**REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE**
**MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE**
**UNIVERSITE MOHAMED BOUDIAF - M'SILA**

**FACULTE DU TECHNOLOGIE**
**DEPARTEMENT D'ELECTRONIQUE**
**N° : 2019/STN…/…/…..**


**DOMAINE : SCIENCES ET TECHNOLOGIE**
**FILIERE : ÉLECTRONIQUE**
**OPTION : INSTUMENTATION**

**Mémoire présenté pour l'obtention**
**Du diplôme de Master Académique**

**Par**: BENYAHIA Abdelkader

**Intitulé**

# Etude et analyse sur les performances des techniques d'identification d'auteurs à partir des documents écrits et des documents transcrits

**Soutenu devant le jury composé de:**

| | | |
|---|---|---|
| Mr. BRIK Youcef | Université M'sila | Président |
| Mr. KHENNOUF Salah | Université M'sila | Rapporteur |
| Mr. DJERIOUI Mohamed | Université M'sila | Examinateur |

**Année universitaire : 2018 /2019**

بسم الله والحمد والشكر لله ...

# *Remerciements*

*Avant tous, il apparait opportun de commencer ce mémoire par des remerciements à ceux qui nous ont beaucoup appris au cours de ce travail.*

*Au terme de ce travail, mes remerciements les plus sincères sont adressés à mes encadreurs  Dr. KHENNOUF Salah et Pr. SAYOUD Halim , pour m'avoir permis de bénéficier de leur savoir dans la matière, pour leur disponibilité, leurs compétences, et leur aide précieux tout au long de ce projet même pendant les moments les plus difficiles. Merci pour une qualité d'encadrement si sérieuse et si consistante.*

*Je remercie également les membres du jury pour honneur qui nous fait accepter de lire et de juger ce travail.*

*Je remercie tous les enseignants du département électronique ainsi que le personnel administratif qui a contribué de près ou de loin à notre formation durant cette année.*

*Enfin je remercie tous ceux qui ont contribués de près ou de loin à la concrétisation de ce travail.*

# *Dédicaces*

*Je dédie cet humble travail à l'esprit de mon cher père, qu'Allah lui fasse miséricorde,*

- *A ma chère mère que je lui souhaite une bonne santé et longévité*
- *A ma femme et mes enfants*
- *A tous mon professeurs du primaire ou moyen ou secondaire ou à l'université et surtout mon encadreurs*

*- A toute ma famille, A tout mes amis,*

*- Enfin à tous ceux et celles qui m'ont encouragé et soutenu.*

# Sommaire

**Chapitre-2 : Transcription des textes à partir des fichiers audio**



**Chapitre-3 : Expériences et résultats obtenus**

# Liste des figures

# Liste des tableaux

# Liste des abréviations

*(IHM)* *Interaction Homme-Machine*

*(IA)* *l'Intelligence Artificielle*

*(PCT'17)* *Parole Convertie en Textes*

*(MLP)* *Multi Layer Perceptron*

*(SVM)* *Support Vector Machins*

*(SMO)* *Sequential Minimal Optimization*

*(SRAP)* *Systèmes de Reconnaissance Automatique de Parole*

*(EI)* *l'Encyclopédie de l'Islam*

*(API)* *L'Alphabet Phonétique Internation*

*(SRM)* *Structural Risk Minimization*

*(AA)* *Attribution Auteur*

*(MCE)* *Most Common Events*

# *Conclusion général*

# INTRODUCTION GÉNÉRALE

Les techniques d'identification d'auteur ont été largement utilisées pour identifier les auteurs actuels des textes anonymes dans la mesure où ces textes ont été écrits à l'origine par ces auteurs. Cependant, à notre connaissance, il y a peu d'expérience dans l'application de ces techniques dans le cas le texte écrit directement par son auteur où le texte n'est pas écrit directement par son auteur mais transcrit à partir de son discours.

Dans ce travail de recherche, on vise à faire une étude et analyse sur les performances des techniques d'identification d'auteur à partir de documents écrits et de documents audio transcrits. Pour cela, plusieurs descripteurs seront utilisés pour modéliser le style de chaque auteur, et un classifieur "Multi Layer Perceptron" (MLP). seront implémenté et deux bases de données textuelles (corpus) seront conçues pour validées les résultats obtenus.

L'écrit a été, et restera, l'un des grands fondements des civilisations et le mode par excellence de conservation et de la transmission du savoir. Ecrire pour communiquer a été depuis tous les temps une préoccupation première de l'homme. En effet, beaucoup d'objets qui nous entourent comportent des traces écrites: les panneaux indicateurs, les notices d'emploi des produits,les journaux, les livres, …etc. [1].

L'attribution d'auteur d'un texte inconnu ou douteux est l'un des plus anciens problèmes de la statistique appliquée à la littérature. Ce type d'étude linguistique consiste à attribuer à un texte anonyme son auteur réel [2]. Dans ce travail de recherche, on s'intéresse à l'identification des locuteurs à travers des textes écrites et à la transcription de leur discours en texte .

L'originalité du présent travail, qui s'inscrit dans le cadre de l'attribution d'auteurs, réside dans la manière d'obtention des fichiers textes utilisés. Cette opération consiste transcrire manuellement le contenu des fichiers audio, comme ils sont prononcés par les interlocuteurs, à l'aide d'une écoute attentive

des enregistrements audio.

Plusieurs objectifs sont fixés pour cette étude, afin de donner les fruits attendus. Ces objectifs sont organisés comme suit :

- Conception d'un Corpus d'évaluation (base de données textuelle) qu'on afin de conduire les expériences de l'identification d'auteurs des textes transcrits
- Faire le point sur le signal de parole et sa transcription en document texte ainsi qu'une présentation générale des différentes techniques (automatiques et manuelles) de la de transcription.
- Réalisation d'un système d'identification d'auteur basé sur les Caractères N-grams comme descripteurs (features) et MLP comme classifieurs.

Ce mémoire est structuré en trois chapitres, comme suit : Dans le premier chapitre on on aborde le domaine de la fouille de textes, ces tâches et ces domaines d'applications. Ensuite on explique le processus de la catégorisation de différents textes et on donne une idée générale sur la notion de représentation numérique de textes. En fin nous présentons les techniques de catégorisation utilisées pour catégoriser les textes.

Le deuxième chapitre on exposera une présentation générale du signal de parole et des différentes techniques de sa transcription en un fichier texte

dans le troisième chapitre on exposera les séries d'expériences d'attribution d'auteur effectuées sur la base de données textuelle (ou Corpus) que nous avons conçu pour cette fin et qui est contient deux catégories de textes. .

Enfin, ce mémoire est achevé par une conclusion générale contenant un résumé du travail réalisé dans ce mémoire, les discussions ainsi que les explications possibles des résultats obtenus et enfin quelques suggestions et perspectives pour les futures recherches.

# *CHAPITRE-1*

# *FOUILLE DE TEXTES ET CATEGORISATION DE DOCUMENTS*

# Chapitre-1

## *Fouille de données et discrimination des documents écrits et transcrits*

### Introduction

Dans ce chapitre on aborde le domaine de la fouille de textes, ces tâches et ces domaines d'applications. Ensuite on explique le processus de la catégorisation de différents textes et on donne une idée générale sur la notion de représentation numérique de textes. En fin nous présentons les techniques de catégorisation utilisées pour catégoriser les textes.

### 1.1 Fouille de textes

La Fouille de Textes (Text-Mining en anglais) est née durant les années 90 et elle a pris ces origines des travaux de la fouille de données (Data-Mining). Son objectif est de tirer les données pour tirer le meilleur profit possible en utilisant des programmes capables de prendre des décisions pertinentes.

La fouille de textes (FT) ou l'extraction de connaissances dans les textes (ECT fait partie du domaine de l'intelligence artificielle, c'est un ensemble de traitements informatiques consistant à extraire des connaissances selon un critère de nouveauté ou de similarité dans des textes.

Dans le domaine de la littérature les critiques font de nombreuses tâches de la fouille de textes mais sur une échelle réduis; avec le développement technologique et le nombre élevés de textes à traiter, on trouve la catégorisation automatique de textes qui s'impose comme une technologie clé dans la recherche et l'extraction d'information.

L'émancipation de l'Internet a bouleversé le domaine, elle a permet aux chercheurs d'accéder une énorme quantité de textes, souvent mal rédigés et portant un potentiel riche et beaucoup d'informations utiles, ce qui a conduit à un intérêt pour la catégorisation automatique de textes dans l'objectif de réduire le travail humain au maximum, et à résoudre des problèmes d'accès à l'information voulue.

### 1.1.1 Définitions

Selon **[10,11]**, la première définition de la fouille de textes, présentée entant que telle, est celle de Feldmanetal.**[12]**.

***Définition 1***: « La fouille de textes est la science qui extrait des motifs caché à partir de grandes collections de textes ».

***Définition2***: Selon Se bastini **[13,14]** « La fouille de textes est de plus en plus employée pour désigner toutes les tâches qui, en analysant de grandes quantités de textes et en détectant des motifs, essaie d'extraire des informations probablement utiles ».

***Définition3***: S'inscrivant dans la tradition de l'ECBD (extraction de connaissances dans des bases de données), Kodratoff **[8],** définit la DCT (découverte des connaissances à partir des textes) comme «la science qui découvre les connaissances dans les textes».

Ces définitions montrent l'intérêt qu'offre la fouille de texte soit dans l'aspect quantitatif d'analyse et traitement de textes, soit dans l'accès aux nouvelles informations. Elles sont assorties des mêmes exigences qu'en ECBD, à savoir que «les connaissances découvertes doivent être ancrées dans le monde réel et doivent modifier le comportement d'un agent humain ou mécanique».

## 1.2 Les différentes tâches de fouille de textes

Parler de fouille de textes c'est parler de text mining, c'est un terme générique, traduction approximative de l'anglais, et l'interprétation la plus immédiate qui pourrait se référer à la recherche d'information, ou à l'extraction de connaissances. En effet dans c'est dans certains thématiques que la fouille de texte a pris naissance.

Ainsi, avec le développement et l'épanouissement dans le domaine scientifique, et surtout ses possibilités d'applications, d'autres thématiques sont venues compléter ce premier ensemble. **[7]**

L'aspect quantitatif est omniprésent dans la fouille de textes comme dans fouille de données, les différentes solutions envisagées peuvent être évaluées et comparées.

L'aspect qualitatif d'un programme se mesurera à sa capacité à s'approcher le plus possible d'une solution de référence validée par un humain.

La fouille de textes peut, par fois, exploiter les statistiques, mais la caractérisation des propriétés d'un texte ou d'un corpus n'est pas sa finalité dernière.

Elle a toujours en vue un autre but, formulé dans cette notion de tâche. Certaines tâches élémentaires joueront en outre le rôle d'unités de base en fouille de textes.

Certaines tâches de la fouille de textes qualifiées dans **[1]** «d'élémentaires» servent de "briques de base " aux autres tâches plus complexes qu'on trouve.

### 1.2.1 La Recherche d'Information (RI):

La fonction « Recherche d'Information » (ou RI, ou IR pour Information Retrie val en anglais). Le but de cette tâche est de retrouver un ou plusieurs document(s) pertinent(s) dans un corpus, à l'aide d'une requête plus ou moins informelle.

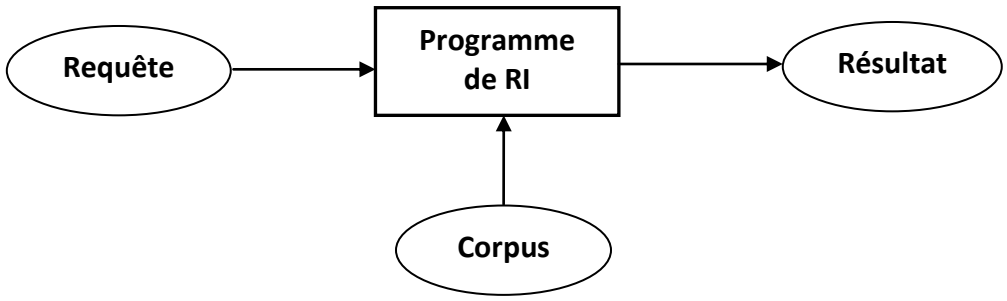


Figure 1.1 Schéma général de la tâche de Recherche d'Information

La tâche de RI est applicable à d'autres données que des textes : il existe des systèmes spécialisés dans la recherche d'images, de vidéos ou de morceaux de musique. D'autres encore qui se fondent sur des distances géographiques (pour les téléphones équipés de géolocalisation) ou dans des réseaux sociaux (pour la recherche de connexions possibles), etc.

La RI est une tâche très connue chez les internautes, elle est utilisée quotidiennement dès qu'ils ouvrent un moteur de recherche.

L'intégration des systèmes de recherche sont aussi au cœur même de chaque ordinateur, dans l'objectif est d'aider l'usager` à fouiller dans son disque dur à la recherche d'un fichier ou d'un mail mal rangé.

### 1.2.2 La classification

L'une des tâches les plus importantes de la fouille de textes est la classification. Elle consiste à lier une "classe" `à chaque donnée d'entrée, comme le montre la figure2.2

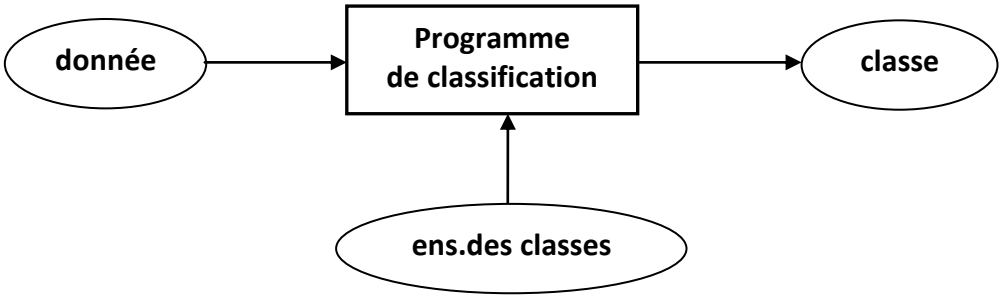


**Figure 1.2 Schéma général de la tâche de Classification**

– la donnée à classer est en principe de type "texte brut" ou " document semi structuré".

– l'ensemble des classes possibles est fini et connu au moment où le programme de classification est sollicité.

La classification binaire est illustrée dans le cas ci-dessus les deux classes sont possibles.

Ainsi la recherche d'information, la classification peut s'appliquer à toutes sortes de données, et pas seulement aux textes:

De multiples et florissantes applications sont nées de la classification des images, des vidéos, des musiques et de tous types de données, qu'il est possible de décrire à l'aide d'attributs.

Parmi ces application, il existe une qui est présente dans la plupart des gestionnaires de courriers électroniques: c'est la reconnaissance automatique des "spams".

Un autre domaine d'application en plein développement, c'est la "fouille d'opinion ", elle cible à identifier les polarités (positives ou négatives) véhiculées par les textes (par exemple les commentaires d'internautes sur les sites marchands ou de loisir), généralement à des fins de marketing, pour mesurer la "e-réputation" d'une société, d'une personne, d'une marque, d'un produit...

### 1.2.3 L'Annotation

L'annotation (ou l'´étiquetage), est une tâche plus spécifiquement linguistique que les précédentes, au sens où elle ne s'applique pas aux données tabulaires et ne relève donc pas de la fouille de données.

La figure 2.3 la présente globalement.

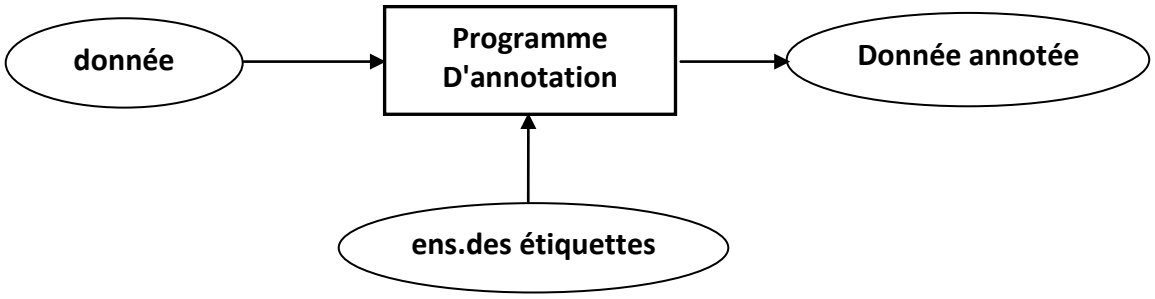


**Figure 1.3 Schéma général de la tâche d'Annotation**

En linguistique l'annotation est une tâche omniprésente. Mais, au lieu d'opérer sur des textes bruts, elle s'applique généralement à des textes segmentés en unités plus grandes. Le découpage le plus courant est celui dans lequel les unités de base sont des tokens (mots, chiffres ou ponctuations). Ainsi un texte est une séquence de tokens qui peut être annotée par une séquence d'´étiquettes.

Les étiquettes caractérisent la nature morphosyntaxique de chaque token. Elles sont les formes les plus traditionnelles pour annoter un texte brut, elles sont appelées "parties du discours" ( "part of speech" abrégé en POS en anglais).

Par exemple, la phrase " Le petit chat est mort." est constituée de 6 tokens et une séquence d'annotations possible est : DET ADJ NC V ADJ PONCT (où DET désigne les

déterminants, ADJ les adjectifs, NC les noms communs, V les verbes et PONCT les ponctuations).

Une des étapes fondamentales d'un programme de traduction automatique est l'alignement de séquences. La figure 4 illustre un tableau d'alignement entre deux séquences qui sont les traductions entre le français et l'anglais.

L'objectif principal des programmes d'alignement est d'annoter chacune des séquences avec les positions des traductions des mots dans l'autre séquence, tel que montre sous le tableau. Les deux séquences annotées visualisent en quelque sorte les projections des cases cochées du tableau suivant ses deux dimensions (horizontale et verticale).

**Tab 1.1–Un alignement bilingue et les deux annotations correspondantes**

| | J' | aime | le | chocolat |
|---|---|---|---|---|
| I | X | | | |
| like | | X | | |
| chocolate | | | | X |

| $J'_1$ | $aime_2$ | $le_3$ | $chocolat_4$ | $I_1$ | $like_2$ | $chocolate_3$ |
|---|---|---|---|---|---|---|
| 1 | 2 | - | 3 | 1 | 2 | 4 |

On a trouvé aussi une autre application intéressante qui divise un texte long en zone thématiques. L'annotation de la page Web peut être destinée par exemple à distinguer ce qui, dans cette page, donne lieu à un titre, un menu de navigation, un en-tête, un pied-de-page, une image, une zone de texte, etc. Pour en extraire le vrai contenu informationnel toute on écartant ses éléments parasites (publicités, etc.).

### 1.2.4 L'Extraction d'Information (EI)

Les schémas de la figure 2.4 montrent l'Extraction d'Information (EI ou Information Extraction en anglais, abrégé en IE). L'objectif de cette tâche, qui relève de l'ingénierie linguistique, est d'extraire automatiquement de documents textuels et des informations factuelles, servant à remplir les champs d'un formulaire prédéfini.

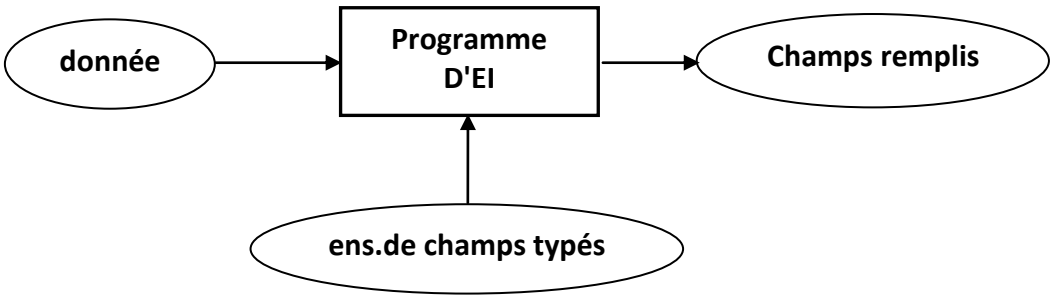


**Figure 1.4 Schéma général de la tâche d'Extraction d'Information**

Lors des conférences MUC ( " Message Under standing Conference " ) qui se sont déroulées entre 1987 et 1998 aux Etats Unis , sous l'impulsion de la Darpa ( l'agence de recherche du département de la Défense américain). L'extraction d'information a vu le jour et l'idée commençait à s'épanouir.

Les intervenants dans ces conférences se voyaient confier des corpus et leurs programmes s'étaient comparés en fonction de leur capacité à remplir à partir de chaque texte les champs d'un formulaire prédéfini. Par exemple, en 1992, il s'agissait d'extraire de dépêches d'agences de presse décrivant des attentats des informations telles que: date, lieu, auteur présumé ou revendiqué, nombre de victimes, etc.

L'intérêt stratégique de ce genre d'applications est clair pour faciliter les investigations

Parmi les applications les plus importantes actuellement est la reconnaissance des *entités nommées*, ces mots et sous-groupes de mots qui identifient soit des noms propres (désignant des personnes, des lieux ou des organisations) soit des quantités mesurables (exprimant notamment des dates, des valeurs numériques ou monétaires). Les célèbres "cinq W" du journalisme anglo-saxon ("who did what, where and when, and why", c'est-à-dire "qui a fait quoi, où, quand et pourquoi" en français) attendent, pour la plus part, une réponse en forme d'entité nommée, la reconnaissance des noms propres et des dates présents dans les textes ou les pages HTML ou XML. L'analyse automatique de CV, ou de sites marchands pour faire de la comparaison de prix, sont encore d'autres applications potentiellement très utiles de l'extraction d'information.

### 1.2.5 La segmentation de textes

Une autre tâche de reconnaissance thématique c'est la segmentation de texte [7].

Toutes les recherches sur la segmentation de texte se partagent les caractéristiques suivantes:

- La détection de la cohésion (thématique, lexicale) dans un texte

- La définition de la limite de segment lors qu'il y a rupture de cohésion: changement lexical **[16]**.

- La capacité à présumer de l'unité du segment par rapport à une unité connexe si elle est cohérente: ainsi, des segments seront constitués de plusieurs phrases adjacentes si toute fois ces dernières maintiennent la cohésion choisie.

Dans le cas où des phrases séparées par plusieurs autres phrases qui ne pourront pas relever d'un même segment, sauf à considérer les phrases intermédiaires comme une forme de remplissage (filler) qui ne rompt pas la chaîne (thématique, lexicale) ainsi créée.

Souvent est – il que les tâches de segmentation sont liées fortement de ce pourquoi elles sont réalisées.

Elles peuvent être par exemple liées:

- A une tâche de recherche d'information, dans laquelle on cherchera à fournir en réponse non seulement un texte (issu d'une URL par exemple) mais plutôt, dans ce texte, le ou les fragments les plus véritablement compatibles avec la question posée **[17].**

- A une tâche d'indexation d'un texte pour des buts de création de métadonnées à usage pédagogique ou documentaire **[18]**.

- A une tâche de résumé automatique ou semi-automatique, dirigé par le thème, où le résumé se fait par extraction des segments les plus appropriés à un thème donné et création d'un nouveau document

- A des travaux d'extraction du plan ou de la structure du document pour diverses fonctions ultérieures.

### 1.2.6 Le profilage

Le profilage est l'opération qui consiste à donner des contours lexicaux, sémantiques ou rhétoriques à un ensemble de textes ou de fragments de textes (7) dans l'objectifest de :

–reconnaître un auteur particulier ou une période donnée dans une masse de documents non datés **[19].**

– à l'inverse, étant donné un profil fait de préférences fournies par des utilisateurs, rechercher l'ensemble des textes obéissant à ce profil.

– détecter des tendances ou des opinions dans des discours **[20]**

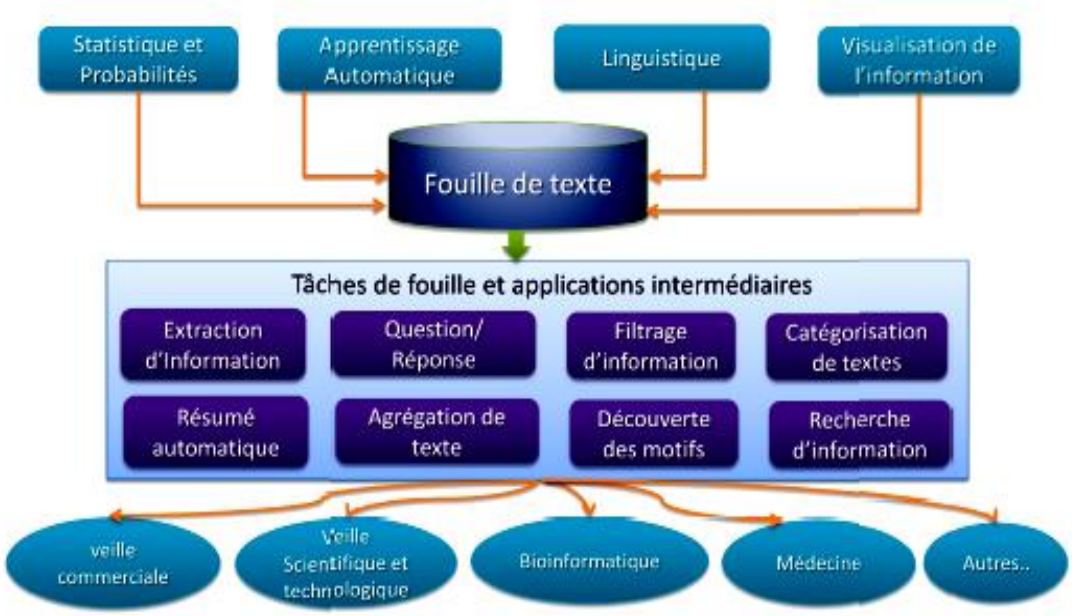


**Figure 1.5 Apports disciplinaires et domaines d'application de la fouille de textes [10].**

### 1.2.7 La reconnaissance d'auteurs

La découverte du véritable auteur d'un écrit (œuvre littéraire, article de presse, lettre, courriel) a donné lieu à de nombreuses études au cours de ces deux dernières décennies **[9].**

La question de l'attribution d'auteur reconnaît de nouveaux prolongements comme la vérification d'auteur pour éviter le plagiat et sécuriser leurs droits.

Dans ce cas, nous voulons savoir si un texte a été écrit ou non par un auteur donné.

Ainsi, au lieu d'obtenir le nom probable de l'auteur, on peut se limiter à déterminer des informations socio-économiques le concernant (profilage) comme le sexe, l'âge, la nationalité, le niveau d'éducation, etc. **(9-Argamonetal., 2009**).

Puis, l'attribution d'auteur fait également partie des sciences forensiques[1], de débats légaux, mais surtout d'un intérêt grandissant sur le Web avec, comme variantes, l'analyse de la crédibilité des auteurs (blogs, Twitter), voire la détection de plagiat.

## 1.3 Les étapes de la fouille de textes

La figure 2.6 illustre les principales étapes de la fouille de textes, depuis la collecte des textes jusqu'à la découverte de nouvelles informations.

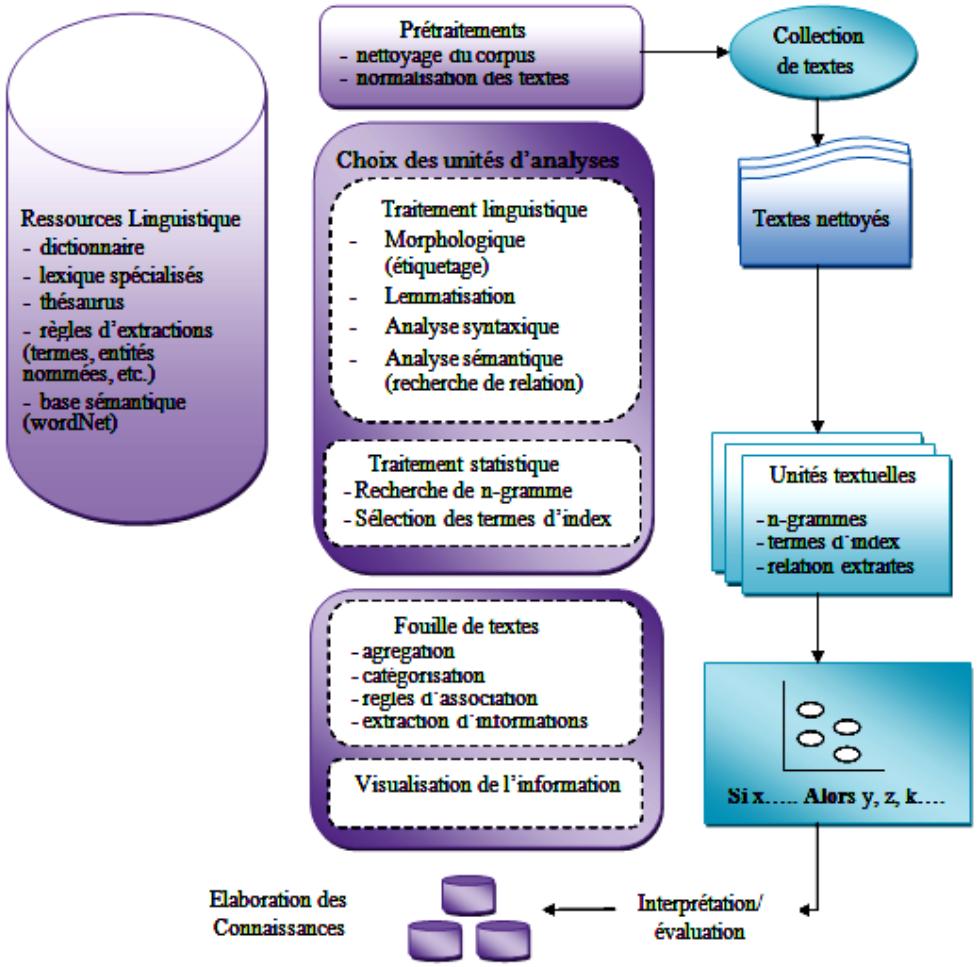


**Figure 1.6 Vue schématique des étapes de la FT [10] .**

Dans cette figure proposée dans **[10]**, les traitements effectués à chaque niveau varient en fonction de l'application et du type de connaissances utilisé. La phase de prétraitements peut être plus ou moins élaborée.

Elle peut éliminer des mots vides (mots grammaticaux) ou une normalisation plus poussée des textes dans le cas d'un corpus très technique (corpus médical, par exemple).

L'étape du choix des unités d'analyse peut faire appel aux connaissances linguistiques (extraction des termes, des relations sémantiques entre eux) ou simplement statistique, avec la recherche des n-grammes dans les textes (séquences de mots adjacents qui se répètent dans le corpus).

Les deux techniques peuvent enfin être combinées lorsqu'il s'agit de choisir, par les unités extraites, celles qui ont un poids discriminant (indexation automatique). Certaines applications de fouille de textes peuvent nécessiter des traitements de nature sémantique avec la recherche de relations entre les unités extraites. La phase de fouille va les puiser dans un panel de techniques pour choisir celle ( s ) adaptée ( s ) à l'application et aux types de résultats attendus.

### 1.4 La categorisation automatique de textes:

L'une des tâches classiques de la recherche d'information est appelée « la catégorisation automatique de textes » a suscité de nombreuses études depuis relativement longtemps **[3,11,13]**.

Dans ce domaine, la recherche est toujours très pertinente, car les résultats obtenus aujourd'hui sont encore sujets à amélioration.

La catégorisation de textes est l'activité du TAL qui consiste à classer de façon automatique des ressources documentaires, généralement en provenance, d'un côté, d'un corpus de textes ou d'une banque de documents textuels et de l'autre, d'un ensemble prédéfini de catégories.

La Catégorisation de Textes (CT) est le processus qui consiste à assigner une ou plusieurs catégories parmi une liste prédéfinie à un document.

L'objectif du processus est d'être capable d'effectuer automatiquement les classes d'un ensemble de nouveaux textes.

La CT consiste à apprendre, à partir d'exemples caractérisant des classes thématiques, un ensemble de descripteurs discriminants pour permettre de ranger un document donné dans la (ou les) classe (s) correspondant à son contenu ( **Brown & Chong, 1998** ).

Les méthodes d'apprentissage constituent l'appui des algorithmes de catégorisation, qui à partir d'un corpus d'apprentissage, permettent de catégoriser de nouveaux textes.

La définition de la catégorisation peut se résumer en une formalisation de la notion de similarité textuelle, soit en d'autres termes à trouver un modèle mathématique capable de représenter la fonction de décision d'appartenance des textes aux catégories.

### 1.4.1 Problématique de la catégorisation des textes

La problématique de la catégorisation peut se récapituler à trouver un prototype ou une fonction mathématique capable d'assigner automatiquement un document à une catégorie avec le plus grand taux de réussite possible **[5,11,14]**, cette fonction se traduit par

$$\Omega: D \times C \rightarrow \{ \text{Vrais} ; \text{Faux} \}$$

Où: **D** représente l'ensemble des documents et **C** représente l'ensemble des catégories. Pour chaque couple ( di, cj ) appartenant à D×C, la fonction de catégorisation Ω renvoi Vrai si le document appartient à la catégorie et Faux si non.

Dans les systèmes de catégorisation basés sur des méthodes d'apprentissage, la fonction de décision sera évaluée à l'aide d'un corpus d'entraînement. Cette fonction peut faire intervenir un grand nombre de valeurs numériques qu'un humain ne peut pas saisir.

La détermination de cette fonction est appelée *phase d'apprentissage*, tandis que l'utilisation de cette fonction pour attribuer une catégorie à un document se fera pendant la *phase de test*.

### 1.4.2 Nécessité de la catégorisation automatique:

On assiste aujourd'hui à un accroissement de la quantité d'information textuelle disponible et accessible d'une manière exponentielle.

D'après les derniers chiffres, on parle de plus de 200 millions de serveurs hôtes sur Internet et plus de 3 milliards de pages, la taille des corpus tests utilisés est passée de quelques mégaoctets à plusieurs Giga-octets.

Les contraintes majeures qui s'opposent au traitement manuel de la catégorisation des documents textuels se résument dans les trois points suivants:

- ✓ La réalisation manuelle de cette tâche par un expert est extrêmement coûteuse en termes de temps car il s'agit de lire attentivement chaque texte, au vu de la quantité phénoménale de textes aujourd'hui accessibles (par le biais durés à Internet en particulier) ( **Moulinier,1996** ).
- ✓ Les traitements manuels sont peu flexibles et leur généralisation à d'autres domaines est quasi-impossible; c'est pourquoi on cherche à mettre au point des méthodes automatiques (Se **bastiani, 2002)**.
- ✓ Cette opération peut être perçue comme subjective puisque basée sur l'interprétation du document, deux experts peuvent classer différemment un même document, ou encore un même expert peut classer différemment un même document soumis à deux instants différents ( **Clech & Zighed, 2004** ).

Ainsi l'intérêt de la recherche d'automatisation de la catégorisation de textes n'est plus à démontrer, etc.' Et dans cette perspective plusieurs travaux de recherche se concentrent ces dernières années.

### 1.4.3 Systèmes de Catégorisation Automatique

L'objectif de la CAT (catégorisation automatique des textes) est de classer de façon automatique les documents dans des catégories qui ont été définies soit préalablement par un expert (catégorisation *supervisée)* ou classification, soit de façon automatique (*catégorisation non supervisée)* ou clustering (partitionnement de données).

#### 1.4.3.1 Catégorisation supervisée (Classification)

Ainsi, la *classification* de textes correspond à la procédure d'affectation d'une ou de plusieurs catégories ou classes prédéfinies à un texte.

Elle correspond à la *catégorisation supervisée* pour l'apprentissage automatique et à la *discrimination* en statistiques alors que la recherche d'informations utilise des termes plus proches de l'application concernée : *filtrage* ou *routage*.

Cette problématique a par ailleurs dernièrement trouvé de nouvelles applications dans les domaines du traitement du langage tels que: l'affectation de sujets en recherche d'information, l'aide de l'utilisateur pour l'indexation de documents ( **Hayes &**

**Weinstein, 1990** ) , la veille technologique, le filtrage personnalisé des documents intéressant ( Lang, 1995 ) et l'amélioration de la recherche sur le web ( **Armstrong & all, 1995** ).

Aujourd'hui, cette problématique utilise largement des méthodes issues de l'apprentissage automatique et beaucoup d'algorithmes d'apprentissage supervisé lui ont été appliqués (Naïve bayes, K-plus proches voisins, arbres de décision, machines à vecteurs support, réseaux de neurones, etc...)

#### 1.4.3.2 Catégorisation non supervisé ( Clustering )

Quand l'ensemble des catégories n'est pas donné au départ, et qu'il s'agit de le créer en regroupant les textes en classes qui possèdent un certain degré de cohérence interne, on est dans un contexte de *catégorisation non supervisée* pour l'apprentissage automatique. Ce type de *catégorisation non supervisée* consiste à trouver de manière automatique une organisation cohérente à un groupe de documents homogènes pour construire des regroupements cohérents ( des classes ou clusters ), elle correspondent au statistiques et au clustering, qui est également le terme utilisé en recherche d'informations.

Le clustering consiste donc, à diviser les objets (dans notre cas des textes) en groupes sans connaître à priori leurs classes d'appartenance.

Les techniques pour réaliser de tels regroupements constituent un domaine d'étude très riche, qui a donné lieu à de multiples propositions dont le recensement n'est pas l'objet de cette étude.

### 1.4.4 Notion de Classe pour les Systèmes de Catégorisation

La notion de classe pour un système de catégorisation a été habituellement synonyme de « thème ».

Dans ce contexte, classer les documents revient à les organiser par différent thématiques.

Aujourd'hui la catégorisation s'intéresse à différentes tâches pour lesquelles les catégories ne sont pas interprétables comme des thèmes tel que classer les documents par auteur, par genre, par style, par langue, ou encore selon que le document exprime un jugement positif ou négatif, etc.

Ainsi la classe va correspondre à un besoin d'information d'un utilisateur ou d'une société et n'est donc pas obligatoirement un thème unique.

Dans la figure2.7, un système de catégorisation d'emails est représenté où les classes peuvent être de différentes natures (thèmes, messages provenant de certaines personnes, messages d'un certain type, etc.)

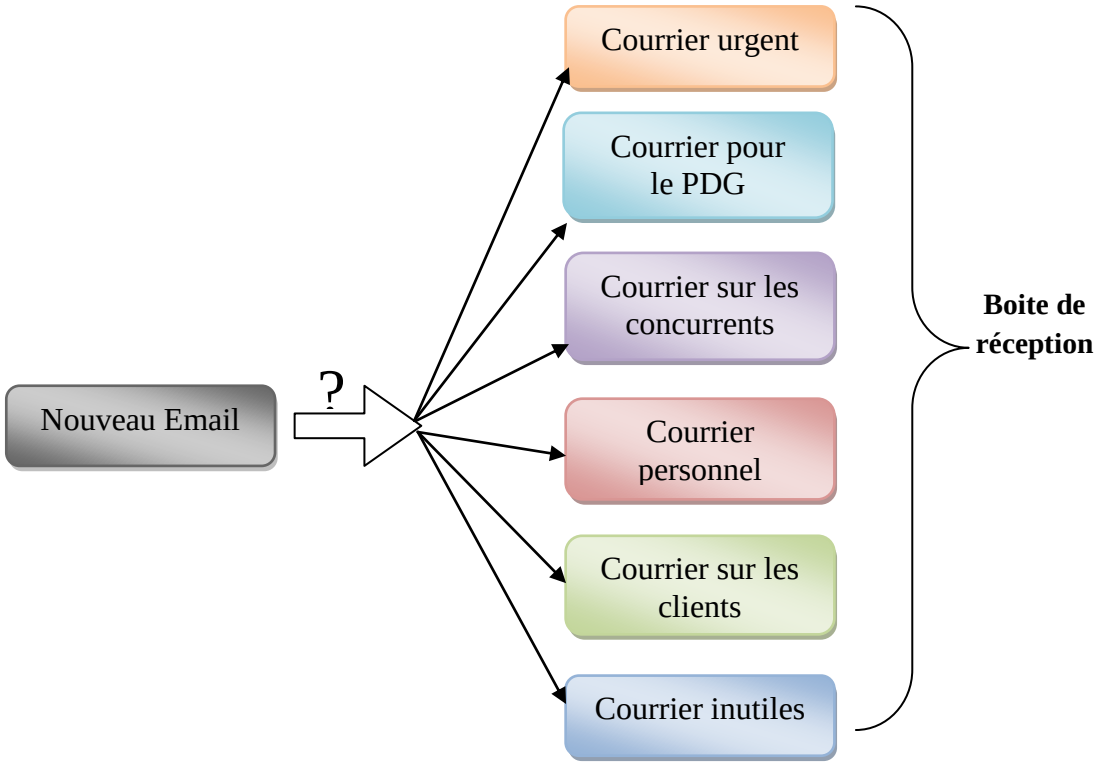


**Figure 1.7 Système de catégorisation d'emails**

### 1.5 Catégorisation de Textes et Text Mining

Le Text Mining est une technique permettant le traitement de gros volumes de textes pour en extraire les principal et les répertorier de manière statistique les différents sujets

évoqués ainsi que découvrir des connaissances et des relations à partir des documents disponibles.

L'outil de Text Mining va générer de l'information sur le contenu du document.

Cette information n'était pas présente, ou implicite, dans le document sous sa forme initiale, elle va être rajoutée, et donc enrichir le document.

Les applications en Text Mining peuvent être:

- Recherche d'information
- Correction orthographique / grammaticale
- Traduction automatique
- Résumé automatique
- Question / réponse (interfaces en langage naturel)
- La veille technologique
- La Catégorisation automatique des documents.

### 1.5.1 Démarche à Suivre Pour la Catégorisation de Textes:

Pour réaliser l'opération de catégorisation automatique de textes, la démarche commune est la suivante: la première phase consiste à formaliser les textes afin qu'ils soient compréhensibles par la machine et utilisables par les algorithmes d'apprentissage.

La catégorisation des documents est la deuxième phase, cette étape est bien entendu décisive car c'est elle qui va permettre ou non aux techniques d'apprentissage de produire une bonne généralisation à partir des couples (Document, Classe).

La démarche d'une approche standard de catégorisation automatique de textes peut être résumée de la manière suivante:

- Eliminer les caractères de séparation, les signes de ponctuations, les mots vides, etc.
- Les termes restants sont tous des attributs
- Un document devient un vecteur < terme, fréquence >
- Entrainer le modèle de catégorisation à partir des couples (Document, Classe).
- Évaluer les résultats du classifieur.

La figure 2.8 illustre la démarche de catégorisation de textes.

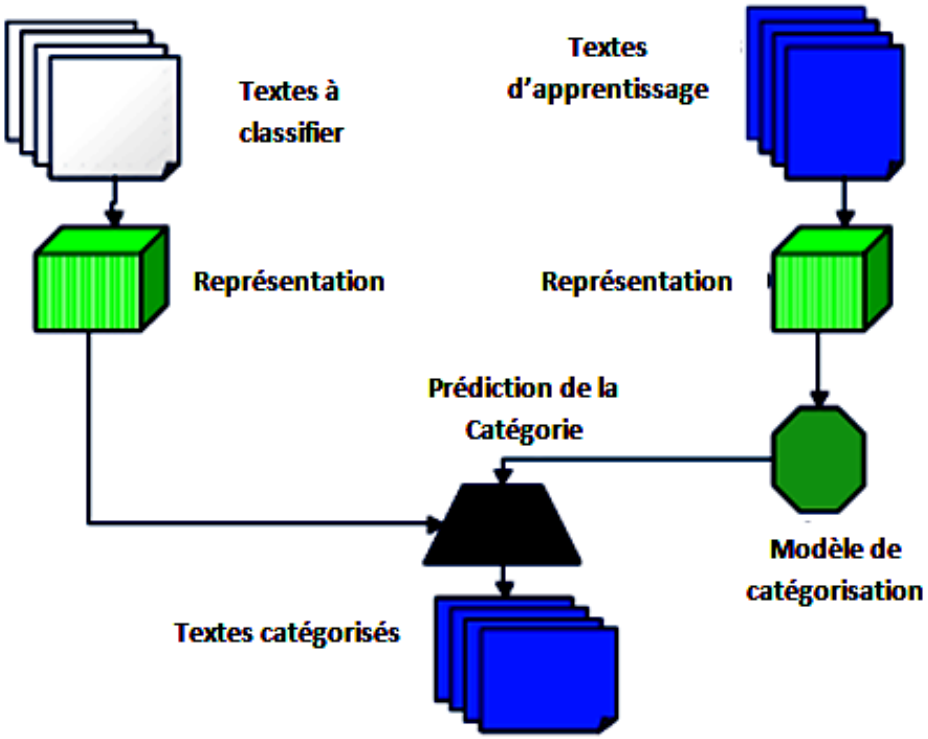


**Figure 1.8 démarche de la catégorisation de textes**

### 1.5.2 Problèmes de la Catégorisation de Textes

Plusieurs difficultés peuvent s'opposer au processus de la catégorisation de textes. Des problèmes liés à l'apprentissage automatique supervisé comme la subjectivité de la décision prise par les experts, les sur-apprentissages, etc.

Mais aussi des problèmes particuliers liés à la nature des données traitées à savoir des données textuelles comme la polysémie, la redondance, Les variations morphologiques ou même L'homographie, etc..

Dans ce qui suit nous allons signaler les difficultés principales qui s'opposent à la catégorisation de textes.

- **Redondance (Synonymie)**

La redondance et la synonymie permettent d'exprimer le même concept par des expressions différentes, plusieurs façons d'exprimer la même chose.

Cette difficulté est liée à la nature des documents traités exprimés en langage naturel contrairement aux données numériques.

Lors d'une représentation vectorielle d'un document, ces termes sont représentés séparément, et les occurrences du concept sont dispersées.

Il est alors important de rassembler ces termes en un groupe sémantique commun.

Pour y remédier, il est alors intéressant de concevoir une ontologie afin de cerner les sens des termes, naturellement, cela engendre des coûts supplémentaires pour sa réalisation.

- **Polysémie (Ambiguïté)**

A la différence des données numériques, les données textuelles sont sémantiquement riches, contrairement des langages informatiques, le langage naturel, autorise des violations des règles grammaticales engendrant plusieurs interprétations d'un même propos.

Un même mot possède, dans différents cas, plus d'un sens et, par conséquent, à cause de la polysémie, les mots seuls sont parfois de mauvais descripteurs.

Le mot *livre* peut désigner une unité monétaire, ou un bouquin ou le verbe livrer (nom: livraison).

Le mot *avocat* peut désigner le fruit, le juriste, ou même au sens figuré, la personne qui défend une cause.

Le mot *table* de cuisine ce n'est pas le même que dans *table* de multiplication.

Le mot *pièce* peut correspondre à une pièce de monnaie par exemple, ou à une pièce dans une maison, de même pour *pavillon*, *bloc*, *glace*, etc.

- **L'homographie**

Deux mots sont dits homographes s'ils s'écrivent de la même façon sans forcément avoir la même prononciation.

L'homographie est une sorte d'ambigüité supplémentaire. (Ex: avocat en tant que fruit et avocat en tant que juriste).

L'homographie et l'ambiguïté génère du bruit qui va causer une dégradation de précision (indicateur nécessaire pour mesurer la performance du classificateur). Il sera alors préférable d'ôter ces ambigüités.

- **La graphie**

Un terme peut comporter des fautes d'orthographe ou de frappes comme il peut s'écrire de plusieurs manières ou s'écrire avec une majuscule.

Ce qui va peser sur la qualité des résultats. Parce que si un terme est orthographié de deux manières dans le même document ( Ghelizane, Relizane ), la simple recherche de ce terme avec une seule forme graphique néglige la présence du même terme sous d'autres graphies, ce qui va influencer les résultats puisque les différentes graphies vont être traitées séparément.

- **Les variations morphologiques**

Les conjugaisons, pluriels, influent négativement sur la qualité des résultats puisque les différentes variations morphologiques vont êtes considérées séparément et chaque une va être prise comme un élément à part comme par exemple les trois termes: maître, maîtresse, maîtriser sont traités indépendamment quoi que en réalité cela pivote sur la même idée.

Pour y remédier soit on applique la lemmatisation ou le stemming, à notre texte soit carrément on opte pour une représentation en n-grammes qui peut nous éviter ces prétraitements.

- **Les mots composés**

La non prise en charge des mots composés comme: comme Arc-en-ciel, peut-être, sauve qui peut, etc.

Dont le nombre est très important dans toutes les langues, et traiter le mot Arc en ciel par exemple en étant 3 termes séparés réduit considérablement les performances d'un système de catégorisation néanmoins l'utilisation de la technique des n-grammes pour le codage des textes atténue considérablement ce problème des mots composés.

- **Présence-Absence de termes**

La présence d'un mot dans le texte indique un propos que l'auteur a voulu exprimer, on a donc une relation d'implication entre le mot et le concept associé, quoi que on sait très bien qu'il y a plusieurs façons d'exprimer les mêmes choses.

Dès lors l'absence d'un mot n'implique pas obligatoirement que le concept qui lui est associé est absent du document.

- **Sur-apprentissage**

Le nombre de termes très important et très varié qui ne se répètent dans tous les textes va causer énormément de creux dans le tableau de grande dimension ( textes / termes ) qui peut provoquer du sur-apprentissage qui s'explique par le fait que le modèle n'arrive pas à bien classer les nouveaux textes, pour tant il l'a bien fait dans la phase d'apprentissage en classant correctement les textes de la base d'apprentissage.

Pour limiter les sur-apprentissages, on doit sélectionner des termes pour réduire la dimensionnalité.

En général le nombre de textes d'apprentissage est limité, c'est pour cela on cherche à agir sur le nombre des termes utilisés en les diminuant, pour éviter ce sur-apprentissage. Sans bien sûr pénaliser le système en supprimant des termes pertinents (Se bastiani, 2002).

- **Subjectivité de la décision**

Après la lecture du texte à classer, l'expert va trancher à quelle (s) catégorie (s) ce texte appartient en se basant sur le contenu sémantique et le contexte du texte et même en consultant d'autres textes préalablement associés à certaines classes, pour valider la décision prise qui ne peut être que subjective.

Les experts humains ne lisent pas de la même manière! Ne réfléchis sent pas de la même manière! Donc ne classent pas de la même manière!

Ainsi un même document peut être classé différemment par deux experts, ou encore un même document peut être classé différemment par le même expert, soumis à deux instants différents ( Clech & Zighed, 2004).

D'après les expériences: Lorsque deux experts humains doivent déterminer les classes d'une collection de textes, il y a souvent désaccord sur plus de 5% des textes.

Il est donc illusoire de rechercher une catégorisation automatique parfaite.

### 1.5.3 Applications de la catégorisation de texte

Depuis les travaux de [ Maron, 1961 ], la catégorisation de textes est utilisée dans de nombreuses applications.

Parmi ces domaines figurent: l'identification de la langue [Cavnar and Trenkle,1994], la reconnaissance d'écrivains [ Forsyth, 1999, Teytaud and Jalam,2001] et la catégorisation de documents multimédia [Sableand Hatzivassiloglou, 2000].

Aussi, le filtrage qui consiste à déterminer si un document est pertinent ou non, par exemple: la détection de spam (les courriers indésirables) pour ensuite les supprimer et le routage qui consiste à affecter un document à une ou plusieurs catégories, comme la diffusion sélective d'information.

## Conclusion

Nous avons présenté dans ce chapitre brièvement l'essentiel des notions liées au domaine de la fouille de texte et notamment la notion de catégorisation des documents. Actuellement la fouille de texte est définie par ces objectifs qui sont divers et visent à extraire des informations utiles et exploitables. Ensuite nous avons approché une des tâches les importantes de la fouille de texte qui est la catégorisation des documents. Pour éclaircir l'utilité de ce travail qui se fait automatiquement, nous avons présenté les techniques utilisées dans la littérature pour effectuer cette tâche. La notion de représentation des textes et techniques de classification fera l'objet du prochain chapitre.

# *CHAPITRE-2*

# *TRANSCRIPTION D'UN SIGNAL DE PAROLE*

# Chapitre-2
# Transcription des textes à partir des fichiers audios

## Introduction

La parole est la pierre angulaire de la communication humaine. Contrairement aux autres moyens de communication, les systèmes utilisant la parole offrent à l'utilisateur un accès simple et naturel. L'importance de la parole augmente dans toute interaction homme-machine qui doit plus ou moins la traverser [4]. La communication vocale élimine tout contact physique avec la machine et permet à l'utilisateur de s'acquitter d'autres tâches. Au-delà des mots associés à une représentation de type phonologique dans le dictionnaire de prononciation, il est nécessaire de modéliser les respirations, les hésitations, les fragments de mots, les brouillons de la parole peu articulés. [5]

Dans ce chapitre, pour comprendre le mécanisme de parole chez l'être humain nous commencerons par un bref préambule concernant le système vocal humain, des mécanismes de production de la parole et des caractéristiques d'un signal de parole. Nous exposerons ensuite les systèmes d'interaction homme-machine et donnerons un état de la technique rapide des systèmes de transcription automatique, présenterons leurs performances et analysons les types d'erreurs les plus représentatifs [6].

## 2.1 Signal de parole

La parole peut être définie comme un son émis par le locuteur, en d'autres termes c'est une variation de pression acoustique plus ou moins rapide et forte qui est captée par un microphone placé à proximité. La possibilité de reconnaître la phrase sera donc très dépendante des conditions d'enregistrement : qualité du microphone lui-même, distance au locuteur et le niveau du bruit environnemental.

L'air contenu dans les poumons est La source d'énergie utilisée pour produire les sons. Le écoulement d'air sous pression parvient à travers la trachée jusqu'au conduit vocal, aux fosses nasales, aux organes d'articulation (langue, lèvres...) qui vont avoir chacun leur rôle dans la production de la parole. La parole sera ainsi très dépendante des caractéristiques physiques du locuteur : âge, taille, sexe, etc…[7].

### 2.1.1 L'appareil vocal humain

L'appareil vocal humain est composé d'un excitateur, le complexe glotte-cordes vocales, et d'un ensemble de résonateur de l'appareil phonatoire : le pharynx, la cavité buccale, la cavité labiale, les fosses nasales. Lorsqu'un excitateur entre en vibration, il fournit un signal, dont le résonateur va amplifier certaines composantes. La présence ou l'absence d'obstacles sur parcours de la colonne d'air modifier la nature de son produit. Le domaine de la phonétique articulatoire distinguée les différentes classe de sons en classant ces obstacles éventuels. [8].

Pour différencier la formation d'une voyelle et d'une consonne, il suffit de déterminer si le passage de l'air se fait librement à partir de la glotte ou non. Si tel est le sac, une voyelle est formée, alors que si le passage est partiellement ou totalement bouché, c'est une consonne qui est prononcée. Une distinction peut être faite entre mode d'articulation et point d'articulation, surtout du point de vue du classement des consonnes. [8].

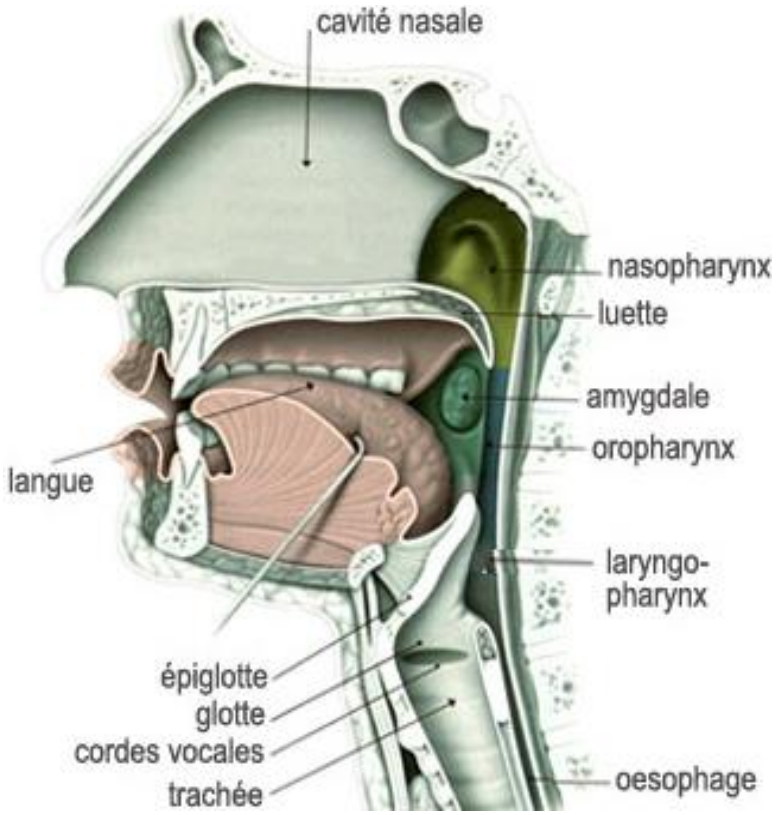


**Figure 1.1 : Système phonatoire**[8].

Il existe nombre d'éléments qui peuvent modifier la nature de l'air expiré et définit le mode d'articulation :

- Libre passage, ou mise en vibration de l'air au niveau de la glotte (sourde ou sonore).
- Libre passage, ou non, en un point quelconque des cavités supra glottiques.
- Passage par voie unique ou deux voies différentes (orale ou nasale).
- Passage dans le conduit buccal, par une voie médiane ou latérale.

L'emplacement d'un obstacle au passage de l'air dans la cavité buccale, est appelée le point d'articulation. En générale, cet obstacle créé par le placement de la langue. Les lèvres, les dents, le palais et la luette peuvent créer des points d'articulation. Lorsqu'un signal est fourni par vibration d'un excitateur, et amplifie par le résonateur, on obtient des formants. Ce sont des facteurs fondamentaux qui forment le timbre de la voix, et caractérisent donc ce dernier.[8]

Le nombre de formants est variable, pouvant passer d'un seul à une infinité. Mais même s'il existe beaucoup, seuls quelques-uns jouent un rôle du point de vue perceptif. Par contre, un formant ne peut jamais être ramené à une fréquence fixe, il s'agit plutôt d'une bande de fréquence. L'étendue spectrale du signal acoustique est comprise entre 80 et 8000 Hz, avec une étendue dynamique de 60 à 70 dB. Il est à noter que la fréquence fondamentale moyenne de vibration des cordes vocales, appelées « pitch » est situé entre 40 – 140 Hz pour les hommes, entre 180 – 300 Hz pour les femmes et entre 300 – 600 Hz pour les enfants. [8]

### 2.1.2 Production de la parole

Deux fonctions mécaniques de bases sont l'origine de la production de la parole à savoir: la phonation et l'articulation. La phonation est la production de signal acoustique par le mouvement du larynx. L'articulation est la modulation de signal acoustique par les articulateurs (les lèvres et la langue) et la résonnance de ce signal dans les cavités (la poche el le nez) [7].

#### 2.1.2.1 Mécanismes de production de la parole

Les étapes de production de la parole est un mécanisme très complexe qui repose sur une interaction entre les systèmes neurologique et physiologique. La parole commence par une activité neurologique. Après que soient survenues l'idée et la volonté de parler, le cerveau dirige les opérations relatives à la mise en action des organes phonatoires. Le fonctionnement de ces organes est bien, quant à lui, de nature physiologique. Une grande

quantité d'organes et de muscles entrent en jeu dans la production des sons des langues naturelles. Le fonctionnement de l'appareil phonatoire humain. [7]

#### 2.1.2.2 Caractéristique d'un signal de parole

Durant les premières années de traitement du signal, on a mélangé les caractéristiques de la musique et de la parole. Bien qu'on puisse profiter beaucoup des caractéristiques de la première en relation avec la compression, les spécifications de la parole sont encore plus fortes et profitables.

Deux limitations fondamentales méritent d'être prises en compte : les limitations du système auditif et celles du système vocal chez l'être humain. Le système auditif humain est surtout sensible dans une gamme de fréquence située entre 800 Hz à 8000 Hz, les limites extrêmes sont respectivement 20 et 20000 Hz.

En résumé, pour des sons vocaliques à des fréquences au-dessus de 4Khz, les hautes fréquences sont plus de 40 dB en dessous du sommet du spectre. Par ailleurs, en ce qui concerne des sons du type fricatif, le spectre ne chute pas nettement avant 8 Khz. C'est pour cela que pour représenter correctement des sons de ce type, il serait nécessaire d'utiliser une fréquence d'échantillonnage égale ou supérieur à20 khz. Cependant, pour beaucoup d'application, il n'est pas nécessaire de recourir à une fréquence d'échantillonnage si élevée, tout simplement parce qu'une qualité parfaite de reproduction n'est pas exigée.

C'est pourquoi, si on applique un filtre passe-bas d'ordre élève à 4 khz avant l'échantillonnage, il est possible d'effectuer un échantillonnage à 8 khz avec une bonne qualité. Mais il y a quelque consonantes de langues étrangères (à 8 ou 10 khz) qui ne seront pas reproduites très fidèlement. C'est pour cela que, si l'on peut, un échantillonnage à la fréquence de 20 khz est à conseiller, toujours de parole précédé par un filtre analogique à 10 khz. Une fois l'échantillonnage fait, le signal de parole résultant est fondamentalement variable. [7]

### 2.2 Interaction homme-machine

Dans le but que l'être humain puisse contrôler et communiquer avec une machine Les Interactions Homme-Machines (IHM) définissent les moyens et outils a été mis en œuvre. Les ingénieurs en ce domaine étudient la façon dont les humains interagissent avec les ordinateurs ou entre eux à l'aide d'ordinateurs, ainsi que la façon de concevoir des systèmes qui soient ergonomiques, efficaces, faciles à utiliser ou plus généralement adaptés à leur contexte d'utilisation.

#### 2.2.1 Techniques d'interaction

De nombreuses manières existent pour qu'un humain puisse interagir avec les machines qui l'entourent. Ces manières sont très dépendantes des dispositifs d'interactions et des forces ou compétences que l'être humain ne peut étendre qu'extérieurement. L'informatique a évolué très rapidement avec ses débuts dans les années 1940 jusqu'à aujourd'hui. [2]

Les premières tentatives sur ordinateur étaient sous forme de traitement par lots et toutes les entrées (programmes et données) étaient alimentées en entrée par des cartes perforées, des rubans perforés ou des bandes magnétiques. Il y avait un clavier pour interagir avec le système (console système). Avec l'arrivée de la micro-informatique on a commencé à utiliser des cassettes audio et des claviers, puis des disquettes et des souris informatique avant de passer aux écrans tactiles. Un système de pointage tel que la souris permet d'utiliser un ordinateur avec le paradigme WIMP qui s'appuie sur les interfaces graphiques pour organiser la présentation d'informations à l'utilisateur

Les premiers éléments de sorties ont été les imprimantes, les perforateurs de cartes et les perforateurs de ruban secondés ensuite par bandes magnétique. La console système était équipée d'une imprimante remplacée par un écran plus tard. Avec l'arrivée de la micro-informatique on a utilisé d'abord des cassettes audio, puis des disquettes avant d'utiliser des CD puis des DVD. Certaines techniques tentent de rendre l'interaction plus naturelle [9]

- La reconnaissance automatique de la parole ou de gestes permettent d'envoyer des informations à un ordinateur :
- La synthèse vocale permet d'envoyer un signal audio compréhensible par l'être humain

- Les gants de données offrent une interaction plus directe que la souris ;
- Les visiocasques essayent d'immerger l'être humain dans une réalité virtuelle, ou d'augmenter la réalité :
- Les tables interactives permettent un couplage fort entre la manipulation directe par l'être humain sur une surface et le retour d'information.

### 2.2.2 Modes d'interaction

On appelle une interaction multimodale si elle met en jeu plusieurs modalités sensorielles motrices [10]. Un système interactif peut contenir un ou plusieurs de ces modes d'interaction **:**

- Mode parlé **:** commandes vocales, guides vocaux...
- Mode écrit **:** entrées par le clavier et la tablette graphique, affichage du texte sur l'écran...
- Mode gestuel : désignation 2D ou 3D (souris, gants de données, écran tactile), retour d'effort...
- Mode visuel **:** graphiques, images, animations...

## 2.3 Transcription du signal de parole

La transcription audio en texte permet de transformer des données audio en fichier texte. Les métiers de l'audiovisuel, des médias, le monde académique et les institutions publiques sont amenés régulièrement à transcrire de l'audio en texte. Le numérique prend une part primordiale et grandissante dans la communication en rendant la parole accessible à tous et pour tous. Très prisé lors des séminaires, colloques, conférences, entretiens ou encore pour les cours en ligne… l'intérêt est de garder une trace écrite de ce qui se dit. Il est tellement utile d'avoir en main ou en format électronique des transcriptions d'enregistrements audio ou vidéo.

### 2.3.1 Conventions de la transcription

L'un des problèmes qui se posent lorsque l'on entreprend d'effectuer une transcription est celui des conventions d'annotation à adopter. Outre le texte lui-même, que veut-on représenter à l'écran, et surtout comment souhaite-t-on le faire ? Le premier aspect à aborder est celui de la représentation textuelle. Comment représenter par écrit des conversations orales? La majorité des corpus crées jusqu'à aujourd'hui ont adopté une

orthographe normalisée, semblable à celle que l'on trouve dans les dictionnaires. Bien que ce choix suive une certaine logique de normalisation, il présente cependant quelques inconvénients. Ainsi, dans la langue parlée, il n'est pas rare que la prononciation de certains mots soit déformée, voire transformée. [11]

### 2.3.2 Structuration de la transcription

L'étape d'annotation vise à structurer les enregistrements, c'est-à-dire à segmenter et à décrire le signal acoustique à différents niveaux jugés pertinents pour le traitement ultérieur. Il s'agit ici principalement de l'identité du locuteur, de l'identification du contenu thématique, ou de la qualité du canal de transmission (acoustique). Actuellement, un document de transcription est structuré de la manière la suivante :

- L'enregistrement correspond à la totalité de l'enregistrement à transcrire.
- L'enregistrement est découpé en sections, délimitant les parties des émissions à transcrire, et les parties non-transcrites.
- A certains points de synchronisation, des changements durables de bruit de fond sont indiqués ; cette segmentation en conditions acoustiques est indépendante de la structuration en tours et sections.

## 2.4 Transcription en langue arabe

Nous allons maintenant développer la transcription d'un fragment audio pour montrer de quelle manière en vertu du principe de disponibilité le niveau de granularité d'une transcription est étroitement associé à des possibilités et donc à des exigences d'analyse. Un discoure est proche d'un texte qui aurait pu être celui préparé par l'orateur en vue de son intervention, à quelques détails près signalant des discontinuités dont on peut faire l'hypothèse qu'elles sont à imputer à la performance en direct face au public. Ce texte permet de faire un certain nombre de constats analytique. On peut souligner la structure rhétorique et argumentative du discoure, qui recourt notamment à une forme récurrente de proposition principale, déclinée avec des variantes. Ainsi transcrire un discours permet donc une première analyse centrée sur la performance de l'orateur et sur les caractéristiques de sa parole. Dans le même ordre d'idée, la réaction finale du public permet d'évaluer son efficacité au regard de la réception qui lui est réservée.

### 2.4.1 Reconnaissance de la langue arabe

La langue arabe est une langue sémitique, elle est parmi les langues les plus anciennes dans le monde [12]. L'arabe classique standard a 34 phonèmes parmi lesquels 6 sont voyelles et 28 sont des consonnes. Les phonèmes arabes se distinguent par la présence de deux classes qui sont appelées pharyngales et emphatiques. Ces deux classes sont caractéristiques des langues sémitiques. Les syllabes permises dans la langue arabe sont : CV, CVC et CVCC. Où le V désigne voyelle courte ou longue et le C représente une consonne [13]. La langue arabe comporte cinq types de syllabes classées selon les trais ouvert/fermé et court/long. Une syllabe est dite ouverte (respectivement fermée) si elle se termine par une voyelle (respectivement une consonne). Toutes les syllabes commencent par une consonne suivie d'une voyelle et elles comportent une seule voyelle. La syllabe CV peut se trouver au début, au milieu ou à la fin du mot [14].

### 2.4.2 Transcription et Translitération en arabe

La transcription scientifique est un effort minutieux exigeant des niveaux de connaissances phonologiques et morphologiques élevés, et les erreurs de traduction et les incohérences ont tendance à s'infiltrer dans les publications les mieux éditées. Pour un certain nombre de raisons, surtout en raison de l'invisibilité des voyelles courtes dans un script arabe, la notation de voyelle courte par rapport à la longue, la représentation des voyelles épenthétiques, les limites de morphèmes, le marquage des cas et les limites des mots. Traditionnellement, une distinction est établie entre la transcription et la translittération. On s'est longtemps appuyé sur les définitions de Charles Ferguson de ces processus.

#### 2.4.2.1 Transcription

La Transcription (ou conversion phonémique) est la représentation écrite d'une langue par des symboles ou des orthographes autres que ceux de l'orthographe standard de la langue. Si la langue est normalement non écrite, tout système d'écriture conçu s'appelle transcription sauf si elle devient l'orthographe acceptée. Si une transcription est basée exclusivement sur les sons de la langue, elle s'appelle une transcription phonétique. Une variété importante de transcription phonétique est la transcription phonémique dans laquelle chaque symbole représente systématiquement un phonème de la langue écrite. Les transcriptions qui ne sont que partiellement phonétiques (ou phonémiques) sont également utilisées pour diverses raisons; Ils sont généralement basés en partie sur des considérations

grammaticales ou sémantiques. Pour les langues parlées, comme les dialectes arabes, 145 systèmes de transcription peuvent être utilisés pour représenter les sons parlés, et les symboles de transcription peuvent être arabes, phonétiques, selon les raisons et le public de la transcription.

#### 2.4.2.2 Translitération

La Translitération (ou conversion graphitique) est l'utilisation systématique des symboles d'un système d'écriture pour représenter ceux d'autrui, l'idéal étant une correspondance individuelle pour que quelque chose écrit par translittération puisse être converti en orthographe originale et vice versa sans ambiguïté. Le terme est le plus souvent utilisé pour désigner des systèmes d'utilisation de l'alphabet romain pour représenter divers alphabets orientaux. Ainsi, la translittération serait l'écriture d'un script romanisé équivalent a écrit script en arabe, représentant tous les éléments orthographiques.

Le problème principal ici est que le script arabe comporte des orthographes «peu profondes» et «profondes», c'est-à-dire qu'il diffère dans la façon dont il «dépeint les relations sonore-symbole». Le script qui comprend toutes les voyelles courtes et les diacritiques (tels que shadda et waṣla) est appelé «peu profond» - c'est-à-dire plus facile à lire, et est donc utilisé pour enseigner aux enfants arabophones comment lire. Le script qui est "profond" manque de ces fonctionnalités, en supposant que les lecteurs adultes savent facilement ce qu'ils sont. Le fait que les voyelles courtes et les diacritiques ne sont pas représentés dans un script profond ne signifie pas qu'ils n'existent pas; La convention orthographique standard en arabe les omit et les prend comme ils l'ont compris. Il s'agit d'un problème de traitement conceptuel et cognitif pour les apprenants d'arabe en tant que langue étrangère, et cela affecte aussi le rendu de l'écrit en arabe en translittération complète.

Les voyelles courtes et les diacritiques sont invisibles dans un script arabe normal, mais sont évidemment prononcés si le texte écrit est lu à haute voix. Par conséquent, pour l'arabe, un système hybride de translittération / transcription - qui tient compte de la prononciation et de l'orthographe - est devenu la norme pour les publications occidentales qui doivent utiliser l'arabe translittéré.

### 2.4.3 Les enjeux de la transcription

Loin d'être un exercice simple et mécanique, consistant à écrire que ce font et disent les participants enregistres, la transformation d'un événement temporal et multimodale en une représentation textuelle est radicalement sélective et «configurant», posant une série de problèmes très divers bien qu'articules, pratique et technologique, théorique représentationnels.[15]

#### 2.4.3.1 La transcription comme action située

La transcription est le produit de pratiques professionnelles d'écoute et de visualisation spécifique, qui dépendaient nouvellement encore manipulation de données analogique (grâce aux fonctions Start et Play des enregistreurs).Elles sont liées aujourd'hui aux divers traitements informatiques possibles de données numérisées association du flux vidéo à la visualisation du flux audio, sans parler des possibilités de coupage et de montage. L'écoute et le visionnement répète des détails sont un fondement de la transcription comme de l'analyse : en ce sens, la répétition 1 est la caractéristique essentielle de cette pratique : c'est une constructive car les données sont soumises à des coupés s électives, au haut-parleur à l'écoute au casque ou d'un type casque à un autre), de rythme de qualité intervient.

Ce sens du détail peut être précisé si l'on tient compte du fait que l'organisation de l'interaction repose de manière cruciale, du point de vue de l'analyse conversationnelle d'inspiration ethno méthodologique, sur la gestion de sa temporalité et de sa séquentiellité. Dans ce sens, le travail de la transcription permet de suivre pas à pas le travail interactif effectue par les participants en temps réel : en cela elle hérite ses contraintes et ses choix d'une vision théorique de l'interaction qui souligne l'impotence de son organisation émergente, temporelle, incrémentale et interactive.

#### 2.4.3.2 La transcription comme problème de représentation

La transcription effectue une transformation radicale des données temporelles et dynamiques en données écrites, spatialisées par conséquent. Représenter le temps en l'inscrivant dans l'espace des conventions écrites constitue le problème fondamental que doit résoudre une convention de transcription. Plusieurs solutions ont été proposées dans la littérature : la spatialisation en colonne la partition, le format liste. Chacune propose une conception spécifique du flux temporel et de son articulation en unités. [15]

En effet, l'écriture se caractérise par un espace vide : tel est le cas du mot qui résulte d'un découpage morphologique du flux verbal voulu par l'orthographe ; tel est le cas de la ligne qui résulte de choix théorique ou de contraintes pratique (aller à la ligne lorsqu'il n'y a plus de place).

#### 2.4.3.3 La transcription comme entité hétéronome

La transcription est une entité hétérogène, un objet intermédiaire relie à de nombreux autres objets. Nous soulignerons ici deux aspects, responsables de cette hétéronomie : le paradoxe qui fonde l'exercice de la transcription, visant à produire une « fixation » et l'ensemble de pratiques dont dépend la transcription.

La transcription est une acte qui tente de résoudre une contradiction central dans l'étude oral et des pratiques langagières et sociales : elle vise à produire une « fixation dynamique», c'est-à-dire une entité qui cherche à documenter par et dans une « inscription » [15]. Autrement dit, elle essaie de rendre saisissable et reproductible un moment labile disparu après son occurrence. Elle partage ainsi avec l'enregistrement la propriété d'être une reconstruction par le biais d'une bande ou d'un texte des caractéristiques formelles et temporelles d'un labile, disparu à jamais.

Cette dimension reconstructive nous invite à la considérer d'ailleurs à l'aide d'autres images conceptuelles que celles de « l'enregistrement » ou du « recueil » de « données » (termes qui tous laissent entendre déjà-là de l'événement qu'il s'agit simplement de « capturer »), par exemple en termes de « fabrication », « production », « construction » d'évidence, de traces, de documents.

## Conclusion

Dans ce chapitre, on s'est focalisé sur les mécanismes de la production du signal de parole (la phonation et l'articulation), les éléments constitutifs du système phonatoire, les principales caractéristiques. Et puis on a montré le développement de la relation de communications être humain machine. Ensuite On a passé en revue l'ensemble des changements qui passent au cours de la transcription d'un signal de parole, en général, et un signal de parole en langue Arabe d'une façon précise. Dans le chapitre suivant nous allons développer plus en détails la notion de style de l'écriture chez l'être humain.

# *Chapitre-3*

# *Expériences et résultats obtenus*

# Chapitre-3
# Expériences et résultats obtenus

## Introduction

Nous exposerons dans ce chapitre les séries d'expériences d'attribution d'auteur effectuées sur la base de données textuelle (ou Corpus) que nous avons conçu pour cette fin et qui est contient deux catégories de textes : la première catégorie comprend les textes qui sont écrits directement par leurs auteurs que nous avons appelés ' textes écrits' et la deuxième catégorie comprend les textes qui sont transcrits à partir des fichiers audio ou vidéo que nous avons appelés 'les textes transcrits'. Par la suite, nous examinerons les résultats obtenus, tout en essayant de donner des interprétations objectives et des conclusions concrètes.

## 3.1 Base de données textuelle (Corpus)

La base de données textuelle (ou Corpus) comporte **36 textes** de chaque catégorie appartenant à **12 auteurs-locuteurs** dont 3 textes pour chacun d'entre eux. Ces textes, ayant une taille moyenne de **300 mots**, ont fait l'objet d'une série d'opérations de préparation (ou de prétraitement) avant leurs utilisation dans les expériences d'attribution d'auteurs.

## 3.2 Opérations de préparation du corpus

Avant de commencer toute expérience quantitative de fouille de données sur une base de données textuelles, il est nécessaire de commencer par des opérations préliminaires ayant pour but de préparer cette dernière. Les opérations les plus importantes qu'on utilise pour préparer un corpus sont illustrées dans le **tableau 3.1** ci-dessous. Il est à noter que selon le type d'analyse, on peut ignorer une étape pour des raisons pratiques ou théoriques.

**Tableau 3.1 Opérations de préparation du corpus**

| Opérations de préparation | Observations |
|---|---|
| **Extraction du corps de document** | Pour les pages Web contenant des menus et autres éléments non-reliés au contenu. Plusieurs algorithmes et outils sont utilisés. |
| **Transformation en format texte brut** | Pour enlever des balises (textes HTML, XML, Latex, etc.). Le résultat n'est pas toujours optimal, notamment pour le PDF et les pages HTML. |
| **Mettre tous les mots en minuscules/majuscules** | |
| **Enlever les ponctuations** | Pour les analyses statistiques seulement. Dans certains cas on les remplace manuellement par des blancs. |
| **Enlever les nombres** | A voir cas par cas. |
| **Enlever les blancs en plus** | Pour éviter d'avoir des termes vides dans certains logiciels |
| **Remplacer des caractères spéciaux** | Par exemple des apostrophes (en Français ou en Allemands) |
| **Remplacer des mots** | Des acronymes afin de préserver une identité à un terme composé ou encore pour s'assurer que des synonymes importants soient remplacés par le même mot. |
| **Enlever les "stop words"** | Mots communs à tous les textes, et mots auxiliaires comme les articles (utilisée en conjonction avec la racinisation, mais pas avec la lemmatisation). |
| **Racinisation (ou Stemming)** | Racinisation ou Désuffixation (stemming) est un procédé de transformation des flexions en leur radical ou racine. |
| **Lemmatisation** | Regroupement des mots d'une même famille dont Chacun des mots d'un contenu se trouve ainsi réduit en une entité appelée lemme (forme canonique). Le nom, son pluriel, verbe à l'infinitif, etc." Elle est une annotation qui peut être utilisée avec un étiquetage "part of speech" (voir ci-dessous). |
| **Enlever d'autres mots** | Lorsqu'on analyse un domaine précis (par ex. des fiches sur le "jeu"), on peut enlever des termes comme "jeu" qui se retrouvent dans chaque document. |

### 3.2.1 Prétraitement des documents écrits

Une attention particulière a été portée quant au choix des textes considérés, qui ont été pris à partir des livres de ces auteurs, et qu'ils traitent tous le même thème. Avant d'entamer nos expériences, notre corpus de documents textuels est passé par les étapes de préparation suivantes :

- Suppression des espaces entre les mots, les réduire en un seul espace.
- Enlèvements des caractères spéciaux tel que : « [ @ )….
- Enlèvement de la numérotation
- Enlèvement des virgules et points d'interrogations
- Un retour ligne est établit à la rencontre des ponctuations suivantes: point (**.**) et deux points (**:)**
- Finalement, le texte ainsi traité, est enregistré en format UTF8.

### 3.2.2 Prétraitement des documents transcrits

Les documents transcrits, qui sont pris à partir des fichiers audio, appartiennent à 12 Auteur-Locuteur, 3 textes pour chacun d'entre eux. Tous les textes sont d'une longueur moyenne de 250 mots. Le tableau 3.2 suivant illustre les détails du corpus audio utilisé pour valider nos expériences :

**Tableau 3.2 Récapitulatif du Corpus Audio (PCT-17)**

| Auteurs | Textes | Nbre de mot | Longueur des fiches audio |
|---|---|---|---|
| Abdelmajid Zandani | ***AZ_txt-1*** | 269 | 4,17 min |
| | ***AZ_txt-2*** | 272 | 3,57 min |
| | ***AZ_txt-3*** | 273 | 4,21 min |
| Lina Elhimsi | ***LE_txt-1*** | 250 | 2,26 min |
| | ***LE_txt-2*** | 252 | 2,17 min |
| | ***LE_txt-3*** | 260 | 2,10 min |
| Mohamed Elarifi | ***ME_txt-1*** | 252 | 2,31 min |
| | ***ME_txt-2*** | 257 | 2,57 min |
| | ***ME_txt-3*** | 261 | 1,52 min |
| Mohamed Hassan | ***MH_txt-1*** | 281 | 2,47 min |
| | ***MH_txt-2*** | 280 | 3,24 min |
| | ***MH_txt-3*** | 278 | 2,50 min |

| Auteurs | Textes | Nbre de mot | Longueur des fiches audio |
|---|---|---|---|
| Omar Abdelkafi | ***OA_txt-1*** | 257 | 2,15 min |
| | ***OA_txt-2*** | 257 | 2,10 min |
| | ***OA_txt-3*** | 260 | 2,28 min |
| Youcef Elkaradwi | ***YE_txt-1*** | 288 | 3,40 min |
| | ***YE_txt-2*** | 287 | 3,28 min |
| | ***YE_txt-3*** | 290 | 2,58 min |
| Ratib Ennabolsi | ***RE_txt-1*** | 284 | 4,12 min |
| | ***RE_txt-2*** | 288 | 4,05 min |
| | ***RE_txt-3*** | 287 | 3,33 min |
| Mohamed Elgazali | ***AA_txt-1*** | 298 | 3,02 min |
| | ***AA_txt-2*** | 300 | 4,56 min |
| | ***AA_txt-3*** | 299 | 3,14 min |
| Ayid ElKarani | ***AE_txt-1*** | 307 | 2,30 min |
| | ***AE_txt-1*** | 287 | 2,25 min |
| | ***AE_txt-1*** | 285 | 2,29 min |
| Amro Khaled | ***AK_txt-1*** | 384 | 2,45 min |
| | ***AK_txt-2*** | 385 | 5,03 min |
| | ***AK_txt-3*** | 385 | 4,27 min |
| Hanane ElKatan | ***HE_txt-1*** | 323 | 2,22 min |
| | ***HE_txt-2*** | 329 | 2,45 min |
| | ***HE_txt-3*** | 323 | 2,30 min |
| Salman Elouda | ***SE_txt-1*** | 274 | 2,50 min |
| | ***SE_txt-2*** | 279 | 2,34 min |
| | ***SE_txt-3*** | 275 | 2,50 min |

### 3.3 Transcriptions des fichiers audio

Les documents transcrits doivent être traités avant leur utilisation pour l'attribution de leurs véritables auteurs. Ce traitement se résume en trois types d'opérations suivantes :

- On écoute attentivement et on écrit ce que l'auteur-locuteur dit ; s'il s'arrêt plus de (3-4 secondes) on revient à la ligne, et on doit écrire les redondances (les mots répétés).
- Après avoir écrit les textes, on doit supprimer leurs préambules communs et les citations qui ne leur appartiennent pas (poèmes, coran, hadith).
- Finalement, le texte ainsi traité, est enregistré en format UTF8.

- Au total, le corpus contient 36 textes ; 24 textes l'ensemble d'apprentissage, et 12 textes l'ensemble de test.

### 3.4 Travail expérimental

Le système d'attribution d'auteurs que nous avons proposé est basé sur l'utilisation des descripteurs connue sous le non de "N-grammes" et d'un classifieur basé sur les réseaux de neurones ou "Multi Layer Perceptron" (MLP).

Une caractéristique principale de notre base de données c'est que les textes choisis, sont de taille réduite environ 300 mots en moyenne par texte. Ce qui représente un véritable test de la robustesse de notre système d'attribution d'auteurs.

Pour nos expérimentations, deux ensembles de documents ont été choisis :

- Le premier ensemble de textes utilisé est l'ensemble d'entrainement et d'apprentissage qui contient deux textes pour chaque auteur.
- L'ensemble de test, quant à lui, est constitué d'un document pour chaque auteur. Donc nous avons au total 32 documents dont 24 pour l'apprentissage et 10 pour le test.

#### 3.4.1 Méthode utilisées pour l'attribution d'auteurs

L'attribution d'auteurs est effectuée en utilisant les descripteurs (features) suivants : N-grammes et un classifieur MLP. Les performances de notre système d'attribution d'auteurs sont définies en utilisant la relation suivante :

$$TAA = \frac{\text{Nombre de documents correctement attribués}}{\text{Nombre de documents testés}}$$

**TAA** est Taux d'attribution d'auteurs**.**

### 3.5 Séries d'expériences réalisées

Dans ce travail expérimental on a réalisé trois séries d'expériences ; la première série concerne l'attribution d'auteurs des textes écrits, la deuxième série est dédie à l'attribution d'auteurs pour les textes transcrits et la troisième série traite l'attribution d'auteurs en utilisant les deux types de textes (écrits et transcrits). Cette dernière est divisée en deux séries d'expériences ; la première utilise les textes écrits dans la phase d'apprentissage et les

textes transcrits dans la phase de test et la deuxième série utilise les textes transcrits dans la phase d'apprentissage et les textes écrits dans la phase de test.

### 3.5.1 Séries d'expériences pour les textes écrits

#### 3.5.1.1 Pour (N=2) – MLP (Multi-layer perceptron)

Dans cette série les expériences ont étés effectuées pour différentes valeurs de N, les résultats obtenus sont illustrés dans les tableaux et les figures suivantes suivies de quelques discussions afin d'expliquer les résultats de chaque expérience.

**Tableau 3.3 : Taux d'Attribution d'Auteurs (TAA) pour (N=2)**

| **Most common Events (MCE)** | 20 | 30 | 40 | 50 | 60 | 70 | 75 |
|---|---|---|---|---|---|---|---|
| **TAA %** | 30 | 60 | 60 | 70 | 70 | 80 | 90 |

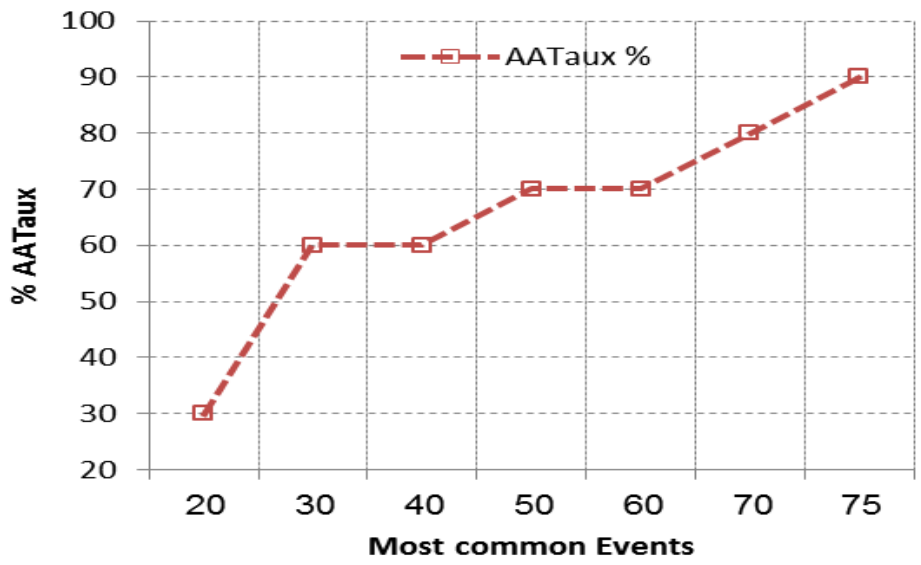


**Figure 3.1 : Taux d'Attribution d'Auteurs (TAA) pour (N=2)**

Dans cette figure on constate une croissance du TAA et une stabilité du système.

### 3.5.1.2 Penta-grammes (N=5) – MLP (Multi-layer perceptron)

**Tableau 3.4 TAA – MLP pour N=5**

| **Most common Events (MCE)** | 100 | 150 | 200 | 250 | 300 | 350 | 400 |
|---|---|---|---|---|---|---|---|
| **TAA %** | 60 | 70 | 80 | 80 | 90 | 90 | 90 |

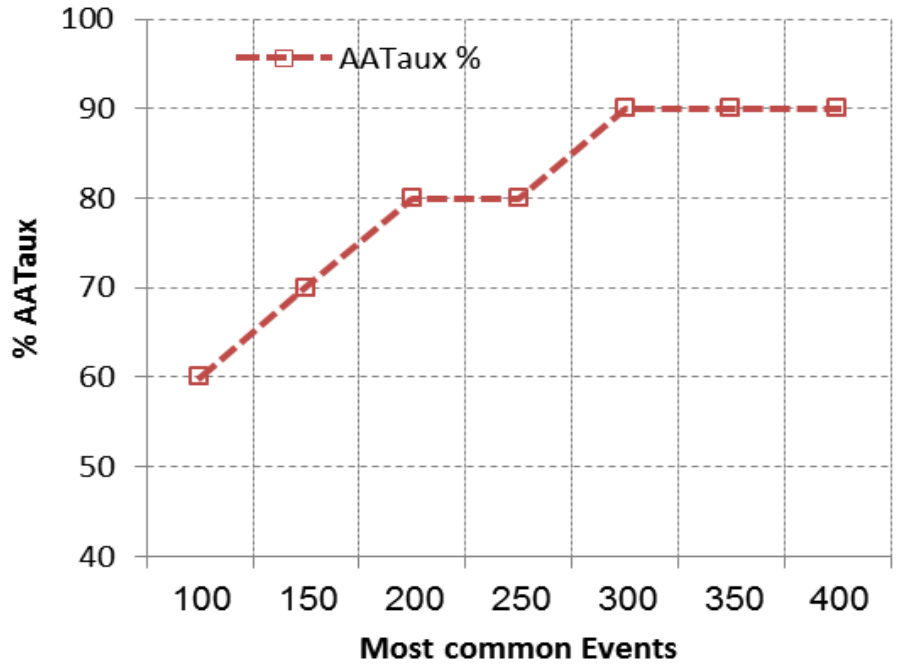


**Figure 3.2 : TAA MLP – penta-grammes (N=5)**

La Figure 3.2 montre des AATaux de plus en plus croissants et un niveau d'attribution de 90% est atteint. Le système montre une très bonne stabilité.

Un tableau récapitulatif des Taux d'Attribution d'Auteurs pour (N=1, 2, …, 5) est illustré ci-dessous :

**Tableau 3.5 TAA pour classifieur MLP avec N-grammes**

| N | AATaux |
|---|---|
| 1 | 58.33 |
| 2 | 65,71 |
| 3 | 85.00 |
| 4 | 76.00 |
| 5 | 80.00 |

Ce tableau récapitule les résultats obtenus avec les différents classifieurs en utilisant un classifieur MLP avec la méthode des N-grammes.

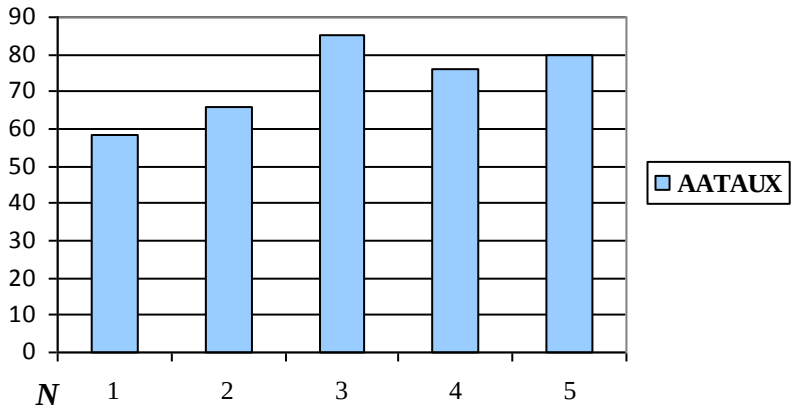


**Figure 3.3 TAA -N-gramme**

La Figure 3.3 représente le comportement de notre système par rapport vaux des différents classifieurs utilisés. En conclusion, on peut dire que la tâche d'attribution d'auteurs pour les documents écrits est effectuée par la méthode basée sur les N-grammes et le classifieurs (MLP). Bien que, les textes choisis sont de petite taille (800 mots environ par texte), les résultats obtenus sont très encourageants. La meilleure performance est obtenue avec le classifieur MLP pour les penta-grammes (N=5=, ce dernier est vivement

conseillé pour les systèmes de catégorisation de documents dont la contenance textuelle est réduite (ou small texts).

### 3.5.2 Séries d'expériences pour les textes transcrits

Dans cette série les expériences Pour effectuer la tâche d'attribution d'auteurs des documents textes transcrits, on a utilisé le N-grams Caractère pour différentes valeurs de N, Les résultats obtenus sont donnés sous formats de tableaux, suivi d'une discussion pour chaque expérience.

#### 3.5.2.1 Pour (N=2) – MLP (Multi-layer perceptron)

**Tableau 3.6 Taux d'Attribution d'Auteurs (TAA) pour (N=2)**

| **Most Common Events (MCE)** | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 |
|---|---|---|---|---|---|---|---|---|
| **TAA %** | 33.33 | 50 | 66.66 | 66.66 | 66.66 | 75 | 66.66 | 66.66 |

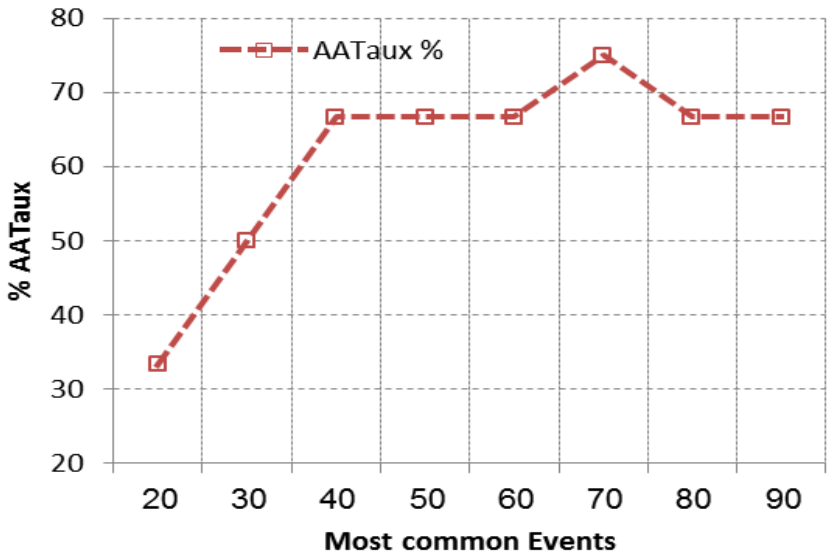


**Figure 3.4 Taux d'Attribution d'Auteurs (TAA) pour (N=2))**

La courbe de la fig.4.2 donne le taux d'Attribution d'Auteurs en fonction du MCE. On remarque que la courbe représentée dans la Figure 3.4 qui augmente rapidement (de

33.33% à 66.66%) pour une variation du MCE entre 20 et 60 avec une valeur de N=2. Ensuite, le taux reste stable pour les autres valeurs de MCE.

### 3.5.2.2 Penta-grammes (N=5) – MLP (Multi-layer perceptron)

**Tableau 3.7 TAA MLP – penta-grammes (N=5)**

| **Most common Events (MCE)** | 100 | 150 | 200 | 250 | 300 | 350 | 400 | 450 |
|---|---|---|---|---|---|---|---|---|
| **AATaux %** | 41.66 | 33.33 | 50 | 41.66 | 50 | 58.33 | 41.66 | 33.33 |

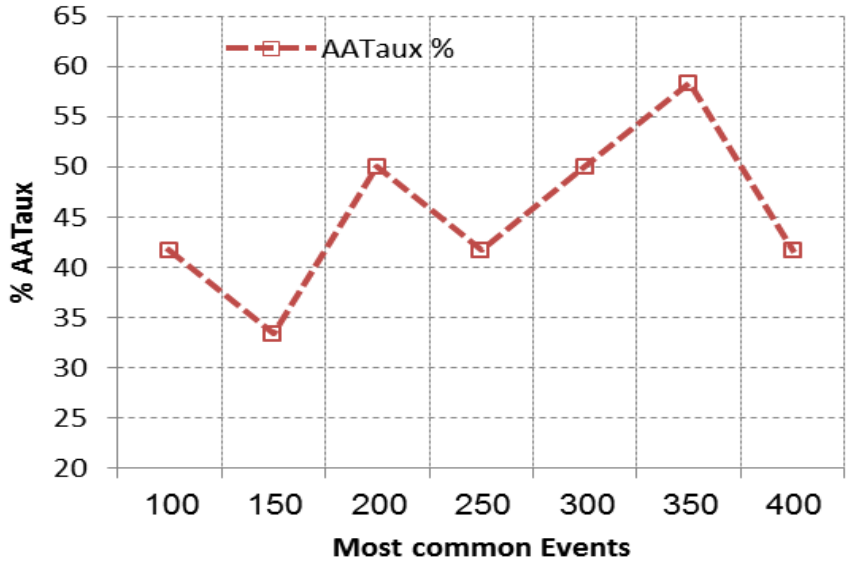


**Figure .3.5 TAA MLP – penta-grammes (n=5)**

La courbe de la fig.4.2 donne la variation de Taux en fonction de (Most Commun Event) MCE avec la valeur de **N=5** dans un intervalle [100; 400] et en utilisant le classifieur MLP, est variant de 33.33% à 58.33%. Dans cette tache nous avons effectué des expériences d'attribution d'auteur des documents textes transcrits à partir des fichiers audio. La méthode utilisée pour l'attribution d'auteurs est basée sur l'utilisation des N-grammes et classifieur (MLP). Les expériences sont effectuées sur une base de données qu'on a appelée "**P**arole **C**onvertis en **T**exte " (**PCT-17).** Bien que, les textes utilisés sont

de taille réduite (250 mots par texte), les résultats obtenus sont encourageants et une moyenne performance est obtenue.

### 3.5.3 Séries d'expériences pour un mélange de textes (écrits et transcrits)

Dans cette série les expériences ont étés effectuées pour différentes valeurs de N, les résultats obtenus sont illustrés dans les tableaux et les figures suivantes suivies de quelques discussions afin d'expliquer les résultats de chaque expérience.

#### 3.5.3.1 Textes écrits pour l'apprentissage et textes transcrits pour le test

Pour nos expérimentations, deux ensembles de documents ont été choisis : Le premier est l'ensemble d'entrainement et d'apprentissage qui contient deux textes écrits pour chaque auteur. L'ensemble de test, quant à lui, est constitué d'un texte transcrit pour chaque auteur. Donc nous avons au total 36 documents dont 24 pour l'apprentissage et 12 pour le test.

**Tableau 3.8 : Taux d'Attribution d'Auteurs (TAA) pour (N=5)**

| **Most common Events (MCE)** | 50 | 100 | 150 | 200 | 250 | 300 | 350 | 400 | 450 |
|---|---|---|---|---|---|---|---|---|---|
| **TAA %** | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 |

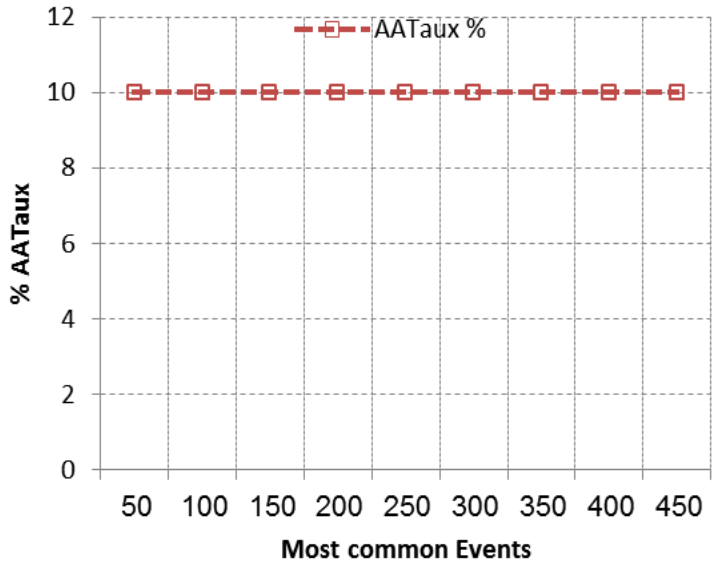


**Figure 3.6 : Taux d'Attribution d'Auteurs (TAA) pour (N=5)**

On remarque que la courbe représentée dans la fig.4.1, donnant la variation du Taux d'Attribution d'Auteur en fonction de (Most Commun Event) MCE avec la valeur de n=5 et en utilisant le classifieur (MLP), est stable variant à 10%.

### 3.5.3.2 Textes transcrits pour l'apprentissage et textes écrits pour le test

Dans cette série d'expérimentations, deux ensembles de documents ont été considérés : Le premier ensemble de textes utilisé est l'ensemble d'entrainement et d'apprentissage qui contient deux textes transcrits pour chaque auteur. L'ensemble de test, quant à lui, est constitué d'un texte écrit pour chaque auteur. Au total, nous avons 36 documents dont 24 pour l'apprentissage et 12 pour le test.

**Tableau 3.9 : Taux d'Attribution d'Auteurs (TAA) pour (N=5)**

| **Most common Events (MCE)** | 50 | 100 | 150 | 200 | 250 | 300 | 350 | 400 | 450 |
|---|---|---|---|---|---|---|---|---|---|
| **TAA %** | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 | 10 |

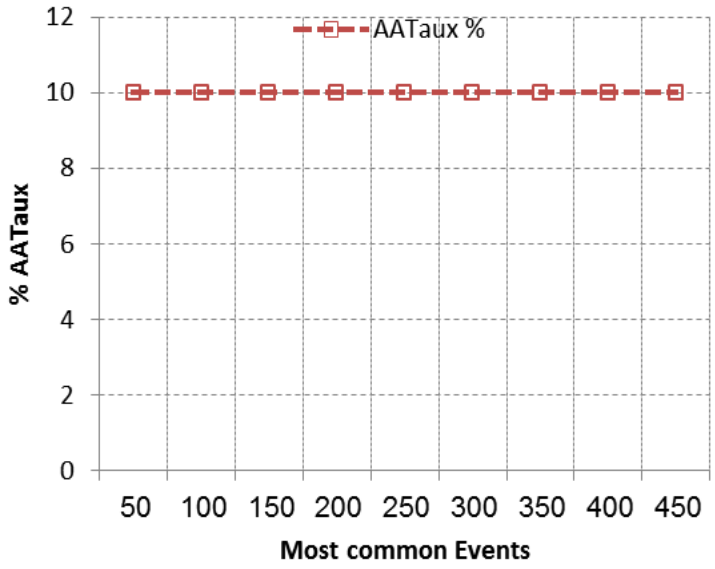


**Figure 3.7 : Taux d'Attribution d'Auteurs (TAA) pour (N=5)**

On remarque que la courbe représentée dans la fig.4.1, donnant la variation du Taux d'Attribution d'Auteur en fonction de (Most Commun Event) MCE avec la valeur de n=5 et en utilisant le classifieur (MLP), est stable variant à 10%.

**Conclusion**

Dans ce chapitre nous avons effectué des expériences d'attribution d'auteur des documents textes écrits et textes transcrits à partir des fichiers audio et un mélange des textes écrits et transcrits. La méthode utilisée pour l'attribution d'auteurs est basée sur l'utilisation des N-grammes et classifieurs (MLP). Les expériences sont effectuées sur une base de données (Corpus). Bien que, les textes utilisés sont de taille réduite (250 mots par texte), les résultats obtenus sont de faibles performances.

# Conclusion général

## ❖ Travail réalisé

Le thème que nous avons abordé dans ce mémoire s'intéresse à l'etude comparative sur les performances des techniques d'identification d'auteurs à partir des documents écrits et des documents transcrits, sur la tâche d'attribution d'auteurs. Le corpus que nous avons conçu pour réaliser nos expériences, est construit autours d'une base de données constituée de 12 Locuteurs-Auteurs et 36 textes écrits et 36textes transcrits (3 textes pour chaque catégorie) d'une taille moyenne d'environ 250 mots pour chaque texte. L'originalité de ce travail de recherche est que la tache d'Attribution d'Auteurs (AA) a été appliquée aux textes écrits et textes transcrits qui ont été reporté fidèlement en écoutant des fichiers audio (discours) et non pas été écrits directement par les auteurs.

La tache principale de notre système est de reconnaitre le véritable Locuteur-Auteur d'un document textes écrits et texte transcrit.

Notre système d'Attribution d'Auteurs est basé sur l'utilisation du descripteur Character N-gram comme caractéristique (feature), qui a démontrée son efficacité dans la tâche de l'AA au cours de nos expériences, et d'un classifieur de type MLP qui a été utilisé pour mener à bien l'opération d'identification du Locuteur Auteur.

## ❖ Discussion des résultats obtenus

Les résultats obtenus étaient très encourageants vu la contrainte liée à la taille des textes choisis (250 mots uniquement), qui. On a vu l'importance et l'efficacité de la représentation en n-grammes pour la tâche de l'AA.

A partir de ces résultats, on a constaté que le classifieur MLP et la représentation en penta-grammes sont les plus appropriés aux tâches d'AA quand il s'agit de textes de taille réduite.

❖ **Suggestion de perspectives**

Afin d'améliorer les performances de notre système, on suggère en perspectives

de compléter le travail réalisé avec les tâches suivantes :

- Combinaison de descripteurs.
- Fusion de classifieurs au niveau du score.

## Références bibliographiques


**[1]** **P. Juola**, Foundations and Trends
« Information Retrieval.Vol. 1, No. 3 » (2006)

**[2]** **Fadoua Drira**
Thèse doctorale « Contribution à Restauration des Images de Documents Anciens ».
Institut National des Sciences Appliquées de Lyon 2007.

**[3]** **Françoise Flieder** et all.
« sauvegarde et conservation »
(Unesco) 1983.

**[4]** **Anis Kricha** et all.
« Exploration des Ondelettes en Pr_etraitement des Documents Anciens ».
Sep 2006, SDN06, pp.157-162. <hal-00115728>

**[5]** **Mohamed Ben Halima** et all.
« Restauration des images couleurs de documents arabes anciens basée sur les EDPs ».
Sep 2006, SDN06, pp.103-108. <hal-00113558>

**[6]** **Fidelia Ibekwe-SanJuan**
« Fouille de textes : méthodes outils et applications »
Livre, ISBN : 978-2-7462-1609-9, Edition Lavoisier, Paris, 2007

**[7]** **Hacène. C**
Thèse doctorale de l'université de Henri Poincaré- Nancy
« Etude et réalisation d'un système d'extraction de connaissances à partir de textes ».
15 Novembre 2004.

**[8]** **Feldman R** et all.
« Trends graph: visualizing the evolution of concept relationships in large document collections ».
Springer Verlag, Berlin, 38-46, 1998.

**[9]** **Sebastiani F**.
« Text Categorization,Text Mining and its Applications to Intelligence,CRM and Knowledge Management »
WIT Press,109-129,2005.

**[10]** **Kodratoff Y**.
« Knowledge discovery in texts: A definition and application »
Springer- Verlag, Berlin, n° 1609, 16 – 29, 1999

**[11]** **Violaine Prince** et all.
« Le Défi fouilles de textes : quels paradigmes pour la reconnaissance d'auteurs? »
Revue Des Nouvelles Technologies De L'information
Cépaduès-Editions, 2007, E (10), pp.001-014. <lirmm-00171291>

**[12]** Introduction à la fouille de textes université de Paris.3, http//www.lattice.cnrs.fr/sites/itellier/poly_fouille_textes/fouille-textes.pdf

**[13]** **F. Y. Y. Choi** et all.
« Latent Semantic Analysis for Text Segmentation, Proceedings of 6th EMNLP »
pp 109-117. 2001

**[14]** **Liopis F**et all.
« Text segmentation for efficient information retrieval Proceedings of CICLing.2002 »
Lecture Notes in Computer Science , vol. 2276, pp. 373-380. 2002

**[15]** **Yang C C** et all.
« A heuristic method based on a statistical approach for chinese text segmentation ».
Journal of the American Society for Information Science and Technology, vol. 56, no13, pp. 1438-1447 2005

**[ 16 ]** **Swan R and D Jensen**.
« TimeMines : Constructing timelines with statistical models of word usage, Proceedings of KDD-2000 »
Workshop on Text Mining, pp 73-80.2000.

**[17]** **Kontosthatis A** et all.
« A Survey of Emerging Trend Detection in Textual Data Mining, In Berry M.W (eds.), Survey of Text Minin »
Springer,NY, 2004, 186-223.

**[18]** **R. Jalam**
« Apprentissage automatique et catégorisation de textes multilingues »

**[19]** **I.Moulinier**
« Une approche de la catégorisation de textes par l'apprentissage symbolique »

**[20]** **F.Sebastiani**
« Machine learning in automated text categorization »

**[21]** J.Clech, D.A.Zighed « Une technique de réétiquetage dans un contexte de catégorisation de textes »

**[22]** **P.Hayes** et all.
« Construe/Tis : A system for content-based indexing of a database of news stories »

**[23]** **K. Lang**
« NewsWeeder : Learning to Filter Netnews »

**[24]** **R.Armstrong** et all.
« WebWatcher : a Learning apprentice for the World Wide Web »

**[25]** **G.Salton & M.McGill**
« Introduction to Modern Information Retrieval»

**[26]** **Y.Yang**

« An evaluation of statistical approch to text categorization »

**[27]** **M.Sahami**

« Using Machine Learning to Improve Information Access »

**[28]** **C.D.Loupy**
« Évaluation de l'Apport de Connaissances Linguistiques en Désambiguïsation Sémantique et Recherche Documentaire »

**[29]** **William B Cavnar** et all.
« N-Gram Text Categorization »

**[30]** **M.F.Porter**
« An algorithm for suffix stripping »

**[31]** **C.Shannon**
« The Mathematical Theory of Communication »

**[32]** **C.D.Manning** et all.
« Introduction to Information Retrieval »
Cambridge University Press 2008.

**[33]** **Lebart and Salem**
« Statistique textuelle » 1994

**[34]** **Hull**
« la régression logistique » 1994

**[35]** **E.D.Wiener**
« A neural network approach to topic spotting in text » 1995

**[36]** **T.Joachims**
« Text categorization with support vector machines: learning with many relevant

features » 1998

**[37] T.Joachims**
« Transductive inference for text classification using support vector machines » 1999

**Résumé**

L'attribution d'auteur d'un texte inconnu ou douteux est l'un des plus anciens problèmes de la statistique appliquée à la littérature. Ce type d'étude linguistique consiste à attribuer à un texte anonyme son auteur réel. Dans ce travail de recherche, on s'intéresse à l'identification des locuteurs à travers des textes écrites et à la transcription de leur discours en texte . notre système d'Attribution d'Auteurs est basé sur l'utilisation du descripteur Character N-gram comme caractéristique (feature), qui a démontrée son efficacité dans la tâche de l'attribution d'auteur au cours de nos expériences, et d'un classifieur de type MLP qui a été utilisé pour mener à bien l'opération d'identification du Locuteur Auteur.

**Most clés:** Attribution d'auteur, identification, transcription de discours en texte

**Abstract**

The attribution of an author of an unknown or doubtful text is one of the most

old problems of statistics applied to the literature. This type of study linguistic consists in attributing to an anonymous text its real author. In this research work, we are interested in the identification of speakers through written texts and the transcription of their speech in text. our Authors attribution system is based on the use of the Character N-gram descriptor as a feature, which has demonstrated its effectiveness in the task of author attribution in our experiments, and an MLP classifier which has been used to carry out the Author Speaker identification operation.

**Keywords:** Author's attribution, identification, speech transcription in text.

ملخص

إن إسناد مؤلف لنص مجهول أو مشكوك فيه هو واحد من أكثر المشاكل القديمة للإحصاءات المطبقة على الأدب. هذا النوع من الدراسة اللغوية لإسناد نص مجهول إلى مؤلفه الحقيقي. في هذا العمل البحثي نحن مهتمون في تحديد المتحدثين من خلال نصوصهم المكتوبة و خطابهم المنسوخ في نصوص. يعتمد نظام الإسناد الخاص بالمؤلفين على استخدام واصف الحرف ن –غرام كميزة ، والذي أثبت فعاليته في مهمة إسناد المؤلف في تجاربنا ، و مصنف م ل ب الذي تم استخدامه لتنفيذ عملية تحديد هوية المؤلف.

**الكلمات المفتاحية :**إسناد المؤلف، تحديد الهوية، النسخ الكلام في نص .