Towards an electronic dictionary of Tamajaq language in Niger

Chantal Enguehard, Issouf Modi

Enguehard, C., Modi I. Towards an electronic dictionary of Tamajaq language in Niger. 12th Conference of the European Chapter of the Association for Computational Linguistics EACL-09. W07 Workshop Language Technologies for African Languages. Athens, Greece, 31 March 2009.

Towards an electronic dictionary of Tamajaq language in Niger

Chantal Enguehard
LINA - UMR CNRS 6241
2, rue de la Houssinière
BP 92208
44322 Nantes Cedex 03
France
chantal.enguehard@univ-nantes.fr

Issouf Modi
Ministère de l'Education Nationale
Direction des Enseignements du Cycle de Base1
Section Tamajaq.
Republique du Niger
modyissouf@yahoo.fr

Abstract

We present the Tamajaq language and the dictionary we used as main linguistic resource in the two first parts. The third part details the complex morphology of this language. In the part 4 we describe the conversion of the dictionary into electronic form, the inflectional rules we wrote and their implementation in the Nooj software. Finally we present a plan for our future work.

1. The Tamajaq language

1.1 Socio-linguistic situation

In Niger, the official language is French and there are eleven national languages. Five are taught in a experimental schools: Fulfulde, Hausa, Kanuri, Tamajaq and Soŋay-Zarma.
According to the last census in 1998, the Tamajaq language is spoken by 8,4% of the 13.5 million people who live in Niger. This language is also spoken in Mali, Burkina-Faso, Algeria and Libya. It is estimated there are around 5 millions Tamajaq-speakers around the world.
The Tamacheq language belongs to the group of Berber languages.

1.2 Tamajaq alphabet

The Tamajaq alphabet used in Niger (Republic of Niger, 1999) uses 41 characters, 14 with diacritical marks that all figure in the Unicode standard (See appendix A). There are 12 vowels: a, â, ă, ə, e, ê, i, î, o, ô, u, û.

1.3 Articulatory phonetics

Consonants		Voiceless	Voiced
Bilabial	Plosive		b
	Nasal		m
	Trill		r
	Semivowel		w
Labiodental	Fricative	f	
Dental	Plosive	t	d
	Fricative	s	z
	Nasal		n
	Lateral		l
Pharyngeal	Plosive	ṭ	ḍ
	Fricative	ṣ	ẓ
	Lateral		ḷ
Palatal	Plosive	c	ǰ

Consonants		Voiceless	Voiced
	Fricative	š	j
	Semivowel		y
Velar	Plosive	k	g, ğ
	Fricative	ɣ	x
	Nasal		ŋ
Glottal	Plosive	q	
	Fricative	h	

Table 1a: Articulatory phonetics of Tamajaq consonants

Vowels	Close	Close-mid	Open-mid	Open
Palatal	i	e		
Central		ə	ă	a
Labial	u	o		

Table 1b: Articulatory phonetics of Tamajaq vowels

1.4 Tools on computers

There are no specific TALN tools for the Tamajaq language.
However characters can be easily typed on French keyboards thanks to the AFRO keyboard layout (Enguehard and al. 2008).

2 Lexicographic resources

We use the school editorial dictionary "dictionnaire Tamajaq-français destiné à l'enseignement du cycle de base 1". It was written by the SOUTEBA[1] project of the DED[2] organisation in 2006. Because it targets children, this dictionary consists only of 5,390 entries. Words have been chosen by compiling school books.

2.1 Structure of an entry

Each entry generally details :
- lemma,
- lexical category,
- translation in French,
- an example,
- gender (for nouns),
- plural form (for nouns).

Examples:

« ăbada$_1$: sn. bas ventre. Daw tədist. Bărar wa yəllûẓăn ad t-yəltəɣ ăbada-net. təmust.: yy. igət: ibadan. »

« ăbada2: sn. flanc. Tasăga meɣ daw ădăg əyyăn. Iməwwəẓla əklăn dăɣ ăbada n əkašwar. Anammelu.: azador. təmust.: yy. ∃səfsəs.: ă. Igət: ibadan. »

Homonyms are described in different entries and followed by a number, as in the above example.

2.2 Lexical categories

The linguistic terms used in the dictionary are written in the Tamajaq language using the abbreviations presented in table 2. In addition, this table gives information about the number of entries of each lexical category.

Lexical category		**Abbreviation**	**Number of entries**
Tamajaq	**English**		
əḍəkuḍ	number	ḍkḍ.	3
ənalkam	deteminant	nlkm.	1
anamal	verb	nml.	1450
samal	adjective	sml.	48
əsəmmadăɣ ən təla	possessive pronoun	smmdɣtl.	5
isən	noun	sn.	3648
isən n ənamal	Verbal noun	snnml.	33
isən an təɣərit	name of shout	sntɣrt.	2
isən xalalan	proper noun	snxln.	29
isən iẓẓəwen	complex noun	snẓwn.	137
əstakar	adverb	stkr.	8

[1]Soutien à l'éducation de base.
[2]DED: Deutscher Entwicklungsdienst.

əsatkar n ădag	adverb of location	stkrdg.	10
əṣatkar n igət	Adverb of quantity	stkrgt.	1
təɣərit	onomato-poeia	tɣrt.	8
tənalkamt	particle	tnlkmt.	2

Table 2: Tamajaq lexical categories

3 Morphology

The Tamajaq language presents a rich morphology (Aghali-Zakara, 1996).

3.1 Verbal morphology

Verbs are classified according to the number of consonants of their lexical root and then in different types. There are monoliteral, biliteral triliteral, quadriliteral verbs...
Three moods are distinguished: imperative, simple injunctive and intense injunctive.
Three aspects present different possible values:
- accomplished: intense or negative;
- non accomplished: simple, intense or negative;
- aorist future: simple or negative.

Examples :

əktəb (to write): triliteral verb, type 1.

əṣṣən (to know): triliteral verb, type 2 (ṣṣn).

əməl (to say): biliteral verb, type 1

akər (to steal): biliteral verb, type 2

awəy (to carry): biliteral verb, type 3

ăšwu (to drink): biliteral verb, type 4

ăru (to love): monoliteral verb, type 2

ăru (to open): monoliteral verb, type 3

Each class of verb has its own rules of conjugation.

3.2 Nominal morphology

a. Simple nouns
Nouns present three characteristics:
- gender: masculine or feminine;
- number: singular or plural;
- annexation state is marked by the change of the first vowel.

Terminology		Abbreviation
təmust	gender	tmt.
yey	masculine	yy.
tənte	feminine	tnt.
awdəkki	singular	wdk.
iget	plural	gt.
əsəfsəs	annexation state	sfss.

Table 3: Tamajaq terminology for nouns

Example :

« aṭrəkka: sn. morceau de sucre. Akku: ablәɣ n°2. təmust.: yy. Ǝsəfsəs.: ə. Igət: əṭrəkkatăn. »

"aṭrəkka" is a masculine noun. Its plural is "əṭrəkkatăn". It becomes "əṭrəkka" when annexation state is expressed.

The plural form of nouns is not regular and has to be specifically listed.

b. Complex nouns
Complex nouns are composed by several lexical units connected together by hyphens. It could include nouns, determiners or prepositions as well as verbs.
Examples:
Noun +determiner + noun

"ejăḍ-n-əjḍăn", literally means "donkey of birds" (this is the name of a bird).

Verb + noun

"awəy-əhuḍ" literally means "it follows harmattan" (kite).

"gazẓăy-təfuk" literally means "it looks at sun" (sunflower).

Preposition + noun

"In-tamaṭ" means "the one of the tree acacia" (of acacia).

Verb + verb

"azəl-azəl" means "run run" (return).

We counted 238 complex nouns in the studied dictionary.

4 Natural Language Processing of Tamajaq

4.1 Nooj software (Silberztein, 2007)

« Nooj is a linguistic development environment that includes tools to create and maintain large-coverage lexical resources, as well as morphological and syntactic grammars. » This software is specifically designed for linguists who can use it to test hypothesis on real corpus. « Dictionaries and grammars are applied to texts in order to locate morphological, lexical and syntactic patterns and tag simple and compound words. » Nooj put all possible tags for each token or group of tokens but does not disambiguate between the multiple possibilities. However, the user can build his own grammar to choose between the multiple possible tags. The analysis can be displayed as a syntactic tree.
This software is supported by Windows.
We chose to construct resources for this software because it is fully compatible with Unicode.

4.2 Construction of the dictionary

We convert the edited dictionary for the Nooj software.
3,463 simple nouns, 128 complex nouns, 46 adjectives and 33 verbo-nouns are given with their plural form. Annexation state is indicated for 987 nouns, 23 complex nouns, 2 adjectives and 7 verbo-nouns.

We created morphological rules that we expressed as Perl regular expressions and also in the Nooj format (with the associated tag).

a. Annexation state rules
Thirteen morphological rules calculate the annexation state.

Examples:
The 'A1ă' rule replaces the first letter of the word by 'ă'.

'A1ă' rule	
Nooj	<LW><S>ă/sfss
Perl	^.(.*)$ → ă$1

Table 4: Rule 'A1ă'

The 'A2ə' rule replaces the second letter of the word by 'ə'.

'A2ə' rule	
Nooj	A2ə=<LW><R><S>ə/sfss
Perl	^(.).(.*)$ → $1ə$2

Table 5: Rule 'A2ə'

b. Plural form rules
We searched formal rules to unify the calculation of plural forms. We found 126 rules that fit from 2 up to 446 words. 2932 words could be associated with, at least, one flexional rule.

Examples:

'I4' rule deletes the last letter, adds "-ăn" at the end and "i-" at the beginning.	
Nooj	I4=ăn<LW><S>i/Iget
Perl	^(.*).$ → i$1ăn
#	446 words

Table 6: Rule 'I4'

'I2' rule deletes the last and the second letters and includes "-en" at the end and "-i-" in the second position.	
Nooj	I2=<B>en<LW><R><S>i/Iget
Perl	^(.).(.*).$ → $1i$2en
#	144 words

Table 7: Rule 'I2'

'I45' rule deletes the final letter and include "-en" at the end.	
Nooj	I45=<B>en/Iget
Perl	^(.*).$ → $1en
#	78 words

Table 8: Rule 'I45'

'I102' rule deletes the two last letters and the second one and includes a final "-a" and a "-i-" in the second position.	
Nooj	I102=<B2>a<LW><R><S>i/Iget
Perl	^(.).(.*)..$ → $1i$2a
#	6 words

Table 9: Rule 'I102'

c. Combined rules

When it was necessary, the above rules have been combined to calculate singular and plural forms with or without annexation state.
We thus finally obtained 319 rules.

Example:
I2RA2ă =
:Rwdk + :I2 + :Rwdk :A2ă + :I2 :A2ă

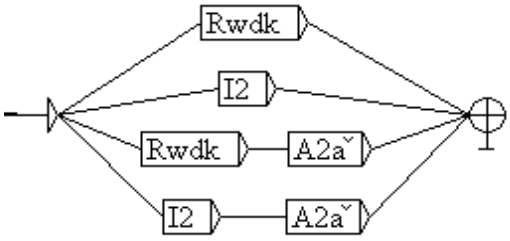

Fig. 1: Rule I2RA2ă

This rule recognizes the singular form (:Rwdk), the plural form (:I2), the singular form with the annexation state (:Rwdk :A2ă) and the plural form with the annexation state (:I2 :A2ă).

25 words meet this rule.

For instance, "taḍləmt" (accusation, provocation), is inflected in:

- taḍləmt,taḍləmt,SN+tnt+wdk
- tiḍləmen,taḍləmt,SN+tnt+Iget
- tăḍləmen,taḍləmt,SN+tnt+Iget+sfss
- tăḍləmt,taḍləmt,SN+tnt+wdk+sfss

d. Conjugaison rules

Verb classes are not indicated in the dictionary. We only describe a few conjugaison rules, just to check the expressivity of the Nooj software

Here is the rule of the verb "əṣṣən" (to know), intense accomplished aspect, represented as a transducer.

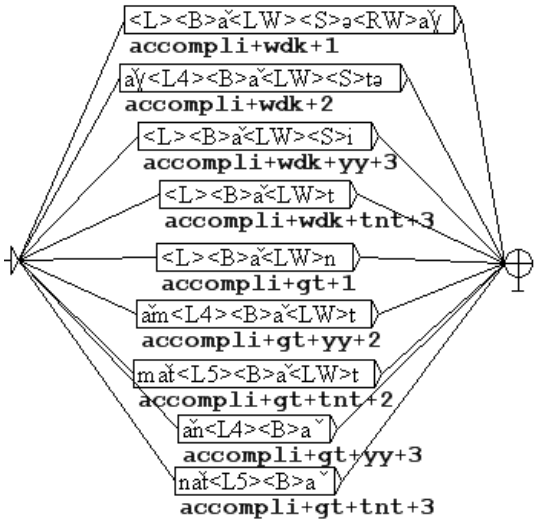

Fig. 2: Verb "əṣṣən", intense accomplished aspect

We obtain, in the inflected dictionary, the correct conjugated forms.

əṣṣănăɣ+əṣṣən,V+accompli+wdk+1

təṣṣănăɣ+əṣṣən,V+accompli+wdk+2

iṣṣăn+əṣṣən,V+accompli+wdk+yy+3

təṣṣăn+əṣṣən,V+accompli+wdk+tnt+3

nəṣṣăn+əṣṣən,V+accompli+gt+1

təṣṣănăm+əṣṣən,V+accompli+gt+yy+2

təṣṣănmăt+əṣṣən,V+accompli+gt+tnt+2

əṣṣănăn+əṣṣən,V+accompli+gt+yy+3

əṣṣănnăt+əṣṣən,V+accompli+gt+tnt+3

e. Irregular words

Finally, the singular and plural forms of 2,457 words were explicitly written in the Nooj dic-

tionary because they do not follow any regular rule.

Examples:

Singular	Plural	Translation
ag-awnaf	kel-awnaf	tourist
amanẓo	imenẓa	young animal
ănaffarešši	inəffərəšša	someboby with bad mood
ănesbehu	inəsbuha	liar
efange	ifangăyan	bank
efajanfăj	ifajanfăɣăn	sling
emagărmăz	imagămăzăn	plant
emazzăle	imazzaletăn	singer
taḍaggalt	tiḍulen	daughter-in-law
tejăṭ	tizḍen	goal (football)

Table 10: Examples of irregular plural forms

f. Result

There are 6,378 entries in the Nooj dictionary. The inflected dictionary, calculated from the above dictionary and with the inflectional and conjugation rules, encounters 11,223 entries.

Nooj is able to use the electronic dictionary we've created to automatically tag a text (see an example in appendix B).

4.3 Future work

a Conversion into XML format

We will convert the inflectional dictionary into the international standard Lexical Markup Framework format (Francopoulo and al., 2006) in order to make it easily usable by other TALN application,.

b Automatic search of rules

Due to the high morphological complexity of the Tamajaq language, we plan to develop a Perl program that would automatically determine the derivational and conjugation rules.

c Completion and correction of the resource

The linguistic resource will be completed during the next months in order to add the class of verbs that are absent for the moment, and also to correct the errors that we noticed during this study.

d Enrichment of the resource

We plan to construct a corpus of school texts to evaluate the out-of-vocabulary rate of this dictionary. This corpus could then be used to enrich the dictionary. The information given by Nooj would be useful to choose the words to add.

Acknowledgement

Special thanks to John Johnson, reviewer of this text.

References

Aghali-Zakara M. 1996. *Éléments de morphosyntaxe touarègue*. Paris : CRB-GETIC, 112 p.

Enguehard C. and Naroua H. 2008. *Evaluation of Virtual Keyboards for West-African Languages*. Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco.

Francopoulo G., George M., Calzolari N., Monachini M., Bel N., Pet M., Soria C. 2006 *Lexical Markup Framework (LMF)*. LREC, Genoa, Italy.

République of Niger. 19 octobre 1999. *Arrêté 214-99* de la République du Niger.

Max Silberztein. 2007. *An Alternative Approach to Tagging*. NLDB 2007: 1-11

APPENDIX A : Tamajaq official alphabet (République of Niger, 1999)

Character	Code	Character	Code
a	U+0061	A	U+0041
â	U+00E1	Â	U+00C2
ă	U+0103	Ă	U+0102
ə	U+01DD	Ǝ	U+018E
b	U+0062	B	U+0042
c	U+0063	C	U+0043
d	U+0064	D	U+0044
ḍ	U+1E0D	Ḍ	U+1E0C
e	U+0065	E	U+0045
ê	U+00EA	Ê	U+00CA
f	U+0066	F	U+0046
g	U+0067	G	U+0047
ǧ	U+01E7	Ǧ	U+01E6
h	U+0068	H	U+0048
i	U+0069	I	U+0049
î	U+00EE	Î	U+00CE
j	U+006A	J	U+004A
ǰ	U+01F0	Ɉ	U+004AU+ 030C
ɣ	U+0263	Ɣ	U+0194
k	U+006B	K	U+004B
l	U+006C	L	U+004C
ḷ	U+1E37	Ḷ	U+1E36
m	U+006D	M	U+004D
n	U+006E	N	U+004E
ŋ	U+014B	Ɲ	U+014A
o	U+006F	O	U+004F
ô	U+00F4	Ô	U+00D4
q	U+0071	Q	U+0051
r	U+0072	R	U+0052
s	U+0073	S	U+0053
ṣ	U+1E63	Ṣ	U+1E62
š	U+0161	Š	U+0160
t	U+0074	T	U+0054
ṭ	U+1E6D	Ṭ	U+1E6C
u	U+0075	U	U+0055
û	U+00FB	Û	U+00DB
w	U+0077	W	U+0057
x	U+0078	X	U+0058
y	U+0079	Y	U+0059
z	U+007A	Z	U+005A
ẓ	U+1E93	Ẓ	U+1E92

APPENDIX B : Nooj tagging Tamajaq text

Nooj perfectly recognizes the four forms of the word "awăqqas" (big cat) in the text:

"awăqqas, iwaɣsan, awaɣsan"

These forms are listed in the inflectional dictionary as:

awăqqas,awăqqas,SN+yy+wdk

awăqqas,awăqqas,SN+yy+wdk+FLX=A1a+sfss

iwaɣsan,awăqqas,SN+yy+iget

awaɣsan,awăqqas,SN+yy+iget+FLX=A1a+sfss

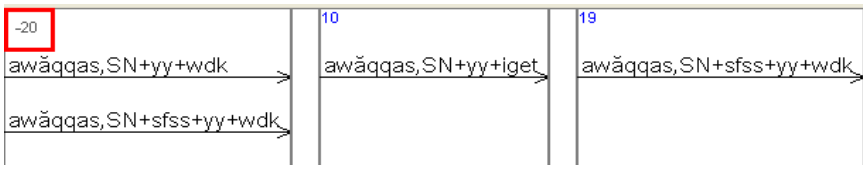

Fig.3: Tags on the text "awăqqas, iwaɣsan, awaɣsan"

On the figure 3, we can see that the first token "awăqqas" gets two tags:
- "awăqqas,SN+yy+wdk" (singular)
- "awăqqas,SN+yy+wdk+sfss" (singular and annexation state).

The second and third tokens get a unique tag because there is no ambiguity.