Iles d Imesli, 9, pp.203-211

Traitements automatiques de manuels scolaires en vue d'applications dictionnairiques

Mohand Akli SALHI[1], Patrice POGNAN[2]
1, Département de Langue et Culture Amazighes,
1, Université Mouloud Mammeri de Tizi-Ouzou.
2, Institut National des Langues et Civilisations Orientales, Paris, France
2, Université Paris Sorbonne, France.

Résumé en tamaziɣt
Ad d-yawi uḍris-agi ɣef yiwen usenfar n usewjed n yisegzawalen s ttawilat n usemkel aselkim. Deg usenfar-agi, nra ad d-nesken amek ara yuɣal wayen yellan d aḍris d amawal ilmend n usbeddi n usegzawal aɣurbiz. Deg-s ad d-nemmel tarrayt i neḍfer.

Abstract

The present text presents a project of realization of dictionary tools from raw corpora using automatic processing. It is precisely a presentation of the methodology of this ongoing project consisting in the production of dictionary documents based on the transformation of the lexical mass of textbooks into dictionary entries.

Keywords: dictionary, automatic processing, dictionaries, textbook, data transformation

Le présent texte expose un projet de réalisation d'outils dictionnairiques à partir de corpus bruts à l'aide de traitements automatiques. Il s'agit précisément d'un exposé de la méthodologie élaborée pour ce projet en cours consistant à réaliser des documents dictionnairiques à partir de la transformation de la masse lexicale des manuels scolaires en entrées de dictionnaires.
Seront décrites ici les grandes étapes prévue dans cette opération de transformation / préparation des données lexicales. Seront donnés avant l'exposé de ces étapes l'utilité de dictionnaires scolaires dans le cas de la langue berbère et les types de lexique contenus dans les manuels scolaires.

Les manuels et l'utilité d'application dictionnairique

Le manuel scolaire est conçu comme une concrétisation d'un programme d'enseignement. Ce dernier s'élabore, dans le cas de la langue berbère en

Algérie, en trois phases : le primaire avec deux années d'enseignement (à partir de la quatrième année de la scolarité de l'élève), le moyen sur quatre années et le secondaire sur trois années. Proposés à partir de 2003, les manuels du berbère sont rédigés dans la variante kabyle et, de facto, sont destinés à des apprenants natifs ayant une pratique solide de l'oral.
Conçu suivant une approche et une organisation précises (approche par compétence avec une organisation en séquences), l'ensemble des manuels proposent des textes, des exercices de compréhension, des exercices structuraux liés à l'explicitation de la grammaire, du lexique et de l'orthographe. Ils proposent également des connaissances diverses (petites présentations biographiques, informations historiques, géographiques, littéraires et textuelles, etc.) relatives essentiellement à l'environnement culturel et social de l'apprenant.
En l'absence presque totale de documents de référence linguistique (dictionnaire unilingue par exemple), un accompagnement dictionnairique de l'enseignement de la langue berbère doit être proposé afin d'assurer à l'apprenant de cette langue une meilleure prise en charge pédagogique et didactique.
Variée, la masse lexicale des manuels est progressive et extensible dans le sens où elle est augmentée d'année en année dans chaque palier d'enseignement.
Cette masse lexicale est composée de quatre types :

- Lexique de la langue quotidienne : essentiellement donné dans les textes support de lecture
- Lexique de l'univers scolaire : qui concerne les objets de la vie et des pratiques scolaires
- Lexique terminologique : il se réfère notamment aux métalangages linguistique et littéraire
- Lexique de la vie moderne : présent notamment dans les textes descriptifs et argumentatifs

Dans le langage de la bibliothéconomie, discipline qui s'occupe du traitement et de la gestion de l'information documentaire, un dictionnaire est considéré comme un usuel, c'est-à-dire qu'il est un document qu'on utilise à chaque fois qu'on est dans des moments de difficultés (difficulté d'accéder au sens par exemple). En règle générale, on ne lit pas un dictionnaire, on le consulte dans le but d'y avoir une information qui nous permettra d'avancer dans notre lecture / compréhension du document de base. Le dictionnaire est donc un document d'accompagnement.
Outre la première utilité d'un dictionnaire (générale de langue et / ou de spécialité), un dictionnaire scolaire en berbère apportera plusieurs solutions à l'apprenant de cette langue. En effet, il peut jouer le rôle de gestionnaire de

la variation, qu'elle soit phonétique, morphologique, sémantique ou syntaxique. Le pointage et le signalement systématiques de cette variation constitueront un moyen de gestion de cette dernière. Le lieu de cette gestion est la notice lexicographique ou dans les différents indexes possibles. Les données de la variation peuvent être dans la caractérisation de l'entrée, dans la stratification des significations, dans les mises en contexte des unités lexicales, etc.
Par ailleurs, étant le fait que le lexique langue berbère est en pleine évolution rapide conséquemment aux changements sociologiques et linguistiques (émigration, urbanisation, scolarisation, etc.), le dictionnaire scolaire aura la tâche d'accompagner l'élève dans l'acquisition de nouvelles unités lexicales (néologismes et archaïsmes). La mise en contexte culturelle et phraséologique (proverbe, chanson, exemples) du mot néologique, du localisme ou du mot archaïque facilitera cette acquisition.
Le dictionnaire aura aussi l'ambition d'élargir et d'appuyer la connaissance de l'univers aspectuel du kabyle par la mise à disposition de l'élève d'un tableau de conjugaison de tous les verbes des manuels (classés par ordre alphabétique et morphologique).
Et enfin, il apportera à l'apprenant de nouvelles connaissances lui permettant d'approfondir et d'élargir sa culture générale en lui proposant des présentations biographiques, des indications de géographie, des informations historiques et de toute autre connaissance présentée brièvement dans les manuels scolaires.

Procédure et Méthodologie

Les transformations de la masse lexicale des manuels scolaires en entrées de dictionnaires nécessitent plusieurs actions. Ces dernières sont présentées ici en six étapes.

Etape 01 : unification du code (ortho) graphique

Signalons de prime abord que le traitement de la typographie et de la ponctuation peut être précédé d'un programme qui assure le codage en Unicode UTF-8 des textes saisis avec d'autres codes, notamment issus de polices de caractères utilisées avant la généralisation de l'Unicode. Le format choisi pour l'ensemble des traitements est « .txt », c'est-à-dire ce que l'on appelle aussi « texte brut ». Si nous obtenons les textes en Word, il est

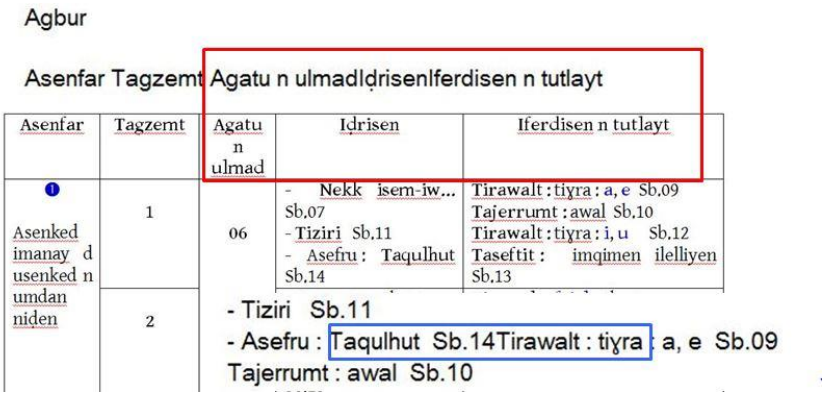
très facile de les transformer en fichier brut grâce au transcodeur de Word d'une remarquable efficacité et que l'on met en action tout simplement en faisant « enregistrer sous… », « texte brut » - « enregistrer » - puis on fait le choix de l'Unicode UTF8.
L'objectif de cette première étape est d'éviter les ambiguïtés et les confusions dues à l'utilisation de plusieurs polices de caractères dans la notation des manuels scolaires.

Etape 02 : Préparation formelle et Nettoyage du texte de base

Nos programmes sont prévus, de manière plus ou moins implicite, pour analyser des textes en continu, qu'ils soient littéraires, techniques, scientifiques ou médicaux. Les manuels scolaires, et de manière spécifique le tout premier, offre un ensemble didactique destiné à enseigner le kabyle à des enfants kabylophones, ensemble dans lequel la proportion de « texte » augmente d'année en année.
Il existe donc tout un appareil qui apporte des connaissances ou contrôle leur acquisition. La partie apprentissage du vocabulaire y est prioritaire, mais accompagnée d'autres éléments d'apprentissage tels que la présentation de l'alphabet ou bien de divers contrôles, souvent de nature ludique.
Nos études portant sur le lexique, une partie importante du contenu de ces manuels fait interférence avec certains éléments que nous ne pourrions pas déterminer automatiquement comme des non-mots sans une analyse linguistique complète de la langue.
La présentation de l'alphabet apporte des éléments qu'il convient d'écarter, p. ex. les lettres en arabe…

Asekkil	Isem-is	Tifinaɣ	Azal-is s taɛrabt	Amedya
a	A / Aɣra	a	ا	Aman
b	Ba	b	ب	Baba, bibb
c	Ca	c	ش	Amcic
č	Yečč	Þ	تش	Ameččim
d	Da	d	د	Dadda, udi

Il en va de même avec les tables des matières présentées sous forme de tableaux où l'on voit que la suppression des colonnes provoque des agglutinations indésirables :


Agbur

Asenfar Tagzemt Agatu n ulmadIḍrisenIferdisen n tutlayt

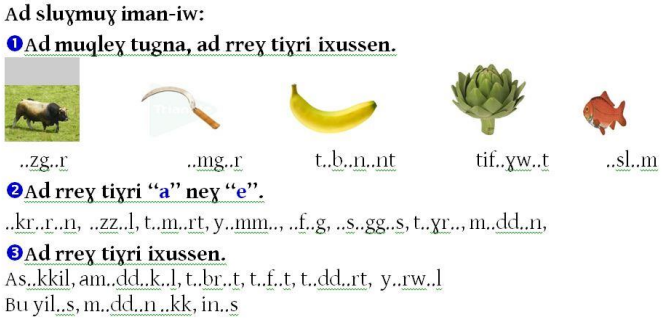
Asenfar	Tagzemt	Agatu n ulmad	Iḍrisen	Iferdisen n tutlayt
❶ Asenked imanay d usenked n umdan niden	1	06	- Nekk isem-iw... Sb.07 - Tiziri Sb.11 - Asefru : Taqulhut Sb.14	Tirawalt : tiɣra : a, e Sb.09 Tajerrumt : awal Sb.10 Tirawalt : tiɣra : i, u Sb.12 Taseftit : imqimen ilelliyen Sb.13
	2			

- Tiziri Sb.11
- Asefru : Taqulhut Sb.14Tirawalt : tiɣra : a, e Sb.09
Tajerrumt : awal Sb.10

Il convient également d'écarter du fichier brut les équivalents français et arabes des mots kabyles du lexique de mots placé en fin d'ouvrage :

116	Tizdit (MZ)	Trait d'union	مطّة
117	Tugna (MW)	Image	صورة شمسيّة
118	Tunẓirt(MW)	Énigme	لغز
119	Tussda (MW)	Tension / redoublement	مضعّف
120	Udad (TG)	Mouflon	ارويّة
121	Udem(MW)	Personne (grammaire)	ضمير (منفصل)

Le manuel de première année du primaire par exemple contient des éléments bien plus pernicieux dont un très bel exemple constitué par les exercices à trou dans lesquels les « trous » représentés par deux points « .. » créent à partir de chaque mot un à plusieurs autres mots n'ayant aucune existence et indétectables automatiquement :


Ad sluɣmuɣ iman-iw:

❶Ad muqleɣ tugna, ad rreɣ tiɣri ixussen.

..zg..r ..mg..r t..b..n..nt tif..ɣw..t ..sl..m

❷Ad rreɣ tiɣri "a" neɣ "e".

..kr..r..n, ..zz..l, t..m..rt, y..mm.., ..f..g, ..s..gg..s, t..ɣr.., m..dd..n,

❸Ad rreɣ tiɣri ixussen.

As..kkil, am..dd..k..l, t..br..t, t..f..t, t..dd..rt, y..rw..l

Bu yil..s, m..dd..n ..kk, in..s

Cette étape doit déboucher sur le constat que les textes bruts de chaque manuel est :

a. débarrassé de tous les éléments qui ne sont pas de nature textuelle et lexicale (illustrations, tableaux, exercices à trous, etc.).
b. débarrassé des éléments autres que kabyles (mots en français et en arabe, références des textes).

Par ailleurs, dans le processus du nettoyage formel, des opérations de transformations des données sont requises comme celle, par exemple, qui consiste à répondre aux questions des exercices à trous.

Etape 03 : Premiers traitements automatiques

La chaine de traitement automatique traite les problèmes de typographie et de ponctuation qui correspondent au plus grand nombre possible de langues écrites à l'aide d'alphabets. Elle possède simultanément 5 sorties spécialisées : 3 pour l'analyse automatique (présentation du texte avec différents types de segmentation), 1 pour la fabrication de concordanciers et 1 pour les études sur le lexique. C'est cette dernière sortie qui présente le texte mot par mot à la verticale que nous avons choisie pour nos travaux sur les manuels scolaires.

L'objectif recherché dans cette étape est double. D'un côté, nous procédons à l'extraction des formes de mots (suivant les morphologies graphiques) et de l'autre, nous entamons l'organisation des premières données obtenues comme le tri, le tassement et le classement alphabétique des formes de mots.

Etape 04 : Traitement automatiques ciblés basée sur une approche linguistique

Nous avons recours à l'informatique pour dégager l'ensemble de vocabulaire pour chacun des manuels, puis un ensemble de manuels d'un niveau donné, par exemple manuels du primaire.

Les premiers traitements réalisés, après avoir trié et tassé le fichier issu de la chaîne de préparation du texte (typographie et ponctuation), se sont focalisés sur le traitement de l'affixation. Nous sommes rapidement arrivés à un ensemble de 3 programmes :

a- un pour la préfixation qui, à de rares exceptions (des termes tels que « mazal » qualifié par Naït-Zerrad « d'autre élément prédicatif ») se révèle être de nature exclusivement verbale où les formes aoristiques prédominent. Nous y trouvons un enchaînement de préfixes dans l'ordre : pronom personnel d'attribution (datif) - pronom personnel objet (accusatif) - particule de direction, chacun de ces trois éléments pouvant être facultatif.

b- un pour la suffixation nominale où nous trouvons un ensemble de suffixes démonstratifs, les suffixes pronoms personnels liés et ceux spécialisés pour les termes de parenté.

c- un pour la suffixation verbale où nous retrouvons dans le même ordre les préfixes verbaux ayant ici un statut de suffixes. Une majorité des verbes concernés sont au prétérit.

Cette étape est déterminante dans la mesure où, en éliminant les petites unités graphiques telles que les prépositions (une à deux syllabes), les particules (une seule lettre ou une seule syllabe), nous pourrons, moyennant une analyse, procéder à la classification morphosyntaxique des unités dégagées.

Par ailleurs, la reconnaissance automatique et la transformation des deux états de la forme nominale sont en cours d'étude.

Etape 05 : Définition dans la mesure du possible des formes susceptibles de constituer des entrées de dictionnaire

Cette étape est dédiée aux différentes préparations des données en vue de leur insertion dans la nomenclature des unités lexicales des différentes applications dictionnairiques. La tâche consiste notamment à convertir ces unités en entrées de dictionnaire. A cet effet, nous devons transformer, suivant la faisabilité, le nom féminin en masculin, le pluriel en singulier et ramener les formes conjuguées des verbes à une forme neutre qui puisse servir à la fois comme premier repère de la racine et comme entrée. Un certain nombre de racines verbales sont reconnues automatiquement. L'analyse du verbe n'est pas totalement terminée, mais nous avons bon espoir de l'automatiser ainsi que les catégories nominales moyennant une analyse fine des correspondances morphologiques.

Etape 06 : Confection du dictionnaire proprement dit

Dernière étape de préparation des données mais néanmoins la plus cruciale car elle nécessite aussi bien du temps que de l'effort. La nature du travail à réaliser à ce niveau concerne au moins les points suivants :

a. La définition de la nomenclature des entrées
b. L'organisation de la notice lexicographique
c. La gestion de l'information dans la notice lexicographique
d. Les types de définition

e. La préparation des annexes et des index (racines, formes attestées dans le manuel mises en relation avec sa transformation morphologique, équivalents linguistiques, etc.)
f. Définition d'un tableau de conjugaison
g. Réseau de relations entre les informations de la notice lexicographique et les index et les annexes.
h. Augmentation de la masse lexicale par des informations complémentaires.

En guise de conclusion à cet exposé, nous souhaitons mentionner que ce projet s'inscrit dans le cadre de l'accord programme algéro-français 14 MDU 925.

Références bibliographiques

Ameur, Meftaha & al., 2014: *Initiation à la langue amazighe*, Rabat, IRCAM.

Chaker, Salem, 1995: *Linguistique berbère. Etudes de syntaxe et de diachronie*, Paris/Louvain, Peeters.

Naït-Zerrad, Kamal, 1994: *Manuel de conjugaison kabyle*, L'Harmattan, Paris. En ligne à l'adresse : https://www.amyag.com/

Naït-Zerrad, Kamal, 1995: « *tajeṛṛumt n tmaziɣt tamirant (taqbaylit)* » « *grammaire du berbère contemporain (kabyle)* », Alger, ENAG.

Naït-Zerrad, Kamal, 2011: *Mémento grammatical et orthographique de berbère. Kabyle – chleuh – rifain.* L'Harmattan, Paris.

Pognan Patrice., Salhi Mohand Akli, Taïfi Miloud, 2012 : Du corpus aux applications, une chaîne de préparation automatique des textes berbères. *7ème Bayreuth-Frankfurt-Leidener Kolloquium zur Berberologie*, Frankfurt.

Pognan Patrice., Salhi Mohand Akli, 2016: Du texte aux données analysables: Lwali n udrar comme corpus. *Iles d Imesli*, n° 8, pp. 173-185. Tizi-Ouzou.

Sabri Malika, Ibri Saliha, 2012 : De la néologie dans les manuels de Tamazight : nécessité d'un dictionnaire scolaire, *Timsal n Tamazight*. Volume 3, Numéro 1, Pages 26-49

Salhi, Mohand Akli,A., Pognan, Patrice, 2013: Pour une indexation raisonnée des corpus littéraires kabyles, 2ème rencontre internationale « La linguistique du corpus : recueil, annotation, exploitation et diffusion », Tizi-Ouzou, La linguistique de corpus : recueil, annotation, exploitation et diffusion, *Iles d Imesli* n° 05.

Miloud Taifi, Pognan, Patrice, 2009: Pour une lexicographie berbère unilingue. Problèmes théoriques et méthodologiques, *Berber Studies* n° 25, 2009.