Correction orthographique pour la langue wolof : état de l'art et perspectives

Alla Lo, El Hadji Mamadou Nguer, N'Diaye Abdoulaye, Cheikh Bamba Dione, Mathieu Mangeot, Mouhamadou Khoule, Sokhna Bao-Diop, Mame-Thierno Cissé

Correction orthographique pour la langue wolof : état de l'art et perspectives

Alla LO[1], El hadji M. NGUER[1], Abdoulaye Y. NDIAYE[1], Cheikh B. DIONE[2], Mathieu MANGEOT[3], Mouhamadou KHOULE[1], Sokhna BAO DIOP [1], Mame T. CISSE[4]

(1) LANI, Université Gaston Berger, Saint Louis, Sénégal
(2) University of Bergen, Norvège
(3) LIG, Université de Grenoble Alpes, 38400 Saint Martin D'HERES, France.
(4) ARCIV, Université Cheikh Anta Diop de Dakar, BP 5005 Dakar-Fann, Sénégal

alcheriawas@yahoo.com,emnguer@ugb.edu.sn,layoussou@yahoo.com, dione.bamba@uib.no,mathieu.mangeot@imag.fr,khoule.mouhamadou@ugb.edu.sn,baosokhna@hotmail.com,thiernoc@gmail.com

RESUME

Les langues nationales des pays d'Afrique de l'ouest sont en général peu dotées d'outils du TAL (Traitement Automatique des Langues). C'est le cas de la langue wolof du Sénégal véhiculaire et majoritairement parlée. Cela constitue un obstacle majeur pour son développement à la hauteur de son utilisation. Ainsi, l'objectif de cet article est de faire l'état de l'art de la correction orthographique et de dégager des perspectives de recherche en vue de mettre en place un correcteur orthographique adapté à cette langue. La mise en place de ce correcteur requiert l'utilisation d'un dictionnaire comme lexique et d'un analyseur morphologique de la langue wolof. Dans la suite du document, nous allons successivement présenter les notions de bases relatives à la correction orthographique, les techniques de détection d'erreurs orthographiques, les techniques de correction d'erreurs orthographiques avant de décrire l'organigramme du correcteur orthographique que nous voulons étudier et mettre en place.

ABSTRACT

The national languages of the West African countries are generally not equipped with tools of NLP (Natural Language Processing). This is the case of the Wolof language of Senegal, which is vehicular predominantly spoken. This is a major obstacle to its development at the height of its use. Thus, the objective of this article is to make the state of the art of spelling and identify research opportunities to develop an adapted to that language. The implementation of this spellchecker requires the use of a dictionary as a lexicon and a morphological analyzer of the Wolof language. In the following document, we will successively introduce the basic concepts related to spelling, orthographic error detection techniques, spelling error correction techniques before describing the flowchart of spellchecker we want to study and implement.

TËNK

Li ëpp ci làkki réewi Afrig gu sowu-jant, dañoo tumránke lool ci jumtukaayi CLO (Càmbar Làkk ak Ordinaatëer). Moo dal làkk bii di wolof bu réewum Senegaal, nga xam moom la ñu fay gën wax te mooy jokkale it askan wépp. Tumránke googu rëq-rëq la bu réy buy tee ag jëfandikoo gu yemook ni ñu koy waxee. Naka noonu, liggéey bii, li ko yékkati mooy wone lépp lu aju ci jubbanti mbind, teg ci leeral naal yees war a gëstu ngir amal jumtukaay bu xarala bu mën di jubbanti mbindum wolof. Boobee jumtukaay nag day laaj baatukaay bu muy xàmmee baati wolof yi, ak jumtukaay buy càmbar meloy baati wolof yi. Ci kanam, dinañu leeral yenn ci xam-xam yi aju ci jubbanti mbind, feemi xàmmee njuumtey mbind, ak feemi jubbanti mbind, laata nuy wone ni, jubbantikaayu mbind bi nu bëgg a amal, war a doxee.

MOTS-CLES: TALN, analyseur morphologique, Correction orthographique, Langue wolof

KEYWORDS: NLP, morphological analyzer, Spell Checking, Wolof language

BAATI SEET: CLO[1], càmbarukaayu meloy baat, Jubbanti mbind, Làkku wolof

1 Introduction

Dans beaucoup de pays francophones d'Afrique de l'ouest à l'instar du Sénégal, l'accès à la formation requiert l'utilisation de la langue française. Cependant, pour des pays comme le Sénégal la grande majorité de la population ne sait ni lire ni écrire le français. Cette situation est à l'origine du manque de formation de la population dans beaucoup de domaines, car la majeure partie des formations sont dispensées dans la langue française.

Par contre, plus de 80% de la population sénégalaise parle le wolof. Aujourd'hui, grâce aux programmes d'alphabétisation mis en place, une frange importante de la population sait lire cette langue en écriture latine ou en écriture Ajami[2]. En plus, les outils de Microsoft, de Firefox ainsi que ceux de Google sont disponibles en wolof. Cela fait de cette langue un atout majeur pour servir d'alternative et de support au français, permettant ainsi de couvrir la quasi-totalité de la population en matière de formation et d'information. Mais, ceci ne pourra se faire réellement qu'en la dotant d'outils efficaces de Traitement Automatique des Langues (TAL) comme les correcteurs orthographiques, etc. Un environnement logiciel adapté, s'appuyant sur des connaissances linguistiques mémorisées dans un lexique, pourrait répondre en partie aux besoins spécifiques en TAL de cette langue. Deux outils sont indispensables à cela : un dictionnaire électronique et un correcteur orthographique. Le premier fait déjà l'objet d'une thèse en cours entre l'université Gaston Berger de Saint-Louis et celle de Grenoble. Il reste le second (correcteur orthographique) dont l'étude présente l'objet de cet article.

Dans ce présent article, nous commençons par faire un petit aperçu de ce qu'est la correction orthographique. Puis, nous décrivons quelques techniques de détection qui en général sont les plus

[1] CLO est le sigle de **C**àmbar **L**àkk ak **O**rdinaatëer duppe ko TAL.

[2] L'alphabet Ajami désigne l'écriture de certaines langues africaines comme l'haoussa, le wolof et le fulfulde avec une variante de l'alphabet arabe.

utilisées. Ensuite, nous parlons des techniques de correction utilisées en correction orthographique. Enfin, nous terminons par l'étude d'un prototype de correcteur pour la langue wolof.

2 Notions relatives à la correction orthographique

Un correcteur orthographique est un logiciel sophistiqué permettant la détection (identification) puis la correction des erreurs orthographiques trouvées dans un texte. Sa mise en place requiert l'utilisation d'un lexique et d'un analyseur morphologique. En linguistique, le lexique d'une langue constitue l'ensemble de ses lemmes ou, d'une manière plus courante mais moins précise, « l'ensemble de ses mots ». Alors que l'analyseur morphologique permet de dériver la construction morphologique d'un mot ou de faire des hypothèses sur un mot inconnu.

Deux approches sont souvent utilisées en correction orthographique : l'approche basée sur un dictionnaire et l'approche stochastique. Suivant les approches, différents outils sont utilisés :

Approche basée sur le dictionnaire : Le correcteur teste chaque mot du texte à traiter dans le lexique du correcteur orthographique, ou dans le cas échéant, vérifie s'il peut être généré à partir des mots du lexique du correcteur. Cette approche considère comme erreur toute forme qui ne corresponde à aucune forme mémorisée dans le lexique du correcteur ou générée à partir des lemmes du lexique. De ce fait, un correcteur orthographique a besoin d'un dictionnaire qui lui sert de lexique, et d'un analyseur morphologique pour vérifier les formes dérivées. Après avoir détecté l'erreur, le correcteur orthographique doit passer à la correction qui consiste à une proposition de mots du lexique du correcteur orthographique les plus proches du mot erroné identifié.

Approche Stochastique (analyse en n-gramme) : dans cette approche le correcteur exploite des données statistiques obtenues à partir de corpus d'apprentissage pour détecter et corriger des erreurs. Il faut aussi noter que la notion d'erreur diffère ici de celle définie dans l'approche basée sur le dictionnaire. Ici, il s'agit d'un seuil et tout mot ayant une probabilité d'erreur qui dépasse ce seuil est considéré comme erroné. L'obtention de cette probabilité sera détaillée dans la suite du document.

Grâce à cette faculté de détection et de correction, les correcteurs orthographiques sont utilisés dans diverses applications par exemple dans la traduction automatique, la recherche documentaire etc.

3 Etat de l'art de la correction orthographique

3.1 Détection

Consultation du dictionnaire : La méthode la plus utilisée en correction orthographique est c'elle basée sur la consultation du dictionnaire (lexique du correcteur). Ce dernier contient l'ensemble des mots de la langue ou du moins les formes de bases. La technique consiste à repérer les mots absents du lexique du correcteur et de les marquer comme erronés. Cependant, garder tous les mots de la langue dans le lexique du correcteur peut rendre la recherche fastidieuse. Dès lors seules les formes de bases sont stockées dans le lexique et d'autres outils tels que les analyseurs morphologiques sont

utilisés pour générer les formes dérivées. Un analyseur morphologique permet de dériver la construction morphologique d'un mot connu ou de faire des hypothèses sur un mot inconnu. La morphologie est traditionnellement divisée en formation des mots (morphologie dérivationnelle ou compositionnelle) et en morphologie flexionnelle (Hacken et al. 2001), (Bouillon et al. 1998), (Nazarenko 2006). La morphologie flexionnelle consiste à décliner un substantif ou un adjectif à partir d'un lexème, tandis que la morphologie dérivationnelle décrit la composition d'un mot à partir de raine et d'affixes, ou encore la création d'un mot à partir d'autres mots. Les outils les plus utilisé pour l'analyse morphologique en correction orthographique sont les lemmatiseurs, les conjugueurs, les déclineurs qui sont généralement réalisés avec des automates.

Analyse par n-gramme : L'analyse en n-grammes est une technique de détection d'erreur qui permet de relever les mots incorrects dans un texte. Il ne compare pas les mots aux mots du dictionnaire mais établit un processus de comparaison au niveau d'une matrice carrée de taille n, qui stocke les n-grammes acceptés qui sont les plus fréquents (E. M. ZAMORA*, J. J. Pollock, A. ZAMORA, 1981) Chaque chaine de caractère du texte est fractionnée dans un ensemble de n-grammes adjacents. Le développement d'un système basé sur les n-grammes suppose avant tout de calculer la probabilité d'erreur de chaque n-gramme pris individuellement. Cette probabilité est tout simplement le nombre de fois qu'il apparait sur des mots erronés sur le nombre total de fois qu'il apparait sur des mots du texte. Il se peut aussi qu'une séquence de n lettres (n-gramme) soit considérée comme valide dans un contexte et comme invalide dans un autre. Cette méthode requiert donc un corpus d'apprentissage. Seul les n-grammes construits à partir des mots incorrects représentent les n-grammes invalides. Un n-gramme, qui est à la fois présent sur un mot et sur sa forme erronée, est considéré comme ne contribuant pas à l'erreur. La probabilité d'erreur (*P*) d'un n-gramme est donnée par la formule suivante $P = \frac{E}{E+V}$.

Où :

- ***E*** est le nombre de fois où il a été classé comme un n-gramme invalide ;
- ***V*** est le nombre de fois où il a été classé comme n-gramme valide.

La méthode d'analyse basée sur les n-grammes suppose de choisir un seuil. Ainsi, tout mot contenant des n-grammes avec une probabilité d'erreur qui dépasse ce seuil, sera marqué comme erroné. Ceci permet de détecter les mots incorrects du texte. Cependant, le choix de ce seuil est crucial, dès lors qu'une erreur affecte plus d'un n-gramme (au moins trois trigrammes contigus sont modifiés par le changement d'une lettre sur une chaine). Un seuil trop petit fait que le système laisse passer des mots incorrects. Par contre, un seuil trop grand entraine beaucoup de fausses alertes. Le choix des n-grammes permet de contrôler la taille du lexique et de la maintenir à un seuil raisonnable. Un lexique obtenu suite à un découpage en n-gramme de caractère ne peut dépasser la taille de l'alphabet à la puissance n. Cependant son développement requiert d'importantes ressources linguistiques.

3.2 Correction

L'étape qui suit la détection d'erreurs dans un processus de correction orthographique est la génération de mots susceptibles d'être la forme correcte du mot mal orthographié. Les techniques de correction étudiées ci-dessous se focalisent sur le mot mais pas sur le contexte dans lequel le mot doit être utilisé. La correction s'exécute en deux étapes :

- La génération de mots susceptibles d'être la bonne orthographe.
- La génération des rangs des mots proposés dans la première étape.

La deuxième étape permet de filtrer et d'ordonnancer les résultats de la première étape.

Alpha-code : Pour retrouver les formes les plus proches du mot inconnu, la plupart des vérificateurs orthographiques utilisent une représentation codée d'un mot. Cette technique par codage de chaines est aussi connue sous le nom d'alpha-code. L'alpha-code d'un mot est la chaîne de caractères constituée de 1'ensemble de ses lettres classées par ordre alphabétique (Garder 1992). Elle détient différentes méthodes de calcul ; (Ndiaye et al. 2003) ont proposé une méthode qui consiste à ranger les consonnes suivies des voyelles par ordre alphabétique, respectivement. L'ensemble des mots relatifs à un alpha-code donné est appelé classe de cet alpha-code. Le système de correction dispose de 1'ensemble des alpha-codes correspondant au lexique et, pour chacun, de la classe de mots. Le correcteur établit tout d'abord 1'alphacode du mot erroné. Si cet alpha-code est répertorié, les mots de la classe correspondante sont retenus comme candidats. Si l'alpha-code est inconnu, le système y ajoute un, puis deux caractères, et vérifie si le nouvel alpha-code obtenu existe. Dans ce cas, la classe est retenue. II procède de même en supprimant puis en substituant un ou deux caractères. Les mots candidats obtenus font alors 1'objet d'un calcul de proximité avec la chaine incorrecte, ce qui autorise un classement de cette liste de mots correctifs. Un mot ne possède qu'un seul alpha-code mais plusieurs mots peuvent partager le même alpha-code. Ainsi, un vérificateur orthographique pourra rechercher les mots qui contiennent le même alpha-code que le mot inconnu.

Distance lexicographique (Levenshtein et Damerau-Levenshtein) : La distance de Levenshtein, encore connue sous le nom de « distance d'édition » ou de « déformation dynamique temporelle », est une métrique entre deux chaines de caractères. Elle est égale au nombre minimal de caractères qu'il faut supprimer, insérer ou remplacer pour passer d'une chaine à l'autre (Levenshtein, 1966). Cette distance est d'autant plus grande que le nombre de différences entre les chaines est grand. La distance de (Damerau-Levenshtein, 1964) est similaire à celle de Levenshtein. C'est-à-dire que c'est une métrique qui calcule le nombre minimal d'opérations pour transformer une chaine en une autre. Cependant celle de Levenshtein calcule la distance d'édition en considérant seulement trois opérations à savoir la substitution, l'insertion et la suppression. Or, d'après notre typologie des erreurs, une faute d'orthographe peut aussi être causée par la transposition de deux caractères successifs. C'est dans cet optique que la distance de Damerau-Levenshtein amène un apport important car non seulement il inclut les opérations citées par Levenshtein (insertion, substitution, suppression) mais ajoute à cela la transposition.

N-gramme : Comme indiqué précédemment, la détection d'erreur par n-gramme permet de connaitre la position de l'erreur dans le mot. Cette information facilite la correction de l'erreur. La méthode de correction proposée par (David Sundby) s'appuie sur cette position pour générer une liste de suggestion pour le mot erroné. Son algorithme est basé sur les quatre opérations d'éditions (Damerau, 1964) à savoir l'insertion, la suppression, la substitution et la transposition pour rechercher des formes correctes pouvant correspondre au mot mal orthographié.

Insertion : C'est un algorithme qui corrige les mots erronés avec un caractère manquant. Le principe de l'algorithme consiste à insérer une lettre de l'alphabet à la position où l'erreur s'est produite et de vérifier si le mot obtenu est correct avant de l'ajouter dans la liste des mots candidats. Le procédé est réitéré pour toutes les lettres de l'alphabet.

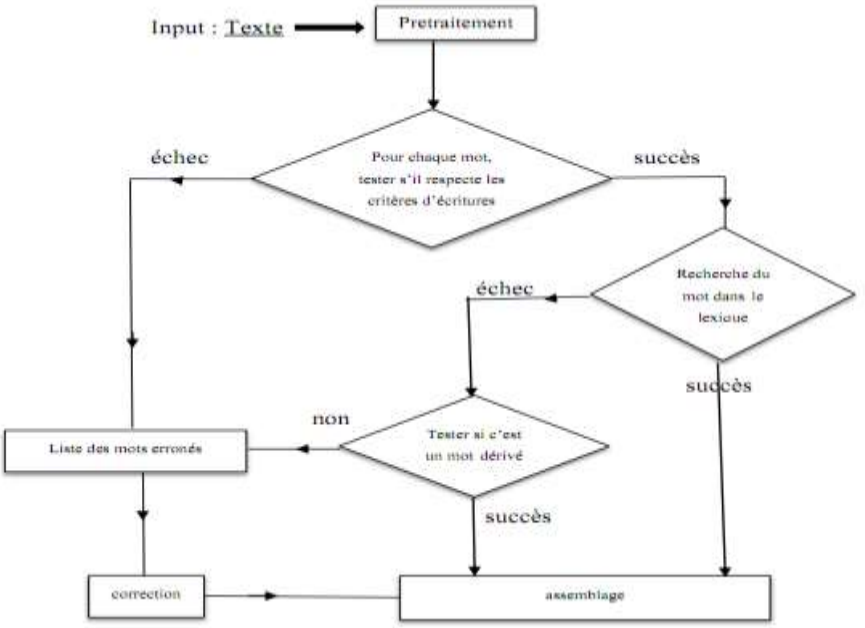
Suppression : C'est un algorithme qui supprime des caractères sur un mot erroné. L'algorithme supprime un caractère du mot et vérifie si le mot formé est correct avant de l'ajouter dans la liste de suggestion. Le procédé est réitéré pour toutes les lettres du mot.

Substitution : L'algorithme de substitution remplace chaque lettre du mot, les unes après les autres par une lettre de l'alphabet et de vérifier si le mot obtenu est correct avant de l'ajouter dans la liste de suggestion. On répète le même procédé pour toutes les lettres de l'alphabet.

Transposition : L'algorithme change la position d'un caractère du mot en le mettant dans toutes les autres positions et à chaque fois vérifie si le mot formé est correct avant de l'ajouter dans la liste de suggestion. Le procédé est réitéré pour toutes les lettres du mot.

4 Etude d'un correcteur orthographique pour la langue wolof

Organigramme du correcteur : Un correcteur orthographique fonctionne en suivant plusieurs étapes : détection des erreurs, sélection des corrections possibles, filtrage et ordonnancement des corrections et proposition à l'utilisateur, correction effective du texte respectant le choix de l'utilisateur. Puisque corriger un texte revient à corriger tous les mots de ce texte, un traitement préliminaire consistant à segmenter le texte en mots (Word Tokenization) est nécessaire. Cette tache de segmentation est assurée par un algorithme.



La détection des erreurs s'effectue souvent en considérant un à un les mots du texte à corriger, de manière isolée et vérifier pour chacun d'eux s'il est dans le dictionnaire ou s'il peut être généré à partir des formes enregistrées dans le lexique du correcteur.

Chacun des mots du texte soumis dans un premier temps à un test du respect des critères d'écritures. En effet, les lexèmes se trouvent souvent sous la forme CVC pour les monosyllabes et CVCV(C) pour les dissyllabes (Robert, 2011). En wolof il y'a 20 consonnes faibles (p, t, c, k, q, b, d, j, g, m, n, ñ, ŋ, f, r, s, x, w, l, y), 25 consonnes fortes (16 géminées : pp, tt, cc, kk, bb, dd, jj, gg, ŋŋ, ww, ll, mm, nn, yy, ññ, qq et 9 prénasalisées : mp, nt, nc, nk, nq, mb, nd, nj, ng), 9 voyelles brèves (a, i, o,

ó, u, e, è, ë, e, à) et 7 voyelles longues (ii, uu, éé, óó, ee, oo, aa). La consonne et la voyelle finale d'une syllabe peuvent être longues mais les deux ne le peuvent pas en même temps. Les consonnes fortes n'apparaissent jamais après une voyelle longue ni à l'initial du mot sauf pour les prénasalisées. Si le test ne passe pas, c'est qu'il y'a erreur, et requiert une correction. Sinon on passe à la seconde phase qui consiste à rechercher le mot dans le lexique du correcteur. Si le mot se trouve dans le lexique, il est accepté et le test du prochain mot s'ensuit. Sinon s'opère une troisième phase. Cette dernière consiste à vérifier si le mot n'est pas une forme dérivée d'un mot du lexique. Si oui le mot est accepté et on passe au prochain mot. Sinon il est considéré comme erroné et sera également soumis à une correction. Lorsqu'une erreur est détectée, le correcteur sélectionne une série de mots susceptibles d'être la version correcte du mot à corriger. Ces mots sont choisis selon différentes techniques (codage de chaine, n-gramme, mesure de distance entre chaines). L'ordonnancement des chaines candidates à la correction prend en compte la mesure utilisée lors de l'étape de sélection, ainsi que des mesures statistiques (comme les fréquences d'apparition des mots, ou bien le mot le plus fréquemment trouvé lors de rencontres préalables avec la même erreur). Enfin, une étape interactive permet à l'utilisateur de superviser la correction. Il peut adopter l'une des trois attitudes suivantes : corriger le mot erroné en sélectionnant un des candidats proposés par le correcteur, modifier le mot erroné, ne pas corriger ; dans ce dernier cas, il peut rajouter ce mot à son dictionnaire personnel.

```
Debut Programme

 Lire (TEXTE)

 Tant que non fin de TEXTE A CORRIGER faire

  Recupérer (MotRechercher)

  Correct <-VerificationCritereDecriture (MotRechercher)

   Si correct

    Correct <- Rechercher(MotRechercher ,Dictionnaire)

     Si non Correct

     Correct <-verficationDeFlexion(MotRecherche)

     Si non Correct

      Corriger(MotRechercher)

     FinSi

   FinSi

   Sinon

    Corriger(MotRechercher)

   FinSi
```

```
  FinTantque

 Fin
```

Une partie importante dans la détection de ces erreurs est l'analyse morphologique. Comme indiqué dans la partie qui concerne l'état de l'art, ne sont enregistrées dans le dictionnaire que les formes de bases des mots de la langue. Ainsi l'outil qui se chargera de vérifier si la chaine soumise au processus de détection est une chaine dérivée d'une forme connue ou non est ce qu'on appelle un analyseur morphologique.

Apres avoir détecté une faute, c'est l'algorithme de correction qui se chargera de trouver les mots susceptibles d'être la bonne correction. La recherche de ces mots se fait d'abord avec les alpha-codes pour trouver tous les anagrammes du mot. Ensuite, on applique l'algorithme de Damerau-Levenshtein (distance d'édition) pour faire un filtrage et un ordonnancement des mots candidats, c'est-à-dire les k (par exemple k=5) mots ayant les distances d'édition les plus petites seront suggérés pour correction, en les rangeant suivant leur distance; les plus petite en premières.

```
Corriger(Mot motACorrige)

  listeDeSugestion<-correction(MotACorrige)

  Afficher(listeDeSuggestion)

   formeCorrecte<-choixUtilisateur(listeDeSuggestion)

    Si (formeCorrecte !=null)

      remplacer(MotACorrige,formeCorrecte)

    FinSi

Fin
```

Après que l'utilisateur ait effectué ses choix, le programme assemblera les mots en texte de sortie en respectant leur ordre. Il faut aussi noter que des modules complémentaires permettront à l'utilisateur d'ajouter des mots dans un dictionnaire personnel. En effet, en montrant la liste de suggestion à l'utilisateur, une zone qui permet d'ajouter le mot dans le dictionnaire personnel sera aménagée dans la boite de dialogue. Ceci permettra au correcteur, dans les utilisations ultérieures de ne plus marquer ce mot comme erroné.

Il faut aussi noter que nous avons déjà débuté à travailler sur l'analyseur morphologique. Cependant nous en sommes à la phase de dés-affixation. Pour cela nous avons mis en œuvre un transducteur qui permet d'enlever des affixes d'un mot wolof. En effet la préfixation est assez rare en wolof mais n'empêche qu'elle existe. Cependant les rares cas observés ne présentent qu'un seul préfixe (exemple : *kaddu* (parole) = *k+addu* (parler)). De l'autre côté, pour la suffixation, il peut y'avoir une concaténation de plusieurs suffixes. Le transducteur cherche si le mot commence par un préfixe, s'il en trouve un, il l'enlève. Ensuite il continue à parcourir le mot jusqu'à la rencontre d'un suffixe, il l'enlève et continue jusqu'à ce qu'il n'en trouve plus. Il faut aussi noter que l'élimination des

suffixes se fait de la droite vers la gauche en enlevant toujours le plus long suffixe trouvé. Ainsi la chaine de sortie du transducteur sera testée dans la base des lemmes du correcteur. L'automate n'est pas déterministe mais il s'agit d'accorder un priorité supérieur à la transition (q0,pref,q1) qu'à celle de (q0,$,q1) (avec $ qui représente le vide); c'est-à-dire il faut vérifier si l'on peut appliquer (q0,pref,q1) avant de chercher à appliquer (q0,$,q1).

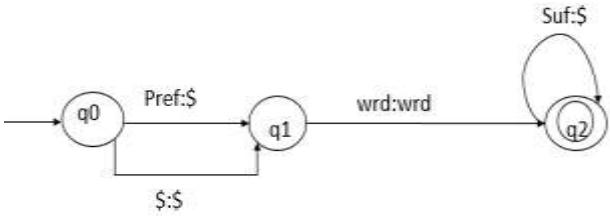


Figure 1 transducteur de la désaffixation

Par exemple pour le mot *dawalkat* (*daw+al+kat*) qui est composé du lemme *daw* suivis des suffixes *al* et *kat*, le transducteur ne trouve pas de préfixe, mais va trouver le suffixe *kat* qu'il enlève en premier, pour ensuite enlever le suffixe *al*. Apres ceci il renvoie la chaine *daw* qui est un lemme reconnu du correcteur.

5 Conclusion

La langue wolof, qui est véhiculaire et majoritairement parlée au Sénégal, est dépourvue d'outils du TAL (Traitement Automatique des Langues) comme les correcteurs orthographiques. Ce qui constitue un obstacle majeur pour son développement à la hauteur de son utilisation. Ainsi, l'objectif de cet article est de faire l'état de l'art de la correction orthographique et de dégager des perspectives de recherche en vue de mettre en place un correcteur orthographique pour cette langue.

Dans ce travail nous avons fait l'état de l'art de la correction orthographique. Précisément nous avons évoqué les notions de bases relatives à la correction orthographique, les techniques de détection d'erreurs orthographiques, les techniques de corrections d'erreurs orthographiques et proposé l'organigramme d'un correcteur orthographique pour la langue wolof.

En perspectives ce travail sera utilisé pour la mise en œuvre d'un correcteur orthographique pour la langue wolof qui requiert l'utilisation d'un dictionnaire comme lexique et d'un analyseur morphologique. Ainsi, il sera utilisé dans la suite l'analyse morphologique (Dione, 2012) et le dictionnaire (Khoulé et al., 2016) en cours de réalisation à travers le projet ibaatukaay de mise en place d'une base lexicale multilingue contributive sur le web pour les langues africaines notamment sénégalaises.

Remerciements

Nous remercions le Centre d'Excellence Africain en Mathématiques, Informatique et TIC pour son soutien à ce projet de recherche.

Référence

BOUILLON, PIERRETTE, VANDOOREN, FRANCOISE ET LEHMANN, SABINE (Eds.)(1998). *Traitement automatique des langues naturelles*. Universites francophones. Champs linguistiques. Recueils. Bruxelles: Duculot.

DAMERAU, FRED J. (1964). *A Technique for Computer Detection and Correction of Spelling Errors*. Communications of the Association for Computing Machinery, 7(3):171–176.

DIONE C. B. (2012). *A Morphological Analyzer For Wolof Using Finite-State Techniques.* In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry De clerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk and Stelios Piperidis (eds.), *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey: ELRA.*

DIONE C.B. (2014). *Formal and Computational Aspects of Wolof Morphosyntax in Lexical Functional Grammar*. Thèse de Nouveau Doctorat. Univesité de Bergen, Norvège.

E. M. ZAMORA,* J. J. POLLOCK and Antinio ZAMORA (1981), *The use of trigram analysis for spelling error detection*. Information Processing & Management Vol. 17 No. 6. pp.305-316.

GARDER NABIL (1992), *Conception et réalisation d'un prototype de correcteur orthographique de l'arabe)*. 83p, 1992.

HACKEN T., PIUS & TSCHICHOLD, CORNELIA (2001): *Word Manager and CALL: Structured access to the lexicon as a tool for enriching learners' vocabulary*, ReCALL 13: 121-131.

LEVENSHTEIN, VLADIMIR I. (1966). *Binary codes capable of correcting deletions, insertions, and reversals.* Soviet Physics Doklady, 10(8):707–710.

NAZARENKO, ADELINE (2006). *Le point sur l'état actuel des connaissances en traitement automatique des langues (TAL). In* Sabah, Gerard (Ed.). *Compréhension des langues et interaction*, Cognition et traitement de l'information, pp. 31–70. Paris: Hermes Science, Lavoisier.

NDIAYE, MAR ET VENDEVENTER F., ANNE (2003) A Spell Checker Tailored to Language Users. Computer Assisted Language Learning (CALL): An international Journal, 16(2-3): 213-232.

ROBERT S. (2011) (remis 2002). In Emilio Bonvini, Joëlle Busuttil & Alain Peyraube (sous la dir.). *Dictionnaire des Langues. Paris* : Quadrige/ P.U.F, 23-30. (Version non corrigée).

Silberztein, Max et Tutin, Agnès (2004*). NooJ: un outil de TAL decorpus pour l'enseignement des langues et de la linguistique. Une application à l'étude des impersonnels. In Journée d'étude de l'ATALA "TAL & Apprentissage des langues" (TAL&AL):* Actes, pp.47–56, Grenoble: LIDILEM: ATALA XRCE.

SUNDBY D. *Spelling correction using n-gram* "Lund institute of technology, Sweden david.sundby@gmail.com"