# Light Morphology Processing for Amazighe Language

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

### Abstract

In the aim to allow the Amazighe language an automatic processing, and integration in the field of Information and Communication Technology, we have opted in the Royal Institute of Amazighe Culture "IRCAM" for an innovative approach of progressive realizations. Thus since 2003, researchers in the Computer Sciences Studies, Information Systems and Communications Center "CEISIC" have paved the way for elaborating linguistic resources, basic natural language processing tools, and other advanced scientific researches by encoding Tifinaghe script and developing typefaces.

In this context, we are trying through this paper to develop a computationally stemming process which is based on analyzing words to their stems. This process consists in splitting Amazighe words into constituent stem part and affix parts without doing complete morphological analysis. This approach of light stemming will conflate word variants into a common stem in order to be used in natural language applications such as indexation, information retrieval systems, and classification.

## 1. Introduction

Stemming has been widely used in several fields of natural language processing such as data mining, information retrieval, machine translation, document summarisation, and text classification, in which the identification of lexical occurrences of words referring to some central idea or 'meaning' is involved. Indeed, the lexical analysis is mainly based on word occurrences, which require some form of morphological conflation that could range from removing affixes to using morphological word structures.

In literature, many strategies of stemming algorithms have been published for different languages, such as English (Lovins 1968; Porter, 1980), French (Savoy, 1993; Paternostre et al., 2002), and Arabic (Larkey et al., 2002; Taghva et al., 2005; Al-Shammari and Lin, 2008). In general, the stemmer structures vary considerably depending on the morphology of languages. For Indo-European languages, most basic techniques consist on removing suffixes; while, for the Afro-Asiatic ones, these techniques are extended to stripping prefixes.

In practice, affixes may alter the meaning of words. So, the fact to remove them would greatly discard vital information. In the Indo-European languages, prefixes modify the word meaning which make their deletion not helpful. While in the Afro-Asiatic languages, the prefixes are also used to fit the word for its syntactic role. Thus, in this paper, we propose an Amazighe stemming algorithm that consists in removing the common inflectional morphemes placed at the beginning and/or the end of words.

The remaining of the paper is organized as follows: in Section 2, we give a brief description of the Moroccan standard Amazighe language. Then, in Section 3, we give an overview about the Amazighe language characteristics. In Section 4, we present our light stemming algorithm. Finally, section 5 gives general conclusions, and draws some perspectives.

## 2. Moroccan Standard Amazighe Language

The Amazighe language, known as Berber or Tamazight, is a branch of the Afro-Asiatic (Hamito-Semitic) language family. 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 into three mean regional varieties: Tarifite in North, Tamazight in Central Morocco and South-East, and Tachelhite in the South-West and the High Atlas. Even though 50% of the Moroccan population are amazighe speakers, the Amazighe language was exclusively reserved for familial and informal domains (Boukous, 1995). However, in the last decade, this language has become institutional.

Since the ancient time, the Amazighe language has its own writing that was adapted by the Royal Institute of the Amazighe Culture (IRCAM) in 2003, to provide an adequate and usable standard alphabetic system called Tifinaghe-IRCAM. This system contains:

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

Furthermore, the IRCAM has recommended the use of the International symbols for punctuation markers: " " (space), ".", ",", ";", ":", "?", "!", "…"; the standard numeral used in Morocco (0, 1, 2, 3, 4, 5, 6, 7, 8, 9); and the horizontal direction from left to right for Tifinaghe writing (Ameur et al., 2004).

# 3. Amazighe Language Characteristics

The purpose of this section is to give an overview of the morphological properties of the main syntactic amazighe categories, which are the noun, the verb, and the particles (Boukhris et al., 2008; Ameur et al., 2004).

## 3.1 Noun

In Amazighe 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 (ⴱⵓⵀⵢⵢⵓⴼ “buhyyuf” *the famine*), or derived one (ⴰⵎⵙⴰⵡⴰⴹ “amsawaḍ ” *the communication*). This unit varies in gender, number and case.

- Gender: Nouns are categorised by grammatical gender: masculine or feminine. Generally, the masculine begin with an initial vowel ⴰ “a”, ⵉ “i”, or ⵓ “u”. While, the feminine, used also to form diminutives and singulatives, is marked with the circumfix ⵜ…ⵜ “t…t” (ⴰⵎⵃⴹⴰⵕ “amḥḍaṛ ”masc., ⵜⴰⵎⵃⴹⴰⵕⵜ “tamḥḍaṛt” fem. *the student*).

- Number: There are two types: singular and plural, which has three forms. The external plural consists in changing the initial vowel, and adding the suffix ⵏ or one of its variants ⵉⵏ “in”, ⴰⵏ “an”, ⵢⵏ “yn”, ⵡⵏ “wn”, ⴰⵡⵏ “awn”, ⵉⵡⵏ “iwn”, ⵜⵏ “tn” (ⵉⵎⵃⴹⴰⵕⵏ “im amḥḍarn” masc., ⵜⵉⵎⵃⴹⴰⵕⵉⵏ “timḥḍarin” fem. *students*). The broken plural involves a change in the vowels of the noun (ⴰⴷⵔⴰⵔ “adrar” *mountain* → ⵉⴷⵓⵔⴰⵔ idurar *mountains*, ⵜⵉⵖⵎⵙⵜ “tiγmst” *tooth* → ⵜⵉⵖⵎⴰⵙ “tiγmas” *teeth*). The mixed plural is formed by the combination of vowels’ change and the use, sometimes of the suffixation ⵏ (ⵉⵣⵉ “izi” *fly* → ⵉⵣⴰⵏ “izan” *flies*, ⴰⵎⴳⴳⴰⵔⵓ “amgguru” *last* → ⵉⵎⴳⴳⵓⵔⴰ “imggura” *lasts*).

- Case: Two cases are distinguished. The free case is unmarked, while the construct one involves a variation of the initial vowel (ⴰⵔⴳⴰⵣ “argaz” *man* → ⵓⵔⴳⴰⵣ, “urgaz” ⵜⴰⵎⵖⴰⵔⵜ “tamγart” *woman* → ⵜⵎⵖⴰⵔⵜ “tmγart”).

## 3.2 Verb

The verb, in Amazighe, has two forms: basic and derived forms. The basic form is composed of a root and a radical (ⴼⴼⵖ “ffγ” *leave*), while the derived one is based on the combination of a basic form and one of the following prefixes morphemes: ⵙ/ⵙⵙ “s/ss”, ⵜⵜ “tt” and ⵎ/ⵎⵎ “m/mm” (ⵙⵙⵓⴼⵖ “ssufγ” *bring out*). Whether basic or derived, the verb is conjugated in four aspects: aorist, imperfective, perfect, and negative perfect. Moreover, it is constructed using the same personal markers for each mood, as represented in Table1.

## 3.3 Particles

In Amazighe 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 Amazighe, some of these particles are flectional, such as the possessive and demonstrative pronouns (ⵜⴰ “ta” *this* (fem.) → ⵜⵉⵏⴰ “tina” *these* (fem.)).

# 4. Light Stemming Algorithm

The light stemming refers to a process of stripping off a small set of prefixes and/or suffixes, without trying to deal with infixes, or recognizing patterns and finding roots (Larkey, 2002). As a first edition of such work in the IRCAM, with regard to the lack of huge digital corpus availability, our method is based only on the composition of words that is usually formed in the Moroccan standard Amazighe language as a sequence of prefix, core, and suffix. We are assuming that we are not making use of any stem dictionary or exception list. Our 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. It is very reasonable to conflate the noun ⵜⴰⵔⴱⴰⵜ “tarbat” *girl* with its masculine form “ⴰⵔⴱⴰ” arba *boy*; while it seems unreasonable, for some application like information retrieval, to conflate the derived verb ⵙⵙⵓⴼⵖ “ssufγ” *bring out* with the simple form ⴼⴼⵖ “ffγ” *leave*.

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

## 4.1 Prefix Set

- One-character: ⴰ, ⵉ, ⵏ, ⵓ, ⵜ.
- Two-character: ⵏⴰ, ⵏⵉ, ⵏⵓ, ⵜⴰ, ⵜⵉ, ⵜⵓ, ⵜⵜ, ⵡⴰ, ⵡⵓ, ⵢⴰ, ⵢⵉ, ⵢⵓ.
- Three-character: ⵉⵜⵜ, ⵏⵜⵜ, ⵜⵜⴰ, ⵜⵜⵉ.
- Four-character: ⵉⵜⵜⴰ, ⵉⵜⵜⵉ, ⵏⵜⵜⴰ, ⵏⵜⵜⵉ, ⵜⵓⵜⵜ.
- Five-character: ⵜⵓⵜⵜⴰ, ⵜⵓⵜⵜⵉ.

## 4.2 Suffix Set

- One-character: ⴰ, ⴷ, ⵉ, ⴽ, ⵎ, ⵏ, ⵖ, ⵙ, ⵜ.
- Two-character: ⴰⵏ, ⴰⵜ, ⵉⴷ, ⵉⵎ, ⵉⵏ, ⵉⵖ, ⵎⵜ, ⵏⵖ, ⵏⵜ, ⵓⵏ, ⵙⵏ, ⵜⵏ, ⵡⵎ, ⵡⵏ, ⵢⵏ.
- Three-character: ⴰⵎⵜ, ⴰⵏⵜ, ⴰⵡⵏ, ⵉⵎⵜ, ⵉⵏⵜ, ⵉⵡⵏ, ⵏⵉⵏ, ⵓⵏⵜ, ⵜⵉⵏ, ⵜⵏⵖ, ⵜⵓⵏ, ⵜⵙⵏ, ⵙⵏⵜ, ⵡⵎⵜ.
- Four-character: ⵜⵓⵏⵜ, ⵜⵙⵏⵜ.

|  | Indicative mood |  |  | Imperative mood |  |  | Participial mood |
|---|---|---|---|---|---|---|---|
|  |  | Masculine | Feminine |  | Masculine | Feminine | Masculine / Feminine |
| Singular | 1st pers.<br>2ndpers.<br>3rdpers. | ... ⵖ<br>ⵜ ... ⴷ<br>ⵉ_..._____ | ... ⵖ<br>ⵜ ... ⴷ<br>ⵜ_..._____ | 2ndpers. | ... Ø | … Ø | ⵉ….ⵏ |
| Plural | 1st pers.<br>2ndpers.<br>3rdpers. | ⵏ ...<br>ⵜ ... ⵎ<br>... ⵏ | ⵏ ...<br>ⵜ ... ⵎⵜ<br>... ⵏⵜ | 2ndpers. | ... ⴰⵜ/ⵜ/ⵎ | ... ⴰⵎⵜ/ ⵎⵜ | ….ⵏⵉⵏ |

Table 1: Personal markers for the indicative, imperative and participial moods

Based on this list of affixes and on theoretical analysis, we notice that the proposed amazighe light stemmer could make two kinds of errors:

- The understemming errors, in which words referring to the same concept are not reduced to the same stem, such the case of the verb ⴼⴼⵖ "ffγ" *leave* that ends with the character ⵖ "γ", which coincides with the 1st singular personal marker. So, the stem ⴼⴼⵖ "ffγ" of the verb when is conjugated in the perfect aspect for the 1st singular person ⴼⴼⵖⵖ "ffγγ" *I left* will not be conflated with stem ⴼⴼ "ff" of the 3rd singular masculine person ⵉⴼⴼⵖ "iffγ" *he left*.
- The overstemming errors, in which words are converted to the same stem even though they refer to distinct concepts, such the example of the verb ⴳ "g" *do* and the noun ⴰⴳⴰ "aga" *bucket*. The stem ⴳ "g" of the verb when is conjugated in the perfect aspect for the 3rd singular masculine person ⵉⴳⴰ "iga" *he did* will be conflated with stem ⴳ "g" of the noun ⴰⴳⴰ "aga".

In general, light stemmers avoid the overstemming errors, especially for the Indo-European languages; however, it is not the case of the Amazighe language. This proves that the Amazighe language constitutes a significant challenge for natural language processing.

## 5. Conclusion

Stemming is an important technique for highly inflected language such as Amazighe. In this work, we have investigated on the Amazighe language characteristics, and have presented a light stemming approach for Amazighe. We should note that the proposed stemming algorithm is primarily for handling inflections – it does not handle derivational suffixes, for which one would need a proper morphological analyzer.

In attempt to improve the amazighe light stemmer, we plan to build a stem dictionary, to elaborate a set of linguistic rules, and to set a list of exceptions to further extend the stemmer.

## 6. Appendix

| Tifinaghe | Latin Correspondence | Tifinaghe | Latin Correspondence |
|---|---|---|---|
| ⴰ | a | ⵍ | l |
| ⴱ | b | ⵎ | m |
| ⴳ | g | ⵏ | n |
| ⴳⵯ | $g^w$ | ⵓ | u |
| ⴷ | d | ⵔ | r |
| ⴹ | ḍ | ⵕ | ṛ |
| ⴻ | e | ⵖ | γ |
| ⴼ | f | ⵙ | s |
| ⴽ | k | ⵚ | ṣ |
| ⴽⵯ | $k^w$ | ⵛ | c |
| ⵀ | h | ⵜ | t |
| ⵃ | ḥ | ⵟ | ṭ |
| ⵄ | ε | ⵡ | w |
| ⵅ | x | ⵢ | y |
| ⵇ | q | ⵣ | z |
| ⵉ | i | ⵥ | ẓ |
| ⵊ | j |  |  |

Table 2: Tifinaghe-Ircam Alphabet

## 7. References

Al-shammari, E. T., Lin, J. (2008). Towards an error-free Arabic stemming. *Actes de the 2nd ACM workshop on improving non English web searching.* pp.9--16.

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

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

Larkey, L. S., Ballesteros, L., Connell, M. (2002). Improving Stemming for Arabic Information Retrieval: Light Stemming and Cooccurrence Analysis. *In Proceedings of the 25th Annual International Conference on Research and Development in Information Retrieval*. Tampere, Finland, pp. 275--282.

Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11(1), pp. 22--31.

Paternostre, M., Francq, P., Lamoral, J., Wartel, D., Saerens, M. (2002). Carry, un algorithme de désuffixation pour le français. Rapport technique du projet Galilei.

Porter, M.F. (1980). An algorithm for suffix stripping. *Program*, 14(3), pp.130--137.

Savoy, J. (1993). Stemming of French words based on grammatical categories. *Journal of the American Society for Information Science*, 44(1), pp.1--9.

Taghva, K., Elkhoury, R., Coombs, J. (2005). Arabic stemming without a root dictionary. *In Proceeding of Information Technology: Coding and Computing*. Las Vegas, pp.152--157.