Amazigh Language Desktop Converter

Fadoua Ataa Allah Jamal Frain Youssef Ait Ouguengay
CEISIC, IRCAM, Morocco
{ataaallah, frain, ouguengay}@ ircam.ma

***Abstract*— Before the creation of the Royal Institute of Amazigh Culture, the Amazigh language has not have the ability to take the advantages of information and communication technologies, even to be written in its native writing system "Tifinaghe". To overcome these limitations, the Amazigh language has undergone a process of standardization and integration into information and communication technologies. This process is passing through several stages, mainly the encoding stage and the development of appropriate standards for the keyboard layout, in addition to the stage of computational linguistics. In this context and in the aim to save the Amazigh cultural heritage, a desktop converter, allowing the ANSI-Unicode transition and Arabic-Latin-Tifinaghe transliteration, is developed.**

Keywords—Amazigh langauge; transcoding; transliteration; ANSI; Unicode; Tifinaghe; Arabic and Latin scripts .

I. INTRODUCTION

The integration of Amazigh language into Information and Communication Technologies (ICT) has become a necessity to promote the Amazigh language. Nevertheless, this integration was confronted by several challenges, especially those related to language planning and computer science standardization.

To let the Amazigh language supporting and conveying knowledge, firstly, a writing form and an alphabetic system have been established. Secondly, based on a linguistic description of the most widely spoken varieties of Amazigh language, a spelling system has been stabilized [1]. Then, a stage of character encoding has been undertaken. However, the difficulty in these steps is to achieve generic solutions in limited time to allow the integration of Amazigh into the Moroccan educational system in 2003. Thus, the native Amazigh writing system encoding went through two steps: ANSI then Unicode encoding [2].

Furthermore, the promotion of the Amazigh culture implies the maintenance and the conservation of the literary heritage, and the diffusion of Tifinaghe script on all media. To this end, an Amazigh desktop converter is developed. This later is based on character transcoder allowing the ANSI-Unicode conversion, and an Arabic-Latin-Tifinaghe transliterator allowing the transmutation of Amazigh document content from the Arabic or Latin alphabets to Tifinaghe.

The remainder of this paper consists of three sections. In Section 2, we present a history overview and the writing systems of the Amazigh language. Then, we describe, in Section 3, the elaborated tool. Finally, in Section 4, we conclude and present some perspectives.

II. AMAZIGH LANGUAGE

A. Amazigh language history

Amazigh is the native language of North Africa. It is also known by the name of "Berber", and the local name "Tamazight". This language is present from Morocco to Egypt passing through Algeria, Tunisia, Libya, Niger and Mali (*cf.* Fig. 1). It was spoken by tens of millions of people as non-standardized dialects [3].

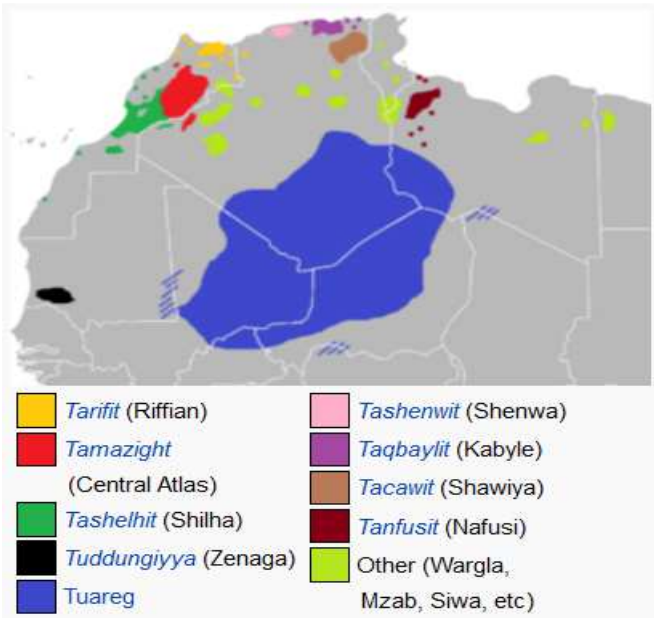


Fig. 1. Amazighophony Map[1]

In Morocco, there are three main varieties of the Amazigh language: Tarifite in the North; Tamazight in the Center, the Middle Atlas and a part of High Atlas; and Tachelhite in the South, South-west of High Atlas, the Anti-Atlas and Sous. These varieties were primarily employed in oral communication. However, in order to preserve the Amazigh language, it was important to transit from orality to literacy and to upgrade the language, then to integrate the Amazigh language into the information and communication technologies.

Although the Amazigh language was primarily an oral tradition, the Amazigh language has, since antiquity, its own writing system called "Libyco-Berber" (Tifinaghe in Amazigh). This system dates back more than 40 centuries [4], [5].

[1] http://en.wikipedia.org/wiki/Berber_languages

However, the appearance form of its signs has undergoing many modifications: since its inception "the Libyan" to the neo-Tifinaghe in the late sixties and Tifinaghe IRCAM-in 2001 [1].

Historically in ancient inscriptions, Amazigh was written horizontally from left to right, from right to left, vertically upward, downward or in boustrophedon (as illustrated in Fig. 2). Nonetheless, the orientation most often adopted in Amazigh language script is horizontal and from left to right, which is also adopted in IRCAM-Tifinaghe writing system.

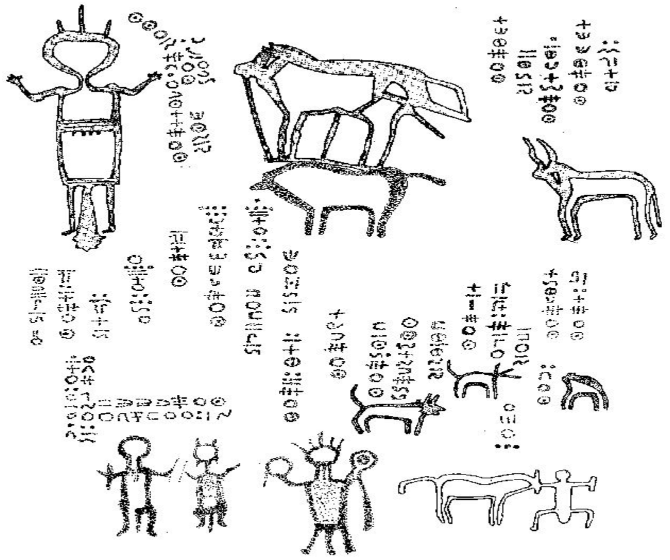


Fig. 2. Plate 9 Anou Elias, Mammanet Valley (Niger). Henri Lhote, Oued Mammanet gravures. Les Nouvelles Editions Africaines. 1979

B. Tifinaghe-IRCAM graphical system

Since 2003, Tifinaghe-IRCAM has become the official graphic system for writing Amazigh in Morocco [1]. This system contains:

- 27 consonants including: the labials (ⴱ, ⴼ, ⵎ), dentals (ⵜ, ⴷ, ⵟ, ⴹ, ⵏ, ⵔ, ⵕ, ⵍ), alveolars (ⵙ, ⵣ, ⵚ, ⵥ), palatals (ⵛ, ⵊ), velar (ⴽ, ⴳ), labiovelars (ⴽⵯ, ⴳⵯ), uvulars (ⵇ, ⵅ, ⵖ), pharyngeals (ⵄ, ⵃ) and the laryngeal (ⵀ);
- 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.

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 system.

C. Amazigh encoding

Before the computer science standardization of Amazigh language, the Websites’ analysis showed that while there is an abundant Amazigh literature, there are very few Websites that use Amazigh language for communication. Furthermore, its presence on the web was especially based either on Arabic alphabet or the Latin alphabet enriched with some special characters borrowed from the International Phonetic Alphabet (IPA). While, Tifinaghe characters appeared occasionally, and generally they were inserted as images. However, to allow the Amazigh people to communicate in their own language and follow at the same time the technological evolution, it was necessary to ensure a wide diffusion and a linguistic analysis of digital document content.

To this end, a digital transcription system has been created, and several efforts have been undertaken to encode Tifinaghe in Unicode/ISO 106461 [6]. Nevertheless, this process was progressing simultaneously with the integration of the Amazigh language into the Moroccan education system. So to resolve the lack of digital transcription system, while producing teaching materials in the official Amazigh script, the ANSI encoding has been used.

1) ANSI encoding: ANSI encoding is a slightly generic term used to refer to the standard code page on a system, usually Windows. It is more properly referred to as Windows-1252. This is essentially one of the American Standard Code for Information Interchange (ASCII) extension that includes all the ASCII characters with an additional 128 character codes. This difference is mainly due to the fact that "ANSI" encoding is 8-bit rather than 7-bit as ASCII is. But since there are many standard that differently use the 8^{th} bit to define characters numbered from 128 to 255, we remark that the non-ASCII characters as accented ones (éÈç) often display incorrectly.

2) Tifinaghe ANSI encoding: While expecting the Tifinaghe Unicode encoding, the Centre of Computer Studies Information Systems and Communication (CEISIC) has chosen to adapt the ANSI encoding for Tifinaghe characters as illustrated in the table of Fig. 3.

In this table, we remark the existence of the same Tifinaghe character in two different locations, which is due to the fact that each Latin character either uppercase or lowercase has been replaced by the same Tifinaghe character, since the Amazigh language does not support capital letter.

Once the encoding table is created, the ANSI encoding has allowed the manipulation of Tifinaghe digital documents, and the edition of publications and Amazigh scholar manuals. Moreover, the phonetic approach adopted by IRCAM in the choice of the graphical system and the correspondence between the Arabic script, Latin and Amazigh (see Table I) have greatly helped in accelerating the production of texts in Amazigh and facilitated the manipulation of multilingual texts with multiple writing systems. However, the scope of this IRCAM private encoding has been limited, because it is not considered as an appropriate Tifinaghe encoding.

3) Unicode encoding: The computer software internationalization and localization has inflected the definition of new character encoding standard that unifies all the existing character sets, and overcomes the limitations of incompatibility of old encoding standards. Thus, Unicode Consortium and the International Organization for

Fig. 3. Tifinaghe ANSI encoding table

Standardization (ISO) have combined their efforts to provide a universal character encoding scheme known as Unicode. This standard is designed to support the worldwide interchange, processing, and display of the technical disciplines of the modern world and the written texts of diverse languages even classical and historical ones.

Unicode can be implemented by different character encodings. The most commonly used encodings are UTF-8, UTF-16 and the now-obsolete UCS-2. UTF-8 uses one byte for any ASCII characters, which have the same code values in both UTF-8 and ASCII encoding, and up to four bytes for other characters. UCS-2 uses a 16-bit code unit (two 8-bit bytes) for each character but cannot encode every character in the current Unicode standard. UTF-16 extends UCS-2, using two 16-bit units (4 × 8 bit) to handle each of the additional characters[2].

4) Tifinaghe Unicode encoding: According to Mr Zenkouar[3] [7], the Unicode standard has constituted the main entrance of the Amazigh native writing system into the world of Information and Communication Technologies. Indeed thanks to this standard, nowdayes, Tifinaghe is used on line based on Tifinaghe Webfonts [8] and internationalized URL [9], and it is integrated into the operating systems Linux and Windows.

Unicode covers almost all scripts in current use today, and reserves to each new writing system a set of blocks. Thus, it has reserved a block for Tifinaghe characters, where each character is currently determined by a single code point. Fig. 4 illustrates the Tifinaghe block and specifies the code points reserved for each of the four subsets of Tifinaghe characters: the basic set of IRCAM, the extended IRCAM set, other Neo-Tifinaghe letters in use, and modern Touareg letters. The first subset constitutes the set of characters chosen by IRCAM to arrange the orthography of the different Moroccan Amazigh varieties while preserving most characters of the historical Tifinaghe script. This subset is classified in accordance in the range U+2D30…U+2D65, U+2D6F with the order specified in Tifinaghe-IRCAM alphabet. While, the Tifinaghe block is the range U+2D30…U+2D7F [6], [10]. For more details on the Unicode and the Information Technology component for adding Tifinaghe to Unicode one can see also [10].

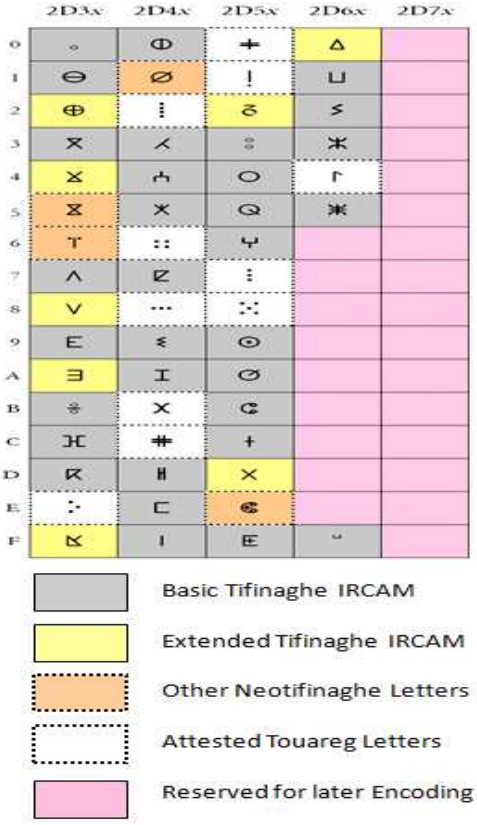


Fig. 4. Tifinaghe Unicode encoding block

[2] http://en.wikipedia.org/wiki/Unicode

[3] The CEISIC's ex-director at IRCAM

5) Amazigh language transcription: Before adopting Tifinaghe as an official script in Morocco, like any oral language, Amazigh was writing by the graphic systems widely used in the country. Thus, the Arabic script was used for religion and rural poetry writing, while Latin supported by the International Phonetic Alphabet was used particularly by berberists since early works of missionaries.

III. AMAZIGH DESKTOP CONVERTER

Through its existence, the Amazigh language has known different forms of writing: Latin supported by the International Phonetic Alphabet, Arabic script, and Tifinaghe character based on ANSI and Unicode encoding. In the aim to allow users to read or write in a suitable form, and to save the Amazigh literature heritage in a standard unique form, a command-line converter has been developed [11]. This tool ensures an automatic shifting from one form to another. However, it has some limitations such as having menu-driven and graphical user interfaces, processing rich text format, and dealing with multilingual text especially when Amazigh language is writing with ANSI script. To overcome these shortcomings a desktop converter is developed.

A. *Desktop converter technical architecture*

The technical architecture of the desktop converter is based on an implementation of the Model-View-Controller (MVC) pattern based on technology .Net, which separates application data model and user interface views in separate components (*cf.* Fig. 5).

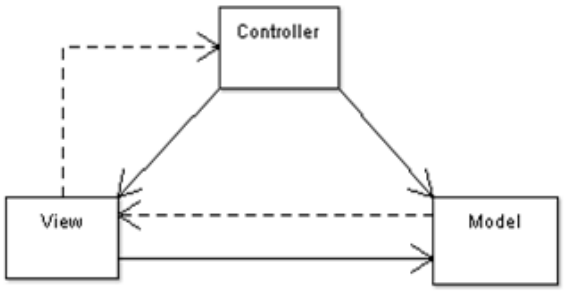


Fig. 5. Technicnal Architecture Diagram

B. *Desktop converter functional architecture*

As summarized in the diagram of Fig. 6, the desktop converter consists of two processes: transcoder and transliterator.

1) Transcoder: This process allows to shift the ANSI representation of Tifinaghe to Unicode representation for the content of a file in one of the following format: .txt, .rtf, .doc, and .docx (*cf.* Fig. 7), or the content of a text area (*cf.* Fig. 8).

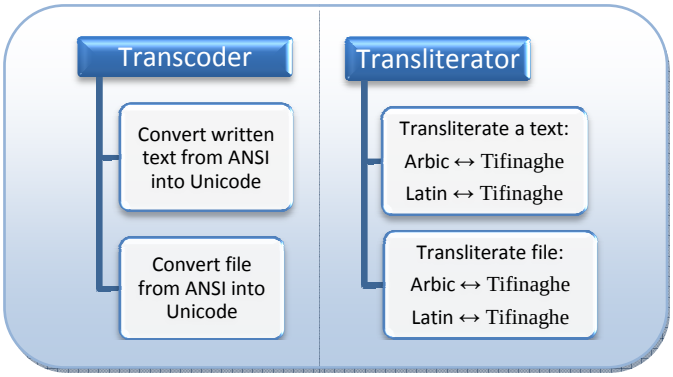


Fig. 6. Functional Architecture Diagram

In the case of file conversion, the transcoder allows browsing hard drives to choose the file to process, and choosing the input ANSI and the output Unicode fonts. Furthermore, it preserves the document layout and text formatting.

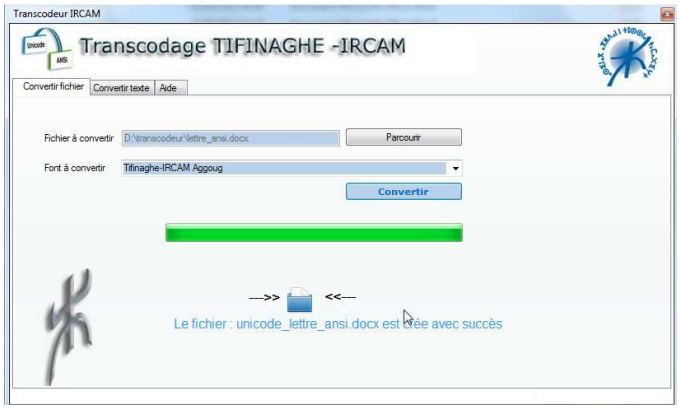


Fig. 7. Transcoder interface for converting a file

In the case of converting the content of a text area, the transcoder enables a real time encoding conversion. The input content could be pasted or typed, and the output could be copied or saved into a text file.

Fig. 8. Transcoder interface for the content of text area

2) Transliterator: The transliterator process aims to substitute the script of a text to another, while conserving the phonetic pronunciation of its words. This process is based on direct mapping between the pairs of scripts (Latin, Tifinaghe Unicode) and (Arabic, Tifinaghe Unicode). In the Latin - Tifinaghe Unicode mapping the IRCAM correspondences illustrated in Table I are used by default. While, phonetic and Stroomer's correspondences [12] are also available.

TABLE I. TIFINAGHE-IRCAM ALPHABET

English appellation	Tifinaghe	Latin correspondence	Arabic correspondence
ya	ⴰ	a	ا
yab	ⴱ	b	ب
yag	ⴳ	g	گ
yagw	ⴳⵯ	gʷ	گُ
yad	ⴷ	d	د
yadd	ⴹ	ḍ	ض
yey	ⴻ	e	
yaf	ⴼ	f	ف
yak	ⴽ	k	ک
yakw	ⴽⵯ	kʷ	کُ
yah	ⵀ	h	ه
yahh	ⵃ	ḥ	ح
yaa	ⵄ	ɛ	ع
yakh	ⵅ	x	خ
yaq	ⵇ	q	ق
yi	ⵉ	i	ي
yazh	ⵊ	j	ج
yal	ⵍ	l	ل
yam	ⵎ	m	م
yan	ⵏ	n	ن
you	ⵓ	u	و
yar	ⵔ	r	ر
yarr	ⵕ	ṛ	رّ
yagh	ⵖ	ɣ	غ
yas	ⵙ	s	س
yass	ⵚ	ṣ	ص
yash	ⵛ	c	ش
yat	ⵜ	t	ت
yatt	ⵟ	ṭ	ط
yaw	ⵡ	w	ۏ
yay	ⵢ	y	ي
yaz	ⵣ	z	ز
yazz	ⵥ	ẓ	ژ

In the Arabic - Tifinaghe Unicode mapping, there are more constrained rules to use the correspondences represented in Table I. These constraints depend mainly on the writing cursivity of the Arabic language, the phonetic pronunciation, and the use of Amazigh and Arabic vowels. Thus, some Arabic - Tifinaghe correspondences have been adapted (*cf.* Table II). Indeed, inspired from the correspondence mapping of Mr Chafik [14], it has been proposed to use the symbol '◌ٜ' in spite of '◌ٚ' to unsure the Arabic writing curisvity, and to use the letter "ڤ" in spite of "ڤٜ" to help the reader to correctly pronounce the letter "ⵅ".

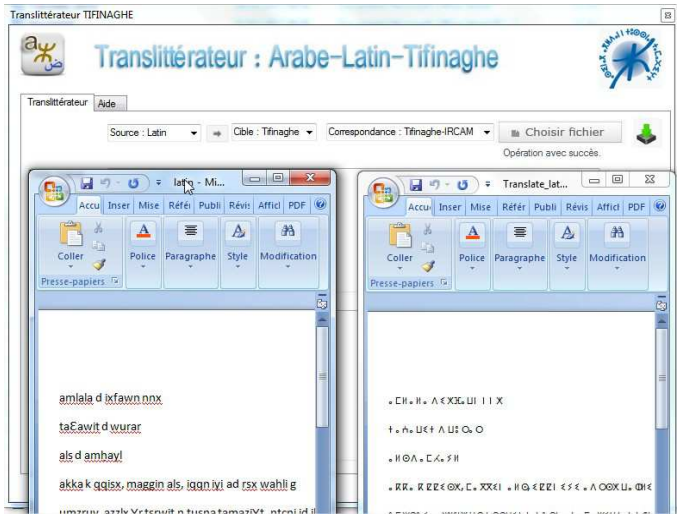
TABLE II. PROPOSED ARABIC CORRESPONDANCE

English appellation	Tifinaghe	New Arabic correspondence
yag	ⵅ	ڤ
yagw	ⵅⵯ	ڤُ
yakw	ⵇⵯ	كُ
yi	ⵉ	ـٕ (in the beginning)
you	ⵓ	ۏ (in the beginning)
yay	ⵢ	ي (in the beginning)

Moreover, some orthographic rules have been specified mainly on the transliteration from Arabic script into Tifinaghe one [14]. These rules are as follow:

- If the word contains any emphatic letter (ⴹ, ⵕ, ⵚ, ⵟ, ⵥ), the letter 'ر' will be represented by 'ⵕ'.
- If the letter 'ي' is preceded by a vowel, it will be represented by the semi-consonants 'ⵢ'. Otherwise, it will be represented by the vowel 'ⵉ'.
- If the letter 'و' is preceded by the vowel 'ا', it will be represented by the semi-consonants 'ⵡ'. If it is preceded by the vowel 'ي' or preceded and succeeded by a consonant, it will be represented by the vowel 'ⵓ'.
- If the word contains a geminated letter represented by the symbol shadda '◌ّ', this later will be replaced by doubling the geminated letter.

The transliterator allows the possibility to convert the content of a file or a text area (*cf.* respectively Fig. 9 -a-, -b-), by choosing the input and the output scripts, in addition to the mapping table in the corresponding dropdown list.



-a-

- b-

Fig. 9. Transliterator interface

Furthermore, the tool enables the user to set his/her proper correspondence mapping table that could be saved for another reuse (*cf.* Fig. 10). Once the new mapping table is saved, it will appear in the dropdown list.

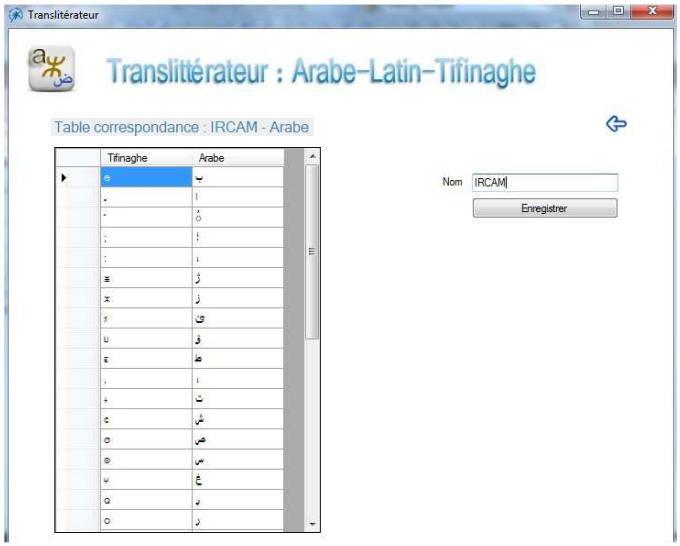


Fig. 10. Correspondence setting layout

IV. CONCLUSION

In the aim to promote the Amazigh language and to preserve its literary heritage, this paper has presented a new desktop converter enabling the ANSI-Unicode transcoding and the Arabic-Latin-Tifinaghe transliteration even in real time for typed text and rich text format for files. This work will be extended to an Amazigh toolbox that could integrate dictionary, conjugator, spelling checker and other computational linguistic tools.

REFERENCES

[1] M. Ameur, A. Bouhjar, F. Boukhris, A. Boukouss, A. Boumalk, M. Elmedlaoui, E. M. Iazzi, H. Souifi, Initiation à la langue amazighe, IRCAM, 2004.

[2] M. Moukhlis, L. Ouniam, Bulletind'information de l'Institut Royal de la Culture Amazighe, n° 5&6, June 2006.

[3] S. Chaker, Le berbère, Actes de Les langues de France, 215-227. 2003.

[4] M. Hachid, "Les premiers berbères", Entre Méditérranée, Tassili et Nili. Aix-en-Provence-Alger : Edisud-Ina-Yas, pp. 173-190.

[5] A. Skounti, A. Lemjidi, E. M. Nami, Tirra aux origines de l'écriture au Maroc, IRCAM,2003.

[6] P. Andries, Unicode 5.0 en pratique : Codage des caractères et internationalisation des logiciels et des documents. Dunod : Collection InfoPro, France, 2008.

[7] L. Zenkouar, "Normes des technologies de l'information pour l'ancrage de l'écriture amazighe", in Etudes et documents berbères, n° 27, pp. 159-172, 2008.

[8] P. Andries, "Demain encore plus de tifinaghes sur Internet", Actes de colloque TICAM 2008.

[9] P. Andries, "Indiquer la langue, l'écriture, le pays dans des documents informatiques , Actes de colloque TICAM 2008.

[10] L. Zenkouar, "L'écriture amazighe tifinaghe et Unicode", Etudes et Documents Berberes, 22, 2004: pp. 175-173.

[11] F. Ataa Allah, and S. Boulaknadel, "Convertisseur pour la langue amazighe : script arabe - latin – tifinaghe", The 2nd Symposium International sur le Traitement Automatique de la Culture Amazighe, SITACAM 2011, 6-7 May 2011, Agadir, Morocco, 2011.

[12] H. Stroomer, Textes berbères des Guedmioua et Goundafa (Haut Atlas, Maroc), basés sur les documents de F. Corjon, J.-M. Franchi et J. Eugène, Edisud, 2001.

[13] F. Ataa Allah, and S. Boulaknadel, "Toward computational processing of less resourced languages: Primarily experiments for Moroccan Amazigh language", in Text Mining. Rijeka: InTech, pp. 197-218, november 2012.

[14] M. Chafik, المعجم العربي الأمازيغي, El Maarif Aljadida, Rabat, Morocco, 1993.