TALN 2010, Montréal, 19-23 juillet 2010

Pseudo-racinisation de la langue amazighe

Fadoua Ataa Allah Siham Boulaknadel

CEISIC, IRCAM
Avenue Allal El Fassi, Madinat Al Irfane, Rabat, Morocco
{ataaallah, boulaknadel}@ircam.ma

Résumé Dans le cadre de la promotion de la langue amazighe, nous avons voulu lui apporter des ressources et outils linguistiques pour son traitement automatique et son intégration dans le domaine des nouvelles technologies de l'information et de la communication. Partant de ce principe, nous avons opté, au sein de l'Institut Royal de la Culture Amazighe, pour une démarche innovante de réalisations progressives de ressources linguistiques et d'outils de base de traitement automatique, qui permettront de préparer le terrain pour d'éventuelles recherches scientifiques.

Dans cette perspective, nous avons entrepris de développer, dans un premier temps, un outil de pseudo-racinisation basé sur une approche relevant du cas de la morphologie flexionnelle et reposant sur l'élimination d'une liste de suffixes et de préfixes de la langue amazighe. Cette approche permettra de regrouper les mots sémantiquement proches à partir de ressemblances afin d'être exploités dans des applications tel que la recherche d'information et la classification.

Abstract In the context of promoting the Amazigh language, we would like to provide this language with linguistic resources and tools in the aim to enable its automatic processing and its integration in the field of Information and Communication Technology. Thus, we have opted, in the Royal Institute of Amazigh Culture, for an innovative approach of progressive realizations of linguistic resources and basic natural language processing tools that will pave the way for further scientific researches.

In this perspective, we are trying initially to develop a light stemmer based on an approach dealing with inflectional morphology, and on stripping a list of Amazigh suffixes and prefixes. This approach will conflate word variants into a common stem that will be used in many applications such as information retrieval and classification.

Mots-clés : Langue amazighe, Pseudo-racinisation, Morphologie flexionnelle.
Keywords: Amazigh language, Light stemming, Inflectional morphology.

1 Introduction

Depuis quelques années, la langue amazighe au Maroc a joui d'un statut institutionnel par la création de l'Institut Royal de la Culture Amazighe (IRCAM). Cette création lui a permis d'avoir une graphie officielle, un codage propre dans le standard Unicode, des normes appropriées pour la disposition d'un clavier amazigh et des structures linguistiques qui sont en phase d'élaboration avec une démarche progressive. Cette démarche a été initiée et entreprise par la construction des lexiques (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) et par l'élaboration des règles de grammaire (Boukhris et al., 2008). Néanmoins, toutes ces étapes de standardisation ne sont pas suffisantes pour qu'une langue peu dotée informatiquement telle que l'amazighe puisse rejoindre ses consoeurs dans le domaine des nouvelles technologies de l'information et de la communication. Dans ce contexte, de nombreuses recherches scientifiques sont lancées au niveau national pour améliorer la situation actuelle. Principalement, elles se focalisent sur la correction orthographique (Es Saady et al., 2009), la traduction automatique (Rachidi et Mammas, 2007) et la reconnaissance optique des caractères (Fakir et al., 2009). Or, celles qui occupent une position prioritaire dans la voie de la conception et la réalisation des ressources et outils linguistiques sont peu nombreuses (Iazzi et Outahajala, 2008 ; Ataa Allah et Jaa, 2009 ; Boulaknadel, 2009).

Cet article s'inscrit dans un mouvement innovant et progressif qui vise l'élaboration des ressources et outils linguistiques, spécialement, la réalisation des racineurs (*stemmers*). En effet, la racinisation (*stemming*) est un processus fortement exploité dans plusieurs domaines basés sur le traitement automatique des langues, tel que la recherche d'information, la traduction automatique et la classification. Il consiste à regrouper les mots sémantiquement proches à partir de mots de forme apparentée relevant de la flexion et de la dérivation. Cette approche a été utilisée dans différentes langues à travers plusieurs algorithmes, tels que l'algorithme de Porter (Porter, 1980) pour l'anglais, l'algorithme de Savoy (Savoy, 1993) pour le français et l'algorithme de Larkey (Larkey et al., 2002) pour l'arabe. Ces divers algorithmes procèdent de manières différentes: ceux basés sur une étape de *désuffixation* qui consiste à ôter aux mots les suffixes les plus longs possibles, suivie par une étape de *recodage* qui ajoute aux racines obtenues des terminaisons prédéfinies. Et ceux qui reposent sur la *désuffixation* et la *dépréfixation*.

Il est un fait que la langue amazighe ayant une morphologie assez complexe, la simple suppression des suffixes ne peut suffire à regrouper des familles de mots. Le plus souvent, les outils s'intéressant aux langues riches morphologiquement font appel à de lourds dictionnaires ou à un étiquetage morphosyntaxique préalable des mots présents dans les documents. Or, étant donné que l'amazighe fait partie des langues peu dotées informatiquement et vu que jusqu'à ce jour nous ne disposons ni de dictionnaire ni d'analyseur morphosyntaxique pour la langue amazighe standard du Maroc, nous avons entrepris de développer un algorithme de pseudo-racinisation (*light stemming*) se basant sur une analyse 'simpliste' prenant en compte le cas relevant de la morphologie flexionnelle. Dans la suite de cet article, nous présentons dans la section 2 un descriptif des caractéristiques de la langue amazighe standard du Maroc. Puis, nous détaillons dans la section 3 l'approche proposée et son évaluation. Alors que nous consacrons la section 4 à la conclusion et aux perspectives.

2 Caractéristiques de la langue amazighe

L'amazighe (berbère) fait partie des langues afro-asiatiques (Greenberg, 1966). Son système d'écriture date de plus de 25 siècles. Il a assimilé des changements afin de fournir à cette langue un système alphabétique adéquat et utilisable pour tous les parlers amazighs actuels. Ainsi en 2003, partant d'un héritage aussi bien ancien que moderne et contemporain, l'IRCAM a développé un système d'alphabet sous le nom de

Tifinaghe-Ircam. Il est orienté horizontalement de gauche à droite et composé de : 27 consonnes, 2 semi-consonnes, 4 voyelles. Par ses propriétés morphologiques, l'amazighe est considérée comme une langue complexe, dont les mots peuvent être classés en trois catégories morpho-syntaxiques:

2.1 Verbe

Le verbe peut apparaître sous une forme simple ou dérivée. La forme simple est composée d'une racine et d'un radical. Alors que la forme dérivée est obtenue à partir des verbes simples par la préfixation d'un morphème ou d'une combinaison de deux morphèmes de valeur différente. Ces morphèmes sont : ⵙ 's' / ⵙⵙ 'ss' qui indique la forme factitive, ⵜⵜ 'tt' qui marque la forme passive et ⵎ 'm' / ⵎⵎ 'mm' qui désigne la forme réciproque. Le verbe, qu'il soit simple ou dérivé, se conjugue selon quatre aspects : l'aoriste, l'accompli, l'inaccompli et l'accompli négatif. En général, le passage du verbe de l'aoriste à l'accompli ne subit pas de changement dans le cas des verbes réguliers. Alors que dans le cas des verbes irréguliers, ce passage est accompagné d'une alternance vocalique qui peut être suivi d'une tension consonantique. Or, l'inaccompli est dérivé de l'aoriste par l'application de la préfixation ⵜⵜ 'tt', la gémination, ou la tension d'une consonne radicale et l'insertion d'une voyelle. Selon le type du verbe, un ou plusieurs procédés peuvent être employés. Par ailleurs, la caractéristique principale de l'accompli négatif est l'apparition de la voyelle ⵉ 'i' devant ou après la consonne finale du radical, en fonction du type du verbe (monolitère, bilitère, etc.). Cependant, les formes de certains verbes à l'accompli positif et négatif sont identiques.

2.2 Nom

Le nom est une unité lexicale formée d'une racine et d'un schème, qui indique un substantif ou un adjectif. Il peut être sous une forme simple, composée ou dérivée. Il varie en genre, en nombre et en état. Le nom existe en deux genres : le masculin et le féminin. Le masculin commence en général par une des voyelles initiales ⴰ 'a', ⵉ 'i' ou ⵓ 'u'. Quant au féminin, il est généralement formé à base du masculin par l'ajout du morphème discontinu ⵜ…ⵜ 't...t'. Cependant, il posséde un singulier et un pluriel qui se réalise selon trois types. Le *pluriel externe* est formé par une alternance vocalique et par la suffixation de l'indice ⵏ 'ⵏ' ou l'une de ses variantes (ⵉⵎⴷⴰⵔⵏ 'imḍarn' *élèves en mas.*, ⵜⵉⵎⴷⴰⵔⵉⵏ 'timḍarin' *élèves en fém.*). Le *pluriel interne* est constitué d'une alternance vocalique accompagnée d'un changement de voyelles internes (ⵜⵉⵖⵎⵙⵜ 'tiγmst' *dent* → ⵜⵉⵖⵎⴰⵙ 'tiγmas' *dents*). Le *pluriel mixte* est formé par l'alternance d'une voyelle interne et/ou finale accompagnée par une suffixation de ⵏ 'n' (ⵉⵣⵉ 'izi' *mouche* → ⵉⵣⴰⵏ 'izan' *mouches*) ; ou une alternance vocalique initiale accompagnée d'un changement vocalique final ⴰ 'a' et d'une alternance interne (ⴰⵎⴳⴳⴰⵔⵓ 'amggaru' *dernier* → ⵉⵎⴳⴳⵓⵔⴰ 'imggura' *derniers*). En outre, le nom a deux états : l'état libre ou il ne subit aucune modification et l'état d'annexion qui se manifeste par une variation des noms à initiale vocalique (ⴰⵔⴳⴰⵣ 'argaz' → ⵓⵔⴳⴰⵣ 'urgaz' *homme*). L'état d'annexion est réalisé quand le nom a la fonction de sujet lexical postposé au verbe et est précédé d'un coordonnant, de termes d'attribution, de termes d'appartenance et de filiation, de nom de nombre ou d'une préposition à l'exception de ⴰⵍ/ⴰⵔ 'al/ar' *jusqu'à* et ⵖⵍⴰ 'bla' *sans*.

2.3 Particules

Les particules sont un ensemble de mots, en général, assez courts qui jouent le rôle d'indicateurs grammaticaux au sein d'une phrase. Cet ensemble contient les particules d'aspect, d'orientation et de négation ; les pronoms indéfinis, démonstratifs, possessifs et interrogatifs ; les pronoms personnels autonomes, affixes sujet, affixes d'objet direct et indirect, compléments du nom ordinaire et de parenté,

compléments de prépositions ; les adverbes de lieu, de temps, de quantité et de manière ; les prépositions ; les subordonnants et les conjonctions (Boukhris et al., 2008). Généralement, les particules sont invariables. Or, dans le cas de l'amazighe similairement au cas du français, il existe des particules flexionnelles telles que les pronoms possessifs (ⵍⵉⵣⵍⵍⵙ 'winns' *le sien* → ⵍⵉⵣⵍⵍⵙⵏ 'winnsn' *le leur*).

3 Pseudo-racinisation de l'amazighe

Malgré l'intérêt accordé à l'amazighe, il n'existe pas jusqu'à ce jour, à notre connaissance, des travaux publiés concernant le traitement automatique de la morphologie de la langue amazighe standard du Maroc. Ainsi, nous proposons à travers ce papier de consacrer plus d'importance à un tel type de travaux, particulièrement la pseudo-racinisation qui consiste à éliminer une liste prédéfinie de préfixes et / ou de suffixes flexionnels, sans tenter d'éliminer les infixes ou de récupérer le schème ou la racine du mot traité.

3.1 L'approche proposée

En général, la structure des racineurs varie considérablement selon les caractéristiques morphologiques d'une langue. Pour les langues indo-européennes, principalement le français et l'anglais, la majorité des techniques exploitées ne passent que par une étape d'élimination de suffixes. Tandis que pour les langues afro-asiatiques, dont la langue amazighe fait partie, ces techniques sont censées passer aussi par une étape de dépréfixation. En effet, dans la pratique les affixes peuvent altérer le sens d'un mot. Par conséquent, le fait de les exclure peut impliquer une grande perte d'information. Dans les langues indo-européennes, les préfixes modifient le signifié des mots ce qui rend leur suppression porteuse d'erreurs. Alors que dans le cas des langues afro-asiatiques, les préfixes sont aussi utilisés pour exprimer la flexion. Ainsi, nous proposons dans ce travail une approche légère de racinisation (pseudo-racinisation, *light stemming*), qui traite la morphologie flexionnelle sans se baser sur un dictionnaire ou un étiqueteur morphosyntaxique, mais en se limitant seulement à la structure du mot amazighe qui est composée d'un préfixe, d'un noyau et d'un suffixe. Cette approche se déroule en deux étapes : dépréfixation et désuffixation. Similairement à la méthode de Larkey (Larkey et al., 2002), notre algorithme reconnaît le préfixe et le suffixe de la liste et il les supprime. Dans le cas où plusieurs affixes sont détectés, c'est toujours le suffixe ou le préfixe le plus long qui est choisi. La liste des préfixes et suffixes est composée des morphèmes flexionnels communs de genre, de nombre et d'états pour les noms ; d'indices de personnes, de l'aspect et de type pour les verbes ; et des pronoms personnels affixes compléments du nom de parenté et de prépositions. Cette liste est regroupée en cinq classes variant d'un à cinq caractères.

- Les préfixes : {{ⴰ 'a', ⵉ 'i', ⵏ 'n', ⵓ 'u', ⵜ 't'} ; {ⵏⴰ 'na', ⵏⵉ 'ni', ⵏⵓ 'nu', ⵜⴰ 'ta', ⵜⵉ 'ti', ⵜⵓ 'tu', ⵜⵜ 'tt', ⵡⴰ 'wa', ⵡⵓ 'wu', ⵢⴰ 'ya', ⵢⵉ 'yi', ⵢⵓ 'yu'} ; {ⵉⵜⵜ 'itt', ⵏⵜⵜ 'ntt', ⵜⵜⴰ 'tta', ⵜⵜⵉ 'tti'} ; {ⵉⵜⵜⴰ 'itta', ⵉⵜⵜⵉ 'itti', ⵏⵜⵜⴰ 'ntta', ⵏⵜⵜⵉ 'ntti', ⵜⵓⵜⵜ 'tett'} ; {ⵜⵓⵜⵜⴰ 'tetta', ⵜⵓⵜⵜⵉ 'tetti'}}.
- Les suffixes : {{ⴰ 'a', ⴷ 'd', ⵉ 'i', ⴽ 'k', ⵎ 'm', ⵏ 'n', ⵖ 'γ', ⵙ 's', ⵜ 't'} ; {ⴰⵏ 'an', ⴰⵜ 'at', ⵉⴷ 'id', ⵉⵎ 'im', ⵉⵏ 'in', ⵉⵖ 'iγ', ⵎⵜ 'mt', ⵏⵖ 'nγ', ⵏⵜ 'nt', ⵓⵏ 'un', ⵙⵏ 'sn', ⵜⵏ 'tn', ⵡⵎ 'wm', ⵡⵏ 'wn', ⵢⵏ 'yn'} ; {ⴰⵎⵜ 'amt', ⴰⵏⵜ 'ant', ⴰⵡⵏ 'awn', ⵉⵎⵜ 'imt', ⵉⵏⵜ 'int', ⵉⵡⵏ 'iwn', ⵏⵉⵏ 'nin', ⵓⵏⵜ 'unt', ⵜⵉⵏ 'tin', ⵜⵏⵖ 'tnγ', ⵜⵓⵏ 'tun', ⵜⵙⵏ 'tsn', ⵙⵏⵜ 'snt', ⵡⵎⵜ 'wmt'} ; {ⵜⵓⵏⵜ 'tunt', ⵜⵙⵏⵜ 'tsnt'}}.

3.2 Evaluation

L'évaluation des algorithmes de pseudo-racinisation ou de racinisation est basée sur le calcul des mesures de précision et de rappel qui sont utilisées particulièrement dans le contexte de la recherche d'information.

Néanmoins, ces mesures ne montrent pas les causes spécifiques des erreurs, ce qui n'aide pas les concepteurs à optimiser leurs algorithmes. Ainsi, Paice a proposé une méthode d'évaluation par l'introduction de deux mesures : la sous-racinisation qui survient quand deux termes apparentés ne sont pas réduits à la même pseudo-racine et la sur-racinisation qui apparaît lorsque deux termes non apparentés sont associés à la même pseudo-racine par erreur (Paice, 1994). Cette méthode nécessite un échantillon de *w* mots classés dans *g* groupes de concepts, dont les n_g attributs de chaque groupe sont morphologiquement et sémantiquement liés les un aux autres. Ainsi, les taux d'erreurs de sous-racinisation et sur-racinisation sont calculés respectivement selon les formules suivantes : $US = {}^{GUMT}/_{GDMT}$ et $OS = {}^{GWMT}/_{GDNT}$; où $GDMT = \sum_g \sum_{n_g} 0.5 n_g (n_g - 1)$, $GDNT = \sum_g \sum_{n_g} 0.5 n_g (w - n_g)$, $GUMT = \sum_g \sum_{s_g} 0.5 u_{s_g} (n_g - u_{s_g})$ et $GWMT = \sum_s \sum_{t_s} 0.5 v_{t_s} (n_s - v_{t_s})$; où s est le nombre total des pseudo-racines distinctes, s_g est le nombre des pseudo-racines dans un groupe, t_s et n_s sont respectivement le nombre de concepts et de mots réduits à la même pseudo-racine, u_{s_g} est le nombre de mots réduits à la même pseudo-racine dans le groupe *g* et v_{t_s} est le nombre des mots réduits à la même pseudo-racine et appartenant au concept *t*.

Dans le but d'évaluer notre approche, nous avons élaboré un échantillon de vocabulaire basé sur la catégorie des noms extrait de l'ouvrage de (Ameur et al., 2009). Le vocabulaire de cet ouvrage est constitué principalement des formes de pluriel, de l'état libre et d'annexion des noms et des formes de verbes aux quatre aspects. Afin d'éviter de biaiser l'évaluation de notre algorithme, nous n'avons utilisé que la catégorie nom, puisque les verbes ne sont représentés que par une seule forme relative à un aspect. Ainsi, notre échantillon se compose de 2171 formes distinctes, que nous avons regroupé manuellement en 750 concepts où chaque groupe contient la forme du masculin et du féminin en plus des formes de pluriel, de l'état libre et d'annexion si elles existent dans l'ouvrage.

Par l'application de notre algorithme à cet échantillon, nous avons établi 1015 pseudo-racines ce qui nous a permis de réduire la taille de l'échantillon élaboré de 53.2%. Par ailleurs, en se basant sur les mesures de Paice pour le calcul des erreurs de sur-racinisation et sous-racinisation, nous avons abouti à un taux d'erreur de 33.7% pour la sous-racinisation et de 0.4% pour la sur-racinisation. Suite à l'analyse des résultats obtenus par notre algorithme nous avons constaté que le taux d'erreur de la sous-racinisation est dû principalement au pourcentage élevé de 64.6% du pluriel irrégulier par rapport aux formes de pluriel, qui constituent à leur tour 40.7% de l'ensemble des mots distincts de l'échantillon. Par contre, le taux d'erreur de la sur-racinisation est dû essentiellement aux problèmes de l'élimination de certains affixes qui font partie du mot. A titre d'exemple, nous citons le cas du mot ⵓⵏⴼⵢⵏ 'anafγan' *dérivé* qui se termine par le morphème ⵏ 'an' du pluriel.

4 Conclusion

Le présent article s'inscrit dans une stratégie progressive qui vise à encourager le traitement automatique de la langue amazighe à travers l'élaboration d'outils de base et de ressources linguistiques. Ainsi, nous avons entrepris de réaliser un pseudo-racineur de la langue amazighe standard du Maroc. L'approche proposée est simple et légère. Elle se base sur l'élimination d'une liste prédéterminée de préfixes et de suffixes flexionnels, sans recours ni à des dictionnaires ni à des étiqueteurs morphosyntaxiques. Elle permet de regrouper les mots liés sémantiquement. Cependant, l'analyse réalisée a permis de remarquer l'existence de deux types d'erreurs la sous-racinisation et la sur-racinisation. Dans la perspective d'améliorer ce travail, nous visons d'étendre l'approche par l'intégration d'un module de recodage basé sur des règles linguistiques.

Références

AMEUR M., BOUHJAR A., BOUKHRIS F., BOUKOUSS A., BOUMALK A., ELMEDLAOUI M., IAZZI E., SOUIFI H. (2004). *Initiation à la langue amazighe*. Rabat, Maroc : IRCAM.

AMEUR M., BOUHJAR A., BOUKHRIS F., BOUKOUSS A., BOUMALK A., ELMEDLAOUI M., IAZZI E. (2006). *Graphie et orthographe de l'amazighe*. Rabat, Maroc : IRCAM.

AMEUR M., BOUHJAR A., BOUMALK A., EL AZRAK N., LAABDELAOUI R. (2009). *Vocabulaire de la langue amazighe (amazighe-arabe)*. Rabat, 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*. 110-119.

BOUKHRIS F., BOUMALK A., ELMOUJAHID E., SOUIFI H. (2008). *La nouvelle grammaire de l'amazighe*. Rabat, Maroc : IRCAM.

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

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*. 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*. 33-47.

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*. 36-39.

KAMEL S. (2006). *Lexique Amazighe de géologie*. Rabat, Maroc : IRCAM.

LARKEY L. S., BALLESTEROS L., CONNELL M. (2002). Improving stemming for arabic information retrieval: light stemming and cooccurrence analysis. Actes de *the 25th annual international conference on research and development in information retrieval*. 275-282.

PAICE C.D. (1994). An evaluation method for stemming algorithms. Actes de *the 17th* ACM *SIGIR conference*. 42-50.

PORTER M.F. (1980). An algorithm for suffix stripping. *Program* 14, 130-137.

RACHIDI A., MAMMAS M. (2007). Vers un système de traduction automatique en ligne des documents amazighs fondé sur les graphes UNL. *La revue électronique des technologies de l'information* 4.

SAVOY, J. (1993). Stemming of French words based on grammatical categories. *The American society for information science* 44, 1-9.