Récital 2009, Senlis, 24-26 juin 2009

La distance intertextuelle pour la classification de textes en langue arabe

AYADI Rami (1), JAOUDI Walid (2)

(1) UTIC (monastir) – ISIMS (sfax) ,Tunisie
ayadi.rami@planet.tn

(2) UTIC (Tunis) , Tunisie
walidjaouadi@yahoo.fr

Résumé : Nos travaux de recherche s'intéressent à l'application de la théorie de la distance intertextuelle sur la langue arabe en tant qu'outil pour la classification de textes. Cette théorie traite de la classification de textes selon des critères de statistique lexicale, se basant sur la notion de connexion lexicale. Notre objectif est d'intégrer cette théorie en tant qu'outil de classification de textes en langue arabe. Ceci nécessite l'intégration d'une métrique pour la classification de textes au niveau d'une base de corpus lemmatisés étiquetés et identifiés comme étant des références d'époques, de genre, de thèmes littéraires et d'auteurs et ceci afin de permettre la classification de textes anonymes.

Abstract: Our researche works are interested in the application of the intertextual distance theory on the Arabic language as a tool for the classification of texts. This theory handles the classification of texts according to criteria of lexical statistics, and it is based on the lexical connection approach. Our objective is to integrate this theory as a tool of classification of texts in Arabic language. It requires the integration of a metrics for the classification of texts using a database of lemmatized and identified corpus which can be considered as a literature reference for times, genres, literary themes and authors and this in order to permit the classification of anonymous texts.

Mots-clés : Distance intertextuelle, arabe, classification, lemmatisation, corpus, statistique lexicale.

Keywords: Intertextual distance, Arabic, classification, lemmatization, corpus, lexical statistics.

1 Introduction

L'abondance de l'information, due au développement de l'Internet, des supports de stockage (contenant de larges corpus textuels) et des encyclopédies numérisées, rend difficile, voir impossible, son exploration et analyse. D'où naît le besoin d'explorer de nouvelles approches automatiques d'aide à l'analyse des textes.

Plusieurs travaux se sont intéressés à la mise en place d'un certain nombre d'outils pour l'aide à l'analyse, bien que suivant différents axes. Nous citons les travaux de Lebart et Salem (Lebart, 1988), Salton (Salton, 1989) et Reinhart (Reinhart,1994) qui ont exploré l'utilisation des classificateurs pour la structuration de l'information au niveau des textes sous forme de réseaux lexicaux et sémantiques. D'autres travaux ont repris les précédents classificateurs pour en exploiter les résultats, tel que les travaux de Deerwester (Deerwester, 1990) traitant de la sémantique latente et plus précisément l'exploration dans la catégorisation automatique des textes en associant ces classes lexicales à des plans généraux de classification. C'est dans cet axe de classification de textes que s'inscrivent nos travaux de recherche. Nous nous proposons dans ce cadre, d'aborder la problématique de catégorisation et d'indexation de textes en langue arabe. Pour ce faire, nous partons des constations et des résultats de classificateurs basés sur des approches statistique et mathématique.

La classification de textes est définie comme une opération qui identifie des classes d'équivalence entre des segments de textes en tenant compte de leur contenu informationnel (mots, n-gram, etc.). De ce fait nous sommes amenés à définir le degré de ressemblance ou de dissemblance entre les segments considérés. En nous situant dans le cadre d'une approche statistique et mathématique, nous avons la possibilité d'expliciter ce degré par des indices numériques. Plusieurs approches permettent de définir chacune une méthode pour calculer ce degré de ressemblance ou dissemblance; Notamment la théorie de la distance intertextuelle qui a été introduite par Charles Muller (Muller, 1977) à la fin des années soixante sous le nom de connexion lexicale pour répondre au besoin de définir une métrique pour la mesure du degré de parenté ou similitude entre textes. Cette théorie se base sur des statistiques relatives aux vocabulaires et expressions employés dans les textes, ce qui crée un besoin de décomposition selon un jeu de classes grammaticales des textes et ceci par le biais de lemmatiseurs.

Nous nous proposons de mettre en place un classificateur doté d'une métrique et d'un outil pour la préparation des textes aux différents traitements (un lemmatiseur) (Jaouadi, 2006). En raison des spécificités de la langue arabe (caractère agglutinatif, richesse des classes grammaticales, etc.), nous sommes amenés à concevoir une métrique et un lemmatiseur ayant un caractère générique, présentant un degré d'abstraction par rapport au choix des classes grammaticales lors de la segmentation ou découpage du texte.

S'ajoute à cela un besoin de mettre en place un référentiel pour la classification des textes sous forme de corpus structurés. Ce référentiel renferme des corpus définis et catégorisés servant comme repère pour la classification. Cette définition du processus nous a permis de réaliser une plateforme de classification de textes en langue arabe. Cette plate-forme présente un lemmatiseur générique à base d'apprentissage qui permet d'étiqueter des textes en laissant à l'utilisateur le choix du jeu d'étiquettes à employer. Elle offre aussi un classificateur utilisant une version améliorée de la formule issue de la théorie de la distance intertextuelle qui permet d'associer au jeu d'étiquetage la notion de poids. Pour finir nous avons associé à cette plate

forme une structure permettant d'accueillir le référentiel de classification sous forme de corpus catégorisés et sélectionnés.

2 Architecture du système de classification

Le système que nous réalisons a pour but essentiel de permettre la classification de textes en langue arabe dans un but de catégorisation et d'indexation. Pour ce faire nous nous proposons de définir un processus de traitement permettant d'avoir en entrée un texte brut et de présenter en sortie la catégorisation de ce dernier. Cette catégorisation peut se faire par rapport à une référence existante ou par rapport à un autre texte en entrée. Notre utilisation de la théorie de la distance intertextuelle pour la mise en place d'une métrique de classification nous contraint à l'intégration d'un processus de lemmatisation des textes (Prétraitement). Cette étape est nécessaire car elle prépare les textes en les décomposant ce qui permet l'exploitation des structures grammaticales dans la détection des classes d'équivalence entre segments de textes. Nous avons exploité la richesse de la grammaire de la langue arabe pour intégrer la notion de classes grammaticales au niveau du lemmatiseur ainsi que dans la métrique, de cette manière le lemmatiseur fonctionne indépendamment du jeu de la structure adopté et nous introduisons des poids associés aux classes grammaticales au niveau de la métrique.

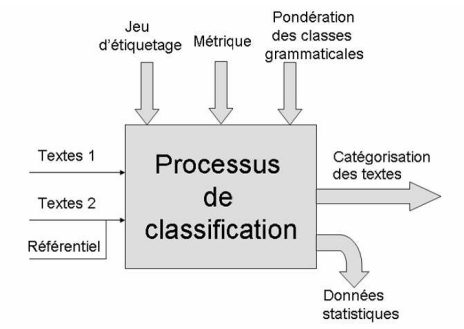

Figure 1 : Définition du processus

2.1 Un lemmatiseur hybride générique à base d'apprentissage

2.1.1 Outil de lemmatisation :

Nous parlerons de système et non d'application en raison de la complexité des interactions entre les différents composants. En effet, notre système est une hybridation de différentes approches hétérogènes collaborant ensemble dans un système homogène et transparent par apport à son environnement.

Nous avons construit notre lemmatiseur en nous inspirant des différentes approches déjà existantes et nous avons essayé de les intégrer pour mettre en place un système performant. Nous avons retenu les notions suivantes :

- Reconnaissance de forme : Chaque mot en cours de traitement subit une recherche au niveau de la base de données de référence pour voir s'il ne représenterait pas :
  - Un cas particulier (nom propre, chiffre, préposition, adverbe,…)
  - Une racine ou un lexique
  - Un cas déjà traité par le lemmatiseur

Ces éléments sont classés dans la base de données sous forme de modèles. Par comparaison, il est possible de rattacher un mot à un modèle et d'extraire ainsi sa décomposition.

- Algorithme de lemmatisation : Nous notons que la structure de la langue impose l'algorithme à utiliser. Dans le cas de la langue arabe, nous nous sommes inspirés de l'algorithme utilisé par Tim Buckwalter pour réaliser des analyses morphologiques sur des transcriptions de textes en langue arabe. Notre algorithme se base sur le principe de la génération à partir d'un mot des différentes possibilités de préfixes, suffixes et corps schématique puis ne garder que le trio qui appartient à une base de référence c'est-à-dire dont le préfixe, le suffixe et le corps schématique appartiennent à la langue.
  Le principe pour des textes écrits en arabe reste plus ou moins le même à part quelques changements et enrichissements que nous avons apportés au niveau des préfixes et suffixes, en effet, par consultation d'un expert nous avons pu reconstruire l'ensemble des préfixes et suffixes de la langue arabe que nous avons implémenté en caractère arabe pour permettre la reconnaissance par rapport à des textes non transcris. Par rapport à la structure, très répandue, en cinq composantes des mots en langue arabe (à savoir antéfixe, préfixe, corps schématique, suffixe, postfixe), nous supposerons dans notre algorithme que la combinaison (antéfixe+ préfixe) forme un élément à savoir le préfixe et la combinaison (suffixe+ postfixe) forme un suffixe. Cette fusion permet de minimiser le nombre de décompositions possibles sans pour autant en diminuer la précision. Le fait de fusionner ainsi des éléments diminue le nombre de possibilités générées à partir d'un seul mot, de plus le fait qu'on connaît les éléments fusionnés on peut après leur détection les décomposer plus facilement. L'exemple de la **figure 2** illustre ce jeu de combinaisons.

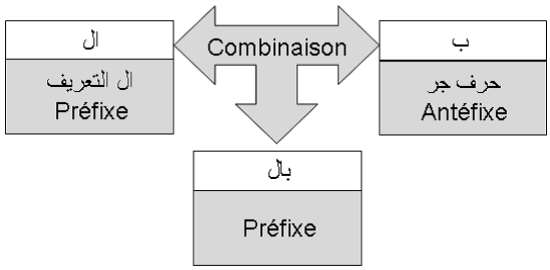

Figure 2 : Combinaison d'éléments

- La lemmatisation manuelle : Il est très difficile d'arriver à automatiser la lemmatisation, l'intervention humaine reste inévitable que ce soit au niveau applicatif ou niveau de la vérification. Pour cela, nous ne manquons pas de noter que la lemmatisation manuelle est le dernier recours de notre système. Un des problèmes posés par les approches précédemment citées est le fait que par leur traitement elles puissent générer plusieurs décompositions possibles. A ce niveau aussi on note l'intervention humaine qui oriente la résolution soit par un choix soit par l'appel à une des autres approches intégrées dans notre système.

Les méthodes de lemmatisation retenues, à savoir la lemmatisation à base de modèles, manuelle et à base d'algorithme, même si différentes en apparence, représentent les piliers de notre système, nous avons mis en place des interfaces permettant la communication entre elles

ainsi qu'un ordonnancement permettant l'optimisation de leur utilisation. L'intégration des trois approches se fait de manière séquentielle l'une après l'autre, lorsqu'une s'endigue l'autre prend le relais. Il est à noter que chacune de ses étapes interagit avec la base de connaissance que ce soit en consultation, ou mise à jour.

2.1.2 **La base de données références :**

En nous basant sur les besoins en données exprimées au niveau du système précédemment détaillé nous avons mis en place la structure de la base de données de référence.
Les éléments de base : Cette base de données constitue le noyau du système que nous avons réalisé en collaboration avec un expert de la langue arabe, elle inclue :
- Un lexique : il englobe deux entités à savoir les mots non décomposables en préfixe, suffixe et corps schématique et les racines (ou corps schématique nécessaire à l'algorithme de décomposition). Notre source initiale au niveau de l'alimentation du lexique est un fichier XML contenant un dictionnaire ayant pour origine les travaux de Belgacem Mohamed et Mars Mourad.
- Les cas particuliers : Nous aurions pu inclure cette classe au niveau du lexique mais il est plus intéressant de les séparer pour faciliter le traitement. Elle inclue les prépositions les noms propres, les articles…Nous avons essayé de déterminer l'ensemble de ces éléments par collaboration avec un expert de la grammaire arabe qui nous a permit de les synthétiser (jaouadi 2006).
- Les cas traités : représente les mots déjà traités que ce soit par décomposition ou par ajout de l'utilisateur. Ils sont stockés avec leur décomposition de façon à être utilisé par le module de reconnaissance de formes.

- Préfixes et suffixes : sont des éléments indispensables à l'algorithme de lemmatisation. Nous avons pu avec l'aide de l'expert déterminer un ensemble de 77 préfixes et de 165 suffixes au niveau de la langue arabe.

2.2 Une formule de calcul basée sur la pondération des classes grammaticales

La distance intertextuelle est la mesure de degré de ressemblance ou de dissemblance entre textes.

Pour pouvoir dire si des textes sont « plutôt proches » ou « plutôt éloignés » par rapport a l'utilisation d'un vocabulaire commun, on cherche à exprimer par un nombre, de propriétés – la ressemblance ou la dissemblance – qui ne sont pas des nombres. De plus, pour pouvoir répéter cette mesure autant de fois que l'on voudra, il faudra encore transformer la mesure absolue en un indice qui doit répondre à l'ensemble des propriétés suivantes (Labbé, 2003):

On recherche donc un indice D :

- insensible aux différences de taille entre les textes comparés : la taille du texte ne doit pas influer sur l'indice.

- applicable à plusieurs textes et, potentiellement, à tous les textes d'une même langue

- variant uniformément : entre 0 (même vocabulaire et fréquence semblable de chacun des mots dans les deux textes) et 1 (aucun vocable en commun) et ceci sans saut ni effet de seuil autour de certaines valeurs.

- symétrique (soit deux textes A et B alors D (A, B) = D (B, A)).

- Vérifiant l'inégalité triangulaire : D (A, B) ≤ D (A, C) + D(C, B).Ce qui revient à dire que le chemin entre deux individus sera toujours plus long si l'on passe par un troisième (a moins que ce troisième soit confondu avec l'un des deux premiers).

- L'indice doit être « transitif » : quand on agrège le vocabulaire de deux textes, les distances de ce nouveau vis-à-vis des autres textes doit refléter l'ordre des distances

  antérieures (si D(A,B) > D(A,C) > D(B,C) alors D(A,B) > D{A,(B*C)]) .

- Aussi "robuste" que possible (i.e. une modification marginale dans le vocabulaire d'un des deux textes doit se traduire par une variation marginale de l'indice)…

Les conditions énoncées précédemment sont une synthèse réalisée par D.Labbé (labbé, 2003) de la théorie énoncée par C.Muler (Muller, 1977) et que les formules qu'on trouve dans la littérature ne répondent pas toutes à toutes les conditions.

2.2.1 Formulation mathématique de la distance intertextuelle

Nous ne tiendrons compte que de la formulation de Labbé étant donné qu'elle représente la dernière amélioration des formules issue de la distance intertextuelle.

Soit deux textes A et B

Va et Vb : nombre de vocables dans A et B (vocabulaire)
Fia : fréquence du vocable i dans A
Fib : fréquence du vocable i dans B
Na et Nb : nombre de mots dans A et B (taille) avec Na = Σ Fia et Nb = Σ Fib
$V'_{b(E)}$: Nombre réduit de vocable du texte B
F_{ia} : Fréquence du $i^{ème}$ terme du vocabulaire du texte A
$E_{ia(u)}$: Fréquence réduite du $i^{ème}$ terme du vocabulaire du texte A dans le texte B
$U_{(a,b)}$: rapport entre les tailles des textes A et B
N'_b : nombre de mots réduits du texte B : $N'_b = \sum_{V_b} E_{ia(u)}$

- **Distance absolue $D_{Va,b(u)}$** : entre le vocabulaire du texte A et B' (B'=la réduction de B à la taille de A)

(1)

$$D_{Va,b(u)} = \sum_{V_a, V'_{b(E)}} \left| F_{ia} - E_{ia(u)} \right|$$

$$E_{ia(u)} = F_{ib} * U_{(a,b)}$$

- **Distance relative $D_{(a,b)}$:** Quand A et B n'ont aucun mot en commun, la distance entre eux sera égale à $N_a + N'_b$. Cette somme sera naturellement placée au dénominateur de la formule de l'indice de la distance (1) : ainsi la valeur maximale sera égale à 1 et, l'indice sera nécessairement inférieur à 1 quand l'intersection des deux textes ne sera pas vide (ce qui sera toujours le cas lorsqu'ils sont écrits dans la même langue).

(2)

$$D_{(a,b)} = \frac{\sum_{V_a, V'_{b(E)}} \left| F_{ia} - E_{ia(u)} \right|}{\sum F_{ia} + \sum E_{ia(u)}} = \frac{\sum_{V_a, V'_{b(E)}} \left| F_{ia} - E_{ia(u)} \right|}{N_a + N'_b}$$

$$E_{ia(u)} = F_{ib} * U_{(a,b)} \qquad N'_b = \sum_{V_b} E_{ia(u)} \qquad U(a,b) = \frac{N_a}{N_b}$$

2.2.2 Motivation et nouvelle orientation

Notre approche de la théorie de la distance intertextuelle est essentiellement motivée par un besoin de définir une métrique pour la classification de textes en langue arabe. Ceci étant notre point de départ, nous avons considéré le problème de ce point de vue pour reformuler la problématique afin qu'elle réponde plus à nos besoins. Nous en avons retenu les points suivants :

- La formule de Labbé est notre point de départ
- La distance intertextuelle se base essentiellement sur les vocabulaires respectifs des textes à comparer.
- Un vocabulaire est l'ensemble de mots utilisés pour générer un texte. Ces mots appartiennent à différentes classes grammaticales
- Le principe de la formule de Labbé est de comparer les fréquences de vocabulaires dans les deux textes indépendamment de la nature de vocabulaire.

Nous nous posons alors une question très importante à nos yeux à savoir : est ce que toutes les classes grammaticales ont la même pertinence dans le vocabulaire ?

Nous pouvons reformuler en :

(i) Est ce qu'elles ont toutes les classes grammaticales ont la même pertinence dans l'écriture de texte ?

(ii) Pourrions-nous définir un certain nombre de classes comme étant un parasite vocabulaire ?

Nous nous sommes aperçu, à travers quelques expériences qu'en éliminant un certain nombre de classes tel que les chiffres et les noms propres, que l'indice de comparaison s'améliorait de plus en plus. D'où cette idée de parasite au niveau du vocabulaire. Et à travers nos lecture nous avons remarqué qu'au niveau d'autres méthode de classification et d'indexation la notion existe déjà comme par exemple dans les travaux sur les n-grammes de Radwan Jalam et Jean-

Hugues Chauchat (JALAM, 2002) ou se terme revient souvent pour définir des mots contenants par hasard un des n-grammes caractéristiques de la classe, sans que le mot lui-même soit intéressant. Dans notre cas nous allons appeler parasite toute composante susceptible de fausser le résultat par sa non pertinence ou par le fait qu'elle n'est pas représentative du style lexical de l'auteur.

Notons par exemple que les noms propres et les chiffres peuvent être considérés comme non représentatif du style lexical d'un auteur. Plus concrètement, en comparant le même texte écrit par la même personne décrivant deux personnages ne donnera pas un zéro ce qui est le résultat logique car les noms et les dates diffèrent ce qui fausse le résultat.

Pour remédier à cette lacune que nous avons rencontrée, nous nous proposons de développer la notion de vocabulaire au niveau d'un texte. Nous considèrerons un vocabulaire V comme la somme de vocabulaires Vi où i désigne les différentes classes grammaticales de la langue.

Pour répondre à l'autre question, à savoir la variation du degré d'importance des classes grammaticales, nous avons intégré une notion de poids relatifs à chaque classe grammaticale. Ainsi nous allons définir une variable qui représentera le poids dans le vocabulaire, cette variable sera rattachée aux classes grammaticales. Soit donc αi le poids de la classe grammaticale i. Il est à noter que la valeur de αi varie entre 0 et 1 et ceci pour nous ramener dans le cas où c'est égal à 1 à l'approche de D. Labbé qui est en fait notre point de départ.

2.2.3 Formulation mathématique

Soit deux textes A et B ayant respectivement les vocabulaires Va et Vb. Nous réalisons l'équilibrage de Labbé pour le mettre à la même taille et obtenir les vocabulaires Va et V'b. Nous intégrons alors les poids au niveau des fréquences des vocabulaires selon la classe grammaticale d'appartenance.

D'où la nouvelle formulation :

Soit αCi le poids de la classe C à laquelle appartient le vocable i

Ce qui donne la distance suivante :

- Distance absolue : entre le vocabulaire du texte A et B' (B'=la réduction de B à la taille de A)

(3)

$$D_{Va,b(u)} = \sum_{V_a, V'_{b(E)}} \alpha_{Ci} \left| F_{ia} - E_{ia(u)} \right|$$

- Distance relative $D_{(a,b)}$:

(4)

$$D_{(a,b)} = \frac{\sum_{V_a, V'_{b(E)}} \alpha_{Ci} \left| F_{ia} - E_{ia(u)} \right|}{\sum F_{ia} + \sum E_{ia(u)}} = \frac{\sum_{V_a, V'_{b(E)}} \alpha_{Ci} \left| F_{ia} - E_{ia(u)} \right|}{N_a + N'_b}$$

Ou $E_{ia(u)} = F_{ib} * U_{(a,b)}$ et $U(a,b) = \frac{N_a}{N_b}$

$N_b^{'} = \sum_{V_b} E_{ia(u)}$

2.2.4 Avantages de l'intégration des pondérations

La nouvelle formule prend en charge les points suivants:

- La diminution du parasitage au niveau des vocabulaires par annulation ou atténuation de l'impact de certaines classes grammaticales au niveau des fréquences des vocabulaires et ceci soit par mise à zéro du poids ou soit par sa diminution.

- La prise en charge d'autres critères au niveau du calcul de la distance intertextuelle comme la pondération des classes grammaticales au niveau d'un certain nombre d'auteurs, de genres littéraires…En effet par attribution d'un certain nombre de pondération à chaque texte référence il est possible de rendre l'indice de distance plus spécifique à une référence et ainsi renforcer sa représentation de la projection du texte à comparer sur la référence retenue.

- Généralisation de la formule écrite par Labbé. En effet, cette nouvelle formulation permet pour une configuration αi=1 de se conformer à l'approche de Labbé.

2.2.5 Problématique posée par la nouvelle approche

La nouvelle approche nécessite de trouver un jeu de pondération adéquat pour optimiser le calcul de la distance ainsi qu'une étude détaillée pour déterminer les classes grammaticales à définir en tant que parasite. Il est possible de penser que le jeu de pondération ainsi que les la définition des classes parasites soit une fonction variable de l'auteur, du genre, du thème de l'époque… ce qui d'un côté rend la tâche plus difficile mais qui au même temps rends la formule plus adaptable au contexte de calcul et aux textes de référence. Pour commencer nous pourrons considérer le côté parasitage sans chercher à définir le jeu de pondération, ainsi les αi prendront soit 0 ou 1 jusqu'à ce qu'on trouve une méthode performante pour la recherche des jeux de pondération qui pourrait résider dans une analyse statistique des lexiques des textes référence. L'utilisation des poids 0 et 1 permet d'exclure les éléments parasites dans le calcul ce qui améliore déjà le résultat numérique.

3 Conclusion

Nos travaux se sont intéressés à la classification de textes en langue arabe basée sur la théorie de la distance intertextuelle. Pour ce faire nous avons réalisé un système permettant à partir d'un certain nombre de composants de réaliser cette tâche.

Nous avons en outre réalisé des batteries de tests pour l'étalonnage et la mise en place des différents composants afin qu'ils soient stables et fiables. En ce qui concerne le lemmatiseur la lemmatisation de texte s'avère une tâche assez lourde au début mais plus nous avançons dans le traitement plus le système acquiert une certaine rapidité.

Une fois notre objectif initial atteint, nous nous proposons dans un troisième volet de notre travail, de mettre en place le référentiel. Ceci par une étude détaillé des différents ouvrages et la sélection, avec l'assistance d'un expert du domaine, de textes en langue arabe représentatifs d'un ensemble de critères à savoir genres, types, registre, thèmes, époques, auteurs. Nous nous proposons de réaliser une étude expérimentale qui nous permettra d'associer à chaque texte référence un jeu d'étiquetage et de pondération de classes grammaticales pouvant mettre en valeur ses spécificités. Cette étude expérimentale permettra de finaliser le système de classification automatique de texte en langue arabe basé sur la théorie de la distance intertextuelle.

Références

MULLER CHARLES (1997), PRINCIPES ET METHODES DE STATISTIQUE LEXICALE, PARIS. HACHETTE UNIVERSITE.

BRUNET E.(1988) Une mesure de la distance intertextuelle : la connexion lexicale, Le nombre et le texte. *Revue informatique et statistique dans les sciences humaines, Université de Liège*.

DEERWESTER S., DUMAIS. S. T., FURNAS, G. LANDAUER. T. K.HARSHMAN (1990). Indexing by latent semantic analysis. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE, 391-407.

JALAM R., CHAUCHAT J. H.(2002), Pourquoi les n-grammes permettent de classer des textes ? recherche de mots-cléfs pertinents à l'aide des n-grammes caractéristiques. JADT 2002 : 6ES JOURNEES INTERNATIONALES D'ANALYSE STATISTIQUE DES DONNEES TEXTUELLES.

JAOUADI W., ZRIGUI M.(2006), Application de la théorie de la distance intertextuelle pour la classification de texte en langue arabe. CONFERENCE REALITER 2006, RIO DE JANEIRO.

LABBE D.(2002), L'attribution d'auteur et la distance intertextuelle. *Revue Corpus.*

LABBE D. ET LABBE C.(2003), La distance intertextuelle. REVUE CORPUS 2, LABORATOIRE "BASES, CORPUS ET LANGAGE", UMR 6039 DU CNRS.

LABBE D.(2003), Réponses à M. J.-M. VIPREY Corneille et Molière.

LABBE D.(2004), CORNEILLE ET MOLIERE. 7E JOURNEES D'ANALYSE DES DONNEES TEXTUELLES.

REINHART M., QUELQUES ASPECTS DU CHOIX DES UNITES D'ANALYSE ET DE LEUR CONTROLE DANS LA METHODE ALCEST. IN L. L. S.

SALTON, G.. AUTOMATIC TEXT PROCESSING, ADDISON WESLEY, 1989.

LEBART L., SALEM, A.. ANALYSE STATISTIQUE DES DONNEES TEXTUELLES. PARIS, DUNOD, 1988.

JALAM R., CHAUCHAT J. H.., POURQUOI LES N-GRAMMES PERMETTENT DE CLASSER DES TEXTES ? RECHERCHE DE MOTS-CLEFS PERTINENTS A L'AIDE DES N-GRAMMES CARACTERISTIQUES. JADT 2002 : 6ES JOURNEES INTERNATIONALES D'ANALYSE STATISTIQUE DES DONNEES TEXTUELLES, 2002.