LES TICS ET LES PROBLEMES DE LEXICOGRAPHIE AMAZIGHE : LE CAS DU DICTIONNAIRE FONDAMENTAL DU KABYLE[1]

Résumé : *La langue Amazighe est exposée aujourd'hui aux exigences imposées par les TICs, sachant bien que ces technologies ne lui sont pas destinées exclusivement. Les logiciels gratuits qui circulent sur le Web et autres outils informatiques telles que les bases de données dont bénéficient actuellement les langues dites de civilisation, ne lui sont pas non plus d'accès facile. Se tailler une place sur et dans le Web n'est pas l'apanage des langues et cultures marginales, car au demeurant elles restent toujours à la périphérie des événements technologiques à l'image du TAL.*

Mots clés : *linguistique, dictionnaire, corpus, Web, TICs.*

Abstract: *Nowadays Amazigh language is exposed to the requirements imposed by ICTs, even though these technologies are not intended exclusively to it. Free software which circulates on the Web and other computer tools such as databases currently enjoyed by the so-called civilisation languages are also easily accessible. Securing an enviable position on the Web is not for marginal cultures and languages which remain still behind technology phenomena such as NLP.*

Keywords*: language, dictionary, corpus, web, ICT.*

0. Introduction

Depuis toujours, les langues des minorités nord africaines ont vécu dans l'ombre de l'oralité. Notre objectif dans la rédaction de cet article est donc d'essayer de combler un manque important qui est celui de doter la langue kabyle d'un outil lexicographique de base : Le dictionnaire fondamental. Cette contribution, s'inscrit dans le cadre de la linguistique du corpus, même si ce travail suit un chemin inhabituel dans la réalisation de ce projet lexicographique. En effet, il a pour but la préparation d'un dictionnaire fondamental du Kabyle à travers les ressources de l'Internet.

Cet outil technologique de pointe offre une multitude de solutions à court et à moyen terme. C'est le meilleur moyen d'accéder à un ensemble de documents rédigés en et dans la langue kabyle. Mais le chemin reste semé d'embûches et nous nous demandons comment on pourrait s'y prendre pour pouvoir construire une base de données pertinente à partir de laquelle pourrait naître cette entreprise. Cette contribution vise à répondre à une demande sociale assez pressante afin de satisfaire les exigences du terrain, surtout celles qui sont relative à l'enseignement.

[1] Rachid **Adjaout,** Université A. Mira de Bejaia, Algérie
adjaoutrachid@yahoo.fr

Afin de parvenir à cerner le sujet, un bon nombre d'interrogations retiennent notre attention à ce sujet, à savoir la manière d'opérer le choix des bons textes dans un domaine aussi étendu que celui du Web. Quelle est la démarche la plus prometteuse pour une bonne récolte de textes rédigés en kabyle ? Que faire de la diversité des textes et de leurs polices de transcriptions une fois les données recueillies ? Et quels sont les moyens techniques qui permettent de constituer une base de données lexicale ?

Pour pouvoir apporter quelques éléments de réponses aux questions posées, à défaut de moyens de mener à terme ce type de recherche en Algérie, nous l'avons réalisé en partie, grâce à un stage effectué à l'Université de Lausanne en 2008 - 2009.

Mais avant d'exposer les données relatives à la réalisation de ce projet, il est utile de souligner un certain nombre de problèmes relatifs à l'adaptation de la langue tamazight aux différents TICs. Il faut rappeler aussi que peu de travaux existent dans le domaine traitant de cette thématique, du moins en ce qui concerne le tamazight. Cependant, les tentatives de proposer des solutions à cette problématique ne sont pas nombreuses du fait que la question d'aménagement n'est pas toujours la priorité de bon nombre de chercheurs en et dans le tamazight.

1. Les TICs, support de l'aménagement linguistique pour le tamazight

Investir dans le domaine des technologies de l'information et de la communication pour le compte de la langue amazighe demande aujourd'hui de la formation et des moyens à moyen et à longs termes. L'emploi de ces technologies comme assistant du processus d'aménagement linguistique devient une nécessité pour différentes raisons, soulignait Nait Zerrad K. (2010 :1) que

« - L'absence des ressources linguistiques fiables et en particulier de dictionnaires ou lexiques ; La diffusion de propositions de normalisation linguistique, faites par des institutions pour combler l'absence de norme instituée à l'image des recommandations faites par l'INALCO en 1996;

La dispersion des différents acteurs intervenant sur la standardisation : chercheurs, universitaires, auteurs, enseignants et étudiants ».

Se mettre à jour est plus que nécessaire, mais il faudrait d'abord faire l'inventaire des besoins en la matière, ensuite tenter de tracer les objectifs que pourraient atteindre les artisans usagers des TICs. Pour cela, l'auteur ajoute en disant qu' « il est nécessaire d'utiliser un format de données assurant leur exploitation par l'utilisateur quel qu'il soit (compatibilité machines et systèmes). Le programme informatique ne peut cependant être viable que si certaines conditions sont réunies, en particulier l'existence d'une norme et une police de caractères adéquate pour écrire la langue » (Ibidem).

Les obstacles qui pourraient se dresser devant la langue amazighe dans l'utilisation des TICs, sont multiples car au demeurant rien n'a été inscrit comme tel dans les programmes des états concernés par la prise en charge telle que la

standardisation de tamazight du moins en ce qui concerne l'Algérie. Pourtant, tous les gouvernements tentent d'offrir à leurs citoyens la meilleure éducation possible, soulignait un rapport publié par les instances (Pelgrum & Law, 2004 :6) de l'UNISCO en 2004. D'ailleurs, c'est ce qui a été préconisé par cette institution dans son annexe A en insistant sur le concept de l'alphabétisation aux TICs.

Ainsi donc, le constat que nous pouvons établir aujourd'hui est accablant du fait que les états maghrébins et surtout l'Algérie n'ont pas su faire profiter à leur population des technologies de l'information et de la communication et cela malgré les moyens mis à leur disposition. Quant à l'issue réservée au tamazight, ceci, est une autre question, car l'absence d'initiative s'inscrivant dans cette optique explique la politique de marginalisation que subissent les minorités amazighes. L'entreprise que nous avons tenté de réaliser dépendait des moyens que pouvait offrir l'état à cette langue fraîchement reconnue.

Afin de mettre sur pied l'idée de construire un dictionnaire fondamental du Tamazight (Kabyle), nous avons été amenés à éplucher les meilleurs travaux réalisés en Europe à l'image (*du Basic English de Charles Kay Ogden et du Français fondamental de G. Gougenheim*). La démarche suivie dans ce travail s'inscrit dans une autre dynamique que celle qui caractérisait les dictionnaires et / ou lexiques produits jadis par rapport aux TICS d'aujourd'hui. Premièrement, nous avons constitué notre corpus à partir du Web. Deuxièmement, nous avons pu le trier grâce à des logiciels de statistique et enfin, nous comptons le rédiger dans un autre outil relevant aussi des TICs.

2. L'idée du vocabulaire simplifié en Europe

Avant de décrire ce nouvel aspect dans la collecte de données via le Web pour la langue kabyle, nous voudrions souligner l'expérience des linguistes Européens qui ont déjà élaboré des dictionnaires fondamentaux, à l'exemple du dictionnaire fondamental du Français réalisé par G. Gougenheim (1958) et celui du *Basic English* réalisé par Odgen (1944). Ces deux ouvrages ont été conçus selon le principe qui repose sur la notion de limitation du vocabulaire et de la grammaire, même si leur démarche était différente. L'idée de restreindre le vocabulaire d'une langue est liée au fait que les locuteurs d'une langue n'utilisent qu'une partie de ce vocabulaire, et cela malgré l'immensité et la diversité de son lexique.

Pour assurer donc la diffusion d'une langue, la contrainte de limiter cet objet complexe qui est le lexique était de ne retenir que l'essentiel. Aussi, l'idée d'une telle démarche a-t-elle été conçue grâce par exemple à des enquêtes de terrain sur le niveau des élèves qui obtenaient des résultats non satisfaisants en termes d'apprentissage lors de leur passage au primaire, exception faite pour les plus intelligents. À partir de ce constat, la conception d'un vocabulaire fondamental a retenu l'attention d'un bon nombre de spécialistes en la matière.

L'essentiel dans une langue, selon la pédagogie traditionnelle, c'est de ne donner lors de l'enseignement des premières années que le vocabulaire le plus

utile. Néanmoins, comment extraire ce type de vocabulaire dans l'extrême diversité du lexique ? Et dans quelle mesure, ceci est-il possible ?

Les spécialistes qui ont amorcé une telle entreprise disposaient d'une démarche assez originale et qui se résume à faire d'abord l'inventaire de tous les écrits réalisés dans la langue et d'établir des enquêtes nécessaires sur la langue orale. Ce recensement du vocabulaire à travers la langue écrite et la langue orale, permet de constituer une base de données de laquelle sera extrait ce qui est utile grâce à la méthode statistique pour les fréquences et à la disponibilité des mots dans l'usage de la langue. C'est dans cette perspective que nous avons orienté ce travail ambitieux sur l'élaboration d'un vocabulaire fondamental du Kabyle. Par conséquent, et pour des raisons d'ordre pratiques étroitement liées au statut de la langue, nous privilégions les sites Web pour la collecte de notre corpus en plus d'une enquête de terrain[1].

Aujourd'hui, beaucoup de travaux exploitent le domaine du Web pour des finalités d'acquisition de données linguistiques comme la constitution de corpus. La plupart des recherches qui se font dans cette perspective sont celles qui reposent sur des corpus. Ces corpus peuvent être de plusieurs types, comme le soulignent F. Duclaye et Al. (2006 :53) : « Construire un corpus qui réponde à des besoins précis en matière d'apprentissage est une étape, certes, déterminante pour la réussite de l'apprentissage, mais très longue [...] C'est pourquoi le Web est devenu une ressource privilégiée, utilisée depuis une dizaine d'années pour en extraire avec succès parfois très variable, tous types de contenus ».

Le travail que nous proposons de décrire dans cet article a comme objectif principal de montrer le procédé avec lequel peut s'élaborer un dictionnaire fondamental à travers des sources générées par le Web. Il se définit comme étant l'une des premières réflexions originales dans la réalisation d'un outil jusqu'ici non conçu pour la langue kabyle. Notre idée part d'un principe qui est celui de collecter un maximum de matériau via l'instrument de l'Internet en passant par les différentes rubriques animées par des internautes kabyles. Ces derniers collaborent par leur création dans différents blogs et expriment des idées en échangeant leur point de vue sur divers sujets (romans, nouvelles, pièces de théâtre, discussions, poésie, chansons, etc.). Cette expérience inédite dans le domaine sera source d'inspiration pour rédiger un tel article.

3. La démarche entreprise dans le moissonnage

La perspective consiste à surfer dans le Web via un moteur de recherche par exemple Google et se mettre à la recherche de tous les sites Internet susceptibles de nous offrir dans ses rubriques des textes rédigés dans le Kabyle selon la graphie

[1] L'enquête de terrain a été réalisée pendant les mois de mars et avril 2009 auprès des étudiants de licence en tamazɨɣt au Département de Langue et de Culture Amazɨɣ de l'Université de Bejaia, Algérie.

latine. Une fois le site trouvé, il faut d'abord juger de sa recevabilité. En effet, d'une part quand un texte se présente sous forme d'image numérisée[1], celui-ci pourrait constituer une contrainte à sa reconversion dans le format word pour qu'il soit traitable au même titre qu'un texte rédigé initialement dans Word. D'autre part, il existe une catégorie de textes qui n'est pas rédigé en format Unicode. Ceux-là rendent la mission encore plus délicate à traiter dans des logiciels de traitements automatiques des données lexicales.

Pour amorcer cette entreprise, nous avons ouvert un fichier dénommé « *Corpus* » avec différents dossiers qui contiennent les textes collectés. Dans la mesure où le texte ne présente pas ces anomalies, le travail s'effectue au fur et à mesure en opérant par la sélection du texte en question et l'enregistrant dans un dossier en format Word sous le nom de texte numéroté. Et sur un autre dossier word nous mettons les métadonnées (l'intitulé du texte, sa source électronique, sa catégorie ou genre, sa première date d'apparition sur le web et le degré de recevabilité de sa transcription).

Durant une période de deux mois[2], nous avons pu moissonner environ 271 dossiers de diverses formes et catégories. Après ce premier travail, nous sommes passés à la seconde phase qui est celle de traiter les textes. Ce traitement a pour mission de régler d'abord certaines anomalies de l'ordre de la notation, puis de mettre chacun d'eux en formât texte brut. Pour enregistrer le texte dans cette forme il faut passer par un codage de l'UTF8[3] afin qu'il puisse garder sa forme initiale sans perte de mots. La longueur des textes pose des problèmes dans leur traitement, alors nous avons fait appel à un logiciel appelé RapidInfo[4].

4. Le matériau utilisé pour le traitement des données

RapidInfo est un logiciel qui permet de gagner du temps et d'éviter les multiples erreurs auxquelles nous sommes confrontés, générées par l'intensité du travail. En fait, il s'agit de programmer des éléments que nous souhaitons changer à titre illustratif : La transcription de certaines lettres n'est pas conforme à l'usage de l'écrit de la langue kabyle comme c'est le cas pour le son « ɣ » qui est représenté dans certains texte par « gh », le « aa» pour le « ɛ » ou « â », le « ḍ » pour le « d' » ou « dh » etc. Nous devons les mettre à disposition de ce logiciel qui se chargera de tous les remplacements nécessaires, et ensuite d'enregistrer les données dans le fichier corpus.

L'étape suivante se résume dans l'application d'un autre logiciel de statistique, Antconc, qui est un logiciel qui nous permet l'élaboration de listes pour chaque texte. Il nous offre plusieurs possibilités de listes et nous avons conçu des

[2] Le travail a été réalisé les mois de décembre 2008 et le mois de Janvier 2009 à l'université de l'UNIL à Lausanne.

[3] C'est l'abréviation du système de codage dénommé UTF8 (Format Texte Unicode).

[4] Ce logiciel de remplacement permet grâce à un programme bien établi de mettre à niveau tous les textes qui ne répondaient pas à la norme requise.

listes par ordre alphabétique croissant et des listes de fréquences enregistrées dans chacun des dossiers.

En dernier lieu, nous nous sommes orientés vers un autre logiciel appelé Toolbox à partir duquel nous avons conçu la construction de notre dictionnaire fondamental du Kabyle. Ce logiciel renferme un certain nombre de caractéristiques permettant de mener à terme notre projet.

5. La constitution du corpus

Notre projet commence donc à se dessiner au fur et à mesure que sont franchies les étapes. À ce stade, les premières données deviennent de plus en plus apparentes. En effet, avec la répartition du corpus en catégories voire en genres, nous avons obtenu de l'ensemble des textes une douzaine de champs qui se répartissent comme suit : les romans, les contes, les nouvelles, la poésie, les chansons, les récits, les textes journalistiques, les discussions, le théâtre, les textes religieux, les enquêtes et les devinettes.

Une fois que les textes sont mis dans la rubrique qui leur correspond, il est utile de reprendre le logiciel Antconc pour établir des listes par catégories. La dernière étape dans le traitement des textes pour obtenir une liste unique qui consiste à soumettre tous les textes traités en une seule fois à Antconc[1]. Nous disposons d'une liste unique à partir de laquelle est constituée la base de données dont dérivera le vocabulaire fondamental du Kabyle. Bien évidement, tous les mots ne seront pas retenus, car nous avons procédé d'abord par l'élimination de ceux à basse fréquence. Nous avons décidé de nous arrêter à la fréquence 5, et le reste a été mis en annexe.

Pour savoir quels sont les mots que partagent quotidiennement les Kabylophones, nous avons fait appel à la notion de disponibilité. Ce facteur est d'une importance capitale pour la suite à donner à l'ensemble du travail. Cependant, différentes lacunes peuvent être relevées et qui sont dues probablement aux manques de l'exhaustivité de notre corpus de base. Il nous semble donc important de revoir toutes les situations qui nécessitent d'être corrigées et complétées par d'autres mots en usage.

Pour ce dernier point, il est à constater que, quelle que soit la rigueur avec laquelle se fait un tel travail, celui-ci ne pourra jamais être complet, car le lexique se présente sous forme de listes ouvertes. C'est parmi les contraintes majeures qui pourraient être circonscrites dans un tel projet via la toile. De plus, un dictionnaire fondamental ne pourra pas retenir tous les mots de la langue mais seulement ceux qui sont essentiels et utiles.

[1] AntConc est un logiciel de traitement statistique des données textuelles et il les affiche en format listes alphabétiques ou par fréquences.

6. L'aspect statistique

Le volet statistique dans l'étude du vocabulaire relève du champ de la logique, car il permet d'obtenir des fréquences objectives comme le souligne G. Gougenheim (1964 :31) : « la statistique des fréquences du vocabulaire offre, semble-t-il, un critère objectif, permettant de déterminer scientifiquement les mots les plus usuels ». L'auteur s'appuie sur une expérience déjà menée en Allemagne en 1897 par J. W. Kading qui a donné des résultats encourageants dans le domaine. Puis l'idée d'élaborer des dictionnaires de fréquences est devenue l'apanage de plusieurs langues en Europe et aux Etats-Unis à cette période.

Les textes que nous avons moissonnés à travers l'outil Internet appartiennent essentiellement à plusieurs catégories alimentant les sites Web animés par les internautes kabyles. On y trouve des textes tirés de romans, d'autres de discussions échangées via la toile, des textes journalistiques, etc. Il s'agit en gros, de textes produits ces dernières années, ce qui est plus en phase avec la nature du projet.

En plus du corpus recueilli, nous disposons aussi d'une enquête de terrain réalisée exclusivement dans la perspective de palier aux carences du corpus obtenu sur le Web. Cette enquête est constituée d'un questionnaire ouvert destiné à être rempli par des étudiants de licence en Tamaziɣt[1]. Il est composé de 19 fiches réparties en champs lexico-sémantiques, au sein desquelles les enquêtés sont sensés donner une vingtaine de noms et de verbes dans la langue Kabyle pour chaque champ de façon spontanée, car l'enquête en question revêt ainsi un caractère psycholinguistique, ce qui implique que l'opération devrait être faite dans un cadre objectif, c'est-à-dire que les enquêtés répondent sur place.

L'aspect relatif à la fréquence dans ce type de travaux est très significatif dans la mesure où l'on constate rapidement que les premiers mots sont tous des mots grammaticaux à l'exception de quelques-uns. Puis viennent les mots lexicaux en seconde position. Une fois le travail de sélection terminé et les anomalies réglées restent à élaguer toutes les formes de vocabulaire qui n'ont pas été retenus. Ainsi donc, le nombre de vocables sera réduit en fonction de l'application de l'indice statistique qui est de 5, i.e. pour qu'un vocable soit retenu, il faut qu'il soit attesté au moins cinq fois ce qui équivaut à cinq régions. En plus du vocabulaire récolté sur la toile, nous avons aussi celui qui est issu de l'enquête de terrain qui nous donnait environ 12 000 mots. Ce dernier, nous l'avons dénommé vocabulaire concret, car issu d'une enquête psycholinguistique.

Parmi les contraintes enregistrées lors de la réalisation de ce travail, nous retenons celle de la confrontation des deux vocabulaires (celui qui est issu du Web et celui résultant de l'enquête de terrain). Une telle opération consiste à établir une liste unique de laquelle découlera par la suite la nomenclature du futur dictionnaire fondamental du Kabyle. À la fin de cette vérification, nous avons pu aboutir à un

[1] Ce sont des étudiants qui viennent presque de tous les coins de la Kabylie , i.e., venant des six Wilayas (Tizi-Ouzou, Bejaia, Bouira, Alger, Bordjbouararidj, Sétif et Jijel).

chiffre de 181 vocables non attestés dans la base de données[1] construite à travers le Web. Nous avons pu arrêter la liste constituant le dictionnaire en question et qui est constituée de 3241 vocables[2]. En principe, l'ensemble des vocables retenus dans la nomenclature du dictionnaire va devoir s'inter définir[3] selon la démarche retenue dans ce genre de dictionnaire.

7. Conclusion

En guise de conclusion pour ce modeste travail, nous tenons à souligner que la langue amazighe en général et celle du kabyle en particulier rencontrent beaucoup d'obstacles sur le chemin des technologies de l'information et de la communication. Cependant, des initiatives se manifestent ici et là à travers diverses expériences à l'image des travaux réalisés par l'IRCAM au Maroc et de ceux qui se font actuellement à l'INALCO à Paris. Introduire le tamazight dans les TICs aujourd'hui est une nécessité et une initiative très encourageante quant au devenir de celle-ci, ce qui signifie lui octroyer une place au même titre que le reste des langues. La multiplicité des travaux ayant trait aux TICs réalisés sur et dans la langue amazighe devrait faciliter, à l'avenir, sa large diffusion via la toile.

Bibliographie

Duclay F. et Al., 2006, « Fouille du Web pour la collecte de données linguistiques : avantages et inconvénients d'un corpus hors normes », In *Acte de l'atelier Fouille du Web des 6èmes journées francophones*, Lille.

Gougenheim G., 1958, *Dictionnaire fondamental de la langue française*, Ed. Librairie Marcel Didier, Paris.

Gougenheim G.., 1964, *L'élaboration du Français fondamental : 1er degré*, Ed. Librairie Marcel Didier, Paris.

Guilbert L., 1963, « De l'utilisation de la statistique en lexicologie appliquée », *Etudes de linguistique appliquée*, Faculté des Lettres en Sciences Humaines, Université de Besançon (Didier Paris), n°2, pp. 12-23.

Guiraud P., 1954, *Caractères statistiques du Vocabulaire*, Paris, Presses Universitaires de France.

Michea R., 1950, *« La culture par la langue », Les langues Modernes*, t. 44, Septembre-octobre 1950, pp. 314-322.

Nait Zerrad K., 2010, « TIC et aménagement linguistique » , In *Revue d'étude berbère*, INALCO, Paris.

Ogden, C. K., 1944, *Basic English*. *A General Introduction with Rules and Grammar*, 9ème édition, London.

[1] Il est à signaler que la vérification de la liste issue de l'enquête pouvait être faite de manière systématique si la langue était dotée d'un logiciel de segmentation et de lemmatisation, d'ailleurs, ceci constitue une carence qui rentre dans les TICs.

[2] La rédaction de ce dictionnaire fondamental du Kabyle est en cours de réalisation.

[3] La démarche telle quelle a été consacrée par l'équipe de Gougengeim à Saint-Cloun lors de la réalisation du vocabulaire fondamental de la langue française.

Pelgrum W.J. et Law N., 2004, *Le TIC et l'éducation dans le monde : tendance, enjeux et perspectives*, Paris, Rapport publié par l'ONU pour l'éducation, la science et la culture.

Rachid **Adjaout** est docteur en langues, littératures et sociétés, obtenu en décembre 2011à l'INALCO, Paris. Spécialisé en linguistique berbère et enseignant au département de langues et cultures berbères de l'université de Bejaia, Algérie. Il a assuré les modules de linguistique générale, de lexicologie/lexicographie, la sémiologie, la méthodologie appliquée à la linguistique et l'analyse du discours. Son domaine de recherche tente de travailler autour des questions de sémantique berbère, lexicologie/lexicographie. Auteur de quelques articles et communications scientifiques.