CIT. Journal of Computing and Information Technology, Vol. 24, No. 1, March 2016, 91–110
doi: 10.20532/cit.2016.1002478

AmAMorph: Finite State Morphological Analyzer for Amazighe

Fatima Zahra Nejme[1], Siham Boulaknadel[2] and Driss Aboutajdine[1]

[1]LRIT, associated unit to the CNRST-URAC n°29, Mohammed V University, Faculty of Science, Rabat, Morocco
[2]CEISIC, Royal Institute of Amazigh Culture, Rabat, Morocco

This paper presents AmAMorph, a morphological analyzer for Amazighe language using a system based on the NooJ linguistic development environment.

The paper begins with the development of Amazighe lexicons with large coverage formalization. The built electronic lexicons, named 'NAmLex', 'VAmLex' and 'PAmLex' which stand for 'Noun Amazighe Lexicon', 'Verb Amazighe Lexicon' and 'Particles Amazighe Lexicon', link inflectional, morphological, and syntactic-semantic information to the list of lemmas. Automated inflectional and derivational routines are applied to each lemma producing over inflected forms.

To our knowledge, AmAMorph is the first morphological analyzer for Amazighe. It identifies the component morphemes of the forms using large coverage morphological grammars. Along with the description of how the analyzer is implemented, this paper gives an evaluation of the analyzer.

ACM CCS (2012) Classification: Computing methodologies ⟶ Artificial intelligence ⟶ Natural language processing ⟶ Phonology / morphology

Keywords: Amazighe language, Natural Language Processing, NooJ, Lexical Analysis, Inflectional Morphology, Derivational Morphology, Unknown words, Recognition Systems

1. Introduction

The Moroccan Amazighe language is considered as a prominent constituent of the Moroccan culture due to its richness and originality. However, it has been long discarded and otherwise neglected as a source of cultural enrichment. Nevertheless, due to the creation of the Royal Institute of Amazighe Culture (IRCAM) (see Appendix J), this language has been introduced in the public domain including administration, media and the educational system. It has enjoyed its proper coding in the Unicode Standard [1][2], an official spelling [3], appropriate standards for keyboard realization and linguistic structures that are being developed with a phased approach [4][5]. This process was initiated by the standardization, vocabularies construction [6][7][8][9], spelling standardization [3] and development of grammar rules [4].

However, all these stages are not sufficient for less-resourced languages as Amazighe to join the group of well-resourced ones in terms of language technologies. In this context, many scientific studies have been undertaken to improve the current situation. These studies can be divided into two categories:

(1) Computational resources: this can be subdivided into three subdomains:

- Tifinagh promotion works [10][11],
- Optical character recognition [12][13][14],
- Amazighe corpora [15][16].

(2) NLP tools: overall, NLP tools for Amazighe have been limited and carried out on light stemmer [17], tagging assistance tool [18], Search engine [19], Concordancer [20] and verb Conjugator [21]. Nevertheless, for the morphological analysis, which presents a core part of NLP applications, there are only a few publications that explicitly discuss the attempts at developing full-fledged morphological analyzers for the Amazighe language [22][23][24][25][26][27].

To this end, this paper presents the continuation of our previous efforts, which are restricted

to the noun processing and describes a step toward the development of full-fledged morphological analyzer tool, using finite state technology within the linguistic development environment NooJ, which will be used as input to higher levels of linguistic analysis such as syntactic parsing.

The remainder of this paper is divided into five main sections: the first presents a brief overview of the Moroccan Amazighe language particularities. The second section exposes and discusses some of Amazighe NLP challenges. The third presents our morphological analyzer system. The fourth section gives an overview of evaluation results, and the last section tries to draw some conclusions and suggests some future directions for our approach.

2. Moroccan Amazighe Language

2.1. Historical Background

The Amazighe language, also called Berber or Tamazight (ⵜⴰⵎⴰⵣⵉⵖⵜ [tamaziɣt]), belongs to the Hamito-Semitic languages [28][29]. It is currently spoken in a dozen countries ranging from Morocco, with 50% of the overall population [30], to Egypt, passing through Algeria with 25%, Tunisia, Mauritania, Libya, Niger and the Mali [31].

In Morocco, we distinguish between three major Amazighe dialects: Tarifit in the North of Morocco, Tamazight in the center and South-East, and Tashelhit in the South-West and the High Atlas.

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

2.2. Moroccan Standard Amazighe Language Particularities

In order to deal with the linguistic problem of Amazighe language due to the fact of using the different dialects (Tamazight, Tachelhit and Tarifit), IRCAM has engaged to achieve a standardization process for Amazighe language [32], which is not based on any of the dialects and aims to unify the three regional dialects into a standard national Amazighe language in order to introduce it in the school system, as well as in the media. This standardization process affects different levels:

- Phonology: adopt a standard script;
- Vocabulary: adopt a common lexicon;
- Morphology, syntax and spelling: apply the same spelling rules, the same morphological and syntactical instructions, and the same neologisms forms.

Thereafter, we give a brief overview of the standard Amazighe characteristics which include: the graphical writing system (see Appendix F), Tifinaghe encoding and morphology properties.

2.2.1. Writing System

Like any language that passes through oral to written mode, the Amazighe language has been in need of a graphic system.

In Morocco, the choice ultimately fell on Tifinaghe for technical, historical and symbolic reasons. Since the Royal declaration on February, 11th, 2003, Tifinaghe has become the official graphic system for writing Amazighe. Thus, IRCAM has developed an alphabet system called Tifinaghe-IRCAM. This alphabet is based on a graphic system towards phonological tendency. This system does not retain all the phonetic realizations produced, but only those that are functional [5]. It is written from left to right and contains 33 graphemes which correspond to:

- 27 consonants including: the labials (ⴼ, ⴱ, ⵎ), dentals (ⵜ, ⴷ, ⵟ, ⵏ, ⵍ, ⵔ, ⵚ, ⵥ), the alveolars (ⵙ, ⵣ, ⵚ, ⵣ), the palatals (ⵛ, ⵊ), the velar (ⴽ, ⵅ), the labiovelars (ⴽⵯ, ⵅⵯ), 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.

2.2.2. Tifinaghe Encoding

Since the adoption of Tifinaghe as an official script in Morocco for the Amazighe language, the Tifinaghe encoding has become a necessity. To this end, considerable efforts have been invested by IRCAM.

Therefore, the Tifinaghe encoding is composed of four Tifinaghe character subsets: the basic set of IRCAM used to arrange the orthography of different Moroccan Amazighe dialects, the extended IRCAM set used for historical and scientific use, The Neo-Tifinaghe letters, and modern Touareg letters. The first two subsets constitute the sets of characters chosen by IRCAM [1].

2.2.3. Amazighe Morphology in Brief

Amazighe is a morphologically rich language. It is highly inflected and also shows derivation to a high degree. The main morpho-syntactic categories in Amazighe language are: nouns, verbs pronouns and function words which include adverbs, prepositions, etc. [4][5]. The following section aims to provide a brief systematic description of Amazighe morphology.

1) Noun Characteristics

In Amazighe language, noun is a lexical unit, composed of a root and a pattern (see Appendix H) [4]. It can appear in several forms: simple form (ⴰⵔⴳⴰⵣ [argaz] "the man"), compound form (ⵜⵓⵜⵎⵣⵔⴰⵢⵜ [bu tmzrayt] "chronicler"), or derived one (ⴰⵎⵙⴰⵡⴰⴹ [amsawaḍ] "the communication").

1a) Simple Noun

The Amazighe simple noun consists of a single word occurring between two blank spaces. There are two major subclasses of Amazighe simple nouns: proper nouns and common ones. This latter constitute the most frequent type.

- Proper noun: the name of a person ⵀⵏⵓ [hnnu], place ⴰⵣⵔⵓ [azru] or date. Such nouns are not inflected.
- Common noun: can be either abstract or concrete. The latter can be either animate (ⵜⴰⵎⵜⵜⵓⴷⵜ [tamṭṭuḍṭ] "woman") or inanimate (ⵜⵉⴳⵎⵎⵉ [tigmmi] "house").
- Adjective: the adjective is a syntactic underclass of the noun; it shares with the nominal all these combinational and functional characteristics: the gender, the number and the state markers, and the complete sentence predicate. As a specific function, they determine the nominal with which they agree in gender and number: ⴰⵚⴱⵃⴰⵏ [aṣbḥan] "beautiful".

1b) Derived Noun

In Amazighe language, we can distinguish between two types of derivation processes: (1) nominal derivation based on noun and (2) nominal derivation based on verb.

- Nominal derivation based on noun: in this case, the derived nouns are obtained by the prefixation of the morphemes ⴰⵎⵙ [ams], ⴰⵏⵙ [ans], or ⴰⵎ [am] to the nominal basis: ⴰⵍⴰⵡ [awal] "word" ⟶ ⴰⵎⴰⵍⴰⵡ [amawal] "lexicon".
- Nominal derivation based on verb: from a verbal root, deverbal nouns are built by the prefixation or suffixation of a derivation morpheme associated with some changes. Thus, four types of nouns are made: the action noun, agent noun, instrument noun and adjective.

Action Noun. The action noun is formed with prefixation associated with some changes. The main processes of this derivation are: (1) for simple verbs we have: (i) prefixation of the vowel ⴰ [a], ⵓ [u] or ⵉ [i] (ⵃⵔⵔⵣ [ḥrrz] "keep" ⟶ ⴰⵃⵔⵔⵣ [aḥrrz] "the fact to supervise his wife (husband) or supervising his wife by someone") or (ii) prefixation and affixation of one of the feminine morpheme: ⵜ…(ⵜ) [t---(t)], ⵜⴰ…(ⵜ) [ta---(t)], ⵜⵉ…(ⵜ) [ti---(t)] or ⵜⵓ…(ⵜ) [tu---(t)] associated with an initial or final vowel variation and a consonant gemination/degimination for some nouns (ⴳⵏⵓ [gnu] "sew" ⟶ ⵜⵉⴳⵏⵉ [tigni] "couture") and (2) for borrowed ones, we have: (i) prefixation of ⵍ [l]: ⵃⵎⵓ [ḥmu] "be hot"⟶ ⵍⵃⵎⵓ [lḥmu] "the heat" or (ii) affixation of vowel ⴰ [a] sometimes associated with some changes (ⴷⵀⵏ [dhn] "paint" ⟶ ⴰⴷⵀⴰⵏ [adhan]).

Agent Noun. This kind of noun is derived from an action verb by the prefixation of one of the following morphemes for simple verbs: ⴰ [a], ⴰⵎ [am]/ⴰⵏ [an], ⵉⵎ [im] or ⵉ [i] associated with an infixation of the ⴰ [a] (ⴰⵣⵍ [azl] "send" ⟶ ⴰⵎⴰⵣⴰⵍ [amazal] "messenger"). For borrowed ones, the derivation process consists of an affixation of

the ⵄ [a] sometimes associated with a gemination of radical consonant and a vowel infixation.

Instrument Noun. This type of noun is extremely scarce, since it is formed from a simple verb by the prefixation of the morphemes ⵄ [a]/ⵄⵙ [as] associated with vowel or consonant changes: ⵔⵖⵍ [rgl] "close"⟶ ⵄⵙⵔⵖⵍ [asrgl] "lid".

Adjective. The adjective is generally derived from quality verbs/stative verbs. The derivation processes of this type of nouns are: (1) prefixation of ⵄ [a] associated with a vowel alternation, (2) prefixation of ⵄⵎ [am]/ⵄⵏ [an] followed sometimes by intra or post-radical changes, (3) prefixation of ⵉ [i] with intraradical change or (4) prefixation of ⵓ [u] sometimes associated with infixation of the vowel ⵉ [i]: ⵥⵎⵔ [qmr] "be narrow" ⟶ ⵓⵥⵎⵉⵔ [uqmir] "close".

1c) Compound Noun

Although it is less productive than derivation, the field of composition is not absent in Amazighe language. A compound noun is a noun that is made of two or more words. The most frequent composition models in Amazighe are presented in the following table (cf. Table 1) [33].

Whether simple, compound or derived, the noun varies in gender (masculine or feminine), number (singular or plural), and state (free or constructed) [4].

Gender. The Amazighe noun is characterized by one grammatical gender: masculine or feminine.

- The masculine noun: begins with one of the initial vowels: ⵄ [a], ⵉ [i] or ⵓ [u]. However, there are some exceptions as: ⵉⵎⵎⵄ [imma] "(my) mother".

- The feminine noun: is marked with the circumfix ⵜ…ⵜ [t....t]. However, there are some exceptions such as nouns which have only the initial or the final ⵜ [t] of feminine morpheme: ⵜⵄⵍⵣⵄ [tadla] "the sheaf", ⵟⵟⵎⵓⵕⵜ [ṛṛmuyt] "the tiredness".

Number. The noun, masculine or feminine, has singular and plural. The latter has four forms.

- The external plural: it is formed by an alternation of the first ⵄ/ⵉ [a/i] associated with a suffixation of ⵏ [n] or one of its variants (ⵉⵏ [in], ⵄⵏ [an], ⵄⵕⵏ [ayn], ⵡⵏ [wn], ⵄⵡⵏ [awn], ⵡⵄⵏ [wan], ⵡⵉⵏ [win], ⵜⵏ [tn], ⵕⵉⵏ [yin]): ⵄⵅⵅⵄⵎ [axxam] "house" ⟶ ⵉⵅⵅⵄⵎⵏ [ixxamn] "houses".

- The broken plural: involves a change of the internal vowels: ⵄⵀⵄⵢⵓⵙ [abaɣus] "monkey" ⟶ ⵉⵀⵓⵢⵄⵙ [ibuɣas] "monkeys".

- The mixed plural: is formed by vowels' change associated sometimes with the suffixation of ⵏ [n]: ⵉⵣⵉⵔⵙ [izikr] "the rope" ⟶ ⵉⵣⵄⴽⵄⵙⵏ [izakarn] "the ropes".

- The plural in ⵉⵍ [id]: this kind of plural is obtained by the noun prefixation with ⵉⵍ [id]. It is applied to a set of nouns includ-

Table 1. Description of most frequent composition models in Amazighe language.

Composition Models	Examples
Noun + Noun (juxtaposition)	ⵀⵄⵀⵄ ⵔⵀⵀⵉ [baba rbbi] "god"
Noun + ⵏ [n] + Noun (genitive)	ⵄⵅⵔⵓ ⵏ ⵡⵀⵓⵔ [agru n lbur] "toad"
Noun + Adjective (juxtaposition)	ⵄⵎⵄⵏ ⵄⵔⵏⵉⵏ [aman ḍrnin] "dew"
Verb + Noun (juxtaposition)	ⵙⵡⵎ ⵄⵅⵅ"ⵔⵏ [slm aggʷrn] "butterfly"
Verb + Verb (juxtaposition)	ⵀⵀⵉ ⵣⵍⵉⵍ [bbi zdi] "bagatelle"
Verb + Adverb (juxtaposition)	ⵊⵓⵍ ⵎⵡⵉⵍ [jud mliḥ] "sycophancy"
Preposition + Noun (juxtaposition)	ⵍⵍⵓ ⵍⵄⴽⵄⵡ [ddu wakal]"the grave"
Moneme (ⵄⵕⵜ [ayt], ⵀⵓ [bu], ⵀⵏ [bn], ⵓ [u], ⵅⵄⵔ [gar], ⵍⵄⵔ [war], ⵀⵡⵄ [bla], ⵜⵄⵔ [tar], ⵎ [m], ⵓⵡⵜ [ult], ⵀⵏⵜ [bnt], ⵉⵔⵜ [ist]) + Noun (affixal)	ⵀⵓ ⵜⵏⵄⵏⵓⵜ [bu tḥanut] "shopkeeper"

ing: nouns with an initial consonant, proper nouns, kinship nouns, compound nouns, numerals, as well as borrowed ones: ⵜⵍⵣⵔⴰⵢⵜ [bu tmzrayt] "chronicler" ⟶ ⵉⵏ ⵓⵙ ⵜⵍⵣⵔⴰⵢⵜ [id bu tmzrayt] "chroniclers".

State. We distinguish between two states: the free state and the constructed one.

- The free state is unmarked. The noun is in free state if it is a single word isolated from any syntactic context, a direct object, or a complement of the predictive particle ⵏ [d].
- The constructed state is marked in the following contexts: (1) when the noun follows the verb in a sentence and it is the subject of the verb, (2) when the noun follows a preposition, and (3) after some numerals. It involves a variation of the initial vowel. In case of masculine nouns, it takes one of the following forms: initial vowel alternation ⴰ [a]/ⵓ [u]; adding of the consonant ⵡ [w] or ⵢ [y]. For the feminine nouns, it consists of dropping or maintaining the initial vowel.

2) Verb

The verb occurs in two forms: simple and derived one. The simple verb is formed through the amalgamation of a root and a pattern, called interdigitation (ⵙⴳⵍ [sgl] "backfill", resulting from the intersection of the form – ⴳⵍ [gl] "collapse" – and the pattern ⵙC [sC]); while the derived one is obtained by the combination of simple verb with one of the following derivational morphemes: ⵙ/ⵙⵙ [s/ss] indicating the factitive form, ⵜⵜ [tt] marking the passive form, and ⵎ/ⵎⵎ [m/mm] designating the reciprocal one. The prefixation of these morphemes can lead to the deletion of the initial vowel/consonant accompanied with some change of the simple form of the verb.

The verb, whether simple or derived, inflects in three moods (indicative, imperative and participial), where in each mood the same personal markers are used (cf. Table 2). It has four aspects (aorist, perfective, negative perfective and imperfective) that are marked with vocalic alternations, prefixation or consonant gemination/degimination. The indicative and the participial moods occur on the four aspects, while the imperative mood has two forms, simple and intensive, that are based on the aorist and the imperfective aspects [4] respectively.

According to a study done by IRCAM [34], verbs are arranged into thirty one conjugation classes, according to the aorist/perfective and the aorist/imperfective conjugation oppositions.

The last class includes irregular verbs. It should be noted that, based on these classification criteria, the Amazighe verb and its derived forms do not necessarily belong to the same class, since they may not use the same morphotactic rules to be conjugated. Furthermore, to deal with regional varieties, a verb can belong to many classes.

3) Pronouns

A pronoun refers to any element that could replace a noun or nominal group. It may represent a nominal group already employed, or designate a person participating in the communication. The paradigm of pronouns includes: personal pronouns (ⵏⴽⴽ [nkk] "me"); the possessives (ⵡⵉⵏⵓ [winu] "mine"); the demonstratives (ⵡⴰⵏ [wad] "that"); the interrogatives (ⵎⴰ [ma] "which") and the indefinite ones (ⴽⵔⴰ [kra] "something").

Table 2. Personal markers for the indicative, imperative and participial moods.

	Indicative mood			**Imperative mood (simple and intensive)**			**Participial mood**	
		Masc.	**Fem.**		**Masc.**	**Fem.**		**Masc./ Fem.**
Singular	1st pers.	...ⵖ	...ⵖ					
	2nd pers.	ⵜ...ⵏ	ⵜ...ⵏ	2nd pers.	...Ø	...Ø	2nd pers.	ⵉ...ⵏ
	3rd pers.	ⵉ...	ⵜ...					
Plural	1st pers.	ⵏ...	ⵏ...					
	2nd pers.	ⵜ...ⵎ	ⵜ...ⵎⵜ	2nd pers.	...ⴰⵜ/ⵜ/ⵎ	...ⴰⵎⵜ/ⵎⵜ	2nd pers.	...ⵏⵉ
	3rd pers.	...ⵏ	...ⵏⵜ					

Generally, some of these pronouns are inflected, such as the possessive and demonstrative ones: ⵡⴰⴷ [wad] "this one" ⟶ ⵡⵉⴷ [wid] "these ones".

4) Function words

Function words are a set of Amazighe words that is assignable neither to a noun nor to a verb. It consists of several elements, namely:

4a) Prepositions

A preposition is a closed paradigm that combines simple and complex forms that express various semantic values. For the first form, it consists of several formats: ⵏ [n], ⵉ [i], ⵙ [s], ⵖ [g], ⴷⵉ [di], ⵣⵖ [zg], ⵅⵀ [xh], ⵖⵔ [gr], ⴰⵍ [al]/ⵄⵔ [ar], ⴱⵍⴰ [bla], ⵢⵔ [ɣr], ⴷⴰⵔ [dar], ⴰⵖⴷ [agd], ⴷ [d]. For the second one, it is composed of two or three prepositions which may be used adverbially: ⵉⵣⴷⴰⵔ ⵏ [izdar n] "below".

4b) Adverbs

Adverbs are the elements that change the meaning of a verb. Generally, they are classified according to their semantics. Thus, we distinguish adverbs of place (ⴷⴰ [da] "here"), time (ⴰⵣⴽⴽⴰ [azkka] "tomorrow"), quality (ⴷⵔⵓⵙ [drus] "little") and manner (ⵎⵍⵉⵃ [mliḥ] "well").

Besides these elements, there also exist: aspectual particles (ⵔⴰⴷ [rad] "that"), orientation particles (ⵏⵏ [nn] "there"), negative particles (ⵓⵔ [ur] "not") and conjunctions (ⵎⵉⵏⵣⵉ [minzi] "because").

3. Amazighe NLP Challenges

Despite its position as second official language in Morocco, Amazighe language still lacks in studies from the computational point of view. This situation is due to its complex and challenging pre-processing tasks. Thus, we can describe three main difficulties which need to be taken into account.

3.1. Amazighe Script

The Amazighe writing system poses three main difficulties:

- Dialectal variations: in linguistic terms, the Amazighe language is characterized by the proliferation of dialects (Tamazight, Tashelhit and Tarifit) due to historical, geographical and sociolinguistic factors. Each of these dialects has its own specificity with a set of sub-dialects. Dialects are still present in a set of resources despite of the standardization process.
- Scripts: Amazighe is among the languages having different scripts, namely: Latin, Arabic and Tifinaghe. Thus, it requires a transliterator to convert all the scripts into the standard "Tifinaghe-Unicode" form. However, this process is confronted with spelling variation related to regional varieties ⵜⴰⴼⵓⴽⵜ [tafuct] "Sun" – ⵜⴰⴼⵓⴽⵜ [tafukt] "Sun".
- Spelling: the Amazighe language has remained essentially an oral language for a long time. Therefore, the Amazighe text does not respect the standard writing convention.

3.2. Amazighe Corpora

Even with the attention paid to the Amazighe language in the recent years, there are still some gaps in the corpora which present a very valuable resource for NLP tasks. The Amazighe lacks such resources, and most of those existing are in paper format which needs to be digitized. To this end, a set of studies has been undertaken to build Amazighe corpora in a progressive way until reaching a large-scale corpus [16].

3.3. Amazighe Morphology

The third and the last reason, as cited in this paper, for Amazighe processing complexity is its rich and complex morphology. Bellow we describe three main difficulties in Amazighe morphology.

Part of speech ambiguity. One of the Amazighe NLP challenges is ambiguity; the same surface form might have different annotations depending on how it has been used in the sentence. To give some examples of different categories, we present the following examples:

- Some stop words such as ⴷ [d] might function as a preposition, a coordination conjunction, a predicate particle or an orientation particle. For instance, in the sentences

below, the word "d" might be: a coordination conjunction: ⵜⴰⵎⴰⵣⵉⵖⵜ ⴷ ⵜⵜⴽⵏⵓⵍⵓⵊⵉⵏ ⵜⵉⵎⴰⵢⵏⵓⵜⵉⵏ [tamaziɣt d tiknulujiyin timaynutin] "Amazighe and new technologies" → ⴷ= and; a preposition: ⵉⵎⴰⵍ ⴷ ⵓⴱⵔⵉⴷ [iman d ubrid] "he went with the road" → ⴷ= with; a predication particle: ⴷ ⴰⵔⴳⴰⵣ [d argaz] "he is a man" → ⴷ= he is; or an orientation particle: ⴰⵙⵉ ⴷ ⵜⵜⴽⵉⵜ ⵜⴰⵎⴰⵊⵊⵉⵜ [asi d tikint tamjahdit] "bring to here large bowl";

- ⵉⵍⵍⵉ [illi] may have many meanings; as a verb in negative perfective, it means "do not exist" when used after a negative particle, while as a noun, it refers to a kinship noun meaning "my daughter".

Inflectional and derivational prosess. As mentioned earlier in this paper (cf. 2.2.3), the Amazighe language is highly inflected and also shows derivation to a high degree. The first process presents more difficulties especially for noun inflection. As the example for the plural, there are four different forms based primarily on both prefix and suffix concatenations, but there are no specific rules which determine the use of one or the other form to inflect a noun. 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, the latter will be altered in the derivational form (ⵇⵇⵉⵎ [qqim] "make sit" → ⵙⵢⵉⵎ [sɣim] "sit").

Contractions. The third reason for processing complexity is contractions. For example, the kinship term ⴱⴰⴱⴰ [baba] "father" followed by a pronoun ⵏⵙ [nns] "his", becomes the single word ⴱⴰⴱⴰⵙ [babas] "his father" [4][3].

4. Amazighe Morphological Analyzer: Development and Evaluation

Morphological analysis is one of important tasks in NLP which presents a core part of several NLP applications, providing information about the possible part of speech and other morpho-syntactic features of words and tokens as they appear in the context. It is a basic step to various higher levels applications including text mining, information retrieval, machine translation, automatic summarization, and the Amazighe learning systems.

Over the last few years, AmNLP has gained increasing importance, and several state of the art systems have been developed for a wide range of applications. These applications had to deal with several complex problems pertinent to the nature and structure of the Amazighe language. The lack of available resources and their limitations have motivated many scholars to follow the rule based approach and rely on hand-constructed linguistic rules in developing their tools, systems, and resources. Furthermore, basic tools and applications such as morphological analyzer system also needs to be developed.

As Amazighe is a morphologically rich language due to its high inflectional and derivational nature, morphological processing plays a key role in developing Amazighe NLP systems and applications. The basic principle of morphological analysis is to breakdown an inflected form into a root and a set of features (lexical category and morpho-syntactic properties). Therefore, the effort spent on creating a reliable, efficient, and flexible Amazighe morphological analyzer is justified by its reuse in many of these applications.

In this paper, we will present our approach to build a highly flexible Amazighe morphological analyzer. Our morphological analyzer will identify the component morphemes of the input word (nouns, verbs, pronouns or function words) on the texts, and also label them with sufficient information to be useful for the other NLP tasks.

4.1. Computational Approaches to Morphology: Brief State-of-the-Art

During the last 25 years the Finite-State Technology (FST) has had a great impact on a variety of NLP applications, as well as in industrial and academic Language Engineering.

Finite State Morphology (FSM) aims at handling morphology within the computational power of finite state automata. This approach is especially attractive in dealing with human language morphologies; among these are the ability to handle concatenative and non-concatenative morphotactics, the high speed and efficiency in handling large automata of lexicons with their derivations, and inflections that can run into millions of paths, modularity of the design, due

to the closure properties of regular languages and relations; the compact representation that is achieved through minimization; and reversibility, resulting from the declarative nature of such devices. The adequacy of this technology for Semitic languages has frequently been challenged.

Although there is now a variety of Finite state tools, the development or adaptation of existing tools to facilitate the creation, annotation, and description of Amazighe corpora remains a major challenge. For morphological analysis, several tools and frameworks have been used. Well known tools include:

- Two-Level Morphology: it depends heavily on finite state methods, which are well known and are often described as elegant [35]. Two-level morphology was "the first general model in the history of computational linguistics for the analysis and generation of morphologically complex languages" [36]. Developed by Koskenniemi [37], this technology facilitates the specification of rules that relate pairs of surface strings through systematic rules. Such rules, however, do not specify how one string is to be derived from another; rather, they specify mutual constraints on those strings. Furthermore, rules do not apply sequentially. Instead, a set of rules, each of which constrains a particular string-pair correspondence, is applied in parallel, such that all the constraints must hold simultaneously. In practice, one of the strings in a pair would be a surface realization, while the other would be an underlying form. One of the greatest advantages of two-level morphology is that rules are entirely declarative: indeed, the original formulation of [37] allows for both analysis and generation within the same grammar. The formalism was later implemented as part of the Xerox tools; two-level rules are compiled to finite-state transducers, which indeed allow for both analysis and generation.
- XFST: stands for Xerox Finite-State Tool, one of the most sophisticated tools for constructing finite state language processing applications [38], developed at the XRCE by Kenneth R. Beesley and Lauri Karttunen [35]. It is "based on solid and innovative finite-state technology", designed for multi-purpose use with explicit support for automata theoretical research. The XFST toolkit provides powerful and elegant linguistic descriptions and treatment of irregularities via different operators, at a high level of abstraction, such as restriction, replacement, and left-to-right longest match replacement.
- SFST: Stuttgart Finite State Transducer is a non-commercial, open-source, freely-available set of finite state tools that support the analysis, generation, and processing of finite-state automata and transducers. SFST tools were primarily developed to provide finite state technology for building two-level morphological analyzers. The tools run under Linux are command-line oriented, and not so easy to use, especially since they could be better documented.

Although these finite state technologies present a number of advantages, they present also several disadvantages. The most common one being that they provide a single formalism (powerful grammar) supposed to be used to describe all linguistic phenomena. Unlike NooJ, which is a linguistic development environment, it provides linguists with one or more formal tools specifically designed to facilitate the description of each linguistic phenomenon, as well as parsing tools designed to be as computationally efficient as possible. Furthermore, it allows linguists to combine in one unified framework Finite-State descriptions such as in XFST (see [39] and [40] for more details). This fact makes NooJ an ideal tool to parse complex phenomena that involve those across all levels of linguistic phenomena, and allows NooJ's parsers to be extremely efficient compared with other NLP parsers.

4.2. NooJ: Linguistic Developmental Framework

NooJ, released in 2002 by Max Silberztein [41] [42][43], is a freeware language-engineering development environment, which runs on different operating systems such as Windows, Linux, Solaris and Mac OSX, and provides a set of tools and methodologies for formalizing and developing a set of NLP applications. It presents a package of finite state tools that integrates a broad spectrum of computational technology from finite state automata to augmented/recursive transition networks. This package allows building

and managing a large coverage of electronic dictionaries and formal grammars in order to formalize the different linguistic phenomena such as: spelling, morphology (inflectional and derivational), vocabulary (simple words, compound words and frozen expressions), syntax (local, structural and transformational), disambiguation, semantics and ontology, and which can be applied to treat texts and large corpora. For each of these levels, NooJ provides linguists with one or more formal frameworks specifically designed to facilitate the description of each phenomenon, as well as parsing, development and debugging tools designed to be as computationally efficient as possible, from Finite-State to Turing machines. This approach distinguishes NooJ from other computational linguistic frameworks providing a unique formalism that is supposed to cover all linguistic phenomena. As of today, NooJ can process 23 languages [44] as Croatian language [45], Armenien [46], including Arabic [47] as a Semitic language like Amazighe.

Given the situation of Amazighe as less resourced language, we recognized the necessity of developing an Amazighe component for NooJ platform, according to its robustness and simplicity, with the aim to enable its automatic processing and its integration in the field of Information and Communication Technology. One of the important and useful features of NooJ, regarding Amazighe as morphologically rich language, is its simple description of morphological phenomena and efficient morphological processing. This component would allow us to process and take advantage of this readily available data. The use of this technology was extremely attractive and allows generating and analyzing several thousands of words per second.

The NooJ lexical module that will be used throughout this paper relies on operators performing transformations inside strings, and morphological graphs describing grammatical rules for morphological analysis. Generally, transformations inside strings are based on the use of some generic predefined commands such as:

- ⟨LW⟩ : position the cursor (I) at the beginning of the form,
- ⟨RW⟩ : position the cursor (I) at the end of the form,
- ⟨R⟩ : keyboard Right arrow,
- ⟨L⟩ : keyboard Left arrow,
- ⟨B⟩ : delete the last character,
- ⟨S⟩ : delete the current character,
- Etc. (see Appendix B).

Our main agenda is to build the morphological analyzer as the combination of several finite state transducers using the NooJ framework. In order to do this, and given that the linguistic resources required by the morphological analyzer include a lexicon and inflection rules for all paradigms, we started by building a morphological Amazighe lexicon.

4.3. Building Amazighe Lexicons: NAmLex, VAmLex and PAmLex

The most basic and yet most needed step in morphological analysis of any language is the development of morphological lexicon. To this end we have created, in the first step and using the NooJ robust dictionary module, three Amazighe lexicons named:

- *NAmLex*, which stands for *Noun Amazighe Lexicon* and contains a list of simple, compound and derived nouns which are represented as a singular form,
- *VAmLex*, which stands for *Verb Amazighe Lexicon* and contains a list of simple, and derived verbs which are represented as a second person, singular, masculine, imperative mode, and
- *PAmLex*, which stands for *Particles Amazighe Lexicon* and contains a list of pronouns and function words.

In order to do this, manual collection of the terms is carried out following the three steps described below:

1) We developed a list of nouns (simple, derived and compound ones) from a set of resources such as Taifi dictionary [48], Amazighe vocabulary [6], Amazighe media vocabulary [7], Amazighe grammatical vocabulary [8], new grammar of Amazighe [4] and school lexicon [49].
2) We collected a list of verbs from Amazighe Manual Conjugation [34].

3) We built a lexicon of pronouns and function words from Amazighe grammar [4].

Each lexical entry presents the following details: the lemmas, lexical category, semantic feature and its translation in French and Arabic languages.

Eventually, to complete the morphological concept of our lexicon, in the second step, we formalized the inflectional and derivational morphology in order to generate from each entry its inflectional and derivational information.

4.4. Amazighe Lexicon Formalization: Inflectional and Derivational Morphology

Both inflection and derivation are standard processes for the Amazighe and they generate large amounts of word forms. Thus, the purpose of this section is the Amazighe categories (i.e. noun, verb, pronouns and function words) formalization. This study presents the implementation of inflectional and derivational rules that will allow the generation of inflected forms from each entry.

4.4.1. Inflectional Morphology

To formalize the inflectional rules, we have relied on the works of [4][50][34][33][51], as well as on heuristic study of our lexicons. Therefore, through graphs integrated in the linguistic development platform NooJ, we have created a set of hand-encoded graphs: 224 inflectional paradigms for nouns (descriptions of noun inflectional paradigms are explained in detail in [22]), 484 inflectional paradigms for verbs and 8 of them for pronouns [23].

By these inflectional descriptions we refer to the set of all possible transformations which allow us to obtain, from a lexical entry, all inflected forms. On average, there are 8 inflected forms per noun entry, 116 inflected forms per verb entry, and 2 to 8 ones per inflected pronouns.

These rules are invoked by the property "+ FLX =" that each inflective word has within its lexicon description.

4.4.2. Derivational Morphology

The morphological derivation of new words is considered to be a higher degree of Amazighe morphological process, in the level below the inflection. Therefore, we have implemented a hierarchical system of morphological paradigms as a tool able to capture different levels of the Amazighe derivational morphology. Our hierarchical system includes three derivation levels: (1) nominal derivation based on noun, (2) nominal derivation based on verb, and (3) verbal derivations.

1) Nominal derivation based on noun

According to the rules described before (cf. 2.2.3.1.b) and on heuristic study of our noun lexicon, we have formalized noun derivation with about 21 rules, generally based on prefixation and suffixation of the nouns.

2) Nominal derivation based on verb

Deverbal nouns are formed by prefixation or suffixation of a derivation morpheme accompanied by intraradical changes. Thus, four types of nouns are made: action noun, agent noun, instrument noun and noun of quality. Each of these branches (nominal pattern) is formed based on verbal patterns [24].

2a) Action nouns

Based on the work of [52] we have formalized the derivation of action nouns based on (1) the pattern and on (2) the verbal root(monoliteral (see Appendix A), biliteral, etc. with or without initial tension). We have raised, in total, 66 rules of which 5 rules are for the monoliteral, 22 for biliteral, 23 for triliteral, and 9 for quadriliteral roots. In the following (cf. Table 3) we show an example of derivation from monoliteral root (see Appendix C).

Table 3. Example of the action nouns derivation.

Verbal class	**Patterns**		**Example**
	Verbal pattern	**Nominal pattern**	
Monoliteral	C̄u	i C̄u	ⵜⵜⵓ [ttu] "forget" → ⵉⵜⵜⵓ [ittu] "oversight".

Amazighe, as well as other languages, has borrowed words. These words are strongly borrowed from Arabic dialect (Moroccan Arabic: Darija). To this end, we have formalized the derivation of action nouns based on this kind of verbs. Given that the triliteral roots are the most productive for borrowed verbs, we have formalized, relied on the works of [52], 7 rules for this morphological type. We show in the following (cf. Table 4) an example of derivation from triliteral root.

Table 4. Example of the action nouns derivation based on borrowed verbs.

Verbal class	Patterns		Example
	Verbal pattern	Nominal pattern	
Triliteral	CaCC	aCaCC	ⵃⴰⵙⴱ [ḥasb] "count" → ⴰⵃⴰⵙⴱ [aḥasb] "counting".

2b) Agent nouns

Similarly to action nouns and based on the work of [53], we formalized the derivation of agent nouns based on (1) the pattern and on (2) the verbal roots: biliteral and triliteral. We have raised, in total, 9 rules of which 4 rules are for the biliteral and 5 for triliteral roots. In the following (cf. Table 5) we show an example of derivation from biliteral root.

Table 5. Example of the agent nouns derivation.

Verbal class	Patterns		Example
	Verbal pattern	Nominal pattern	
Biliteral	CC	amCaC	ⵏⵢ [ny] "mount" → ⴰⵎⵏⴰⵢ [amnay] "rider".

2c) Instrument nouns

For this nominal type, there are no regular patterns. Based on a list of verbs, we have formalized a set of cases for biliteral and triliteral roots. Thus, we have raised, in total, 8 rules of which 3 rules are for the biliteral and 5 for triliteral roots. In the following (cf. Table 6) we show an example of derivation from triliteral root.

2d) Nouns of quality

Table 6. Example of the instrument nouns derivation.

Verbal class	Patterns		Example
	Verbal pattern	Nominal pattern	
Triliteral	CCC	asCCC	ⵔⴳⵍ [rgl] "close" → ⴰⵙⵔⴳⵍ [asrgl] "lid".

To formalize the nouns of quality, and based on the work of [53], we have raised 14 rules of which 4 rules are for the biliteral, 9 for triliteral and 1 rule for quadriliteral roots. In the following (cf. Table 7) we show an example of derivation from biliteral root.

Table 7. Example of the nouns of quality derivation.

Verbal class	Patterns		Example
	Verbal pattern	Nominal pattern	
Biliteral	CC	amCCu	ⵔⵣ [rz] "break" → ⴰⵎⵔⵣⵓ [amrzu] "breaker".

Then, like the simple form, each derived noun is associated with the corresponding inflectional paradigm (cf. Figure 1) in order to recognize all the corresponding inflected terms. The latter may be the same as the basic noun and may also be different.

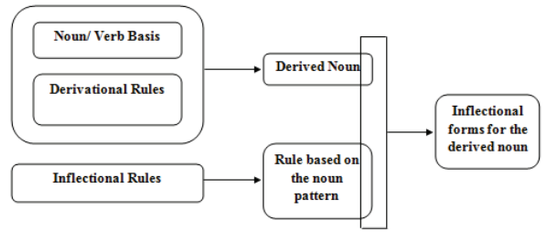

Figure 1. Processing of derived noun.

3) Verbal derivations

According to the rules described in (cf. 2.2.3.2), we have formalized verb derivation within 50 derivational paradigms for the totality of the verbs (ⴽⴽⵙ [kks] "remove" ⟶ ⵎⵢⵓⴽⴽⴰⵙ [myukkas] "remove mutually").

All the descriptions inside the derivational grammars are given in the form of graphs. These rules are invoked by the property "+ DRV =" that each derivative word has within its lexicon description.

To give an overview of all these morphological transformations (inflectional and derivational ones), we take as an example the verb ⊏ⵎ [mun] "to accompany".

⊏ⵎ,V+Simple+C2+FLX=T2_16+DRV=Forme1_s:
T8_10+DRV=NAC_Pf1:PES:Ta_cT1
+FR=accompagner+AR=استشار+شاور+صاحب

This example illustrates two processes:

The inflectional process. Inflectional grammar is looking for the paradigm named T2_16 (+ FLX = T2_16 (see Appendix G) in order to generate all forms. Among the 116 inflectional transformations which are described in the flexional paradigm "T2_16", here is one:

ⵖ<LW>ⵜⵜ/ Inacc+1+m+s

This NooJ paradigm, written in NooJ graphic editors, illustrates the imperfective form. The first part of this paradigm describes a change on the word (e.g. ⟨ LW⟩ / – position the cursor (I) at the beginning of the form), while the second part describes features that the newly made word is given (e.g. /Inacc + 1 + m + s – word is added description that it is in imperfective form (Inacc), first person (1), masculine (m) and singular (s)). The meaning of the transformation is to insert the consonant ⵖ [ɣ] at the end of the form and ⵜⵜ [tt] into the head of it. These operations, applied in succession, generate the form: ⵜⵜ⊏ⵎⵖ [tmunɣ] "I accompanied".

The derivation process. Derivational grammar is looking for the paradigm Forme1_s (+ DRV = Forme1_s). In this paradigm we have the following derivational transformations:

Forme1_s = <LW>⊙
/V+Dérivé+Factitive+Préf_⊙

Like the inflectional transformation, the derivational one is written in NooJ graphic editors and illustrates the factitive derivation. The first part of this transformation describes a change on the word, while the second part describes features that the newly made word is given (e.g. /V + Dérivé + Factitive + Préf_⊙ – word is added description that it is in derived form (+ Dérivé), in factitive type (+ Factitive) formed by the prefixation of the morpheme ⊙[s] (+ Préf_⊙)). The meaning of this transformation is to insert the consonant ⊙ [s] at the beginning of the form. This operation generates the form: ⊙⊏ⵎ [smun] "gather". Given that this derived form has a different inflection in the imperfective mood (imperfective = aorist), we have associated the correspondent paradigm in (:T8_10).

The deverbal noun derivation and its flexion. Moreover, this verb presents also nominal derivation. The derivational grammar is looking for the paradigm NAC_Pf1 (+ DRV = NAC_Pf1: PES:Ta_cT1). The + DRV feature first invokes the derivation pattern named NAC_Pf1 after which the newly derived word is inflected as the paradigm named PES:Ta_cT1. For the first paradigm "NAC_Pf1" we have the following derivational transformations:

NAC_Pf1= ⵜ<LW>ⵜₒ/N+Dérivé+Nom_Action

This NooJ transformation consists in the insertion of the consonant ⵜ [t] at the end of the form, and ⵜₒ [ta] at the beginning. After the derivation, the newly derived noun (ⵜₒ⊏ⵎⵜ [tamunt] "union") is marked as derived (+ Dérivé) noun (N) with the type: action noun (+ Nom_Action). The new word (ⵜₒ⊏ⵎⵜ) is then inflected using the inflectional pattern name that follows the name of the derivational one (:PES:Ta_cT1). Among the 8 inflectional transformations described in this inflectional paradigm, we can cite the following model which allows to obtain the plural form in constructed state:

<B>ⵉ<LW><R2><B> /EA+f+p

The meaning of this transformation is to delete: (1) the last ⵜ [t] and (2) the vowel after the first one, and then insert the variant ⵉ [in] at the end of the form. These operations, applied in succession, generate the form: ⵜ⊏ⵎⵉ [tmunin].

Since Amazighe language is highly inflectional, the number of word forms in the compiled lexicons is much larger than the number of lemmas that are in the main lexicons. We present below an overview of the final lexicons:

Table 8. Lexical entries.

Grammatical categories	Forms	
	Simple forms	Inflectional and derivational forms
Nouns	15 325	105 258
Verbs	4109	250 214
Pronouns + Function words	1642	1878
Total	21 076	357 350

4.5. Evaluation of Amazighe Lexicons and Lexical Rules

The test of the lexical coverage of our Amazighe lexicons is evaluated on lexical analysis of our manually-constructed corpus. This corpus is a collection of school texts and contains 96 031 tokens. Automatic evaluation of our Amazighe transducer is a challenge due to the lack of standard annotated corpora.

The main objective from this analysis is to evaluate how many words in our texts can be recognized and annotated after applying our compiled lexicons. It consists of testing membership of each word of the text to our lexicons and then labeling them with the corresponding information. To better illustrate our analysis, we show in the following figure (cf. Figure 2) an example analysis of the noun +ⵜⵎⵓⵏⵉⵏ [timunin] “unions, associations”:

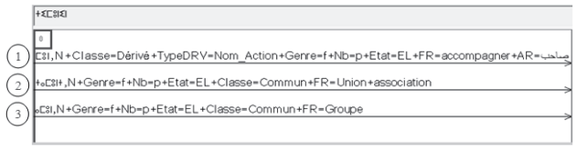

Figure 2. Schematic illustration of derived noun analysis (see Appendix D and I).

- +ⵜⵎⵓⵏⵉⵏ is the plural form of the action noun +ⵜⵎⵓⵏⵜ [tamunt] derived from the verb ⵎⵓⵏ [mun] “accompany” (cf. Figure 3) (For further guidance see the example cited in the previous section of the verb ⵎⵓⵏ [mun]),
- +ⵜⵎⵓⵏⵉⵏ is the plural form of the feminine noun +ⵜⵎⵓⵏⵜ[tamunt] “union, association”,
- +ⵜⵎⵓⵏⵉⵏ is the plural form of the feminine noun +ⵜⵎⵓⵏⵜ [tamunt] derived from the masculine one ⴰⵎⵓⵏ [amun] “community” (cf. Figure 3).

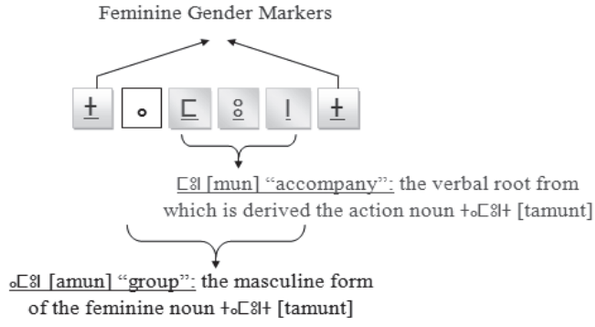

Figure 3. A template describing a feminine noun derived from a masculine one and from a verb.

Lexical analysis of our corpora shows the results presented below:

Table 9. Result of our lexical analysis.

Results	Number of recognized forms		Number of unrecognized forms	
Lexicons	Number	%	Number	%
	88 897	92.57%	7134	7.42%

Lexical analysis of these corpora shows coverage of about 93% by our lexical and morphological resources.

However, despite the efforts on improving a complete lexicon, unknown words typically represent 7.42% of the words in our corpus that are not contained in our lexicons. Therefore, we propose a new perspective to handle the out-of-vocabulary words that are considered important, using rules and patterns. This approach presents an alternative to investing manual effort in improving the lexicon.

4.6. The Unknown Words Processing: Recognition Systems (RS)

Unknown words are one of the key factors which drastically impact the analysis quality. In order to increase the robustness of our analyzer, we have undertaken to propose a new method for handling such unavoidable lack of information. To this end, there have been two possibilities: the first way is to attempt to construct a complete lexicon, and the second way is to attempt to analyze the word by discovering its part of speech and feature information, and storing that information in the lexicon.

Given that the second approach is less expensive, in this paper we describe a morphological analysis method based on morphological and syntactic grammars. This method uses a model that cannot only consult the lexicons, but also estimates how likely it is that a word can be a noun or verb according to the information at hand.

The following sections will cover the various concepts used by our recognition systems (RS) to help in the processing of unknown words. First, the architecture system will be discussed. Second, the morphological RS are detailed. Finally, an evaluation of our approach is described.

4.6.1. System Architecture

Our morphological analysis involves 4 steps (see Figure 4).

For convenience, we split the field of our morphology into three different areas: morphological generation, morphological reconstruction and morphological recognition. Morphological generation involves creating lexical entries based on the inflectional rules (cf. 4.5.1). Morphological reconstruction produces the new word through derivational rules (cf. 4.5.2). These two morphological areas are applied directly, after the application of the NooJ segmentation system to the corpus (step 1), then we identify

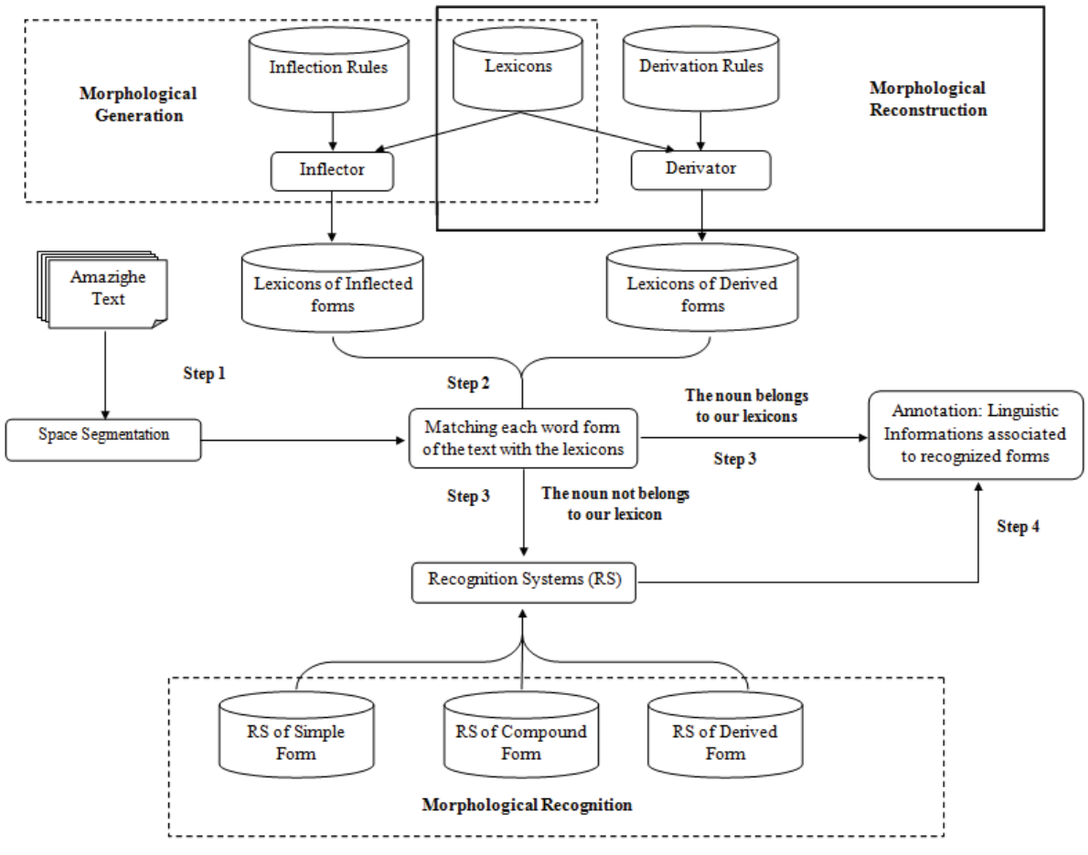

Figure 4. Amazighe Morphological Analyzer Architecture.

the component morphemes of the input words, matching them with the lexicons (step 2) and labeling (step 3) the known words with sufficient information to be useful for the tasks at hand. And finally, if the input word does not belong to our lexicons (step 4), we apply the morphological recognition systems which use knowledge about affixes and other features of the word in order to provide the possible annotation, without using any direct information about the words stem. The following section describes the methods employed by our recognition systems.

4.6.2. Morphological Recognition: Recognition Systems (RS)

No matter how big the applied lexicon is, there will always remain unknown words. Moreover, natural language is dynamic and it is impossible to compile huge dictionaries which will contain all the words that could appear in real-life texts: new words are constantly added to a language; other words get less frequent or are dropped out, while some of the existing ones get lost or change, etc. To remedy to this situation, we introduce a system analysis without lexicon in order to increase the robustness of our analyzer. This system will empirically investigate how well syntactic and morphological rules can analyze sentences containing unknown words.

Given that the unknown word can be a noun or a verb, we proposed a two-level mechanism. The first level is developed for the nouns (in simple, derived and compound forms) and the second one is developed for the verbs (in simple and derived forms). For these two levels, we are based on syntactic knowledge.

1) Noun Recognition System

Our goal was the design and implementation of a system for identification and distribution of possible tags that can serve as the analysis of the unknown Amazighe words from texts. To this end, we have developed a set of five morphological and syntactic grammars presented below:

- Syntactic grammar for simple nouns recognition: according to a set of rules that we have developed, this grammar is used to extract the morphological information for the unrecognized nouns. These rules are based, in the first step, on the morphemes which present the gender, number and state marks, and, in the second step, it is based on the previous word (e.g. a word directly following the predicative particle ⵏ [d] is typically a noun in free state form).
- Syntactic grammar for derived nouns recognition: following the same method as simple nouns, we have developed a set of rules based on the previous word and on the derivation patterns (suffixes, prefixes and infixes).
- Syntactic grammar for compound nouns recognition: we have developed, in the first step, two syntactic grammars allowing recognizing a compound noun in the most frequent composition models (cf. 2.2.3.1.c). Many compound nouns, which are formed on the moneme (ⴰⵙ, ⵜⴰ, etc.) + Noun (cf. 2.2.3.1.c), present a challenge when they come attached (e.g. ⵜⴰⵒⵓⵙⵙⴰ [buhyyuf] “the famine”). Thus, as a second step, we have developed a morphological grammar to handle these words. To illustrate our proposition we show an example in Figure 5 (cf. Figure 5). In our graphical presentations, brackets refer to variables, dollar sign ($) is used for

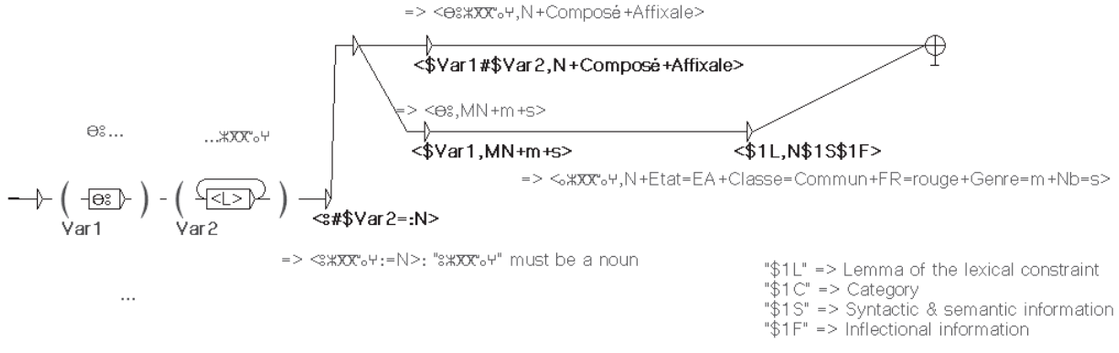

Figure 5. Example of compound noun recognition.

the variable names ($Var1), etc. These and the remaining symbols are explained in more detail in [42].

This grammar decomposes the noun in morpheme ⵙⵓ [bu] + word (⟨ L ⟩). Given that any word following these morphemes is in constructed state, we combine this latter (⟨ L ⟩) with constructed state prefix and then check if the word belongs to the lexicons (required by the lexical constraint ⟨ ⵙ#$Var2=:N ⟩ , the word must be a noun). If such a word exists, it imports all relevant morphological information from the lexicon (required by the lexical constraint ⟨ $1L,N$1S$1F ⟩ , the N correspond to Noun). Thus, if the lexicon contains the word ⵙⵣⵅⵅᵘⴰⵢ [uzggʷaɣ] "red", the grammar will annotate the word ⵙⵓⵣⵅⵅᵘⴰⵢ [buzggʷaɣ] "measles" as follows (cf. Figure 6).

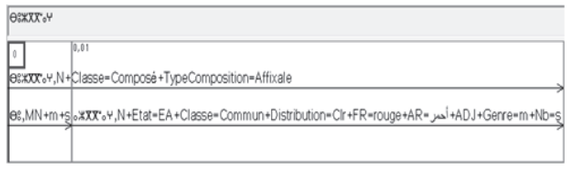

Figure 6. Example of compound noun recognition.

The noun will be annotated as a noun (N), in compound form (Classe=Composé) in the Affixal model (TypeComposition = Affixale).

Then, an accurate annotation will be extracted: ⵙⵓ [bu] will be annotated as morpheme in singular masculine form (MN + m + s), and the information of the noun corresponding to the form ⵙⵣⵅⵅᵘⴰⵢ [uzggʷaɣ] will be extracted from the lexicon.

- Morphological grammar for kinship nouns (see Appendix E): this grammar recognizes kinship nouns which form one entity with affixal pronouns (e.g. ⵜⴰⵜⴰⵙ [babas] "his father"). To fully illustrate our proposition, we present below (cf. Figure 7) an example of kinship nouns recognition system.

  This morphological grammar recognizes the word forms like ⵜⴰⵜⴰⴽ [babak] "your father", ⵅⵎⵎⴰⵜⵏⵖ [immatnɣ] "our mother", etc. Each of these forms will be computed from the root formed by letter succession (⟨ L ⟩) stored in the variable Pref ($Pref) which needs to be listed as a kinship noun in our lexicon (required by the lexical constraint ⟨ $Pref=:N + Parenté ⟩). If such a word exists, we identify the second variable which correspond to affixal pronouns (ⴽ [k], ⵎ [m], ⵙ [s] etc.) and then produce the annotations. We use the special tag ⟨ INFO ⟩ , associated to the affixal pronouns to be concatenated at the end of the annotation (e.g. ⵜⴰⵜⴰⴽ, N + Parenté + ⴽ).

2) Verb Recognition System

Following the same method of the recognition systems for nouns, we have developed two syntactic grammars presented below:

- Syntactic grammar for simple verbs recognition: we proposed a set of rules used to extract the morphological information associated to the unrecognized verb (the conjugation aspects, the gender (masculine or feminine), the number (singular, plural), and the person (first, second or third)).

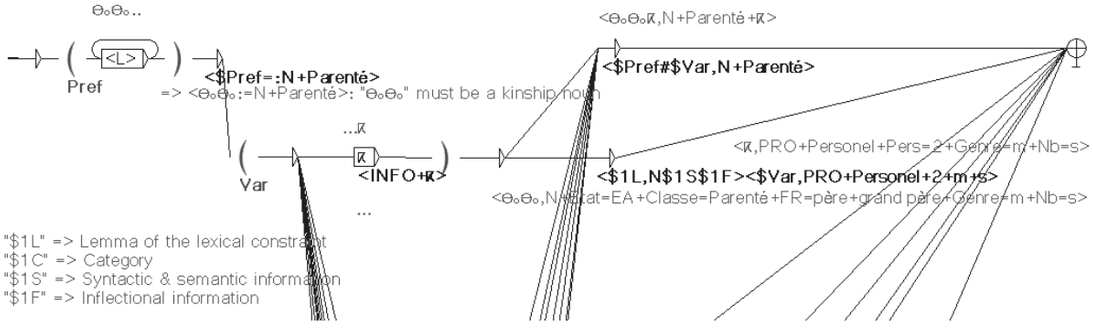

Figure 7. Example of recognition system for kinship nouns.

- Syntactic grammar for derived verbs recognition: we have developed a set of rules based on the derivation patterns (cf. 4.5.2) allowing testing on an input word whether it is a derived verb or not.

4.7. Experiments

The main goal of this evaluation is to prove the flexibility of our approach and to prove that it can satisfy the morphological analysis of the most unknown words remaining after the lexical analysis. The experiment reported here is very simple: it consists of parsing the technical corpus before and after integration of the recognition systems in the preprocessing components, and comparing the results obtained. To do this, we use the corpus of school texts cited at the lexicons evaluation (cf. 4.6). The results of full analysis can be seen in the following table (cf. Table 10).

Table 10. Amazighe morphological analyzer results.

Results	**Number of recognized forms**		**Number of unrecognized forms**	
	Number	**%**	**Number**	**%**
Lexicons	88 897	92.57%	7134	7.42%
Lexicons + RS	95 753	99.71%	278	0.29%

After the application of our resources, we have undertaken a manual analysis of the outputs to evaluate the performance of our rules. The evaluation results shown in Table (10) indicate the strong impact of using RS component in word analysis. The unrecognized forms present a ratio of 0.29%.

The unanalyzed words are mostly classified in three sets:

- Named entities: foreign proper nouns of person such as ⵎⵓⵃⴰⵏⴷ [muḥand] "Muhand", cities such as ⵎⵉⵚⵔ [misr] "Egypt" and numbers such as ⴽⵔⴰⴹ [kṛaḍ] "four".
- Spelling mistakes: The most common mistake is replacing the prefix of the constructed state such as the form ⵡⴰⴹⵉⵍ [waḍil] "grapes" where the noun is written ⵡⵡⴰⴹⵉⵍ [wwaḍil] according to its pronunciation.
- Foreign words (nouns, verbs, pronouns and function words): Amazighe has imported foreign words from Arabic (ⴱⵣⵣⴰⴼ [bzzaf] "a lot").

5. Conclusion

Morphological analysis caters to the needs of variety of applications like machine translation, information retrieval and spell-checking. Therefore, in this paper we have proposed a module of Amazighe morphological analyzer which can recognize lexical units from texts and also label them with sufficient information to be useful for the other NLP tasks. The finite-state approach has proven to be very successful in the consistent description of the Amazighe morphology, consisting of a network of lexicons and RS.

In the near future we plan to pursue a set of paths, such as: (1) implementing a module of named entity recognition using the NooJ syntactic module, (2) developing morphological grammars to correct the spelling mistakes with disambiguation problems, and (3) enlarge our corpus with the number of inputs of our lexicons.

References

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

[2] L. Zenkouar, "Normes des technologies de l'information pour l'ancrage de l'écriture Amazighe", *Etudes et documents berbères*, vol. 27, pp. 159–172, 2008.

[3] M. Ameur *et al.*, *Graphie et orthographe de l'Amazighe*. Rabat, Maroco: IRCAM, 2006.

[4] F. Boukhris *et al.*, *La nouvelle grammaire de l'Amazighe*. Rabat, Maroc: IRCAM, 2008.

[5] M. Ameur *et al.*, *Initiation à la langue Amazighe*. Rabat, Maroco: IRCAM. 2004.

[6] M. Ameur *et al.*, "Vocabulaire de la langue Amazighe (Français-Amazighe)", *Lexiques*, no. 1, IRCAM, Rabat, Maroco, 2006.

[7] M. Ameur *et al.*, "Vocabulaire des médias (Français-Amazighe-Anglais-Arabe)", Lexiques no. 3, IRCAM, Rabat, Maroco, 2009.

[8] M. Ameur *et al.*, "Vocabulaire grammatical", *Lexiques*, no. 5, IRCAM, Rabat, Maroco, 2009.

[9] S. Kamel, *Lexique Amazighe de gèologie*. Rabat, Maroco: IRCAM, 2006.

[10] IRCAM. (2003). *Conception et mise au point des polices tifinaghe. Centre des Etudes Informatiques, Systèmes d'Information et Communication, plan d'action*. [Online]. Available: http://www.ircam.ma/fr/index.php?soc=telec&rd=1

[11] IRCAM. (2004). *Polices et Claviers UNICODE. Centre des Etudes Informatiques, Systèmes d'Information et Communication*. [Online]. Available: http://www.ircam.ma/fr/index.php?soc=telec &rd=3

[12] M. Amrouch *et al.*, "Handwritten Amazighe Character Recognition Based On Hidden Markov Models", *International Journal on Graphics, Vision and Image Processing*, vol. 10, no. 5, pp. 11–18, 2010.

[13] M. Fakir *et al.*, "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, 2009, pp. 33–47.

[14] Y. Es Saady *et al.*, "Printed Amazighe Character Recognition by a Syntactic Approach Using Finite Automata", *International Journal on Graphics, Vision and Image Processing*, vol. 10 no. 2, pp. 1–8, 2010.

[15] M. Outahajala *et al.*, "Using Confidence And Informativeness Criteria To Improve POS Tagging In Amazigh", *Journal of Intelligence and Fuzzy Systems*, vol. 28, no. 3, pp. 1319–1330, 2015. http://dx.doi.org/10.3233/IFS-141417

[16] S. Boulaknadel and F. Ataa Allah, "Building a standard Amazighe corpus", *In Proceedings of the International Conference on Intelligent Human Computer Interaction*, Prague, Tchec, 2011.

[17] F. Ataa Allah and S. Boulaknadel, "Pseudoracinisation de la langue Amazighe", *In Proceeding of Traitement Automatique des Langues Naturelles*, Montréal, Canada, 2010.

[18] F. Ataa Allah and H. Jaa, "Etiquetage morphosyntaxique: Outil d'assistance dédié à la langueAmazighe", *In Proceedings of the 1er Symposium international sur le traitement automatique de la culture Amazighe*, Agadir, Moroco, pp. 110–119, 2009.

[19] F. Ataa Allah and S. Boulaknadel, "Amazighe Search Engine: Tifinaghe Character Based Approach", *In Proceeding of the International Conference on Information and Knowledge Engineering*, Las Vegas, Nevada, USA, 2010, pp. 255–259.

[20] S. Boulaknadel and F. Ataa Allah, "Online Amazighe Concordancer", *In Proceedings of the International Symposium on Image Video Communications and Mobile Networks*, Rabat, Maroco, 2010. http://dx.doi.org/10.1109/isvc.2010.5656272

[21] F. Ataa Allah and S. Boulaknadel, "Amazigh Verb Conjugator", *In Proceedings of the 9th edition of the Language Resources and Evaluation Conference*, Reykjavik, Iceland, 2014.

[22] F. Nejme *et al.*, "Analyse Automatique de la Morphologie Nominale Amazighe", *Actes de la conférence du Traitement Automatique du Langage Naturel (TALN)*, Les Sables d'Olonne, France, vol. 1, Taln, 2013.

[23] F. Nejme *et al.*, "Finite State Morphology for Amazighe Language", *In Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, Samos, Greece, 2013. http://dx.doi.org/10.1007/978-3-642-37247-616

[24] F. Nejme *et al.*, "Toward a noun morphological analyser of standard Amazinge", *In Proceedings of the InternationalConferenceon Systems and Applications (AICCSA)*, Fes, Maroco, 2013. http://dx .doi.org/10.1109/AICCSA.2013.6616457

[25] F. Nejme *et al.*, "Toward an amazigh language processing", *In Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP)*, COLING, December, Mumbai 2012, pp. 173–180.

[26] F. Nejme and S. Boulaknadel, "Formalisation de l'Amazighe standard avec NooJ", *Actes de la conférence JEP-TALN-RECITAL*, Grenoble, France, 2012.

[27] F. Nejme *et al.*, "Vers un dictionnaire électronique de l'Amazighe", *Actes de la Conférence Internationale sur les Technologies d'Information et de Communication pour l'AMazighe (TICAM)*, Rabat, Maroco, 2012.

[28] J. Greenberg, *The Languages of Africa*. The Hague, 1966.

[29] O. Ouakrim, "Fonetica y fonologıá del Bereber", Survey at the University of Autònoma de Barcelona, 1995.

[30] A. Boukous, "Société, langues et cultures auMaroc: Enjeux symboliques", Casablanca, Najah El Jadida, 1995.

[31] S. Chaker, "Le berbère", *Les langues de France (sous la direction de Bernard Cerquiglini)*, Paris, PUF, pp. 215–227, 2003.

[32] M. Ameur and A. Boumalk, "Standardisation de l'Amazighe. Actes du séminaire organisé par le Centre de l'Aménagement Linguistique", *Publication de l'Institut Royal de la Culture Amazighe*, Rabat, Maroco, 2004.

[33] H. Moujahid, "La classe du Nom dans un parler de la langue tamazighete, le tachelhiyt d'Ighrem (Souss-Maroc)", *Thèse de 3ème cycle*, Paris V, Université René Descartes, 1981.

[34] R. Laabdelaoui *et al.*, *Manuel de conjugaison de l'Amazighe*, IRCAM, Rabat, Morocco, 2012.

[35] K. Beesley and L. Karttunen, "Finite State Morphology: CSLI Studies in Computational Linguistics", *Stanford University, CA: CSLI Publications*, 2003.

[36] L. Karttunen and K. R. Beesley, "A short history of two–level morphology", In: *Talk given at the ESSLLI workshop on finite state methods in natural language processing*, 2001.

[37] K. Koskenniemi, "Two-level morphology: a general computationalmodel forword-formrecognition and production", *The Department of General Linguistics*, University of Helsinki, 1983.

[38] L. Karttunen *et al.*, "Regular Expressions for Language Engineering", *Journal of Natural Language Engineering*, vol. 2, no. 4, pp. 307–330, 1997.

[39] A. Donabédian *et al.*, "EDS, NooJ Computational Devices", In *Formalising Natural Languages with NooJ*, Cambridge Scholars, Cambridge, 2013.

[40] M. Silberztein, "Open Source Multiplatform NooJ for NLP: CESAR Project", In *Proceedings of COLING 2012*. Demonstration Papers, Mumbai, December, 2012, pp. 401–408.

[41] M. Silberztein, "NooJ: The Lexical Module In NooJ pour le Traitement Automatique des Langues", *S. Koeva, D. Maurel, M. Silberztein EDS, Cahiers de la MSH Ledoux*, Presses Universitaires de Franche-Comté. 2005.

[42] M. Silberztein, (2006). *NooJ Manual*. Available: http://www.nooj4nlp.net

[43] M. Silberztein, "An Alternative Approach to Tagging", *NLDB 2007*, pp. 1–11, 2007. http://dx.doi.org/10.1007/978-3-540-73351-5_1

[44] M. Silberztein, (2015, March, 17), *NOAJ: A Linguistic Development Environment* [Online]. Available: http://www.nooj-association.org/index.php?option=com_nooj&controller=module&task=display_module_fo&Itemid=529

[45] A. Donabédian and N. Boyacioglu, "La lemmatisation de l'arménien occidental avec Nooj S. Koeva, D. Maurel, M. Silberztein, Formaliser les langues avec l'ordinateur, de INTEX 'a NooJ" *Presses Universitaires de Franche-Comté*, pp. 55–75, 2007.

[46] K. Vučkovic *et al.*, "Croatian Language Resources for NooJ", *CIT. Journal of Computing and Information Technology*, vol. 18, no. 4, pp. 295–30, 2010. http://dx.doi.org/10.2498/cit.1001914

[47] S. Mesfar, "Analyse Morpho-syntaxique Automatique et Reconnaissance Des Entités Nommées En Arabe Standard", Thesis, Graduate School-Languages, Paris, France, 2008.

[48] M. Taifi, *Dictionnaire tamazight-francÜfais: parlers du Maroc central Paris*. L'Harmattan-Awal, 1991.

[49] F. Agnaou *et al.*, *Lexique Scolaire*. IRCAM, Rabat, Maroco, 2011.

[50] L. Oulhaj, *Grammaire du Tamazight*. Imprimerie Najah El Jadida, 2000.

[51] A. Berkai, "Lexique de la linguistique Francais-Anglais Berbère: précédé d'un essai de typologie des procédés neologiques", 2007.

[52] A. Lhousni, "Dérivation des nominaux en Tamazight", *Mémoire de Licence, Faculté des Lettres et des Sciences Humaines*, Université Sidi Mohamed Ben Abdellah, Maroco, 1990.

[53] M. Azougarh, "Lexique berbère: structures et signification", Thèse, Faculté des Lettres et des Sciences Humaines, Oujda, 1992.

Appendix

Appendix A
Monoliteral root represents one consonant like ⵕⵓ [ru] "to cry", Biliteral root stands for two consonants like ⴰⵖⵍ [agl] "suspend", etc.

Appendix B
Arguments for commands ⟨ B⟩ , ⟨ L⟩ , ⟨ R⟩ , ⟨ S⟩ : xx number: repeat xx times.

Appendix C
C̄ is used when a consonant is reduplicated.

Appendix D
Figure 2: In the aim to census the maximum rules to recognize and generate all possible inflected forms of each noun, we have made a study based on works of [7] [16] [37] [13] [1] and on heuristic study of our lexicons. Thus, for each pattern, we have extracted the most relevant rules (e.g. for the form ⵜⴰ...Cⵜ [ta. . . Ct] the rule will be an alternation of the first vowel accompanied by suffixing of ⵉⵏ [in] and removing the last ⵜ [t]). Our study includes also the exceptions.

Appendix E
Kinship noun: the noun designing the relationship between members of the same family.

Appendix F
Writing System: Scripts.

Appendix G
+ FLX introduces the functionality which describes all the potential forms from a lemma.

Appendix H
A root is a sequence of one or many consonants and the pattern is a template of vowels (V) and consonants.

Appendix I
Genre = Gender, Nb = Number, EL = Free State and EA = Constructed one.

Appendix J
IRCAM: an institution created in 2001 to preserve, promote and endorse Amazighe culture in all its dimensions.

Received: September, 2014
Revised: July, 2015
Accepted: January, 2016

Contact addresses:
Fatima Zahra Nejme
LRIT
Mohammed V University
Faculty of Science
Rabat
Morocco
e-mail: fatimazahra.nejme@gmail.com

Siham Boulaknadel
CEISIC
Royal Institute of Amazigh Culture
Rabat
Morocco
e-mail: boulaknade@icram.ma

Driss Aboutajdine
LRIT
Mohammed V University
Faculty of Science
Rabat
Morocco
e-mail: aboutaj@hotmail.com

FATIMA ZAHRA NEJME is a PhD student in Engineering Sciences at Mohammed V-Agdal University, Rabat, Morocco. She received the Master's degree from the same University in Computer Sciences and Telecommunications in 2011. Her research interests focus on the development of natural language processing tools for Amazighe language.

SIHAM BOULAKNADEL is researcher at Royal Institute of Amazighe Culture, Morocco. She obtained her PhD in Computer Sciences from the Faculty of Science of Rabat in collaboration with the University of Nantes in 2008. Her research interests focus on natural language processing, information retrieval, artificial intelligence, and e-learning. She is currently involved in several national projects dealing with developing linguistics resources and natural language processing tools for less resourced languages, particularly the Amazighe language. She has also largely contributed to the supervision of young researchers in various topics, especially in developing numerical learning resources of the Amazighe language. In addition, she is the author or co-author of numerous national and international publications.

DRISS ABOUTAJDINE received the PhD degrees in Signal Processing from the Mohammed V-Agdal University, Rabat, Morocco, in 1985. He joined Mohammed V-Agdal University in 1978, first as an assistant professor and since 1990, he is professor at the Faculty of Science, heading the LRIT laboratory. Actually, he is the national coordinator of the National Information Technology Network of Excellence. Over 30 years, he has developed research activities covering various topics of signal and image processing, wireless communication, pattern recognition and natural language processing which allowed him to publish over 300 journal papers and conference communications. He was elected member of the Hassan II Moroccan Academy of Science and Technology in May 2006 and he was Vice President in charge of research, cooperation and partnership of the Mohammed V Souissi University from September 2008 to December 2010. Actually, he is the Director of the National Center for Scientific and Technical Research.