Tagging Amazigh with AnCoraPipe

Mohamed Outahajala, Lahbib Zekouar, Paolo Rosso, M. Antònia Martí

Institut Royal de la Culture Amazighe, Ecole Mohammadia des Ingénieurs, Natural Language Engineering Lab – EliRF(DSIC), CLiC - Centre de Llenguatge i Computació

Avenue Allal El Fassi, Madinat Al Irfane - Rabat - Instituts Adresse postale : BP 2055 Hay Riad Rabat Morocco, Avenue Ibnsina B.P. 765 Agdal Rabat Morocco, Universidad Politécnica de Valencia, Spain, Universitat de Barcelona 08007 Barcelona, Spain

E-mail: outahajala@ircam.ma, zenkouar@emi.ac.ma, prosso@dsic.upv.es, amarti@ub.edu

Abstract

Over the last few years, Moroccan society has known a lot of debate about the Amazigh language and culture. The creation of a new governmental institution, namely IRCAM, has made it possible for the Amazigh language and culture to reclaim their rightful place in many domains. Taking into consideration the situation of the Amazigh language which needs more tools and scientific work to achieve its automatic processing, the aim of this paper is to present the Amazigh language features for a morphology annotation purpose. Put in another way, the paper is meant to address the issue of Amazigh's tagging with the multilevel annotation tool AnCora Pipe. This tool is adapted to use a specific tagset to annotate Amazigh corpora with a new defined writing system. This step may well be viewed as the first step for an automatic processing of the Amazigh language; the main aim at very beginning being to achieve a part of speech tagger.

Introduction

Amazigh (Berber) is spoken in Morocco, Algeria, Tunisia, Libya, and Siwa (an Egyptian Oasis); it is also spoken by many other communities in parts of Niger and Mali. It is a composite of dialects of which none have been considered the national standard used by tens of millions of people in North Africa mainly for oral communication.

With the emergence of an increasing sense of identity, Amazigh speakers would very much like to see their language and culture rich and developed. To achieve such a goal, some Maghreb states have created specialized institutions, such as the Royal Institute for Amazigh Culture (IRCAM, henceforth) in Morocco and the High Commission for Amazigh (HCA) in Algeria. In Morocco, Amazigh has been introduced in mass media and in the educational system in collaboration with relevant ministries. Accordingly, a new Amazigh television channel was launched in first mars 2010 and it has become common practice to find Amazigh taught in various Moroccan schools as a subject.

Over the last 7 years, IRCAM has published more than 140 books related to the Amazigh language and culture, a number which exceeds the whole amount of Amazigh publications in the 20th century, showing the importance of an institution such as IRCAM. However, in Natural Language Processing (NLP) terms, Amazigh, like most non-European languages, still suffers from the scarcity of language processing tools and resources.

In this sense, since morphosyntactic tagging is an important and basic step in the processing of any given language, the main objective of this paper is to explain how we propose to supply the Amazigh language with this important tool.

For clarity reasons, this paper is organized as follows: in the first part we present an overview of the Amazigh language features. Then, we provide a brief retrospective on Amazigh morphology as conceived by IRCAM linguists. Next we give an overview on Amazigh corpora. The fourth section describes how to tag with AnCoraPipe and the fifth section deals with Amazigh tagset.

2. The Amazigh language

Amazigh belongs to the Hamito-Semitic/"Afro-Asiatic" languages (Cohen 2007, Chaker 1989) with rich templatic morphology. In linguistic terms, the language is characterized by the proliferation of dialects due to historical, geographical and sociolinguistic factors. In Morocco, one may distinguish three major dialects: Tarifit in the North, Tamazight in the center and Tashlhiyt in the southern parts of the country; 50% of the Moroccan population speak Amazigh (Boukouss, 1995), but according to the last governmental demolinguisitc data of 2004, the Amazigh language was spoken only by some 28% of the Moroccan population (around 10 Million inhabitants), showing an important decrease of its use.

Amazigh standardization cannot be achieved without adopting a realistic strategy that takes into consideration its linguistic diversity (Ameur et al., 2006a; Ameur et al. 2006b). As far as the alphabet is concerned, and because of historical and cultural reasons, Tifinaghe has become the official graphic system for writing Amazigh. IRCAM kept only pertinent phonemes for Tamazight, so the number of the alphabetical phonetic entities is 33, but Unicode codes only 31 letters plus a modifier letter to form the two phonetic units: ⵗⵯ(g^w) and ⴽⵯ(k^w). The whole range of Tifinagh letters is subdivided into four subsets: the letters used by IRCAM, an extended set used also by IRCAM, other neo-tifinaghe letters in use and some attested modern Touareg letters. The number reaches 55 characters (Zenkouar 2004, Andries 2004). In order to rank strings and to create keyboard layouts for Amazigh in accordance with international standards, two other standards have been adapted (Outahajala and Zenkouar, 2004):

- ISO/IEC14651 standard related to international string

ordering and comparison method for comparing character strings and description of the common template tailorable ordering;
- Part 1: general principles governing keyboard layouts of the standard ISO/IEC 9995 related to keyboard layouts for text and office systems.
Most Amazigh words may be conceived of as having consonantal roots. They can have one, two, three or four consonants, and may sometimes extend to five. Words are made out of these roots by following a pattern (Chafiq 1991). For example the word 'aslmad' is built up from the root *lmd* "study" by following the pattern as12a3, where the number 1 is replaced by the first consonant of the root, number 2 is replaced by the second consonant of the root and number 3 is replaced by the 3rd consonant of the root.
Concerning spelling, the system put by IRCAM is based on a set of rules and principles applied to "words" along which the parsing of pronounced speech into written separated words is effected. A grapheme, a written word, according to the spelling system is a succession of letters which can sometimes be one letter delimited by whitespace or punctuation.
The graphic rules for Amazigh words are set out as follows (Ameur et al 2006a, 2006b, Boukhris et al 2008):
- Nouns consist of a single word occurring between two blank spaces. To the noun are attached the morphological affixes of gender (masculine/ feminine), number (singular/plural) and state (free/construct) as it is shown in the following examples: ⵄⵎⴰⵢ/ⵜⵄⵎⴰⵢⵜ (amzday (masc.)/**t**amzday**t**(fem.)) "a dweller", ⵄⵎⴰⵢ/ⵉⵎⴰⵢⵏ (amzday (sing.)/**i**mzday**n**(plr.)) "dweller/dwellers", and ⵄⵎⴰⵢ/ⵓⵎⴰⵢ (amzday (free state)/umzday (construct state)). Kinship names constitute a special class since they are necessarily determined by possessive markers which form with them one word, for example: ⴱⴰⴱⴾ (babak) which means "your father";
- Quality names/adjectives constitute a single word along with the morphological indicators of gender (masculine/ feminine), number (singular/plural), and state (free/construct);
- Verbs are single graphic words along with its inflectional (person, number, aspect) or derivational morphemes. For example: ⵜⵜⴰⵣⵣⵍ /ttazzl/which means "you run (imperfective)". The verb is separated by a blank space from its predecessor and successor pronouns, i.e.: ⵢⴰⵙⵉ ⵜⵍ / ⴰⵏ ⵜⵍ ⵢⴰⵙⵉ ("yasi tn / ad tn yasi" which means "he took them / he will take them");
- Pronouns are isolated from the words they refer to. Pronouns in Amazigh are demonstrative, exclamative, indefinite, interrogative, personal, possessive, or relative. For instance, ⴰⵏ (ad) in the phrase ⴰⴱⵓⵉⵏ ⴰⵏ (abrid ad), which means "this way", is an example of a demonstrative pronoun;
- An adverb consists of one word which occurs between two blank spaces. Adverbs are divided into adverbs of place, time, quantity, manner, and interrogative adverbs.
- Focus mechanisms, interjections and conjunctions are written in the form of single words occurring between two blank spaces. An example of a conjunction is: ⵎⵓ (mr) which means "if";
- Prepositions are always an independent set of characters with respect to the noun they precede; however, if the preposition is followed by a pronoun, both the preposition and the noun make a single whitespace-delimited string. For example: ⵖⵓ (γr) "to, at" + ⵉ (i) "me" possessive pronoun gives ⵖⴰⵓⵉ/ⵖⵯⵓⵉ (γari/γuri) "to me, at me, with me";
- Particles are always isolated. There are aspect particles such as ⴰⵣⵣⴰ (aqqa), ⴰⵓ (ar), ⴰⵏ (ad), particles of negation such as ⵯⵓ (ur), orientation particles like ⵍⵍ in ⴰⵍⵉ ⵍⵍ!(awi nn) "take it there" and a predicative particle ⵏ (d);
- Determinants take always the form of single between two blank spaces. Determiners are divided into articles, demonstratives, exclamatives, indefinite articles, interrogatives, numerals, ordinals, possessives, presentatives, quantifiers. ⴽⵯⵍⵍⵯ (kullu) "all" is a quantifier for instance;
- Amazigh punctuation marks are similar to the punctuation marks adopted in international languages and have the same functions. Capital letters, nonetheless, do not occur neither at the beginning of sentences nor at the initial of proper names.
The English terminology used above was extracted form (Boumalk and Naït-Zerrad, 2009).

3. Amazigh corpora

3.1 Amazigh corpora features

Amazigh corpora have the following characteristics:
- They are extracted from geographically circumscribed dialects;
- Some varieties are less represented than others, or not studied at all;
- There is special need for a more general type of work whose goal is to collect the data of all dialects;
- Existing publications are scattered and inaccessible in most cases. Some of them go back to the XIXth century and the beginning of the XXth century. The few existing copies of those references are only available in specialized libraries, mainly in France;
- General documents containing the data of all Amazigh dialects do not exist (phonetics, semantics, morphology, phraseology…etc.).
- Some existing texts need revision because of segmentation problems.
To constitute an annotated corpus, we have chosen a list of corpora extracted from the Amazigh version of IRCAM's web site[1], the periodical *Inghmisn n usinag*[2] (IRCAM newsletter) and school textbooks. We were able to reach a total number of words superior to 20k words. A comparative quantity of corpora was used in tagging other languages, for example (Allauzen and Bonneau-Maynard, 2008).

[1] www.ircam.ma
[2] Freely downloadable from http://www.ircam.ma/amz/index.php?soc=bulle

3.2 Writing systems

Amazigh corpora produced up to now are written on the basis of different writing systems, most of them use Tifinaghe-IRCAM (Tifinaghe-IRCAM makes use of Tifinaghe glyphs but Latin characters) and Tifinaghe Unicode. It is important to say that the texts written in Tifinaghe Unicode are increasingly used.

Even though, we have decided to use a specific writing system based on ASCII characters for the following reasons:

- To have a common set of characters for annotated corpora;
- To facilitate texts treatment for annotators since ASCII characters are known by all systems;
- To handle its use due to the fact that people are still more familiar with Arabic and Latin writing systems.

In Table 1 of correspondences between the different writing systems and transliteration correspondences is shown

Tifinaghe Unicode		Transliteration		Used characters in Tifinaghe IRCAM		Chosen characters for tagging
Code	Character	Latin	Arabic	characters	codes	
U+2D30	ⴰ	a	ا	A, a	65, 97	a
U+2D31	ⴱ	b	ب	B, b	66, 98	b
U+2D33	ⴳ	g	گ	G, g	71, 103	g
U+2D33& U+2D6F	ⴳ ⵯ	g^w	گ + ◌ٗ	Å, å	197, 229	g°
U+2D37	ⴷ	d	د	D, d	68, 100	d
U+2D39	ⴹ	ḍ	ض	Ä, ä	196, 228	D
U+2D3B	ⴻ	e^3	يـــ	E, e	69, 101	e
U+2D3C	ⴼ	f	ف	F, f	70, 102	f
U+2D3D	ⴽ	k	ک	K, k	75, 107	k
U+2D3D& U+2D6F	ⴽ ⵯ	k^w	گ + ◌ٗ	Æ, æ	198, 230	k
U+2D40	ⵀ	h	ه	H, h	72,104	h
U+2D40	ⵃ	ḥ	ح	P, p	80,112	H
U+2D44	ⵄ	ɛ	ع	O, o	79, 111	E
U+2D45	ⵅ	x	خ	X, x	88, 120	x
U+2D47	ⵇ	q	ق	Q, q	81, 113	q
U+2D49	ⵉ	i	ي	I, i	73, 105	i
U+2D4A	ⵊ	j	ج	J, j	74, 106	j
U+2D4D	ⵍ	l	ل	L, l	76, 108	l
U+2D4E	ⵎ	m	م	M, m	77, 109	m
U+2D4F	ⵏ	n	ن	N, n	78, 110	n
U+2D53	ⵓ	u	و	W, w	87, 119	u
U+2D54	ⵔ	r	ر	R, r	82, 114	r
U+2D55	ⵕ	ṛ	ڕ	Ë, ë	203, 235	R
U+2D56	ⵖ	ɣ	غ	V, v	86, 118	G
U+2D59	ⵙ	s	س	S, s	83, 115	s
U+2D5A	ⵚ	ṣ	ص	Ã, ã	195, 227	S
U+2D5B	ⵛ	c	ش	C, c	67, 99	c
U+2D5C	ⵜ	t	ت	T, t	84, 116	t
U+2D5F	ⵟ	ṭ	ط	Ï, ï	207, 239	T
U+2D61	ⵡ	w	ۏ	W, w	87, 119	w
U+2D62	ⵢ	◌ٔ + ي	ي	Y, y	89, 121	y
U+2D63	ⵣ	z	ز	Z, z	90, 122	z
U+2D65	ⵥ	ẓ	ژ	Ç, ç	199, 231	Z
U+2D6F	ⵯ	ʷ	◌ٗ	No correspondant in Tifinaghe-IRCAM		°

Table1: The mapping from existing writing systems and the chosen writing system.

A transliteration tool was build in order to handle transliteration to and from the chosen writing system and to correct some elements such as the character "^" which exists in some texts due to input error in entring some Tifinaghe letters. So the sentence portion "ⵄⵙⵙ ⵉ ⵜⵍⵢⵓⵄ" using Tifinaghe Unicode or "ass n tmvra" using Tifinaghe-IRCAM will be transliterated as "ass n tmGra" ("When the day of the wedding arrives").

4. AnCoraPipe tool

AnCoraPipe (Bertran et al. 2008) is a corpus annotation tool which allows different linguistic levels to be annotated efficiently, since it uses the same format for all stages. The tool reduces the annotation time and makes easy the integration of the different annotators and the different annotation levels.

The input documents may have a standard XML format, allowing to represent tree structures (specially usefull at syntactic anotation stages). As XML is a wide spread standard, there are many tools available for its analysis, transformation and management.

AnCoraPipe includes an integrated search engine based on XPath language (http://www.w3.org/TR/xpath/), which allows to find structures of all kinds among the documents. For corpus analysis, an export tool can summarize the attributes of all nodes in the corpus in a grid that can easily be imported to basic analysis tools (such as Excel or OpenOffice calc), statistical software (SPSS) or Machine Learning tools (Weka).

A default tagset is provided in the standard installation. It has been designed as generic as possible in order to match the requisites of a wide amount of languages. In spite of that, if the generic tagset is not useful, the interface is fully customizable to allow different tagsets defined by the user.

In order to allow AnCoraPipe usable in a full variety of languages, the user can change the visualization font. This may help viewing non-latin scripts such as Chinese, Arabic or Amazigh.

AnCoraPipe is currently an Eclipse Plugin. Eclipse is an extendable integrated development environment. With this plugin, all features included in Eclipse are made available

[3] note : different use in the IPA which uses the letter ə

for corpus annotation and developing. In particular, the Eclipse's collaboration and team plugins can be used to organize the work of a group of annotators.

5. AnCoraPipe for Amazigh

AnCoraPipe allows the definition of different tagsets. We have decided to work with a set of ASCII characters for the following reasons:
- Amazigh text corpora are written in different writing systems;
- Amazigh linguists are still familiar with Latin alphabets;
- the default tagset is a multilevel tagset;
- to simplify the interface for linguists;
- to avoid adding some tags which are not currently needed as co-reference tags, syntactic tags...etc.

Based on the Amazigh language features presented above, Amazigh tagset may be viewed to contain 13 nodes with two common attributes to each node: "wd" for "word" and "lem" for "lemma", whose values depend on the lexical item they accompany.

Amazigh nodes and their attributes are set out in what follows:

PoS	**attributes and subattributes with number of values**
Noun	gender(3), number(3), state(2), derivative(2),PoSsubclassification(4), person(3), possessornum(2), possessorgen(2)
Adjective/ name of quality	gender(3), number(3), state(2), derivative(2), PoS subclassification(3)
Verb	gender(3), number(3), form(5), aspect(3), negative(2), form(2)
Pronoun	gender(3), number(3), PoS subclassification(7), deictic(3), autonome(2), person(3), possessornum(2), possessorgen(2)
Determiner	gender(3), number(3), PoS subclassification(11)
Adverb	PoS subclassification(5)
Preposition	gender(3), number(3), PoS subclassification(6), person(3), possessornum(2), possessorgen(2)
Conjunction	PoS subclassification(2)
Interjection	
Particle	PoS subclassification(5)
Focus	
Residual	PoS subclassification(5), gender(3), number(3)
Punctuation	punctuation mark type(16)

Table2: A synopsis of the features of the Amazigh PoS tagset with their attributes and values

In Table 2 the node Residual stands for attributes like currency, number, date, math marks and other unknown residual words.

Manual annotation is being carried out by a team of linguists. Technically, manual annotation proceeds along the requirements of the tool presented above.

A sample of annotated Corpora as presented in Section 3:

Here follows the annotation of a sentence extracted from a text about a wedding ceremony:

"ass n tmGra, iwsn asn ayt tqbilt. illa ma issnwan, illa ma yakkan i inbgiwn ad ssirdn"

[English translation: "When the day of the wedding arrives, the people of the tribe help them. Some of them cook; some other help the guests get their hands washed "]

```
<sentence>
<n gen="m" lem="ass" num="s" state="free" wd="ass"/>
<prep wd="n"/>
<n gen="f" lem="tamGra" num="s" state="construct" wd="tmGra"/>
<pu punct="comma" wd=","/>
<v aspect="perfective" gen="m" lem="aws" num="p" person="3" wd="iwsn"/>
<p gen="m" num="p" person="3" postype="personal" wd="asn"/>
<d gen="m" num="p" postype="indefinite" wd="ayt"/>
<n gen="f" lem="taqbilt" num="s" postype="common" state="construct" wd="tqbilt"/>
<pu punct="period" wd="."/>
<v aspect="perfective" gen="m" lem="ili" num="s" person="3" wd="illa"/>
<p postype="relative" wd="ma"/>
<v aspect="imperfective" gen="m" lem="ssnw" num="s" person="3" form="participle" wd="issnwan"/>
<pu punct="comma" wd=","/>
<v aspect=" perfective" gen="m" lem="ili" num="s" person="3" wd="illa"/>
<p postype="relative" wd="ma"/>
<v aspect="imperfective" form="participle" gen="m" lem="fk" num="s" person="3" wd="yakkan"/>
<prep wd="i"/>
<n gen="m" lem="anbgi" num="p" state="construct" wd="inbgiwn"/>
<pr postype="aspect" wd="ad"/>
<v aspect="aorist" gen="m" lem="ssird" num="p" person="3" wd="ssirdn"/>
```

The main aim of this corpus is to achieve a part of speech tagger based on Support Vector Machines (SVM) and Conditional Random Fields (CRF) because they have been proved to give good results for sequence classification (Kudo and Matsumoto, 2000, Lafferty et al. 2001). We are planning to use freely available tools like Yamcha and

CRF++ toolkits[4].

6. Conclusion and future works

In this paper, after a brief description about social and linguistic characteristics of the Amazigh language, we have addressed the basic principles we followed for tagging Amazigh written corpora with AnCoraPipe: the tagset used, the transliteration and the annotation tool.

In the future, it is our goal to tag more corpora to constitute a reference corpus for works on Amazigh NLP and we plan also to work on Amazigh Base Phrase Chunking.

Acknowledgments

We would like to thank Manuel Bertran for improving the AnCora Pipe tool to support Amazigh features, all IRCAM researchers and Professor Iazzi El Mehdi from Ibn Zohr University, Agadir for their explanations and precious help. The work of the last two authors was carried out thanks to AECID-PCI C/026728/09 and TIN2009-13391-C04-03/04 research projects.

References

Allauzen, A. Bonneau-Maynard, H. (2008).Training and evaluation of POS taggers on the French MULTITAG corpus. In proceedings of LREC 08.

Ameur, M., Boujajar, A., Boukhris, F. Boukouss, A., Boumaled, A., Elmedlaoui, M., Iazzi, E., Souifi, H. (2006a), *Initiation à la langue Amazighe*. Publications de l'IRCAM. pp. 45—77.

Ameur, M., Boujajar, A., Boukhris, F. Boukouss, A., Boumaled, A., Elmedlaoui, M., Iazzi, E. (2006b) *Graphie et orthographe de l'Amazighe*. Publications de l'IRCAM.

Andries, P. (2004). La police open type Hapax berbère. In proceedings of the workshop : *la typographie entre les domaines de l'art et l'informatique*, pp. 183—196.

Bertran, M., Borrega, O., Recasens, M., Soriano, B. (2008). AnCoraPipe: A tool for multilevel annotation. *Procesamiento del lenguaje Natural*, nº 41. Madrid (Spain).

Boukhris, F. Boumalk, A. El moujahid, E., Souifi, H. (2008). *La nouvelle grammaire de l'Amazighe*. Publications de l'IRCAM.

Boukhris, F. (2006). Structure morphologique de la préposition en Amazighe. In proceedings of the workshop: *Structures morphologiques de l'Amazighe*. Publications de l'IRCAM. pp. 46-56.

Boukouss, A. (1995). Société, langues et cultures au Maroc: Enjeux symboliques, *publications de la Faculté des Lettres de Rabat*.

Boumalk, A., Naït-Zerrad, K. (2009). *Amawal n tjrrumt -Vocabulaire grammatical*. Publications de l'IRCAM.

Chafiq, M. (1991) .أربعة وأربعون درسا في الأمازيغية. éd. Arabo-africaines.

Chaker, S. (1989). Textes en linguistique berbère - introduction au domaine berbère, éditions du CNRS, 1984. P 232-242.

Cohen, D. (2007). Chamito-sémitiques (langues). In *Encyclopædia Universalis*.

Iazzi, E., Outahajala,M. (2008), Amazigh Data Base. In proceedings of LREC 08.

Kudo, T., Yuji Matsumoto, Y. (2000). Use of Support Vector Learning for Chunk Identification.

Lafferty, J. McCallum, A. Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In proceedings of ICML-01 282-289

Outahajala, M., Zenkouar, L. (2005). La norme du tri, du clavier et Unicode. In proceedings of the workshop : *la typographie entre les domaines de l'art et l'informatique*, pp. 223—238.

Saa, F. (2006). Les thèmes verbaux de l'Amazighe. In proceedings of the workshop: *Structures morphologiques de l'Amazighe*, pp.102--111.

Zenkouar, L. (2004). L'écriture Amazighe Tifinaghe et Unicode, in *Etudes et documents berbères*. Paris (France). n° 22, pp. 175—192.

Zenkouar, L. (2008). Normes des technologies de l'information pour l'ancrage de l'écriture Amazighe, in *Etudes et documents berbères*. Paris (France), n° 27, pp. 159—172.

[4] Freely downloadable from http://chasen.org/~taku/software/YamCha/ and http://crfpp.sourceforge.net/