Chapter 9

Toward Computational Processing of Less Resourced Languages: Primarily Experiments for Moroccan Amazigh Language

Fadoua Ataa Allah and Siham Boulaknadel

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/51069

1. Introduction

The world is undergoing a huge transformation from industrial economies into an information economy, in which the indices of value are shifting from material to non-material resources. This transformation has been rightly described as a revolution that is accompanied by considerable dangers for the future and the survival of many languages and their associated cultures. The last years have seen a growing tendency in investigating applying language processing methods to other languages than English. However, most of tools and methods' development on language processing has so far concentrated on a fairly small and limited number of languages, mainly European and East-Asian languages.

Nevertheless, there is a mandatory requirement for all people over the world to be able to employ their own language when accessing information on the Internet or using computers. To this end, a variety of applications is needed, and lots funds are involved. But the fact that the most of the research sponsored around the world has focused only on the economically and politically important languages makes the language technology gap between the languages of the developed countries and those of the less developed ones leading up to a larger and a largest gap.

According to some linguists' estimations in 1995, half of the 6000 world's languages is being disappearing, 2000 among the 3000 remaining will be threatened in the next century [1]. This means that if there are no efforts put in place to reduce the technology gap and to preserve these languages so many of them will disappear completely by the end of this century. Unfortunately, there are numerous obstacles to advance in language processing for this kind of languages. In the first hand, the language features themselves might impose specif-

ic strategies to be undertaken. In the second hand, the lack of previously existing language resources produces a vicious circle: having resources makes creating electronic ones and processing tools easier, but not having resources makes the development and testing of new ones more difficult and time-consuming. Furthermore, there is usually a disturbing lack of interest that people needs to be able to employ their own language in computer applications.

In the aim to help in revitalizing endangered languages, that are generally under or less resourced languages, many efforts need to be made. One way is to encourage younger generations to use their mother tongue by building e-learning platforms, and creating instructive games. Oral documenting can be used to preserve the culture of endangered languages; especially that many of these languages are only spoken. They have rich oral cultures with stories, sayings, songs, chants and histories, but no written forms. So, the extinction of such language will quickly lead to the annihilation of its culture. Machine translation system can also be employed to produce translations from other languages, in order to extend the use of these languages from familiar and home use to more formal social contexts such as media, administration, and commercial relations. Another way to contribute in preserving endangered languages is the use of Internet. This later might be handy to raise awareness about the issues of language extinction and language preservation.

In this context, this paper presents the key strategies for improving endangered languages on human language technologies. It describes the experiments currently underway on Amazigh at Computer Science Studies, Information Systems and Communications Center (CEISIC) in the Royal Institute of the Amazigh Culture (IRCAM), in order to let this language becoming more intercommunicated and widely used in the community.

2. Strategies for enhancing under and less resourced languages

Recently, several private companies, technology centers, and public institutes have begun to get interested and to invest in developing technology for under and less resourced languages. To successfully deal with this task some studies have focused on studying the main strategies that could be taken in order to promote and develop this set of languages.

2.1. Linguistic contribution

Generally, the computational processing of a language involves linguistic contributions that consist on matching or modeling language competence by discovering and presenting formally the rules governing this language. These linguistic contributions can be efficiently shared by a collaborative work on the web [2], substituting a local development team with potentially bigger distributed team. This idea avoids reduplication and wastage of efforts and resources. It has been investigated in an early Montaigne project (1996), and has been implemented at GETA for the Lao language. It has also been applied by Oki to the Japanese language and by NII/NECTEC to a Japanese-Thai dictionary [3].

http://dx.doi.org/10.5772/51069

2.2. Resource recycling

Building electronic resources is indispensable parts of any computational language process. However, this task requires time and valuable human competence. An alternative solution for developing such resources is to get firstly electronic files by using Optical Character Recognition (OCR) [4], then to generate from these files a standardized suitable format of resources that will be exploitable for automated task.

The resource standardization is an important step in the process of resource building. It allows the reuse of resources in different researches, tools and applications. Furthermore, it facilitates the maintenance of a coherent document life cycle through various processing stages, and enables the enrichment of existing data with new information.

2.3. Adapting CLP techniques

Adapting Computational Language Processing (CLP) techniques is an interesting way to build new tools for a specific language while taking the advantages of the similarity between languages' features. Such experiment has been particularly applied in machine translation projects. One of these project is the 'MAJO system', where the investment of syntactical and morphological similarities between Japanese and Uighur has helped sufficiently to obtain good results [5].

2.4. Extensibility focused

The philosophy of this direction suggests that the conception of any project should be made in such a way that others can easily come and extend the work to another level. This means that the project's development should not focus only on getting results, but looking for others to be able to continue the work [6]. In this context, there are several examples: The 'Acabit system' has been developed firstly for the extraction of French multiword. Then, it has been extended to Japanese and Arabic languages [7]. Similarly, the 'NOOJ framework' has been built for European languages. Whereas, the work is still continuing on this framework for other languages such as Arabic [8], and Amazigh [9].

2.5. Open source focused

In general, the under and less resourced languages are economically poor. Whereas, doing computational language processing involves lots funds. To get around this obstacle and to cut down on the financial issues, it is suggested to adopt the open source strategy. Furthermore, this strategy will allow the adoption of the two previous directions (adapting CLP techniques and extensibility focused).

2.6. Professional documentation

Documentation will also greatly help in the continuation and the extension of projects. This documentation could be in terms of manuals or Websites, assisting people who may be interested in the use of a project, or allowing them to access to any phase of the work and continue its development.

2.7. Evaluation system

The evaluation system can be defined as a process allowing measuring the gap between fixed objectives and attained results. The choice of the time of the evaluation depends on the aim of the evaluation. Generally, the evaluation of a project could be done before its realization, to make a diagnostic that determines the objectives of this project and its prerequisites; during the development, to make a progressive evaluation that pilots and directs the progress of the development; and after the implementation, to make a final evaluation which yields the results of the level of satisfaction, relevance, durability of the project, and finally of the continuity and the extensibility of the project.

2.8. Road map

Conscious that search engine, machine translation, human-machine dialogue, and e-learning play a key role in the survival of under and less resourced languages, in manner that they will strongly help these languages to find their way into our daily lives by extending their use from familiar use to social one, we have organized and prepared a clear vision to realize these specific projects in a progressive approach.

While studying these projects, we have noted that:

- Search engine is designed to look for the information needed of the user by understanding his/her query, retrieving the relevant information related to the given query independently of the used language, and presenting a list of ranked search results. To this end, most of the search engines are based on automatic web crawlers, ranking algorithm, and relevance techniques of automatic indexing. These later either rely on keyword-based indexing, linguistic analysis indexing, concept-based indexing, or multilingual indexing.
- Machine translation objective is to allow translating text with roughly the skill of a human, by ensuring a high quality of hierarchical phrase-based translation. In this aim, most of the machine translation systems combine the strengths of rule-based and statistical approaches to reduce the amount of the required linguistic information and training data, and also reduce the size of the statistical models while maintaining high performance. The rule-based machine translation (RBMT) approach is described as interlingual or transfer-based machine translation. It is based on lexicons with morphological, syntactic, and semantic information, and sets of rules. While the statistical machine translation (SMT) approach is based on parallel corpora to properly train the translation system. The two approaches can be merged in different ways: either translation is performed using a rules based engine, then statistics are used in an attempt to adjust/correct the output from the rules engine; rules are used to pre-process data in an attempt to better guide the statistical engine; or rules are used to post-process the statistical output.
- Human-machine dialogue aims to support interactions between users and machines by designing receptive products to the user's needs. The Human-machine dialogue systems can be represented as a four process: Speech recognition process to transcribe sentences spoken into written text, natural language understanding process to extract the meaning from the text, execution process to perform actions on the conversation meaning, and response generation process to give feedback to the user.

http://dx.doi.org/10.5772/51069

- E-learning increases access to learning opportunities, by offering knowledge and skills' online transfer that can be ensured anytime and anywhere through a variety of electronic learning and teaching solutions such as Web-based courseware, online discussion groups, live virtual classes, video and audio streaming. Nowadays, modern technology, especially computational language processing, is strongly used in e-learning to assist reading, writing, and speaking a language. While a person writes a sentence or reads a text aloud, the system can correct and monitor which words are not right or even analyze and tutor particular problems.

From this study we have noticed that these projects are mutually related to each other, and one can act as a part of the other. Furthermore, they are based on various processing which requires a large amount of specialized knowledge. Therefore, we have identified a list of the necessary processes needed to ensure the functionality of these projects, and we have suggested arranging them in a road map chronologically for short, medium, and long term *according to the availability of resources and the level of functionality expected within each term* [10].

As discussed, the achievement of our goal requires a large amount of specialized knowledge that is mainly encoded in complex systems of linguistic rules and descriptions, such as grammars and lexicons, which will in turn involve a considerable amount of specialized manpower. Thus depending on the availability of linguistic expertise and resources, we have estimated that short term phase will necessitate at least 5-year to 10-year plan to establish the low level processing and resources, and pave the way for medium and long terms applications. While, based on the undertaken studies for well resourced language, we have gauged that the two other phases will demand only 5-year plan. Figure 1 represents the road map structured on these three phases.

2.8.1. Short term phase

This phase is considered as an initial step. It consists mainly on the identification of the language encoding, and the foundation of the primarily resources namely keyboard, fonts, basic lexical database (list of lemmas and affixes), and the elementary corpora that serve in the elaboration of most computational language processing applications (text raw corpus, corpus for evaluating search engine and information retrieval systems, manually part of speech tagged corpus, and speech corpus). Furthermore, basic tools and applications such encoding converter, sentence and token splitter, basic concordancer and web search engine, morphological generator, and optical character recognition system also need to be developed in this phase.

2.8.2. Medium term phase

After paving the way by the elaboration of the fundamental and the basic resources in the first phase, this one needs to be focused on advanced tools and applications. Based on the size and the representativity of the elaborated resources, the processing tools of this phase could be even rule-based or statistical. The most important processing to undertake, in this step, are stemming or lemmatization (depending on the morphological features of the stud-

ied language), part of speech tagging, morphological analyzer, chunker, syntactical analyzer, and speech recognition. These processing tools will enable to build a spell checker, a terminology extractor, a text generator, and human-machine dialogue. Furthermore, they will allow the enhancement of the first phase tools and applications.

The medium term phase represents also the time to prepare the necessary resources for the next step, including multilingual dictionaries, multilingual aligned corpora, and semantic annotated corpora.

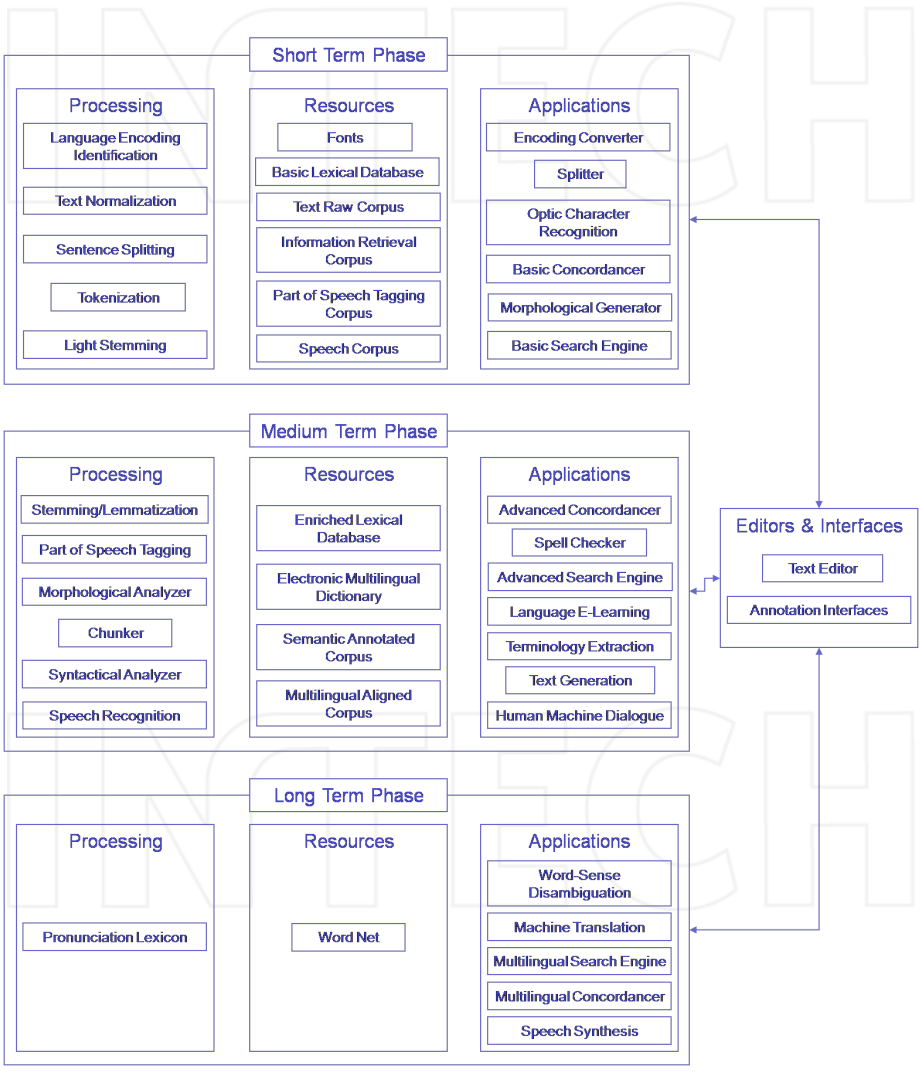

Figure 1. Road map for under and less resourced languages.

http://dx.doi.org/10.5772/51069

2.8.3. Long term phase

The third step of the road map could be considered as the synthesis phase of the realized work. Beside the elaboration of a pronunciation lexicon, Word Net, word-sense disambiguation and speech synthesis, this phase is also focused on the multilingualism applications, mainly machine translation system.

3. Amazigh language features

The Amazigh language, known as Berber or Tamazight, is a branch of the Afro-Asiatic (Hamito-Semitic) languages [11, 12]. 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.

3.1. Sociolinguistic context

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.

The Amazigh is spoken approximately by the half of Moroccan population, either as a first language or bilingually with the spoken Arabic dialect. However, it was until 1994 reserved only to family domain [13]. But in 2001, thanks to the King Mohammed VI Speech, which has established by a Dahir the creation of the Royal Institute of the Amazigh Culture, the Amazigh language has become an institutional language nationally recognized; and in July 2011, it has become an official language beside the classical Arabic.

3.2. Tifinaghe-IRCAM graphical system

Since the ancient time, the Amazigh language has its own script called Tifinaghe. It is found engraved in stones and tombs in some historical sites attested from 25 centuries. Its writing form has continued to change from the traditional Tuareg writing to the Neo-Tifinaghe in the end of the sixties, and to the Tifinaghe-IRCAM in 2003.

The Tifinaghe-IRCAM graphical system has been adapted, and computerized, in order to provide the Amazigh language an adequate and usable standard writing system. While, it has been chosen to represent to the best all the Moroccan Amazigh varieties, it tends to be phonological [14].

However, before adopting Tifinaghe-IRCAM as an official graphic system in Morocco, the Arabic script was widely used for religion and rural poetry writing, and the Latin script supported by the International Phonetic Alphabet (IPA) was used particularly in missionaries' works.

The Tifinaghe-IRCAM graphical system contains:

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

3.3. 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 the Tifinaghe system writing.

3.4. Directionality

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

3.5. Amazigh morphological properties

The main syntactic categories of the Amazigh language are the noun, the verb, and the particles [14, 15, 16].

3.5.1. Noun

In the Amazigh language, noun is a lexical unit, formed from a root and a pattern. It could occur in a simple form (ⴰⵔⴳⴰⵣ 'argaz' *the man*), compound form (ⵜⵓⴱⵓⵢⵢⵓⵀ 'buhyyuf' *the famine*), or derived one (ⴰⵎⵢⴰⵡⴰⵟ 'amyawaṭ' *the communication*). This unit varies in gender (masculine, feminine), number (singular, plural) and case (free case, construct case).

3.5.2. Verb

The verb, in Amazigh, has two forms: basic and derived forms. The basic form is composed of a root and a radical, while the derived one is based on the combination of a basic form and one of the following prefixes morphemes: ⵙ 's' / ⵙⵙ 'ss' indicating the factitive form, ⵜⵜ 'tt' marking the passive form, and ⵎ 'm' / ⵎⵎ 'mm' designating the reciprocal form. Whether basic or derived, the verb is conjugated in four aspects: aorist, imperfective, perfect, and negative perfect.

http://dx.doi.org/10.5772/51069

3.5.3. Particles

In the Amazigh language, particle is a function word that is not assignable to noun neither to verb. It contains pronouns, conjunctions, prepositions, aspectual, orientation and negative particles, adverbs, and subordinates. Generally, particles are uninflected words. However in Amazigh, some of these particles are flectional, such as the possessive and demonstrative pronouns (ⵜⴰ 'ta' *this* (fem.) ▪ ⵜⵉⵏⴰ 'tina' *these* (fem.)).

4. The complexity of Amazigh in CLP

Amazigh is an official language in Morocco. However, it has been less studied from the computational point of view for many years. Moreover, it is among the languages having rich morphology and different writing forms. Below we describe the difficulties that the Amazigh language confronts in developing computational language applications.

4.1. 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 CLP 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)).

4.2. Phonetic and phonology

The Amazigh phonetic and phonological problems depend particularly on the regional varieties. These problems consist on allophones and two kinds of correlations: the contrast between constrictive and occlusive consonants, and that between lax and tense ones.

- The allophone problems concern single phonemes that realized in different ways, such as /ll/ and /k/ that are pronounced respectively as [dž] and [š] in the North.
- The contrast between constrictive and occlusive consonants concern particularly the Riffian and the Central varieties. Those have a strong tendency to plosive spirantization, where b, t, d, ḍ, k, g become respectively b̲, t̲, d̲, k̲, g̲.

- In the phonological Amazigh system, all phonemes can alternate from lax to tense, which is characterized by greater articulator energy and often a longer duration. Some phonetic and phonological evidence consider the opposition lax versus tense as a tense correlation and not a gemination [17], while others consider this opposition as gemination [14]. Moreover, the realization of this opposition varies from region to region and from consonant to consonant.

4.3. 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 (ⴳⴳⵉⵎ 'qqim' ▪ ⵙⵎⵉⵎ 'svim' (make sit)).

5. Primarily experiments for the Amazigh language

For many decades the Amazigh language was solely oral, exclusively reserved for familial and informal domains, although 50% of the Moroccan population are Amazigh speakers [14]. Since the colonial period, many studies have been undertaken, but most of them have contributed to the collection of the Amazigh oral tradition or have focused on linguistic features. Whereas, the computational studies have been neglect until the creation of the IRCAM in 2001. This creation has enabled the Amazigh language to get an official spelling [18], proper encoding in the Unicode Standard [19], appropriate standards for keyboard realization, and linguistic structures [15, 18].

Nevertheless, this is not sufficient for a less-resourced language as Amazigh to join the well-resourced languages in information technology. In this context, many researches, based on the approaches used for well-resourced languages, are undertaken at national level to improve the current situation [20, 21, 22]. In the remainder of this paper we present existing systems and resources built for Amazigh languages.

5.1. Amazigh encoding

5.1.1. Tifinaghe encoding

Over several years, the Amazigh language has been writing in Latin alphabet supported by the IPA, or in Arabic script. While after adopting Tifinaghe as an official script in Morocco, the Unicode encoding of this script has become 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.

http://dx.doi.org/10.5772/51069

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 varieties, 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 Unicode Consortium accepts the 55 Tifinaghe characters for encoding in the range U+2D30..U+2D65, U+2D6F, with Tifinaghe block at U+2D30..U+2D7F [19].

5.1.2. Keyboard and fonts

Amazigh integration in international standard prescription keyboards ISO/IEC 9995 has fixed two keyboards: a basic one containing the Tifinaghe characters recommended by IRCAM, and an enhanced keyboard including all characters adopted by ISO. To facilitate keyboarding, Tifinaghe characters' position was chosen in a manner that matches their Latin correspondent position of Azerty keyboard (see Table 2).

In order to integrate the Amazigh language in the Moroccan education system in 2003, eight Tifinaghe fonts, associated with the ANSI encoding, were developed. Then a new Unicode font generation was succeeded [23].

5.2. 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. focused on isolated printed characters recognition based on a syntactic approach using finite automata [24]. Amrouch et al. proposed a global approach based on Hidden Markov Models for recognizing handwritten characters [20]. El Ayachi et al. presented a method using invariant moments for recognizing printed script [25]. Ait Ouguengay et al. proposed an artificial neural network approach to recognize printed characters [26].

5.3. Fundamental processing tools

5.3.1. Transliterator

The Amazigh language has known through its existence 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 an automatically passage from one form to another, a transliterator tool has been developed [27]. This later allows users to read or write in a suitable form, and converts all writing prescriptions into a standard unique form for text pre-processing tasks.

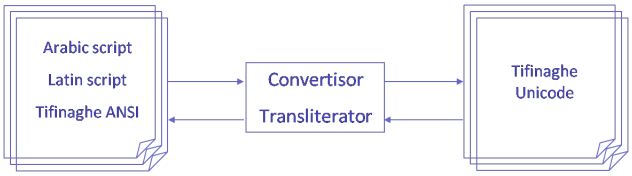

Figure 2. Transliterator tool conception.

The Amazigh transliterator consists of two processes: Convertisor and transliterator.

- Convertisor: This process allows the passage from ANSI representation of Tifinaghe into Unicode representation and vice versa.
- Transliterator: The transliterator process aims to transcribe words from one script to another, while conserving the phonetic pronunciation of the 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 correspondences illustrated in Table 2 are used. While in the Arabic - Tifinaghe Unicode mapping, there are more constrained rules to use the correspondences represented in Table 2. 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 (c.f. Table 1), and orthographic rules have been specified mainly on the transliteration from Arabic script into Tifinaghe one.

These rules are as follow:

- If the word contains any emphatic letter (ⴹ, ⵇ, ⵚ, ⵟ, ⵠ), the letter 'ر' will be represented by 'ⵔ'.
- 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'.

English appellation	Tifinaghe	Latin correspondence	New Arabic correspondence
yag	ⵅ	g	ڭ
yagw	ⵅ	g^w	ڭْ
yakw	ⴽ	k^w	كُ
yi	ⵉ	i	ـئـ (in the beginning of a word)
you	ⵓ	u	ؤ (in the beginning of a word)
yarr	ⵔ	ɣ	ر
yaw	ⵡ	w	و
yay	ⵢ	y	ي

Table 1. Tifinaghe – Arabic adapted correspondences.

http://dx.doi.org/10.5772/51069

5.3.2. Stemmer

To enhance the performance of information retrieval systems for the Amazigh language a computational stemming process was realized. This process consists in splitting Amazigh words into constituent parts (stem, prefix, suffix) without doing complete morphological analysis, in order to conflate word variants into a common stem [28].

The algorithm is merely based on an explicit list of prefixes and suffixes that need to be stripped in a certain order. This list is derived from the common inflectional morphemes of gender, number and case for nouns; personal markers, aspect and mood for verbs; and affix pronouns for kinship nouns and prepositions. While, the derivational morphemes are not included in order to keep the semantic meaning of words. See Figure 3 for more details.

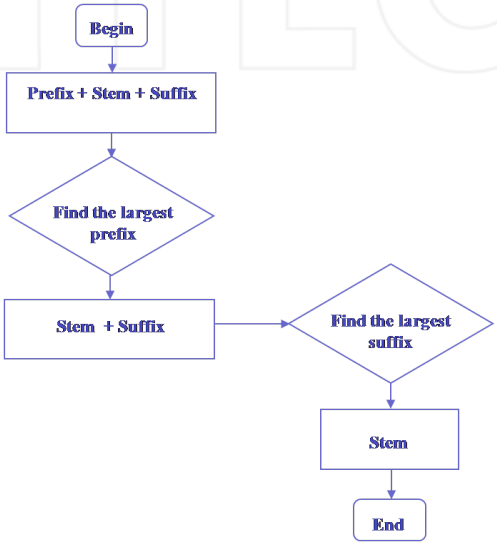

Figure 3. Amazigh light stemming process.

The set of prefixes and suffixes, that has been identified, are classified to five groups ranged from one character to five characters.

Prefix Set

- One-character: ⴰ ‘a’, ⵉ ‘i’, ⵏ ‘n’, ⵓ ‘u’, ⵜ ‘t’.
- Two-character: ⵏⴰ ‘na’, ⵏⵉ ‘ni’, ⵏⵓ ‘nu’, ⵜⴰ ‘ta’, ⵜⵉ ‘ti’, ⵜⵓ ‘tu’, ⵜⵜ ‘tt’, ⵍⴰ ‘wa’, ⵍⵓ ‘wu’, ⵢⴰ ‘ya’, ⵢⵉ ‘yi’, ⵢⵓ ‘yu’.
- Three-character: ⵉⵜⵜ ‘itt’, ⵏⵜⵜ ‘ntt’, ⵜⵜⴰ ‘tta’, ⵜⵜⵉ ‘tti’.
- Four-character: ⵉⵜⵜⴰ ‘itta’, ⵉⵜⵜⵉ ‘itti’, ⵏⵜⵜⴰ ‘ntta’, ⵏⵜⵜⵉ ‘ntti’, ⵜⵓⵜⵜ ‘tett’.
- Five-character: ⵜⵓⵜⵜⴰ ‘tetta’, ⵜⵓⵜⵜⵉ ‘tetti’.

Suffix Set

- One-character: ⵄ ‘a’, ⴷ ‘d’, ⵉ ‘i’, ⴽ ‘k’, ⵎ ‘m’, ⵏ ‘n’, ⵖ ‘γ’, ⵙ ‘s’, ⵜ ‘t’.
- Two-character: ⴰⵏ ‘an’, ⴰⵜ ‘at’, ⵉⴷ ‘id’, ⵉⵎ ‘im’, ⵉⵏ ‘in’, ⵉⵖ ‘iγ’, ⵎⵜ ‘mt’, ⵏⵖ ‘nγ’, ⵏⵜ ‘nt’, ⵓⵏ ‘un’, ⵙⵏ ‘sn’, ⵜⵏ ‘tn’, ⵡⵎ ‘wm’, ⵡⵏ ‘wn’, ⵢⵏ ‘yn’.
- Three-character: ⴰⵎⵜ ‘amt’, ⴰⵏⵜ ‘ant’, ⴰⵡⵏ ‘awn’, ⵉⵎⵜ ‘imt’, ⵉⵏⵜ ‘int’, ⵉⵡⵏ ‘iwn’, ⵏⵉⵏ ‘nin’, ⵓⵏⵜ ‘unt’, ⵜⵉⵏ ‘tin’, ⵜⵏⵖ ‘tnγ’, ⵜⵓⵏ ‘tun’, ⵜⵙⵏ ‘tsn’, ⵙⵏⵜ ‘snt’, ⵡⵎⵜ ‘wmt’.
- Four-character: ⵜⵓⵏⵜ ‘tunt’, ⵜⵙⵏⵜ ‘tsnt’.

5.3.3. Search engine

As the number of Amazigh documents grew, searching algorithms have become one of the most essential tools for managing information. Thus, Ataa Allah and Boulaknadel have been proposed a first tentative in order to develop a search engine that could support the Amazigh language characteristics [21]. They have tried to develop an Amazigh search engine that is mainly structured on three parts: data crawling, indexing, and searching.

- Data crawling: Data crawling is a process behind the search engine, based on spiders or web robots, to fetch web pages. In this context, they have developed a spider that collects automatically and daily all the pages containing Tifinaghe script from the IRCAM Website. All the fetched pages are stored in a web pages’ repository, and associated to an ID number. To this ID, called also docID, a web page is assigned whenever a new URL is parsed.
- Data indexing: Based on vector space model [29], the data indexing system creates an index for a set of documents and assigns the Okapi BM-25 weight to each term-document association. The Okapi formulas, especially the BM-25 scheme, attack the problem of higher term frequencies in long documents, and the chances of retrieving long documents over shorter ones [30]. Considering the common use of infixes for both Arabic and Amazigh languages, and based on the study undertaken in Arabic information retrieval [31], the data indexing system of this search engine performs four steps on each document: word identification, stop-word removal, light stemming, and indexing. First, it identifies individual words in the document. Second, all the “stop words” in a document are removed based on a pre-defined list. Then, it reduces the words to their light stem. Finally, it records the information about the relationships between these words and the documents to support searching.
- Data searching: The data searching system includes query engine and a web user interface. The query engine accepts search queries from users, applies the light stemming, represents the query as a vector in term-document space, and assigns a weight to each term-query. Then, it calculates scores between the user query and the set of documents. After retrieving search results, the query engine ranks the search results according to content analysis scores, generates a summary for each search result, based on the web pages’ repository, and renders its link. Whereas, the web user interface allows submitting queries and view the search results. When a user performs a search through the web interface, the query is passed to the query engine, which retrieves the search results and passes them back to the user, who can specify the number of retrieved web pages per each result page.

http://dx.doi.org/10.5772/51069

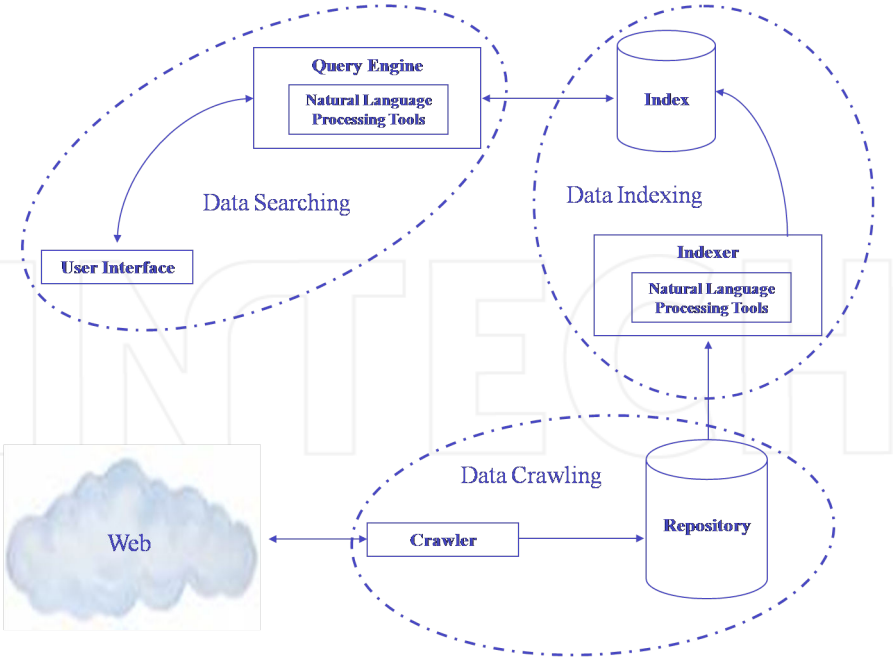

Figure 4. Amazigh search engine architecture.

5.3.4. 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 displaying a specified target word in its context. However, obtaining one that can reliably cope with the Amazigh language characteristics has proved an extreme difficulty. Therefore, an online concordancer that supports all Moroccan Amazigh language scripts was developed [32].

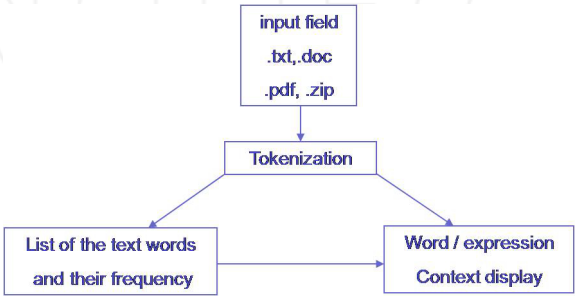

Figure 5. Amazigh concordancer conception.

This concordancer is relatively simple to get up and running. Its interface ensures:

- Corpus selection: The Amazigh concordancer accepts compressed file format and most common text file formats including proprietary and non-proprietary formats such as Microsoft Word, RTF, and PDF files.
- Query keyboarding: The goal of the Amazigh concordance system is to allow a user to look for instances of specific words or expressions. For this purpose, the system opens up two ways for the user to enter his/her query in any one of the scripts involved. The first way is by specifying one of the following kinds of query: exact single word query, or exact string query. Whereas, the second way allows choosing a word from a list in the frequency panel.
- Word frequency panel: The Amazigh concordancer displays a list of the corpus words and their frequency.
- Display of results: Concordances generated by the Amazigh concordancer can be sent either to screen, to printer as hard copy, or to file for future manipulation.
- Context size choice.

5.3.5. Tagging assistance tool

The use of corpora in CLP, especially those annotated morphosyntactically, has become an indispensable step in the language tools' production and in the process of language computerization. In this context, an initiative has been taken to build a morphosyntactic corpus, which has elicited the development of a tool providing support and linguists' assistance [33].

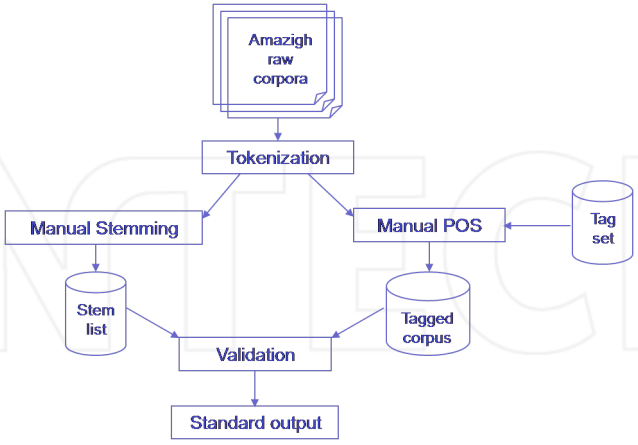

Figure 6. Tagging assistance tool conception.

This tool is structured in three-step process: management, annotation, and validation processes.

http://dx.doi.org/10.5772/51069

- Management: Management is a process for managing user's account and electronic documents. It allows creating or deleting user account, and modifying user information or privileges. Also, it provides to store different range of data, namely original documents, transliterated or corrected ones. Furthermore, it provides a standardized format of the annotated corpus.
- Annotation: This process assists linguists to grammatically annotate a text, by providing them sentences segmented into single words, and allowing them the ability to select for each word its morphological tag from a list of tags that has been elaborated in collaboration with some linguists. Moreover, it requires them to specify the appropriate lemma of each word. Then, the process stores the triples (word, tag, lemma) in the database for further processing.
- Validation: Validation process consists on comparing tags and lemmas specified by all linguists working on the same document, and extracting a list of anomalies. Then, the process returns the context of each word in the list and the information related to, namely the name of linguists that have annotated the word along with the selected tags and the specified lemmas. After, the process allows the correction by updating inconsistent annotations.

5.4. Language resources

Human language technologies are 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.

5.4.1. Corpora

Corpora are a very valuable resource for CLP tasks, but Amazigh lacks such resources. Therefore, researchers at IRCAM have tried to build Amazigh corpora in progressive way until reaching a large-scale corpus that follows TREC's standards. Thus, three parallel works are undertaking [[34], [35], [36]]. 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; whereas the third work is dealing with the construction of a corpus for evaluating information retrieval system, where the queries' creation and the relevance assessment are manually performed by narrative speakers.

5.4.2. 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 [22]. This application has provided all necessary information such as definition, Arabic French and English equivalent words, synonyms, classification by domains, and derivational families. Moreover, it provides the possibility to generate a word copy of the dictionary.

5.4.3. 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 [37].

6. Conclusion and future directions

In this paper, we have outlined the key strategies to uphold under and less resourced languages on human language technologies. We have specified a road map based on the elementary and necessary processing, resources and applications needed to ensure the survival of under and less resourced languages in "information society". Moreover, we have discussed the main challenges in processing the Amazigh language, and we have attempted to survey the research work on Amazigh CLP in Morocco.

In the aim to convert Amazigh language from a less resourced language into a resourced and well studied language from computational point of view we need to expedite the basic research on Amazigh CLP 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.
- Re-use the existing language technologies developed over the years for resourced languages and particularly for Arabic that shares with Amazigh a number of linguistic properties, which will allow saving time and funds.
- Elaborating an E-learning application to ensure the language and culture transmission to young generation.
- 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 CLP by offering scholarship for higher degrees and attracting young researchers with attractive salary.

However, there are other research tracks, especially those related to recent technologies, that need to be taken into account to attract young generations, namely:

- Investing on media social contents will strongly contribute in the salvation of many less resourced languages.
- Developing mobile phone and tablet applications will also keep language alive for future generations and help foreign visitors to discover the region for better understanding of the local culture.
- Elaborating special applications in Amazigh language for people with disabilities, in order to offer them opportunities to access information and services in their native language.

http://dx.doi.org/10.5772/51069

Apendix

English appellation	Tifinaghe	Latin correspondence	Arabic correspondence
ya	ⵄ	a	ا
yab	ⴱ	b	ب
yag	ⴳ	g	گ
yagw	ⴳⵯ	g^w	گ‘
yad	ⴷ	d	د
yadd	ⴹ	ḍ	ض
yey	ⵉ	e	
yaf	ⴼ	f	ف
yak	ⴽ	k	ک
yakw	ⴽⵯ	k^w	ک‘
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	ڕ
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	ۆ
yay	ⵢ	y	يˆ
yaz	ⵣ	z	ز
yazz	ⵥ	ẓ	ژ

Table 2. Tifinaghe-Ircam Alphabet.

Author details

Fadoua Ataa Allah and Siham Boulaknadel

Computer Science Studies, Information Systems and Communications Center, The Royal Institute of the Amazigh Culture, Rabat, Morocco

References

[1] Brickley, D. A. (1995). The Conservation of Endangered Languages. In: Seminar at the centre for theories of language and learning. *University of Bristol department of philosophy, Bristol, UK*.

[2] Boitet, C. (1999). A research perspective on how to democratize machine translation and translation aids aiming at high quality final output. *In: proceeding of VII Summit of Machine Translation in the Great Translation Era*, MT VII, 13-17 September, Singapore.

[3] Shimohata, S., Kitamura, M., Sukehiro, T., & Murata, T. (2001). Collaborative translation environment on the Web., *In: proceeding of VIII Summit of Machine Translation in the Information Age*, MT, 18-22 September, Santiago de Compostela, Spain.

[4] Nguyen, H. D. (1998). Techniques génériques d'accumulation d'ensembles lexicaux structurés à partir de ressources dictionnairiques informatisées multilingues hétérogènes. PhD thesis. *University of Joseph Fourier*.

[5] Mahsut, M., Ogawa, Y., Sugino, K., & Inagaki, Y. (2001). Utilizing agglutinative features in Japanese-Uighur machine translation. *In: proceeding of VIII Summit of Machine Translation in the Information Age*, MT, 18-22 September, Santiago de Compostela, Spain.

[6] Muhirwe, J. (2007). Towards Human Language Technologies for Under-resourced languages. *In Joseph Kizza et al. (ed.) Series in Computing and ICT Research*, 2Fountain Publishers, 123-128.

[7] Boulaknadel, S. (2008). TAL et recherche d'information dans un domaine de spécialité : apport des connaissances morphologiques et syntaxique pour l'arabe. *Ph.D. Thesis. University Mohamed V*.

[8] Mesfar, S. (2008). Etude linguistique de la langue arabe. *Ph.D. Thesis. University of Franche Comté*.

[9] Nejm, F. Z., & Boulaknadel, S. (2012). Formalisation de l'amazighe standard avec NooJ. *In: proceeding of atelier de traitement automatique des langues africaines*, TALAF, 04-08 June, Grenoble, France.

http://dx.doi.org/10.5772/51069

[10] Ataa Allah, F., & Boulaknadel, S. (2009). Note méthodologique pour la réalisation des outils et ressources linguistiques de la langue amazighe. *Intern Report. IRCAM, Rabat, Morocco*.

[11] Greenberg, J. (1966). The Languages of Africa. *The Hague*.

[12] Ouakrim, O. (1995). Fonética y fonología del Bereber. *Survey, University of Autònoma de Barcelona*.

[13] Boukous, A. (1995). Société, langues et cultures au Maroc : Enjeux symboliques. *Najah El Jadida*.

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

[15] Boukhris, F., Boumalk, A., Elmoujahid, E., & Souifi, H. (2008). La nouvelle grammaire de l'amazighe. *IRCAM, Rabat, Morocco*.

[16] Ataa Allah, F., & Boulaknadel, S. (2010). Light Morphology Processing for Amazighe Language. *In: proceeding of the Workshop on Language Resources and Human Language Technology for Semitic Languages*, 17 May, Valletta, Malta.

[17] Galand, L. (1953). La phonétique en dialectologie berbère. *In : Bulletin Internationale de Documentation Linguistique. Orbis*, 225-233.

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

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

[20] 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), 11-18.

[21] Ataa Allah, F., & Boulaknadel, S. (2010). Amazigh Search Engine: Tifinaghe Character Based Approach. *In: proceeding of International Conference on Information and Knowledge Engineering, IKE*, 14-16 July, Las Vegas, Nevada, USA.

[22] Iazzi, E., & Outahajala, M. (2008). Amazigh Data Base. *In: proceeding of HLT & NLP Workshop within the Arabic world: Arabic language and local languages processing status updates and prospects*, 31 May, Marrakech, Morocco.

[23] Ait Ouguengay, Y. (2007). Quelques aspects de la numérisation des polices de caractères : Cas de Tifinaghe. *In : Zenkouar L. (ed.) La typographie entre les domaines de l'art et de l'informatique. IRCAM*, 159-181.

[24] 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), 1-8.

[25] 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), 61-66.

[26] Ait Ouguengay, Y., & Taalabi, M. (2009). Elaboration d'un réseau de neurones artificiels pour la reconnaissance optique de la graphie amazighe : Phase d'apprentissage. *In : Bellafkih M., Ramdani M., Zreik K. (ed.) Systèmes intelligents-Théories et applications. Europia productions*.

[27] Ataa Allah, F., & Boulaknadel, S. (2011). Convertisseur pour la langue amazighe : script arabe- latin- tifinaghe. *In: proceeding of the 2ème Symposium International sur le Traitement Automatique de la Culture Amazighe, SITACAM*, 6-7 May, Agadir, Morocco.

[28] Ataa Allah, F., & Boulaknadel, S. (2010). Pseudo-racinisation de la langue amazighe. *In: proceeding of Traitement Automatique des Langues Naturelles, TALN*, 19-23 July, Montreal, Canada.

[29] Gerard, Salton. (1968). Automatic Information Organization and Retrieval. *McGraw-Hill*.

[30] Stephen, E. R., Steve, W., Susan, J., Micheline, H. B., & Mike, G. (1994). Okapi at TREC-3. *In: proceeding of the 3rd Text Retrieval Conference, TREC*, November, Gaithersburg, Maryland, USA.

[31] Ataa Allah, F., Boulaknadel, S., El Qadi, A., & Aboutajdine, D. (2008). Evaluation de l'Analyse Sémantique Latente et du Modèle Vectoriel Standard Appliqués à la Langue Arabe. *Revue de Technique et Science Informatiques*, 27(7), 851-877.

[32] Ataa Allah, F., & Boulaknadel, S. (2010). Online Amazigh Concordancer. *In: proceeding of the international Symposium on Image Video Communications and Mobile Networks, ISIVC*, 30 September-2 October, Rabat, Morocco.

[33] Ataa Allah, F., & Jaa, H. (2009). Etiquetage morphosyntaxique : outil d'assistance dédié à la langue amazighe. *In: proceeding of the 1ère Symposium International sur le Traitement Automatique de la Culture Amazighe, SITACAM*, 12-13 December, Agadir, Morocco.

[34] Boulaknadel, S., & Ataa Allah, F. (2011). Building a standard Amazigh corpus. *In: proceeding of the International Conference on Intelligent Human Computer Interaction, IHCI*, 29-31 August, Prague, Tchec.

[35] 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*, 17 May, Valletta, Malta.

[36] Ataa Allah, F., & Boulaknadel, S. (2011). Les ressources langagières pour la recherche d'information textuelle: Cas de la langue amazighe. *In: proceeding of colloque sur l'Amazighe et les Nouvelles Technologies de l'Information et de Communication, NTIC*, 24-25 February, Rabat, Morocco.

[37] EL Azrak, N., & EL Hamdaoui, A. (2011). Référentiel de la Terminologie Amazighe : Outil d'aide à l'aménagement linguistique. *In: proceeding of colloque sur l'Amazighe et les Nouvelles Technologies de l'Information et de Communication, NTIC*, 24-25 February, Rabat, Morocco.