Towards an electronic dictionary of Tamajaq language in Niger
Chantal Enguehard, Issouf Modi

To cite this version:
Chantal Enguehard, Issouf Modi. 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., Mar 2009, Athènes, Greece. publication électronique, 2009. <halshs-00409455>

HAL Id: halshs-00409455
https://halshs.archives-ouvertes.fr/halshs-00409455
Submitted on 7 Aug 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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.
République du Niger
modyissouf@yahoo.fr

Abstract

We present the Tamajaq language and the dictionary we used as the 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 experimental schools: Fulfulde, Hausa, Kanuri, Tamajaq and Sonjaj-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, 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      |
|                 | Fricative | s      |
|                 | Nasal     | n      |
|                 | Lateral   | l      |
| Pharyngeal       | Plosive   | t      |
|                 | Fricative | š      |
|                 | Lateral   | l      |
| Palatal          | Plosive   | c      |
**Consonants**

| Voiceless | Voiced |
|-----------|--------|
| Fricative | š j    |
| Semivowel | y      |

| Velar     |
|-----------|
| Plosive   | k g ģ |
| Fricative | γ x    |

| Nasal     |
|-----------|
| Fricative | h      |

| Glottal   |
|-----------|
| Plosive   | q      |

Table 1a: Articulatory phonetics of Tamajaq consonants

**Vowels**

| Close | Close-mid | Open-mid | Open |
|-------|-----------|----------|------|
| Palatal | i e       |
| Central | ā ā 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 SOUTEBA¹ project of the DED² 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:

« ābada1: sn. bas ventre. Daw ābada. Bārā wa yallūzan ad t-yallūzā ābada-net. tamust.: yy. īget: ibadan. »

« ābada2: sn. flanc. Tasāga meywā dāy ābada n ooyān. Imewwazā qālān dāy ābada n okašwar. Anammel.: azador. tamust.: yy. Esoso.: īget: 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 |
| adəkud | number | dkd. | 3 |
| anaalkam | determinant | nlkm. | 1 |
| anamal | verb | nml. | 1450 |
| samal | adjective | sml. | 48 |
| asmmadāγ an təla | possessive pronoun | smmdyt. | 5 |
| isən | noun | sn. | 3648 |
| isən n anamal | Verbal noun | snnml. | 33 |
| isən an təɣərit | name of shout | sntryt. | 2 |
| isən xalalan | proper noun | snxln. | 29 |
| isən izzəwen | complex noun | snzwn. | 137 |
| ēstakar | adverb | stkr. | 8 |

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

| øsatkar n ådag | adverb of location | stkrdg. | 10 |
|----------------|--------------------|---------|----|
| øsatkar n igət | Adverb of quantity  | stkrgt. | 1  |
| təyərit        | onomatopoemia       | tyrt.   | 8  |
| tənalkamt      | particle            | tnlktm. | 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:
- øktəb (to write): triliteral verb, type 1.
- øşşən (to know): triliteral verb, type 2 (şşn).
- øməl (to say): biliteral verb, type 1
- akər (to steal): biliteral verb, type 2
- awəy (to carry): biliteral verb, type 3
- åswu (to drink): biliteral verb, type 4
- åru (to love): monoliteral verb, type 2
- åru (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 |
|---------------|--------------|
| tamust        | gender       | tmt.      |
| yey           | masculine    | yy.       |
| tante         | feminine     | tnt.      |
| awdəkki       | singular     | wdk.      |
| igət          | plural       | gt.       |
| øsəfsəs       | annexation state | sfss. |

Table 3: Tamajaq terminology for nouns

Example:
- "øtrəkkə: sn. morceau de sucre. Akku: abløy n°2. tamust.: yy. Øsəfsəs.: ø. Igət: øtrəkkətän. »

"øtrəkka" is a masculine noun. Its plural is "øtrəkkətän". It becomes "øtrəkka" 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
  - "ejəd-n-øjədän", literally means "donkey of birds" (this is the name of a bird).
- Verb + noun
  - "awəy-əhud" literally means "it follows harmattan" (kite).
  - "gazzäy-əfuk" literally means "it looks at sun" (sunflower).
- Preposition + noun
  - "In-təmət" means "the one of the tree acacia" (of acacia).

Verb + verb

3
"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` |

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:

The 'A2o' rule replaces the second letter of the word by 'o'.

| 'A2o' rule |
|---|
| Nooj: `A2=<LW><R><S>/sfss ö/ıget` |
| Perl: `^(.*)\.(.*)$ → $1ı$2` |

| 'I4' rule deletes the last letter, adds "-ān" at the end and "-i-" at the beginning. |
|---|
| Nooj: `I4=än<LW><S>/ıget` |
| Perl: `^(.*)s → ı$1än` |
| # 446 words |

| '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>/ıget` |
| Perl: `^(.).(.*).$ → $1ı$2en` |
| # 144 words |

| 'I45' rule deletes the final letter and include "-en" at the end. |
|---|
| Nooj: `I45=<B>en/ıget` |
| Perl: `^(.*)s → $1en` |
| # 78 words |
Enguehard, C., Modi I. Towards an electronic dictionary of Tamajeq 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.

'II02' rule deletes the two last letters and the second one and includes a final "-a" and a "-" in the second position.

| Nooj   | Perl       | Count |
|--------|------------|-------|
| I102=<B2>a<LW><R><S>i/Iget | ^(\.).(\.*).S → $1i$2a | 6 words |

Table 9: Rule 'II02'

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:
I2RA2ā =
:Rwdk + :I2 + :Rwdk :A2ā + :I2 :A2ā

Fig. 1: Rule I2RA2ā

This rule recognizes the singular form (:Rwdk), the plural form (:I2), the singular form with the annexation state (:Rwdk :A2ā) and the plural form with the annexation state (:I2 :A2ā).

25 words meet this rule.

For instance, "tdlmt" (accusation, provocation), is inflected in:
- tdlmt,tdlmt,SN+tnt+wdk
- tdlen,tdlmt,SN+tnt+Iget
- tdlen,tdlmt,SN+tnt+Iget+sfss
- tdlmt,tdlmt,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.

Verb "əṣṣən", intense accomplished aspect

We obtain, in the inflected dictionary, the correct conjugated forms.
- əṣṣən+əṣṣən,V+accompli+wdk+1
- əṣṣən+əṣṣən,V+accompli+wdk+2
- əṣṣən+əṣṣən,V+accompli+wdk+yy+3
- əṣṣən+əṣṣən,V+accompli+wdk+tnt+3
- əṣṣən+əṣṣən,V+accompli+gt+1
- əṣṣən+əṣṣən,V+accompli+gt+yy+2
- əṣṣən+əṣṣən,V+accompli+gt+tnt+2
- əṣṣən+əṣṣən,V+accompli+gt+yy+3
- əṣṣən+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           |
| amanzo         | menza        | young animal      |
| ánaffarešši    | inaffarešša  | somebody with bad mood |
| ánesbehu       | inasbuha     | liar              |
| efange         | ifangâyan    | bank              |
| efajanfâj      | ifajanfâyan  | sling             |
| emâgâmâz       | imagâmâzan   | plant             |
| emazzâle       | imazzâletân  | singer            |
| tadaggalt      | tidulien     | daughter-in-law   |
| tejâṭ           | tizâden      | goal (football)   |

Table 10: Examples of irregular plural forms

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 |
| ä         | U+0103 | Ä         | U+0102 |
| ø         | U+01DD | Ø         | U+018E |
| b         | U+0062 | B         | U+0042 |
| c         | U+0063 | C         | U+0043 |
| d         | U+0064 | D         | U+0044 |
| d         | U+1E0D | D         | 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 |
| ġ         | U+01E7 | Ġ         | U+01E6 |
| h         | U+0068 | H         | U+0048 |
| i         | U+0069 | I         | U+0049 |
| î         | U+00EE | Í         | U+00CE |
| j         | U+006A | J         | U+004A |
| ğ         | U+01F0 | Ģ         | U+004AU+030C |
| k         | U+006B | K         | U+004B |
| l         | U+006C | L         | U+004C |
| l         | U+1E37 | Ł         | U+1E36 |
| m         | U+006D | M         | U+004D |
| n         | U+006E | N         | U+004E |
| ñ         | U+014B | Đ         | U+014A |
| o         | U+006F | O         | U+004F |
| ô         | U+00F4 | Ō         | U+00D4 |
| q         | U+0071 | Q         | U+0051 |
| r         | U+0072 | R         | U+0052 |
| s         | U+0073 | S         | U+0053 |
| š         | U+1E63 | Š         | U+1E62 |
| š̃         | U+0161 | Š         | U+0160 |
| t         | U+0074 | T         | U+0054 |
| ẗ         | U+1E6D | Ṭ         | U+1E6C |
| u         | U+0075 | U         | U+0055 |
| ū         | U+00FB | Ū         | 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 |
| z̃         | U+1E93 | Z̃        | U+1E92 |
APPENDIX B : Nooj tagging Tamajaq text

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

"awâqqas, iwaɣsan, awâyyan" 

These forms are listed in the inflectional dictionary as:

awâqqas,awâqqas,SN+yy+wdk
awâqqas,awâqqas,SN+yy+wdk+FLX=A1a+sfss
iwaɣsan,awâqqas,SN+yy+iget
awâyyan,awâqqas,SN+yy+iget+FLX=A1a+sfss

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

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