Natural Language Processing for Amazigh Language: Challenges and Future Directions

Ataa Allah Fadoua, Boulaknadel Siham
CEISIC, IRCAM
Avenue Allal El Fassi, Madinat Al Irfane, Rabat, Morocco
E-mail: {ataaallah, boulaknadel}@ircam.ma

Abstract

Amazigh language, as one of the indo-European languages, poses many challenges on natural language processing. The writing system, the morphology based on unique word formation process of roots and patterns, and the lack of linguistic corpora make computational approaches to Amazigh language challenging.
In this paper, we give an overview of the current state of the art in Natural Language Processing for Amazigh language in Morocco, and we suggest the development of other technologies needed for the Amazigh language to live in "information society".

Keywords: Amazigh Language, Natural Language Processing, Less-resourced language

1. Introduction

During the last few decades, most researches have focused on automatic natural language processing in European and East Asian languages at the expense of native languages of many countries that are under or less resourced. The Moroccan Amazigh language is a part of this list. For many years, it has been neglected and less studied from computational point of view.
However, the official status and the institutional one have enabled Amazigh language to get an official spelling, proper coding in Unicode Standard, appropriate standards for keyboard realization, and linguistic structures that are being developed with a phased approach. This process was initiated and undertaken by spelling standardization and establishment of segmentation rules of the spoken chain (Ameur et al., 2006), character encoding specified by extended ASCII, Alphabetical Arrangement (Outahajala, 2007), incorporation into Unicode standard (Andries, 2008; Zenkouar, 2008), implementation of a standard keyboard layout , building new Tifinaghe fonts (Ait Ouguengay, 2007), vocabularies' construction (Ameur et al., 2006-a; Kamel, 2006; Ameur et al., 2009-a; Ameur et al., 2009-b), and elaboration of grammar rules (Boukhris et al., 2008).
Nevertheless, all these stages of standardization are not sufficient for a less-resourced language as Amazigh to join the well-resourced languages in information technology. In this context, many scientific researches are undertaken at national level to improve the current situation. Primarily, they focus on optical character recognition (Amrouch et al., 2010; Es Saady et al., 2010; Fakir et al., 2009). But those concentrated on natural language processing are limited (Iazzi and Outahajala, 2008; Ataa Allah and Jaa, 2009; Boulaknadel, 2009, Es Saady et al., 2009; Ataa Allah and Boulaknadel, 2010; Outahajala et al., 2010; Boulaknadel and Ataa Allah, 2011).
The remainder of this paper is divided into four main sections. In the first, we give a brief overview of the Moroccan Amazigh language features. In the second section, we present and discuss some of Amazigh linguistic challenges. In the third section, we survey existing systems and resources built for Amazigh languages at the Royal Institute of Amazigh Culture (IRCAM). While in the fourth section, we try to identify needs and suggest some future directions on Amazigh natural language processing.

2. Moroccan Amazigh language features

The Amazigh language, known as Berber or Tamazight, is a branch of the Afro-Asiatic (Hamito-Semitic) languages (Greenberg, 1966; Ouakrim, 1995). Nowadays, it covers the Northern part of Africa which extends from the Red Sea to the Canary Isles, and from the Niger in the Sahara to the Mediterranean Sea.
In Morocco, this language is divided, due to historical, geographical and sociolinguistic factors, into three main regional varieties, depending on the area and the communities: Tarifite in North, Tamazight in Central Morocco and South-East, and Tachelhite in the South-West and the High Atlas.
Since the ancient time, the Amazigh language has its own writing system that has been undergoing many slight modifications. In 2003, it has also been changed, adapted, and computerized by IRCAM, in order to provide the Amazigh language an adequate and usable standard writing system. This system is called Tifinaghe-IRCAM (Ameur et al., 2004).

2.1 Tifinaghe-IRCAM graphical system

Since 2003, Tifinaghe-IRCAM has become the official graphic system for writing Amazigh in Morocco. This system contains:

- 27 consonants including: the labials (ⵀ, ⵃ, ⵇ), the dentals (ⵜ, ⵟ, ⴹ, ⴷ, ⵍ, ⵏ, ⵕ, ⵝ), the alveolars (ⵙ, ⵣ, ⵚ, ⵥ), the palatals (ⵛ, ⵊ), the velar (ⴽ, ⵅ), the labiovelars (ⴽⵯ, ⵅⵯ), the uvulars (ⵇ, ⵅ, ⵖ), the pharyngeals (ⵄ, ⵃ) and the laryngeal (ⴼ);
- 2 semi-consonants: ⵢ and ⵡ;
- 4 vowels: three full vowels ⴰ, ⵉ, ⵓ and neutral vowel (or schwa) ⴻ which has a rather special status in Amazigh phonology.

2.2 Punctuation and numeral

No particular punctuation is known for Tifinaghe. IRCAM has recommended the use of the international symbols: " " (space), ".", ",", ";", ":", "?", "!", "…", for punctuation markers; and the standard numeral used in Morocco (0, 1, 2, 3, 4, 5, 6, 7, 8, 9) for Tifinaghe writing.

2.3 Directionality

Historically, in ancient inscriptions, the Amazigh language was written horizontally from left to right, from right to left, vertically upwards, downwards or in boustrophedon (as illustrated in Figure 1) . However, the orientation most often adopted in Amazigh language script is horizontal and from left to right, which is also adopted in IRCAM-Tifinaghe writing.

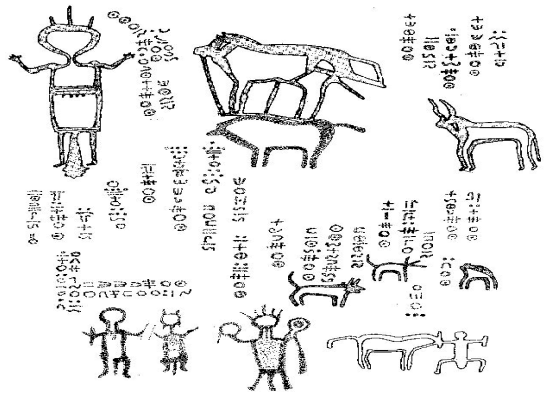

Figure 1: Plate 9 Anou Elias Valley Mammanet (Niger). Henri Lhote, The engravings of Wadi Mammanet. Les Nouvelles Editions Africaines. 1979

3. The complexity of Amazigh in Natural Language Processing

Amazigh is the second official language of Morocco. However, it has been less studied from computational point of view for many years. It has only a limited set of users over the world which makes the interest in developing NLP applications less attractive for foreign developers. Moreover, Amazigh is among the languages having rich morphology and different forms of writing. Below we describe the difficulties that the Amazigh language confronts in developing natural language applications.

3.1 Amazigh script

Amazigh is one of the languages with complex and challenging pre-processing tasks. Its writing system poses three main difficulties:

-Writing forms' variation that requires a transliterator to convert all writing prescriptions into the standard form 'Tifinaghe – Unicode'. This process is confronted with spelling variation related to regional varieties ([tfucht] [tafukt] (sun)), and transcription systems ([tafuct] [tafukt]), especially when Latin or Arabic alphabet is used.

-The standard form adopted 'Tifinaghe – Unicode' requires special consideration even in simple applications. Most of the existed NLP applications were developed for Latin script. Therefore, those that will be used for Tifinaghe – Unicode require localization and adjustment.

-Different prescriptions differ in the style of writing words using or elimination of spaces within or between words ([tadartino] [tadart ino] (my house)).

3.2 Phonology and phonetic

The main problem of Amazigh phonology and phonetic consists on allophones. This problem depends particularly on the regional varieties, when a single phoneme realized in different ways, such as /ll/ that is realized as [dj] in the North.

3.3 Amazigh morphology

An additional reason for the difficulties of computational processing of the Amazigh language is its rich and complex morphology. Inflectional processes in Amazigh are based primarily on both prefix and suffix concatenations. Furthermore, the base form itself can be modified in different paradigms such as the derivational one. Where in case of the presence of geminated letter in the base form, this later will be altered in the derivational form (qqim → svim (make sit)).

4. The State of Amazigh language technology

In the context of promoting Amazigh language, many works have been done to provide this language with linguistic resources and tools in the aim to enable its automatic processing and its integration in the field of Information and Communication Technology. In this section, we describe existing works on Amazigh language

processing.

4.1 Tifinaghe Encoding

Over several years, the Amazigh language has been writing in Latin alphabet supported by diacritics and phonetic symbols, or in Arabic script. While after adopting Tifinaghe as an official script in Morocco, the Unicode encoding of this script was became a necessity. To this end considerable efforts have been invested. However, this process took ample time to be done, which required the use of ANSI encoding as a first step to integrate the Amazigh language into the educational system at time.

Considering Tifinaghe variants used in all parts of the Amazigh world, the Unicode encoding is composed of four character subsets: the basic set of IRCAM, the extended IRCAM set, other Neo-Tifinaghe letters in use, and modern Touareg letters. The two first subsets constitute the sets of characters chosen by IRCAM. While, the first is used to arrange the orthography of different Moroccan Amazigh dialects, the second subset is used for historical and scientific use. The letters are classified in accordance with the order specified by IRCAM. Other Neo-Tifinaghe and Touareg letters are interspersed according to their pronunciation. Thus, the UTC accepts the 55 Tifinaghe characters for encoding in the range U+2D30..U+2D65, U+2D6F, with Tifinaghe block at U+2D30..U+2D7F (Andries, 2008).

4.2 Optical character recognition

In the aim to achieve perfection on Amazigh optical character recognition systems many studies have been undertaken using different approaches. Most of these approaches have achieved a recognition rate around 92%. In the following, we present briefly some Amazigh optical character recognition systems. (Es Saady et al., 2011) focused on isolated printed characters recognition based on a syntactic approach using finite automata. (Amrouch et al., 2010) proposed a global approach based on Hidden Markov Models for recognizing handwritten characters. (El Ayachi et al., 2010) presented a method using invariant moments for recognizing printed script. (Ait Ouguengay et al., 2009) proposed an artificial neural network approach to recognize printed characters.

4.3 Fundamental processing tools

In this section, we describe natural language processing systems that have been developed for the Amazigh language at the Royal Institute of Amazigh Culture.

- **Transliterator:** The Amazigh language has known through its existence different forms of writing: Latin alphabet supported by diacritics and phonetic symbols, Arabic script, and Tifinaghe character based on ANSI and Unicode encoding. In the aim to facilitate the passage from one form to another, and to convert all writing prescriptions into a standard unique form in order to simplify the text processing a transliterator tool has been developed (Ataa Allah and Boulaknadel, 2011).

- **Tagging assistance tool:** The use of corpora in natural language processing, especially those annotated morphosyntactically, has become an indispensable step in the language tools' production and in the process of language computerization. In this context, we have lead to build a morphosyntactic corpus; which has elicited the development of a tool, providing support and linguists' assistance (Ataa Allah and Jaa, 2009).

- **Stemmer:** To enhance the performance of information retrieval systems for the Amazigh language a computationally stemming process was realized. This process consists in splitting Amazigh words into constituent stem part and affix parts without doing complete morphological analysis, in order to conflate word variants into a common stem (Ataa Allah and Boulaknadel, 2010-a).

- **Search engine:** As the number of Amazigh documents grew, searching algorithms have become one of the most essential tools for managing information of Amazigh documents. Thus, a first attempt has been proposed in order to develop a search engine that could support the Amazigh language characteristics. The proposed search engine is designed to crawl and index the Amazigh web pages written in Tifinaghe. Moreover, it is based on some natural language processing such as stop words removal and light stemming in retrieval task (Ataa Allah and Boulaknadel, 2010-b).

- **Concordancer:** Amazigh linguistics corpora are currently enjoying a surge activity. As the growth in the number of available Amazigh corpora continues, there is an increased need for robust tools that can process this data, whether it is for research or teaching. One such tool that is useful for both groups is the concordancer, which is a simple tool for displaying a specified target word in its context. However, obtaining one that can reliably support all Moroccan Amazigh language scripts has proved an extreme difficulty. In this aim, an online concordancer was developed (Ataa Allah and Boulaknadel, 2010-c).

4.4 Language resources

Natural language processing is showing more interest in the Amazigh language in recent years. Suitable resources for Amazigh are becoming a vital necessity for the progress of this research. In this context some efforts are currently underway.

- **Corpora:** Corpora are a very valuable resource for NLP tasks, but the Amazigh language lacks such resources. Therefore, researchers at IRCAM have tried to build an Amazigh corpora in progressive way until reaching a large-scale corpus that follows TREC's standards. Thus, two parallel works are undertaking (Outahajala et al., 2010; Boulaknadel and Ataa Allah, 2011).

The first consists in building a general corpus based on texts dealing with different literary genres: novels, poems, stories, newspaper articles, and covering various topics. While the second is based on POS tagged data that was collected from IRCAM's newspapers, websites and

pedagogical supports.

- **Dictionary:** Although many paper dictionaries are available for the Amazigh language, none of them is computational. To deal with this lack, an application that is helping in collecting and accessing Amazigh words has been elaborated (Iazzi and Outahajala, 2008). This application has provided all necessary information such as definition, Arabic French and English equivalent words, synonyms, classification by domains, and derivational families.

- **Terminology database:** While the Amazigh language is given new status, it becomes necessary, even inevitable to own a terminology covering the largest number of lexical fields. Thus, a tool managing terminology database has been developed to facilitate the work of researchers allowing an efficient exploitation of users. This tool allows the processing of new terminology data, the compilation, and the management of existing terminology (El Azrak and El Hamdaoui, 2011).

5. Conclusion and Future Directions

In this paper, we discussed the main challenges in processing the Amazigh language, and we attempted to survey the research work on Amazigh NLP in Morocco. In the aim to convert Amazigh language from a less resourced language into a resourced, studied language from computational point of view, we need to expedite the basic research on Amazigh NLP tools development by addressing the following issues:

-Building a large and representative Amazigh corpus which will be helpful for spelling and grammar checking, speech generation, and many other related topics.

-Developing a machine translation system which will immensely contribute to promote and disseminate the Amazigh language.

-Creating a pool of competent human resources to carry out research work on Amazigh NLP by offering scholarship for higher degrees and attracting young researchers with attractive salary.

6. References

Ait Ouguengay Y. (2007). Quelques aspects de la numérisation des polices de caractères : Cas de Tifinaghe. *La typographie entre les domaines de l'art et de l'informatique*. Rabat, Maroc, pp. 159--181.

Ait Ouguengay Y., Taalabi M. (2009). Elaboration d'un réseau de neurones artificiels pour la reconnaissance optique de la graphie amazighe: Phase d'apprentissage, *Systèmes intelligents-Théories et applications.* Europia productions.

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

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

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

Ameur M., Bouhjar A., Boukhris F., Elmedlaoui M., Iazzi E. (2006). *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. (2009). *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. (2009). *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.

Ataa Allah F., Boulaknadel S. (2010). Amazigh Search Engine: Tifinaghe Character Based Approach. In *Proceeding of International Conference on Information and Knowledge Engineering*, Las Vegas, Nevada, USA, pp. 255-259.

Ataa Allah F., Boulaknadel S. (2010). Pseudo-racinisation de la langue amazighe. In *Proceeding of Traitement Automatique des Langues Naturelles*. Montréal, Canada.

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

Ataa Allah F., Boulaknadel S. (2011). Convertisseur pour la langue amazighe : script arabe - latin – tifinaghe. In *Proceedings of the 2ème Symposium International sur le Traitement Automatique de la Culture Amazighe*. Agadir, Marocco, pp. 3--10.

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.

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

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.

EL Azrak N., EL Hamdaoui A. (2011). Référentiel de la Terminologie Amazighe : Outil d'aide à l'aménagement linguistique. In *Proceedings of theème atelier international sur l'amazighe et les TICs*, Rabat, Morocco.

El Yachi R., Moro K., Fakir M., Bouikhalene B. (2010). On the Recognition of Tifinaghe Scripts. *Journal of*

Theoretical and Applied Information Technology, 20(2), pp. 61--66.

Es Saady Y., Ait Ouguengay Y., Rachidi A., El Yassa M., Mammass D. (2009). Adaptation d'un correcteur orthographique existant à la langue Amazighe : cas du correcteur Hunspell. In *Proceedings of the 1er Symposium International sur le Traitement Automatique de la Culture Amazighe*. Agadir, Morocco, pp. 149--158.

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.

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

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.

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