Initiative pour le développement d'un corpus de la langue amazighe

Siham Boulaknadel Fadoua Ataa Allah
Centre des Etudes Informatiques, des Systèmes d'Information et de Communication
Institut Royal de la Culture Amazighe, Rabat, Maroc
{boulaknadel,ataaallah}@ircam.ma

Résumé

Les corpus électroniques constituent de nos jours un élément essentiel et une base référentielle élémentaire, présentant les faits d'une langue, pour bien mener des recherches linguistiques, philologiques et informatiques. Conscientes de ce fait et dans la perspective de promouvoir la langue et la culture amazighes, nous avons opté au sein de l'Institut Royal de la Culture Amazighe de doter la langue amazighe d'un corpus de référence à visée exhaustive. Ainsi, nous avons entrepris l'élaboration d'un corpus constitué de textes de la langue amazighe permettant de mettre à la disposition de nos chercheurs une ressource représentant toutes les variantes et offrant une information en profondeur sur la langue amazighe.
Le présent article décrit les étapes entreprises au cours de la construction du corpus de la langue amazighe, et fait l'objet d'une évaluation sur la collection actuelle des textes.

1. Introduction

Depuis la création de l'Institut Royal de la Culture Amazighe (IRCAM), la langue amazighe au Maroc a bénéficié d'un statut institutionnel lui permettant d'avoir une graphie officielle, un codage propre dans le standard Unicode, des normes appropriées pour la disposition d'un clavier amazighe et des structures linguistiques qui sont en cours d'élaboration en empruntant une stratégie innovante et progressive. Cette stratégie a été initiée par la construction des lexiques (Ameur et al., 2006-b ; Kamel, 2006 ; Ameur et al., 2009), l'homogénéisation de l'orthographe et la mise en place des règles de segmentation de la chaîne parlée (Ameur et al., 2006-a) et par l'élaboration des règles de grammaire (Boukhris et al., 2008). Certes ces étapes de standardisation sont élémentaires et essentielles mais ne sont pas suffisantes pour qu'une langue peu dotée informatiquement telle que l'amazighe puisse franchir le seuil de la mondialisation informatique et de rejoindre ses consœurs dans ce domaine.

Dans cette perspective s'inscrit de nombreuses recherches scientifiques, principalement celles se focalisant sur la correction orthographique (Es Saady et al., 2009), la traduction automatique (Rachidi et Mammas, 2007) la reconnaissance optique des caractères (Fakir et al., 2009), et celles s'occupant de la conception et la réalisation des ressources et outils linguistiques (Iazzi et Outahajala, 2008 ; Ataa Allah et Jaa, 2009 ; Boulaknadel, 2009 ; Ataa Allah et Boulaknadel, 2010 -a- ; Ataa Allah et Boulaknadel, 2010 -b-; Outahajala et al., 2010 ; Boulaknadel et Ataa Allah, 2011). Or, pour mieux mener ce chantier de construction de ressources et outils linguistiques qui s'est ouvert à la langue amazighe, il s'avère primordial de doter la langue amazighe de corpus essentiel à son traitement automatique.

Dans ce travail, nous nous sommes intéressées à la construction d'un corpus dédié à la langue amazighe. Cette tâche s'inscrit dans le cadre d'un projet mené au sein de l'Institut Royal de la Culture Amazighe visant à fournir à la langue amazighe des ressources linguistiques riches et exploitables. Le but de cette initiative est d'encourager les linguistes à mener des recherches sur la langue amazighe, en créant une ressource utilisable pour cet objet. L'existence d'un tel rassemblement de textes en langue amazighe fournira également aux chercheurs intéressés par la langue amazighe un accès à des données actuellement dispersées ou non disponibles, et ce quel que soit leur domaine académique (linguistique, anthropologie, sociologie, littérature...). Grâce à la possibilité d'accès rapide aux données, et aux nouvelles possibilités de traitement de celles-ci, ce corpus informatisé permettra de nouveaux genres de recherches, auparavant non envisagés. Nous espérons que cette entreprise contribuera à donner une place à la langue amazighe dans la recherche linguistique du 21ème siècle.

Dans la suite de cet article, nous présentons dans la section 2 un descriptif des particularités de la langue amazighe standard du Maroc. Puis, nous détaillons dans la section 3 les étapes de construction de notre corpus ainsi qu'une batterie de mesures statistiques pour son analyse. Alors que nous consacrons la section 4 à la conclusion et aux perspectives.

2. Particularités de la langue amazighe

2.1. Historique

La langue amazighe connue aussi par le berbère est considérée comme la langue autochtone de l'Afrique du Nord (Hachid, 2000; Charles-André, 1978). Elle couvre toute l'Afrique du nord, le Sahara et une partie du Sahel ouest africain. Au Maroc, l'amazighe se répartit en trois grandes régions dialectales qui couvrent l'ensemble des régions montagneuses : au nord-est, le Rif avec le dialecte Tarifite ; au centre, le Moyen-Atlas et une partie du Haut-atlas avec le dialecte Tamazighte ; au sud et sud-ouest, le Haut-Atlas, l'Anti-Atlas et Souss, le domaine chleuh avec le dialecte Tachelhite.

Jusqu'à 1994 l'amazighe a été exclusivement réservée au domaine familial (Boukous, 1995). Cependant, suite au discours royal en 2001, l'amazighe est devenue une langue institutionnelle par la création de l'IRCAM. Et grâce à la constitution de 2011, l'amazighe a jouit auprès de sa consoeur l'arabe d'un statut d'une langue officielle.

2.2. Alphabet amazighe

L'amazighe fait partie des langues afro-asiatiques (Greenberg, 1966). Son système d'écriture, le « Libyco-berbère » (Tifinaghe en amazighe), date de plus de 25 siècles. Il est de nature alphabétique, à tendance phonologique fondé sur des signes à valeur consonantique, à usages traditionnellement assez restreints (funéraires, symboliques et ludiques). Cependant, les formats d'apparition de ses signes n'ont cessé de se développer : depuis son origine, le Libyque, jusqu'à le néo-tifinaghe, à la fin des années soixante, et le Tifinaghe-IRCAM, en 2001(Ameur et al., 2004). Ce dernier est orienté horizontalement de gauche à droite et composé de : 27 consonnes, 2 semi-consonnes, 4 voyelles :

- 27 consonnes dont : les labiales (ⵀ, ⵃ, ⵇ), les dentales (ⵜ, ⵟ, ⴷ, ⴹ, ⵍ, ⵏ, ⵕ, ⵙ), les

alvéolaires (ⵟ, ⵙ, ⵚ, ⵣ), les palatales (ⵛ, ⵊ), les vélaires (ⴽ, ⵅ), les labiovélaires (ⴽⵯ, ⵅⵯ), les uvulaires (ⵇ, ⵅ, ⵖ), les pharyngales (ⵄ, ⵃ) et la laryngale (ⵀ);

- 2 semi-consonnes : ⵢ et ⵡ ;
- 4 voyelles : trois voyelles pleines ⴰ, ⵉ, ⵓ et la voyelle neutre (ou schwa) ⴻ qui a un statut assez particulier en phonologie amazighe.

3. Construction et analyse du corpus

Malgré l'intérêt accordé à l'amazighe, il existe peu de travaux publiés concernant l'évaluation de corpus élaborés pour la langue amazighe standard du Maroc. Dans ce contexte, nous proposons à travers cet article de consacrer plus d'importance à ce genre de travaux, en partant de l'idée que tout traitement automatique de la langue amazighe ne peut se faire sans que cette dernière soit dotée d'un corpus de référence qui fera l'objet de recherches sur la langue. Ainsi, nous avons constitué et analysé un corpus de textes amazighes composé de 160 textes amazighes représentant différents genres littéraires (romans, poésie, contes, articles journalistiques) et couvrant différents thèmes.

3.1. Etapes de construction de corpus

L'élaboration de notre corpus se déroule en quatre étapes, dont la première consiste à collecter les documents écrits en langue amazighe, principalement ceux édités par l'IRCAM ou publiés dans son site officiel. La deuxième étape sert à la normalisation du format des documents collectés, où nous avons procédé à un dé-balisage des sources HTML et à une conversion des formats PDF et WORD en un format texte brut. Cette dernière sera succédée par une troisième étape qui consiste à convertir tous les textes en tifinaghe Unicode, en exploitant le convertisseur et translittérateur de la langue amazighe (Ataa allah et Boulaknadel, 2011). Par la suite, nous procédons à l'identification et la classification de chaque document selon sa thématique.

3.2. Propriétés statistiques du corpus

Dans le cadre de notre projet d'élaboration de corpus électronique pour la langue amazighe, nous avons collecté un corpus composé de 160 textes amazighes représentant différents genres littéraires, à savoir conte, conte pour enfants, poésie et articles de presse contenant les sous-genres journal, magasine et Net. Le Tableau 1 présente les caractéristiques statistiques du corpus collecté.

Statistiques	Conte	Conte pour enfants	Poésie	Net	Magasine	Journal	Global
Nombre de documents	4	21	10	78	11	36	160
Nombre de mots distincts	8,591	6,263	2,792	2,327	6,242	7,624	23,174
Nombre de mots total	37,363	29,741	8,534	9,519	23,086	27,850	136,093

Tableau 1 : Statistiques du corpus amazighe

Par ailleurs, nous avons eu recours à un ensemble de mesures statistiques afin d'évaluer et d'analyser notre corpus. Ainsi, nous nous sommes basées sur la loi Zipf-Mandelbort (Manning et Schütze, 1999), l'usage des caractères tifinaghes dans le corpus (Darrudi et Hejazi, 2004), les mesures de la richesse lexicale, et la représentativité des documents dans le corpus (Abdelali et al., 2005).

3.2.1. Loi de Zipf-Mandelbort

La loi de Zipf-Mandelbort est une distribution de probabilité discrète, connue également sous le nom de la loi de Pareto-Zipf (Zipf, 1949), qui est la forme continue de la loi de Zipf. Cette dernière prédit que si dans un texte de longueur N où les mots sont rangés dans l'ordre décroissant de leur fréquence d'apparition, la fréquence f(r) du mot de rang r est approximativement de forme $f(r) = k/r$, où k est une constante. Cette loi a été élargi par Mandelbrot en : $f(r) = \frac{A}{(B+r)^C}$, où A, B et C sont des constantes.

En utilisant l'outil DataFit pour l'ajustement des courbes, nous avons établi pour chaque catégorie de textes les graphes illustrés sur la Figure 1 qui présentent les occurrences des mots par rapport à leur rang.

D'après ces graphes, nous constatons que le comportement des mots de notre corpus représente bien la diversité et la nature du contenu de chaque catégorie de notre corpus. Généralement, la distribution de fréquence d'un corpus est séparée en 3 zones, à savoir :

- Zones à hautes fréquences dont la nature de ses mots sont essentiellement de type anti-dictionnaires. D'où l'inutilité d'étudier leur comportement vu que leur fréquence dépend mutuellement de la taille des documents du corpus.
- Zones à moyennes fréquences contenant, globalement, les termes représentant les différentes thématiques traitées par le corpus.

  Suite aux courbes de la Figure 1, nous constatons que pour les catégories conte, conte pour enfants et magasine la distribution de la fréquence des termes appartenant à cette zone est supérieure à celle de la loi de Zipf-Mandelbort. Tandis que nous remarquons l'inverse à l'égard des catégories poésie, journal et Net. Ce qui s'explique par le fait que les textes collectés pour les catégories poésie, journal et Net traitent plusieurs thématiques. Cependant, les catégories conte et conte pour enfants se composent principalement d'histoires destinées respectivement aux adultes et aux enfants. Ceux de la catégorie magasine introduisent les productions de l'IRCAM, en particulier les travaux en relation avec les contes et les contes pour enfants. Ce qui induit la mono-thématique de ces catégories.

  Par ailleurs, nous remarquons que la courbe du corpus global suit la même allure que celles des catégories conte, conte pour enfants et magasine. Ceci est dû au nombre des termes distincts de ces 3 dernières catégories qui représente 72.4% des termes du corpus global.

- Zones à basses fréquences dont la largeur dépend principalement de la variété du vocabulaire utilisé qui est liée à la qualité du style d'écriture.

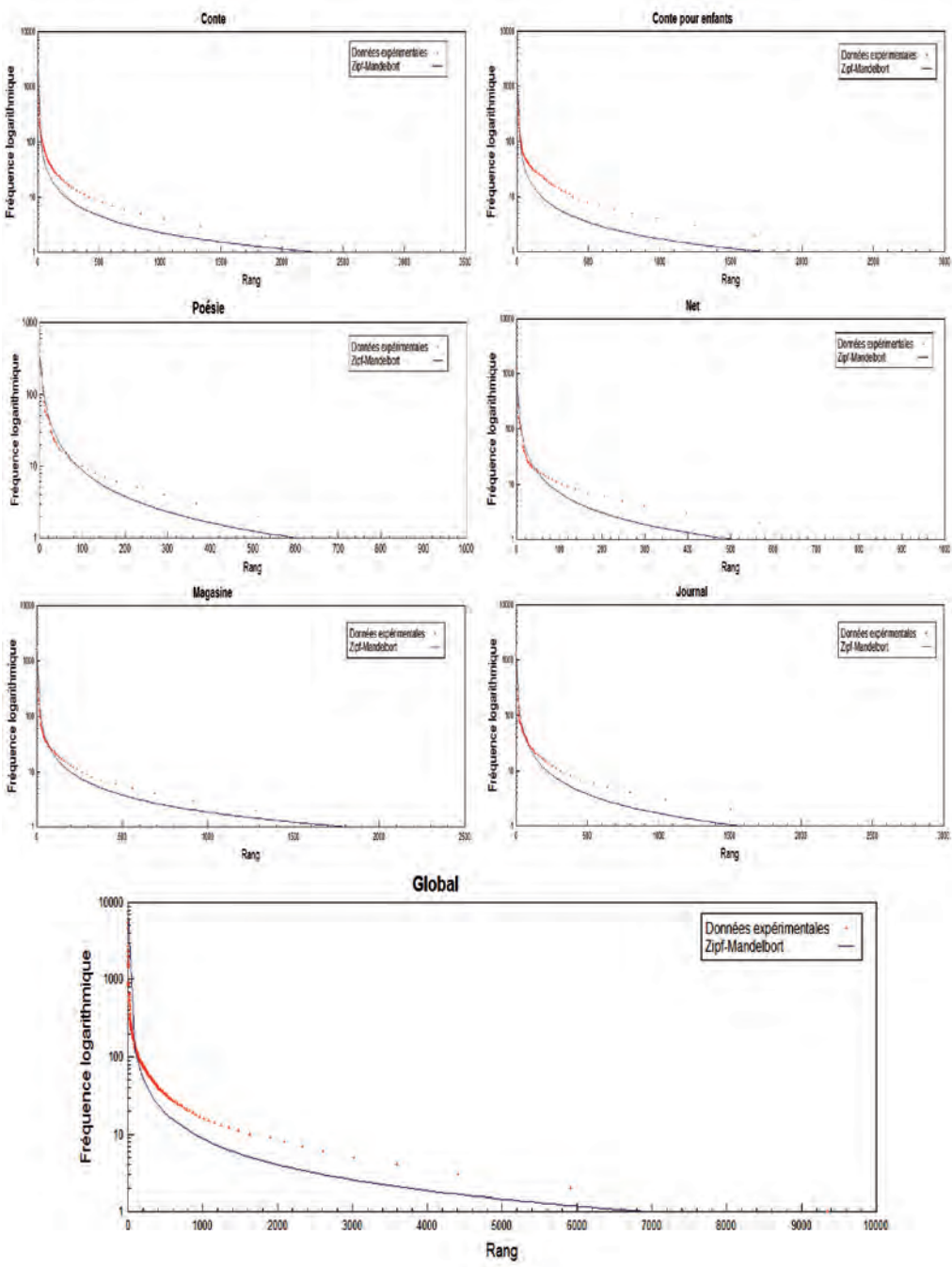

Figure 1 : Représentation schématique de la distribution des mots du corpus selon la loi de Zipf-Mandelbort

3.2.2. Usage des caractères tifinaghes

L'usage des caractères est une méthode d'évaluation qui consiste à calculer le pourcentage d'apparition de chaque lettre de l'alphabet de la langue étudiée dans le corpus élaboré, dans l'objectif de mesurer la richesse du corpus en terme de caractères. Dans ce contexte, nous avons procédé par le calcul de l'usage relatif des caractères tifinaghes, où la Figure 2 montre le pourcentage des occurrences de chaque lettre dans le corpus.

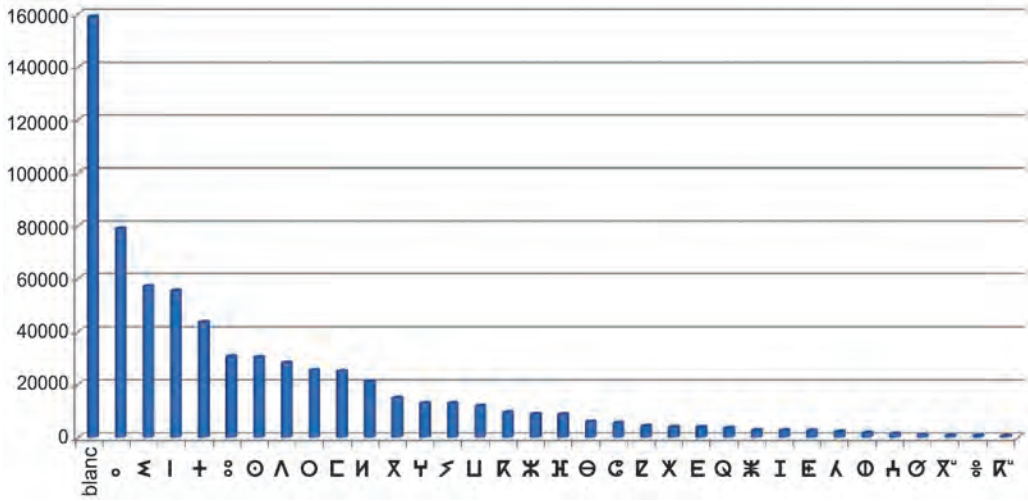

Figure 2 : Pourcentage des occurrences des caractères tifinaghes dans le corpus

D'après cette figure, nous pouvons remarquer que les caractères «ⵯ » « ⵖ » et « ⵇ » ne sont pas très utilisés dans le corpus par rapport aux autres caractères et que le caractère « espace » constitue 23% du corpus ce qui implique, à partir des occurrences des mots de notre corpus, que la moyenne de nombre de caractères par mot est à peu prés égale à 4,26.

3.2.3 Mesures de la richesse lexicale

La richesse lexicale est une notion intuitive et très subjective. Cependant, les mesures de la richesse lexicale cherchent à apporter une solution objective, mathématique, à un problème auquel les réponses n'ont été, pendant longtemps, que approximatives et impressionnistes. Nous allons dans ce qui suit présenter et commenter les résultats obtenus suite à l'application de deux méthodes de mesure de la richesse lexicale retenues.

i. TTR

Afin d'affiner nos analyses, nous procédons à une méthode de calcul de la richesse lexicale à partir d'un indicateur lexical, qui est le TTR ou Type Token Ratio :

TTR = V/N, où V est le nombre de mots distincts et N est le nombre total des mots ou l'étendue du texte.

Une lecture attentive des statistiques du Tableau 1 et du graphique de la Figure 3, représentant le classement des catégories de notre corpus selon la méthode TTR, nous montre une sensibilité de la richesse lexicale selon cette formule aux valeurs de l'étendue des textes.

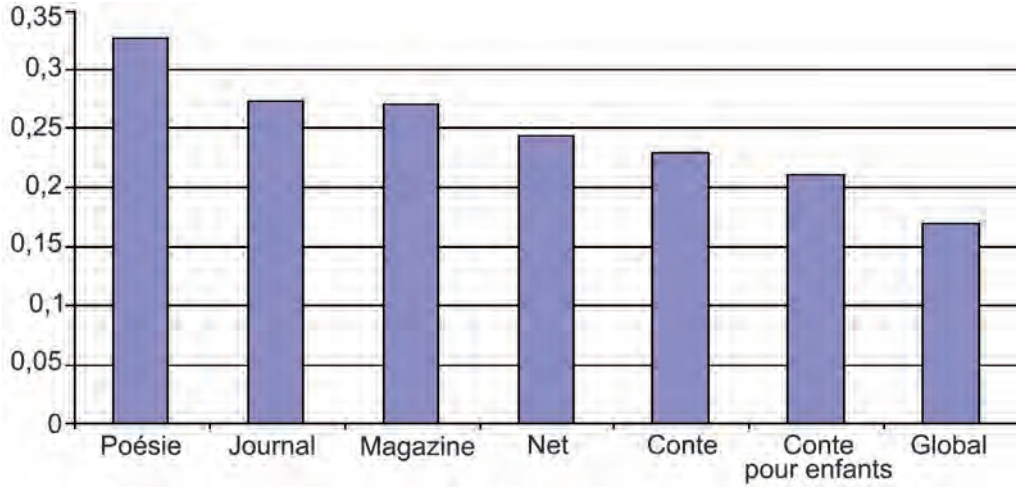

Figure 3 : Richesse lexicale selon la mesure TTR

ii. Indice W de Brunet

Dans le but de minimiser l'influence de l'étendue des textes sur la valeur de la richesse lexicale, Brunet (Brunet, 1978) fait jouer le rôle du facteur de réduction du nombre des mots distincts N à la réciproque du nombre total des mots V, une fois que ce dernier sera convenablement réduit à son tour par l'exposant fractionnaire α=0,172. La formule de l'indice W de Brunet s'écrit donc ainsi :

$$W = N^{V^{-\alpha}}$$

	N	V	V^{α}	$W = N^{V^{-\alpha}}$
Global	136,093	23,174	5,63	8,15
Journal	27,850	7,624	4,65	9,02
Conte	37,363	8,591	4,75	9,18
Magazine	23,086	6,242	4,5	9,34
Conte pour enfants	29,741	6,263	4,5	9,87
Poésie	8,534	2,792	3,91	10,1
Net	9,519	2,327	3,79	11,18

Tableau 2 : Richesse lexicale selon l'indice W de Brunet

En analysant les résultats de cette formule représentés dans le Tableau 2, nous constatons que la catégorie journal qui se trouve en tête du classement des différentes catégories de notre corpus est la plus riche lexicalement, ce qui est dû principalement à la variété des sujets traités par cette dernière.

3.2.4 Représentativité des documents dans le corpus

La représentativité des documents dans le corpus permet d'évaluer l'apport de chaque document en terme de nouveaux mots ajoutés à la collection. Ainsi, nous avons entrepris cette évaluation par l'intégration d'un document par document à l'ensemble traité et la vérification de sa contribution à la construction de notre corpus, en calculant le nombre de mots distincts ajoutés suite à l'adjonction de ce document.

Nombre de documents	Nombre de mots	Nombre de mots distincts
10	29,693	6,930
20	39,421	8,995
40	44,761	9,986
80	88,119	16,824
120	100,424	18,539
160	136,093	23,174

Tableau 3 : Contribution des documents dans la richesse du corpus

En comparant le nombre des mots distincts pour chaque ensemble de documents représentés par le Tableau 3, nous constatons que la quantité de données ajoutées à chaque reprise contribue de manière significative à l'enrichissement du vocabulaire du corpus. Ainsi, nous pouvons conclure qu'à ce stade notre corpus n'a pas encore représenté toutes les variétés lexicales de la langue amazighe.

4. Conclusion

Cet article s'inscrit dans une démarche fondatrice d'élaboration de ressources et d'outils de traitement automatique de la langue amazighe, qui contribue à la promotion et le développement de cette langue. A ce titre, nous avons élaboré un corpus de textes à visée exhaustive pour la langue amazighe, que nous avons analysé et évalué en exploitant une batterie de mesures. A la base de ces mesures, nous retenons que ce corpus nécessite un enrichissement d'ordre horizontal et vertical. Dans le sens d'augmenter respectivement la richesse lexicale des catégories existantes, et la variété thématique de notre corpus.

Références

Abdelali A., Cowie J., Soliman H. S. (2005). Building a modern standard Arabic corpus. *Actes du computational modeling of lexical acquisition workshop,* pp. 1-7.

Ameur M., Bouhjar A., Boukhris F., Boukouss A., Boumalk A., Elmedlaoui M., Iazzi E. (2006). *Graphie et orthographe de l'amazighe*. Maroc : IRCAM.

Ameur M., Bouhjar A., Elmedlaoui M., Iazzi E. (2006). *Vocabulaire de la langue amazighe (français-amazighe)*. Maroc : IRCAM.

Ameur M., Bouhjar A., Boumalk A., El Azrak N., Laabdelaoui R. (2009). *Vocabulaire de la langue amazighe* (amazighe-arabe). Maroc : IRCAM.

Ataa Allah F., Jaa H. (2009). Etiquetage morphosyntaxique : Outil d'assistance dédié à la langue Amazighe. Actes du 1er symposium international sur le traitement automatique de la culture amazighe. pp. 110-119.

Ataa Allah F., Boulaknadel S. (2010). Pseudo-racinisation de la langue amazighe. *Actes du Traitement Automatiques des Langues Naturelles*.

Ataa Allah F., Boulaknadel S. (2010). Online Amazigh Concordancer. *Proceedings of International Symposium on Image Video Communications and Mobile Networks*. Rabat, Maroc.

Ataa Allah F., Boulaknadel S. (2011). Convertisseur pour la langue amazighe : script arabe - latin - tifinaghe. Actes du *2ème symposium international sur le traitement automatique de la culture amazighe*. Agadir, Maroc, pp. 3-10.

Boukhris F., Boumalk A., Elmoujahid E., Souifi H. (2008). *La nouvelle grammaire de l'amazighe*. Rabat, Maroc : IRCAM.

Boukous A. (1995), *Société, langues et cultures au Maroc : Enjeux symboliques,* Casablanca, Najah El Jadida.

Boulaknadel S. (2009). Amazigh ConCorde: an appropriate concordance for Amazigh. Actes du *1er symposium international sur le traitement automatique de la culture amazighe*. pp. 176-182.

Boulaknadel S., Ataa Allah F. (2011). Building a standard Amazigh corpus. *Proceedings of the International Conference on Intelligent Human Computer Interaction*. Prague, Tchec.

Brunet E. 1978. *Le vocabulaire de Jean Giraudoux. Structure et évolution. Statistique et informatique appliquées à l'étude des textes à partir des données du Trésor de la Langue Française*. Genève : Slatkine.

Charles-André J. (1978). *Histoire de l'Afrique du nord des origines à la conquête arabe: Tunisie - Algérie – Maroc*. France : Editions Payot.

Darrudi E., Hejazi M.R. (2004). Assessment of a Modern Farsi Corpus. Actes de *2nd Workshop on Information Technology & its Disciplines*.

Es Saady Y., Ait Ouguengay Y., Rachidi A., El Yassa M., Mammass D. (2009). Adaptation d'un correcteur orthographique existant à la langue Amazighe : cas du correcteur Hunspell. Actes du *1er symposium international sur le traitement automatique de la culture amazighe*. pp. 149-158.

Fakir M., Bouikhalene B., Moro K. (2009). Skeletonization methods evaluation for the recognition of printed tifinaghe characters. Actes du *1er symposium international sur le traitement automatique de la culture amazighe*. pp. 33-47.

Hachid M. (2000), *Les premiers berbères: entre méditerranée, Tassili et Nil*. France : Edisud.

Greenberg J. (1966). *The Languages of Africa*. Mouton, USA: The Hague.

Iazzi E., Outahajala M. (2008). Amazigh Data Base. *Actes de l'atelier HLT & NLP within the Arabic world: Arabic language and local languages processing status updates and prospects*. pp. 36-39.

Kamel S. (2006). *Lexique Amazighe de géologie*. Maroc : IRCAM.

Manning C., Schütze H. (1999). *Foundations of Statistical Natural Language Processing*. USA: MIT Press.

Outahajala M., Zenkouar L., Rosso P., Martí M. (2010). Tagging Amazigh with AncoraPipe. *Actes de Semitic Languages Workshop, 7thInternational Conference on Language Resources and Evaluation*. pp. 52-56.

Zipf G. K. (1949). *Human behaviour and the principal of least-effort*. USA: Addison-Wesley.