these advantages, we have undertaken to adopt NooJ for formalization, description and analysis of Amazigh language for building a module for that language. We begin our work by the formalization of the Amazigh language vocabulary. This formalization is described and stored into inflectional grammars, and can recognize all the corresponding inflected forms. To test these grammars, we built an electronic dictionary in which the lexical entries are attached to a set of linguistic information automatically generate using inflectional grammars which will be used for lexical analysis of texts.

Development of the lexicon

As part of developing a NooJ module for the Amazigh language, we elaborate our dictionary for Amazigh nouns based on a set of lexicons: Taifi dictionary (Taifi, 1988), amazigh vocabulary (Ameur et al., 2006b), and vocabulary of media (Ameur et al., 2009a).Our dictionary contains, currently, 5210 lexical entries which consist of: 4542 simple nouns, 424 proper nouns, 200 Non-inflected nouns and 44 numerals which are given with their plural form, feminine correspondent and annexation state. Our inflected dictionary, calculated after the compilation of the dictionary, encounters 19,597 entries. Thus, we get from each lexical entry all forms related to it.

Morphological rules implementation

This study presents the formalization of the noun category in the NooJ platform. For this, a set of rules has been defined allowing to generate from a each entry, its inflectional information: gender, number and state.

The formalization is based on the use of certain generic commands predefined such as:

- <LW> move at the beginning of Lemma,
- <RW> move at the end of Lemma,
- <S> delete current character,

[3] *See http://www.nooj4nlp.net/ for information of NooJ.*

- <B> delete last character,
- <L> go left,
- <R> go right,

- Gender

To formalize the gender we built this rule that generates from a masculine entry its feminine correspondent. The rule is to add the discontinuous morpheme ⵜ [t] at the beginning and at the end of the noun.

The rules in Nooj	Explanation	Examples
<LW>ⵜ <RW>ⵜ /f+s	This rule adds "ⵜ " at the beginning and at the end.	ⵙ ⵔ ⵍ ⵉ ⵙ [isli] “married” -> ⵜ ⵙ ⵔ ⵍ ⵉ ⵙ ⵜ [tislit] “married”.

TABLE 1 – Example of a gender rule.

- Number

For the amazigh plural, we have many plural forms which are generally unpredictable due to Amazigh complex morphology. To formalize these plural types, we have have relied on the works of Boukhris (Boukhris et al., 2008) and those of Oulhaj (Oulhaj, 2000). We searched formal rules to unify the calculation of plural forms. According to these works and to an heuristic study of the nouns in the Taifi dictionary and those of amazigh language vocabulary, we have raised, at this moment, 303 classes which 97 classes is for the external plural, 99 for the broken plural, 104 for the mixed plural and 3 classes for the plural in ⵙ ⴰ [id]. Each word could be associated with, at least, one flexional class. Thereafter, we provided some examples of rules for each of plural types.

- The external plural

The rules in Nooj	Explanation	Examples
<LW>ⵙ <S><RW>ⵜ ⵏ /m+p	The initial vowel is transformed into ⵙ and the suffix ⵜ ⵏ [tn] is add at the end of the noun.	ⴰ ⵙ ⵉ ⵔ ⴰ [asira] “desk” -> ⵙ ⵉ ⵙ ⵉ ⵔ ⴰ ⵜ ⵏ [isiratn]

TABLE 2 – Plural forms for the masculine nouns beginning and ends with ⴰ [a].

- The broken plural

The rules in Nooj	Explanation	Examples
<LW>ⵙ <S><RW><L>ⴰ /m+p	The rule changes the initial vowel into ⵙ [i] and include ⴰ [a] before the final consonant.	ⴰ ⵣ ⵗ ⵣ ⵍ [azgzl] abbreviation” -> ⵙ ⵣ ⵗ ⵣ ⴰ ⵍ [izgzal]

TABLE 3 – Plural forms for the nouns in VCn form.

- The mixed plural

The rules in Nooj	Explanation	Examples
<LW>ⵉ <S><RW>ⵏ <L2>ⵉ /m+p	The rule change the initial vowel into ⵉ [i], include the vowel ⵉ [i] before the last consonant and add a suffix ⵏ [n] at the end of the noun.	ⴰ ⵔ ⵛ ∧ ⵔ [aḥudr] “fait de se pencher” -> ⵉ ⵔ ⵛ ∧ ⵉ ⵔ ⵏ [iḥudirn]

TABLE 4 – Example of plural forms for the masculine nouns.

- The plural in ⵉ ∧ [id]

The rules in Nooj	Explanation	Examples
<LW>ⵉ ∧ " "/m+p	The rule adds ⵉ∧ [id] before the noun.	ⵒ ⵜ ⵣ ⵔ ⴰ [butgra] “tortoise” -> ⵉ ∧ ⵒ ⵜ ⵣ ⵔ ⴰ [id butgra]

TABLE 5 – Example of plural in ⵉ ∧ [id].

- State

The rules in Nooj	Explanation	Examples
<LW><R><B>ⵓ /EA+m	The rule deletes the initial vowel and adds ⵓ [u] at the beginig of the noun.	ⴰ ⵀ ⵉ ⵔ ⴰ ⵙ [afiras] “pear” -> ⵓ ⵀ ⵉ ⵔ ⴰ ⵙ [ufiras]

- TABLE 6 – Example of plural in ⵉ ∧ [id].

Evaluation

We conducted our experimentation on a sample of texts for story children. By applying our morphological grammars and our dictionary on text, we obtained the following results:

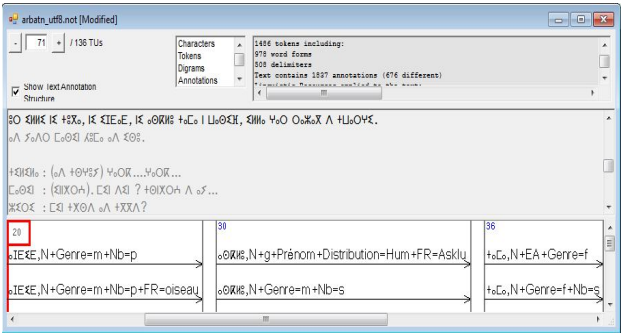

FIGURE 11 – Example of lexical analysis of Amazigh text

The text contains 778 nouns. After performing lexical analysis we identified a total of 686 occurrences of nouns recognized and well annotated (1837 annotations). However, a total of 92 occurrences expressing unknowns nouns. This experiment shows that only 8% of the unknowns nouns that do not belong to our dictionary.

Conclusion and future works

In this paper, we try to restore the Amazigh language and culture and give it more visibility nationally and internationally through developing tools and resources necessary for its computational processing. Our aim to work on a morphological analyzer for Amazigh came from this scarcity of computational framework, a morphological analyzer being one of the fundamental tools in many NLP tasks. However, we build a morphological analyzer for Amazigh nouns, implemented using the Finite State Technology within the linguistic developmental environment Nooj. Amazigh morphological analyzer for nouns is an underway work and further development must be performed to make it a complete one. However, our analyzer achieves over 92% correct results in the analysis of 778 nouns extracted from the corpus.

For future work we planed to:

- Enlarge the lexicon to include nouns from other dialects,
- Include other part of speech in the morphological analyzer,
- Construct a corpus of texts to evaluate the out-of-vocabulary rate of our dictionary.

References

Ameur M., Boumalk A. (DIR) (2004a). *Standardisation de l'amazighe*, Actes du séminaire organisé par le Centre de l'Aménagement Linguistique à Rabat, 8-9 décembre 2003, Publication de l'Institut Royal de la Culture Amazighe, Série : Colloques et séminaires.

Ameur M., Bouhjar A., Boukhris F., Boukouss A., Boumalk A., Elmedlaoui M., Iazzi E., Souifi H. (2004b). *Initiation à la langue amazighe*. Rabat, Maroc: IRCAM.

Ameur M., Bouhjar A., Boukhris F., Boukouss A., Boumalk A., Elmedlaoui M., Iazzi E. (2006a). *Graphie et orthographe de l'amazighe*. Rabat, Maroc : IRCAM.

Ameur M., Bouhjar A., Boukhris F., Elmedlaoui M., Iazzi E. (2006b). *Vocabulaire de la langue amazighe (Français-Amazighe)*. série : Lexiques N°1, IRCAM, Rabat, Maroc.

Ameur M., Bouhjar A., Boumalk A., El Azrak N., Laabdelaoui R. (2009a). *Vocabulaire des médias (Français-Amazighe-Anglais-Arabe)*. série : Lexiques N°3, IRCAM, Rabat, Maroc.

Ameur M., Bouhjar A., Boumalk A., El Azrak N., Laabdelaoui R. (2009b). *Vocabulaire grammatical*. série : Lexiques N°5, IRCAM, Rabat, Maroc.

Amrouch M., Rachidi A., El Yassa M., Mammass D. (2010). *Handwritten Amazigh Character Recognition Based On Hidden Markov Models*. International Journal on Graphics, Vision and Image Processing. 10(5), pp.11--18.

Andries P. (2008). Unicode 5.0 en pratique, *Codage des caractères et internationalisation des logiciels et des documents*. Dunod, France, Collection InfoPro.

Ataa Allah F., Jaa H,. (2009). *Etiquetage morphosyntaxique : Outil d'assistance dédié à la langue amazighe*. In Proceedings of the 1er Symposium international sur le traitement automatique de la culture amazighe, Agadir, Morocco, pp. 110- -119.

Ataa Allah F., Boulaknadel S. (2010). *Online Amazigh Concordancer*. In Proceedings of International Symposium on Image Video Communications and Mobile Networks. Rabat, Maroc.

Berment V. (2004). *Méthodes pour informatiser des langues et des groupes de langues peu dotées*, Thèse de doctorat de l'Université J. Fourier - Grenoble I, France.

Boukhris F., Boumalk A., Elmoujahid E., Souifi H. (2008). *La nouvelle grammaire de l'amazighe*. Rabat, Maroc: IRCAM.

Boukous A. (1995), Société, langues et cultures au Maroc: Enjeux symboliques, Casablanca, Najah El Jadida.

Boulaknadel S. (2009). *Amazigh ConCorde: an appropriate concordance for Amazigh*. In Proceedings of the 1er Symposium international sur le traitement automatique de la culture amazighe, Agadir, Morocco, pp. 176--182.

Boulaknadel S., Ataa Allah F. (2011). *Building a standard Amazigh corpus*. In Proceedings of the International Conference on Intelligent Human Computer Interaction. Prague, Tchec.

Chaker S. (2003), Le berbère, Actes des langues de France, 215-227.

Es Saady Y., Rachidi A., El Yassa M., Mammas D. (2010). *Printed Amazigh Character Recognition by a Syntactic Approach using Finite Automata*. International Journal on Graphics, Vision and Image Processing, 10(2), pp.1--8.

Fakir M., Bouikhalene B., Moro K. (2009). *Skeletonization methods evaluation for the recognition of printed tifinaghe characters.* In Proceedings of the 1er Symposium International sur le Traitement Automatique de la Culture Amazighe. Agadir, Morocco, pp. 33--47.

Greenberg J. (1966). *The Languages of Africa*. The Hague.

Iazzi E., Outahajala M. (2008). *Amazigh Data Base*. In Proceedings of HLT & NLP Workshop within the Arabic world: Arabic language and local languages processing status updates and prospects. Marrakech, Morocco, pp. 36--39.

Kamel S. (2006). *Lexique Amazighe de géologie.* Rabat, Maroc: IRCAM.

Max S. (2007). *An Alternative Approach to Tagging*. NLDB 2007: 1-11

Ouakrim O. (1995). *Fonética y fonología del Bereber*, Survey at the University of Autònoma de Barcelona.

Oulhaj L. (2000). *Grammaire du Tamazight.* Imprimerie Najah El Jadida.

Outahajala M. (2007). *Les normes de tri, Du clavier et Unicode*. La typographie entre les domaines de l'art et de l'informatique. Rabat, Morocco, pp. 223--237.

Outahajala M., Zekouar L., Rosso P., Martí M.A. (2010). *Tagging Amazigh with AnCoraPipe*. In Proceeding of the Workshop on Language Resources and Human Language Technology for Semitic Languages. Valletta, Malta, pp. 52--56.

Taifi M. (1988). *Le lexique berbère (parlers du Maroc central).*

Zenkouar L. (2008). *Normes des technologies de l'information pour l'ancrage de l'écriture amazighe*. Etudes et documents berbères. 27, pp. 159--172.