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

FACULTE DE TECHNOLOGIE
DEPARTEMENT D'ELECTRONIQUE
N°: … 19 ... /INST/ 2022

DOMAINE: SCIENCES ET TECHNOLOGIE
FILIERE: ÉLECTRONIQUE
OPTION: INSTUMENTATION

Mémoire présenté pour l'obtention
du diplôme de Master Académique

Par : HERIZI Sohaib et MIMOUNE Soheyb

Intitulé

Acquisition des données textuelles à l'aide de transcripteurs automatiques en vue d'une identification d'auteurs basée sur la fréquence des mots et l'intelligence artificielle.

Soutenu publiquement le : 22 /06/ 2022 devant le jury composé de:

Dr. DJERIOUI Mohamed	Université M'sila	Président
Dr. KHENNOUF Salah	Université M'sila	Encadreur
Pr. SAYOUD Halim	Université USTHB	Co-Encadreur
Dr. OUALI Mohamed Assam	Université M'sila	Examinateur

Année universitaire : 2021 /2022

REMERCIEMENTS

Avant tout, nous remercions Dieu le Tout-Puissant pour son aide et ses bénédictions et pour tout ce qu 'il nous a donné.

Merci beaucoup à nos superviseurs, Dr. Salah KHANOUFA et Pr. Halim SAYOUD qui nous devons beaucoup, pour l'intérêt, la disponibilité et les conseils qui nous ont donnés.

Nous remercions également les membres du jury, chacun en son nom, d 'avoir accepté de faire partie du jury pour ce modeste travail.

Nous tenons à remercier tous les enseignants du département d 'électronique qui ont contribué à notre formation ainsi que tous les membres du cadre administratif.

Nous tenons à remercier, enfin, tous ceux qui ont aidé de près ou de loin durant ce projet de fin d'études.

S. HERIZI & S. MIMOUNE

DEDICACE

Je dédie ce modeste travail à :

Mes chers parents qui m'ont aidé et encouragé tout au long de mon parcours scolaire

Et d'être la source de mon bonheur et de ma réussite.

Mes chers frères et sœurs.

Chers amis.

Mon cher partenaire et ami avec qui je partage cet humble travail…

HERIZI Sohaib

DEDICACE

Je dédie ce modeste travail à :

Mes chers parents qui m'ont aidé et encouragé tout au long de mon parcours scolaire

Et d'être la source de mon bonheur et de ma réussite.

Mes chers frères et sœurs.

Chers amis.

Mon cher partenaire et ami avec qui je partage cet humble travail…

MIMOUNE Soheyb

Liste des abréviations

AA : Attribution d'Auteur

ADN : Acide désoxyribonucléique

API : application programming interface

ARN : Acide ribonucléique

CAT : Conversion de conférences audio en texte

EPAC : École polytechnique d'Abomey-Calavi Text to Speech

HMM : Hidden Markov model

IA : Intelligence artificielle

OOV : Our Own Voice

RVA : Rendez-Vous Aérien

SAMPA : Speech Assessment Methods Phonetic Alphabet

SVM: Support Victor Machine

TAA : Taux d'Attribution d'Auteurs

TA : Traducteur Automatique

TTS : DNN : Digital News Network

WER : Word Error Rate

Liste des figures

Liste des tableaux

Table de la matière

CHAPITRE-2

FONCTIONNEMENT DU TRANSCRIPTEUR AUTOMATIQUE

CHAPITRE-2

EXPERIENCES ET RESULTATS

INTRODUCTION GENERALE

INTRODUCTION GENERALE

Notre motivation

L'attribution de l'auteur (AA) est le processus qui consiste à deviner l'auteur d'un texte à partir d'un groupe de candidats. Lorsque les objets d'étude proviennent d'Internet, où plusieurs genres textuels, styles et langues coexistent, le défi augmente. Par conséquent, la recherche des AA peut se concentrer sur certaines de ces difficultés, telles que la mise à l'échelle lorsque l'on considère un nombre élevé d'auteurs candidats ou l'indépendance linguistique lorsque les ressources linguistiques sont limitées ou introuvable.

Le développement technologique, y compris l'informatique et l'intelligence artificielle, a dû résoudre de nombreux problèmes différents de l'attribuer à l'auteur. La mention de l'auteur n'y est pas alors systématiquement présente. La fouille de données textuelles permet de classer les auteurs par catégorie (par genre, âge ou par opinion politique) ou en tant qu'individu. Ce dernier cas de figure est appelé le problème d'Attribution d'Auteur (AA). Cela consiste à deviner l'auteur de textes à partir d'un ensemble de candidats. Ainsi, cette tâche peut être vue comme un sous-domaine de l'apprentissage automatique supervisé. Techniquement cela consiste à définir une nouvelle paire reliant un texte à un auteur. Ces méthodes peuvent aussi être utilisées pour savoir si un auteur est facilement détectable via ses productions dans un flux de textes. Ce domaine est aussi connu sous le nom de writeprint, en référence aux termes anglais « écriture » (write) et « empreinte digitale » (fingerprint).

La menace pour l'auteur aujourd'hui est l'attribution de ses idées aux mémoires et aux écrits d'autrui, ou ce qu'on appelle le vol scientifique, qui tente de prendre les citations et les œuvres de l'écrivain et de les lui attribuer.

La technique la plus courante pour la tâche AA est la stylométrie (ou étude du style). Selon la théorie, un auteur laisse involontairement dans son message textuel des indications pouvant mener à son identité. Il spécifie un ensemble de caractéristiques (numériques) qui sont cohérentes pour un certain auteur et distinguent suffisamment son style d'écriture de ceux des autres auteurs.

Nos objectifs

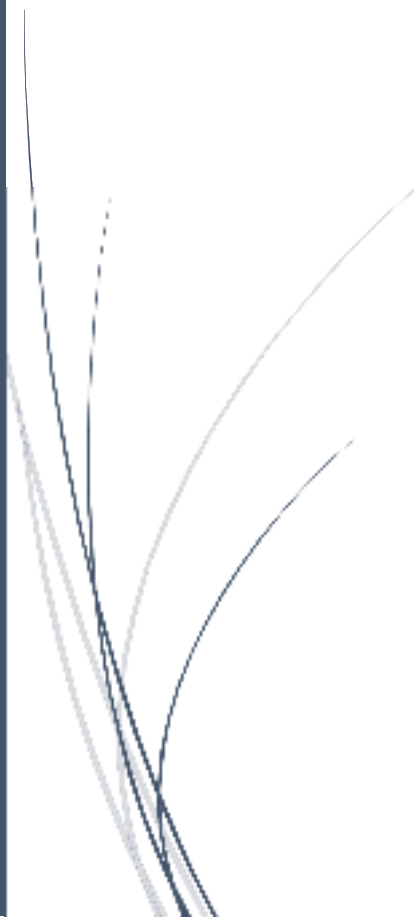
Le but de cette étude est d'examiner et d'évaluer la performance des approches d'identification d'auteur lors de la conversion de données audio en documents écrits. Pour voir si l'algorithme peut détecter les textes volés, les scripts sont rectifiés, et de nombreux descripteurs seront utilisés pour décrire le style de chaque auteur et classificateur ACD, Centroid Driver et JWCross Entropy, et pour choisir la meilleure technologie. Pour les textes à étudier, des transcriptions sont utilisées, c'est-à-dire le processus de conversion de fichiers audio en textes.

Structure du mémoire

Ce mémoire est structuré en trois chapitres, comme suit : Le premier chapitre donne des généralités sur l'attribution d'auteurs ainsi que ses différents types, les notions fondamentales de l'exploration de texte, la stylométrie dans la langue arabe. Le deuxième chapitre aborde la méthodologie de recherche qui a été adopté dans ce mémoire ainsi que les approches et techniques proposées pour l'attribution d'auteurs des documents textuelles. Dans le troisième chapitre on exposera les séries d'expériences d'attribution d'auteurs effectuées sur la base de données textuelle (ou Corpus) que nous avons conçu pour cette fin et les résultats obtenus.

CHAPITRE-1

GENERALITIES SUR L'ATTRIBUTION D'AUTEURS



CHAPITRE-1
GENERALITES SUR L'ATTRIBUTION D'AUTEURS

1.1 Introduction

L'Attribution d'Auteur d'un texte inconnu ou douteux est l'un des plus anciens problèmes de la statistique appliquée à la littérature. Dans ce chapitre, nous présentons des généralités sur l'attribution d'auteur, ensuite la stylométrie, puis catégorisation automatique de textes. Enfin, on a présente des définitions et quelques types du plagiat.

1.2 Attribution d'Auteur

La technique d'identification de la paternité probable d'un document étant donné une collection de documents dont la paternité est connue est connue sous le nom d'attribution d'auteur (AA). L'attribution de la paternité devient un gros problème car la quantité de données anonymes augmente parallèlement au développement rapide de l'utilisation d'Internet dans le monde. Détecter le plagiat, déduire la paternité des communications illégales faites de manière anonyme ou sous un pseudonyme, et résoudre les problèmes de paternité historique peu clairs ou contestés font tous partie de l'application de la paternité de l'attribution.

Lorsqu'il n'est pas clair qui a écrit un texte, l'attribution de paternité est utilisée pour déterminer qui l'a écrit. C'est utile lorsque deux personnes ou plus prétendent avoir écrit quelque chose, ou lorsque personne ne veut (ou ne peut) admettre qu'elle ou il a écrit l'article.

1.2.1 Historique d'Attribution d'Auteur

Qualifié par la loi française de « propriété », le droit d'auteur est composé de deux sortes de droits : les droits d'exploitation qui, de nature patrimoniale, ont pour fonction de garantir pendant une certaine durée les revenus des créations intellectuelles, et le droit moral, inaliénable et perpétuel, qui protège le lien intime qui unit l'œuvre de l'esprit à son auteur. La conquête des droits d'exploitation marque la première étape de l'histoire du droit d'auteur : elle est acquise à la fin du 18ème siècle. Après la Révolution française, s'ouvre une seconde étape qui s'étend jusqu'au milieu du vingt siècle : c'est celle du déploiement du droit d'auteur avec la reconnaissance du droit moral, l'extension des droits à de nouvelles œuvres et de nouveaux modes d'exploitation, l'internationalisation de la protection.

Au-delà des transformations qu'a connues le droit d'auteur, son histoire est ponctuée par la recherche d'un équilibre entre les intérêts des auteurs, ceux des exploitants et ceux du public. Au seize siècle, avec l'essor de l'imprimerie, apparaît pour la première fois en France et ailleurs en Europe un système juridique qui permet de réserver à un seul l'exploitation d'une œuvre de l'esprit. Ce système repose sur des privilèges exclusifs d'imprimer et de vendre des œuvres, concédés par le roi aux sujets qui les sollicitent et protégeant ces derniers contre la concurrence déloyale que leur causent les contrefaçons [1].

1.2.2 Définition des traits d'un auteur

Les traits utilisés en AA peuvent être séparés en différents groupes :

- Valeurs numériques associées à des mots (nombre de mots dans les textes, nombre de caractères par mot, nombre de bi-grammes/trigrammes de caractères au sein de ces mots) autrement dit des traits lexicaux.
- Valeurs associées à la syntaxe des phrases (effectifs des mots outils, des monogrammes, bi-grammes, et tri-grammes de ces mots outils ou des séquences de parties du discours).
- Valeurs numériques associées à des unités plus grandes (nombre de paragraphes ou encore longueur moyenne des paragraphes), autrement dit des traits structurels.
- Valeurs associées avec le contenu thématique (des sacs de mots, des n-grammes de mots clefs).
- Particularités en rapport avec les pratiques individuelles (telles que les fautes d'orthographe ou de frappe).

Sun et al. (2012) défendent l'utilisation d'une valeur fixe de n ne peut mener qu'à l'extraction d'informations lexicales (pour de petites valeurs de n), contextuelles ou thématiques (pour des plus grandes valeurs), mais n'expliquent pas pourquoi ou si cela est valide pour le chinois ou toutes les langues. Les auteurs soutiennent que cet inconvénient est évitable en exploitant des n-grammes de longueurs variables (des sous-chaînes de longueur entre 1 et n), donc en capturant des informations de types différents (lexicales, contextuelles et thématiques). Des sous-chaînes de longueurs variables sont également exploitées dans cette étude pour voir l'impact de ce paramètre sur les résultats en français et en anglais [3].

1.2.3 Catégorisation de textes fondée sur les traits

Différentes techniques pour exploiter les traits extraits des textes ont été proposées. SVM (Support Victor Machine ou Séparateur à Vaste Marge) et les réseaux de neurones (neural

network) sont des approches efficaces pour mener la tâche d'AA suivant le paradigme d'apprentissage automatique supervisé (Kacmarcik & Gamon, 2006; Tweedie et al., 1996). Quand l'ensemble des auteurs candidats est extrêmement grand ou incomplet, d'autres approches comparent les textes comme des ensembles de traits avec des fonctions spécifiques pour calculer les similarités entre ces ensembles (Koppel et al., 2011).

D'autres approches utilisent des ensembles de traits individuels via apprentissage automatique pour construire un classifieur par auteur. Chaque classifieur agit tel un expert pour traiter un sous-ensemble de l'espace de recherche lors de la classification d'un corpus, chaque classifieur étant spécialisé dans la détection d'un auteur spécifique. Les expériences décrites dans cet article utilisent un unique classifieur SVM pour l'ensemble des auteurs en gardant les mêmes paramètres pour chaque expérience, en vue d'analyser finement l'influence du choix des traits sur le traitement. Cette analyse sur les traits est alors en principe valide, même pour d'autres méthodes se basant sur ces mêmes traits [4].

1.3 Les étapes de l'attribution d'auteur

Un processus complet d'attribution de l'auteur consiste en :

- Rassemblement des textes qui sont les observations à classer
- Un méthode d'extraction de caractéristiques qui calcule les informations numériques ou symboliques issues de ces observations.
- Un système de classification ou de catégorisation qui fait le classement à partir de ces observations.

1.4 Stylométrie

La stylométrie est un domaine de la linguistique qui utilise la statistique pour décrire les propriétés stylistiques d'un texte. Elle est utilisée pour identifier le style d'un auteur, pour identifier un auteur de textes anciens, pour identifier un auteur anonyme dans le domaine judiciaire [2].

1.4.1 Principe de la stylométrie

En général, on n'a pas besoin de recourir à l'étude du style d'un auteur pour savoir l'ordre dans lequel il a composé ses œuvres. La plupart des auteurs ont eu soin d'indiquer eux-mêmes la relation de chaque écrit avec les écrits précédents, ce qui nous permet de nous rendre compte de leur progrès et de la voie qu'ils ont suivie pour arriver à leurs dernières conclusions.

Cependant, Platon a traité avec un art si parfait chacune de ses œuvres, a daigné nous renseigner si peu sur sa personne, qu'il est très difficile de trouver le commencement et la fin du cercle admirable formé par ses dialogues. De là est né le célèbre problème de la chronologie platonicienne, réputé insoluble par beaucoup d'historiens et traité de plusieurs manières contra victoires par d'autres. Comme il s'agit non seulement d'un artiste, mais aussi d'un penseur, l'ordre des œuvres de Platon est encore plus important à savoir que celui des drames de Shakespeare, sur lesquels on a dépensé tant d'érudition et de patientes recherches. On désire se rendre compte du Deve lapement de la forme aussi bien que de la pensée quand il s'agit d'un philosophe qui fut aussi un grand écrivain. La grande difficulté consiste dans l'absence à peu près complète de témoin [5].

1.4.2 Un peu de l'histoire

La stylométrie a traversé les époques et les siècles peut être inspirée par Pythagore (Tout est nombre). Elle a véritablement débuté avec le logicien anglais Auguste de Morgan, avant de se développer grâce à l'informatique moderne. La première utilisation du terme stylométrie serait due à Lutoslawski.

La stylométrie prend sa forme moderne en 1963 avec Frederick Moselle de l'université de Harvard et David Wallace, de l'université de Chicago qui publient un article fondateur dans le journal de l'American Statistique Association.

La stylométrie s'est également développée en France depuis, notamment avec Jean-Paul Benz cri, Charles Bernait, Étienne Brunet, Charles Muller et Jean-Marie Vire. Elle est enseignée notamment à l'École des Chartes [2].

1.4.3 Caractéristiques des auteurs

Chaque individu possède son propre vocabulaire, parfois riche, parfois limité. Bien qu'un vocabulaire étendu soit généralement associé à une littérature de qualité, ce n'est pas toujours le cas. Certaines personnes écrivent en phrases courtes, tandis que d'autres préfèrent les phrases complexes comportant plusieurs propositions. Il n'y a pas deux auteurs qui utilisent les points-virgules, les tirets et autres signes de ponctuation exactement pareil.

Cependant, l'identification de l'auteur d'un texte anonyme constitue l'une des applications les plus courantes de la stylométrie. Il est parfois possible de découvrir l'identité de l'auteur d'un texte en mesurant certaines caractéristiques de ce texte, comme la longueur moyenne des phrases ou le rapport entre le nombre d'articles définis et indéfinis [6].

Ces mesures sont ensuite comparées avec celles observées dans textes dont les auteurs sont connus Lorsque l'on parle de trame non –contextuelle, il s'agit de mots qui sont souvent interchangeables ou qui peuvent même être omis perte de la signification générale du texte. Ces mots contribuent peu à l'information contextuelle et sont souvent ignorés consciemment, aussi bien par le lecteur que par l'auteur. Ces mots constituent typiquement 20a 45 % du texte total, Ce qui permet d'avoir un nombre important de choix statistique, et plus les mesures statistiques sont nombreuses, plus leurs résultats sont fiables [7].

1.5 Catégorisation automatique de textes

Comme il vient d'être mentionné, le but de la catégorisation automatique de textes est d'apprendre à une machine à classer un texte dans la bonne catégorie en se basant sur son contenu. Habituellement, les catégories font référence aux sujets des textes, mais pour des applications particulières, elles peuvent prendre d'autres formes. En effet, on peut résoudre, par des techniques de catégorisation, des problèmes tels que l'identification de la langue d'un document, le filtrage de courrier électronique pertinent ou indésirable, ou encore la désambiguïsation de termes.

Un autre aspect du problème qui varie selon les applications est la présence ou non d'une contrainte concernant le nombre de catégories assignables à un document donné. Il se peut qu'on désire qu'un même texte ne soit associé qu'à une seule catégorie ou bien on peut permettre que plusieurs catégories accueillent un même document. Aussi, une précision supplémentaire est à faire : dans le cadre de la catégorisation de textes, l'ensemble de catégories possibles est déterminé à l'avance.

Il est à noter que le problème consistant à regrouper des documents selon leur similarité, mais lorsque les groupes à former sont a priori inconnus, en est un à part entière connu sous le nom de regroupement (clustering) de textes. Il n'en sera pas question ici. Arrêtons-nous au cas où les catégories font référence aux sujets des textes : la classification s'apparente alors au problème de l'extraction de la sémantique d'un texte, puisque l'appartenance d'un document à une catégorie est étroitement liée à la signification de ce texte. C'est en partie ce qui rend la tâche difficile puisque le traitement de la sémantique d'un document écrit en langage naturel n'est pas encore solutionné.

Une observation mérite aussi d'être faite concernant le fait que la nature des textes à traiter influence significativement la difficulté de la tâche de classification. Prenons l'exemple d'articles de journaux écrits généralement dans un style direct et contenant de l'information

purement factuelle. Le vocabulaire utilisé s'avère précis et souvent relativement restreint pour faciliter la compréhension. À l'opposé, imaginons un corpus de textes d'un style plus littéraire, utilisant un vocabulaire très varié et imagé. On peut aisément prévoir que la classification automatique de ce dernier corpus présentera plus de difficultés que pour l'autre [SM99]. Entre ces deux extrêmes, on peut aussi retrouver des textes scientifiques (où chaque catégorie aura potentiellement un vocabulaire caractéristique), des pages Web, du courrier électronique, etc. Chacun de ces types de textes possède des particularités qui rendent la tâche de classification plus ou moins ardue. De façon générale, on distingue deux manières d'aborder le problème de la classification automatique.

Jusqu'à la fin des années 1980, l'approche dominante pour le résoudre s'inscrivait dans une optique d'ingénierie des connaissances (knowledge engineering). On construisait un système expert comportant un ensemble de règles définies manuellement, par des experts du domaine, et qui ensuite pouvait procéder automatiquement à la classification. Ces règles prenaient généralement la forme d'une implication logique où l'antécédent portait, typiquement, sur la présence ou l'absence de certains mots, et où le conséquent désignait la catégorie d'appartenance du texte. Cette approche peut s'avérer très efficace. Entre autres, le système CONSTRUE a démontré une précision de classification impressionnante, mais sur un corpus de textes relativement petit, ce qui ne permet pas de tirer des conclusions favorables [Yan99].

En fait, malgré de bons résultats sur des banques de textes bien précises, l'inconvénient de cette approche est que l'édition des règles de décision peut s'avérer très longue. Et surtout, si des catégories s'ajoutent ou si on désire utiliser le classificateur dans un autre domaine, on doit répéter l'exercice. C'est donc la pertinence de cet ensemble de règles, qui évolue dans le temps, qui mine l'intérêt envers cette façon de faire. Avec le développement de techniques d'apprentissage automatique (machine Learning), on voit le problème d'un autre œil. En fait, même en 1961, avait proposé un classificateur bayésien qui se distinguait de l'édition de règles en se basant plutôt sur un calcul de probabilités.

Cependant, ce n'est pas avant le début des années 1990 que cette approche a pris son envol. Dans l'esprit de l'apprentissage automatique, on vise à construire des systèmes qui vont apprendre par eux-mêmes à classer les documents. On met l'accent sur l'automatisation, par apprentissage, de la création du classificateur. À l'aide d'un ensemble de textes déjà associés à des catégories, on entraîne le système.

Cette banque de textes libellés doit être préalablement construite par un humain. Ensuite, la machine tente d'apprendre la tâche de classification en observant le travail fait par l'humain. Elle essaie de généraliser les liens entre les textes et les catégories en analysant des exemples. Après la phase d'entraînement, le classificateur peut procéder lui-même au classement de nouveaux textes. Souvent, on s'inspire de la stratégie bien connue « diviser-pour-régner » et on découpe le problème en plusieurs sous-problèmes plus faciles à résoudre.

Plus précisément, on construit un classificateur binaire pour chaque catégorie, qui va déterminer si, oui ou non, un document en fait partie. En fait, l'apprentissage de la distinction de deux catégories s'avère plus facile que l'apprentissage sur l'ensemble des catégories. De plus, comme les catégories peuvent éventuellement se chevaucher, le document soumis à plusieurs classificateurs sera jugé par chacun d'eux et pourra être associé à plus d'une catégorie. En résumé, l'objectif global devient donc de créer un constructeur automatique de classificateurs. À prime abord, la tâche est plus complexe qu'avec la première approche, où elle relève d'un expert chargé de produire les règles de classification.

Cependant, en bout de ligne, lorsqu'un tel constructeur de classificateurs est mis au point, il peut être transposé d'un corpus de textes à un autre, d'un domaine à un autre, sans nécessiter beaucoup de travail supplémentaire, mis à part peut-être l'ajustement de quelques paramètres. L'effort en recherche des dernières années semble avoir été mis plutôt de ce côté. La figure 1.1 présente le processus général d'entraînement d'un tel système, tandis que la figure 1.2 illustre le processus de classification d'un nouveau document. Les étapes de représentation des documents et de sélection d'attributs, préalables à l'algorithme d'apprentissage en tant que tel, seront expliquées dans les sections qui suivent [8].

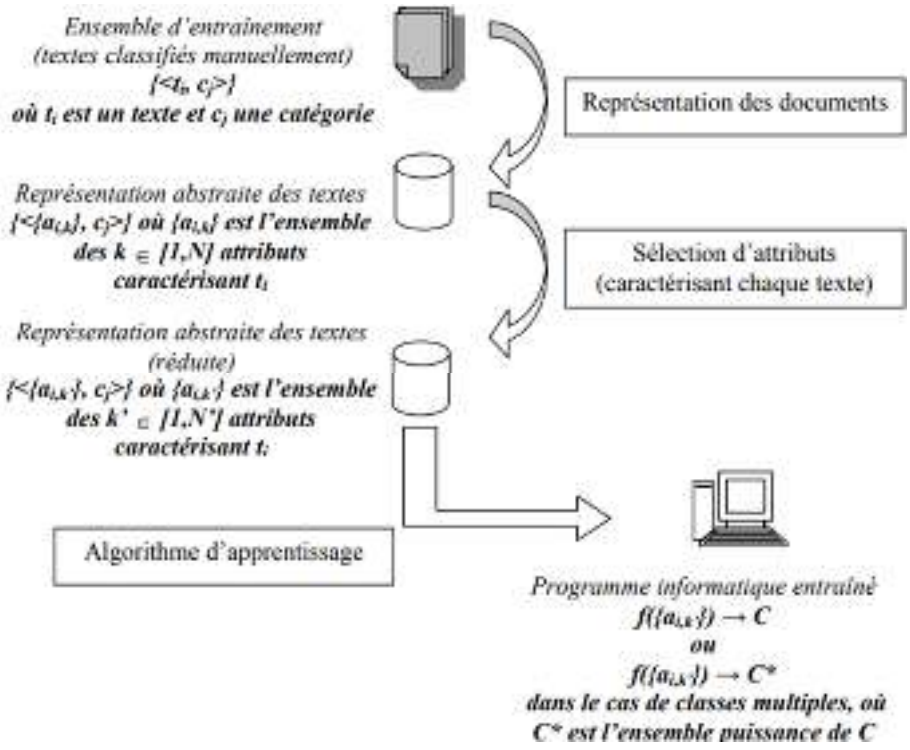


Figure 1.1 - Entraînement d'un système de classification automatique de textes

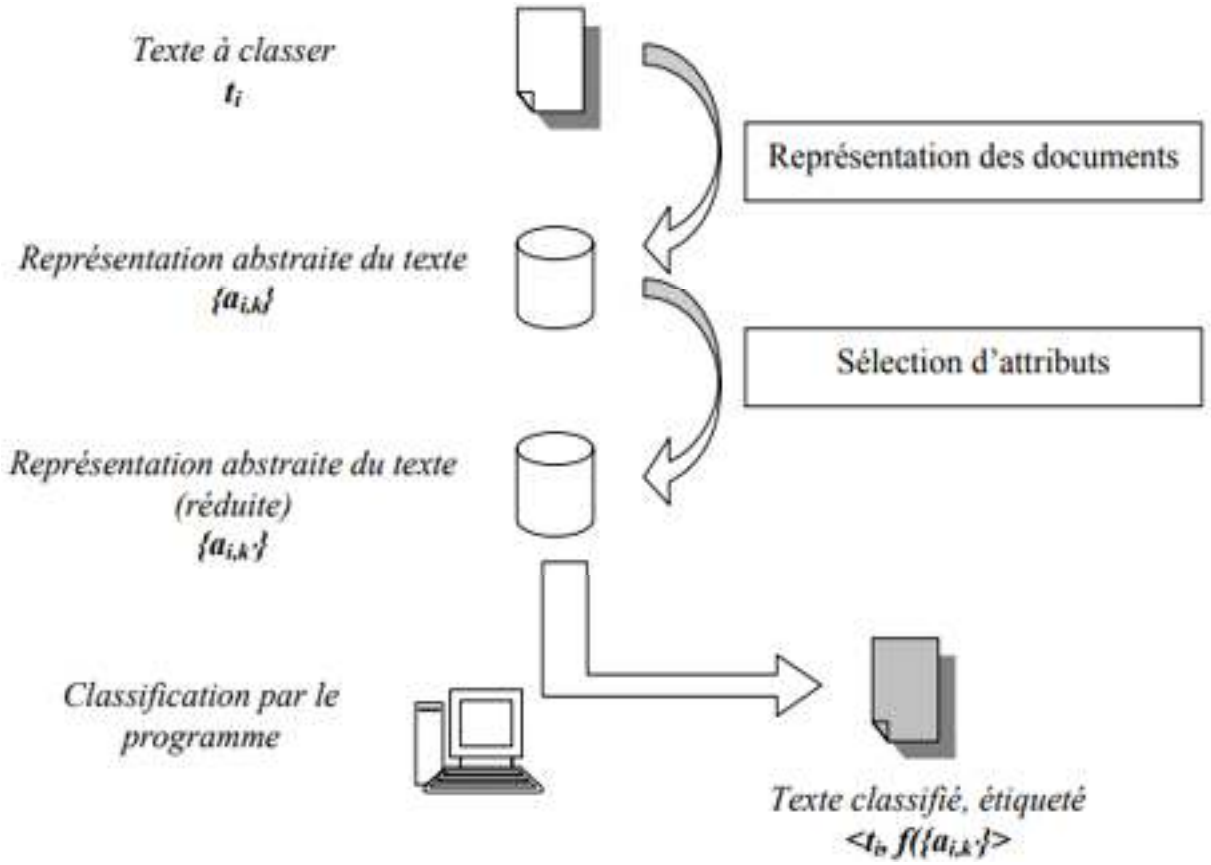


Figure 1.2- Classification d'un nouveau document

1.5.1 Les applications de la catégorisation des textes

La catégorisation de textes peut être un support pour différentes applications parmi lesquelles [9] :

- L'identification de la langue
- La reconnaissance d'écrivains et la catégorisation de documents multimédia - l'étiquetage
- Le filtrage (consistant à déterminer si un document est pertinent ou non (décision binaire)
- E routage (consistant à affecter un document à une ou plusieurs catégories parmi n

1.5.2 Démarche à suivre pour la catégorisation de textes

Pour réaliser l'opération de catégorisation automatique de textes comme nous l'avons défini, la démarche commune est la suivante : la première phase consiste donc 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). Pour améliorer la performance des modèles, une évaluation de la qualité des classifieurs et la comparaison des résultats fournis par les différents modèles est effectuée en fin de cycle. La démarche d'une approche standard de classification 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,
- Entrainer le modèle de classification à partir des couples (Document, Classe),
- Évaluer les résultats du classifieur,

La figure 1.3 illustre la démarche de catégorisation de textes avec ses trois étapes qui peuvent être schématisées comme suit [10] :

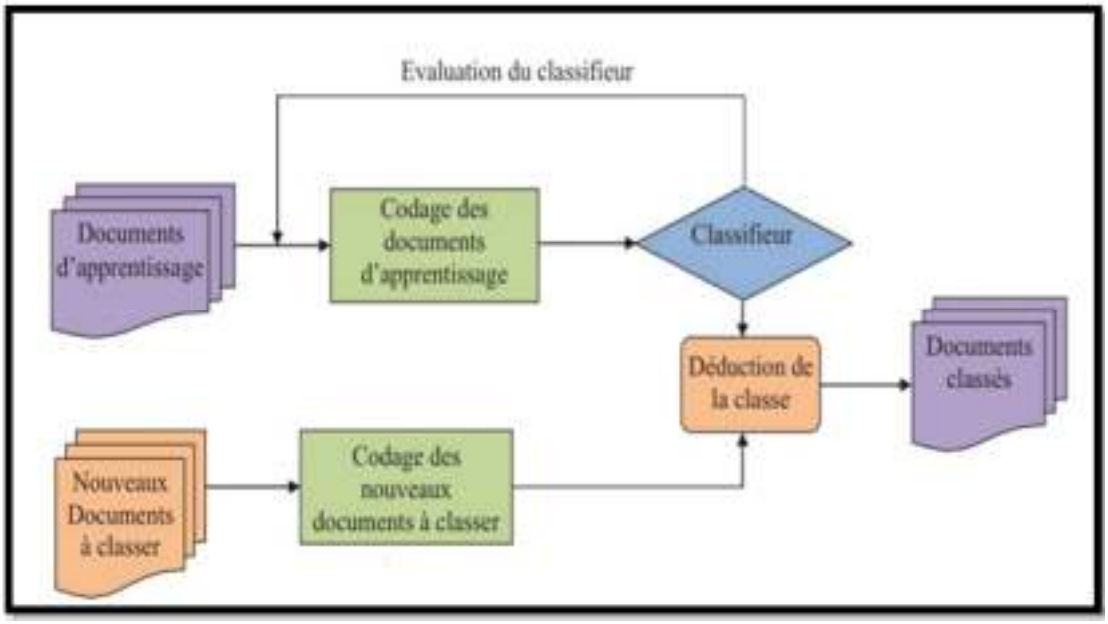


Figure-1.3 démarche de la catégorisation de textes

1.6 Plagiat

1.6.1 Définition du plagiat

Le plagiat est un terme à connotation morale et esthétique, par lequel on désigne en littérature le fait qu'un texte reprend, de façon non avouée et plus ou moins fidèlement, un élément textuel provenant d'un autre auteur. Ce terme n'a pas cours en droit, où l'on parlera plutôt de *contrefaçon* et d'infraction à la loi du droit d'auteur (*copyright*).

Dans son sens initial, le plagiat implique chez l'écrivain qui s'y adonne une volonté de dissimulation visant à donner le change sur ses talents réels et à accréditer la prétention selon laquelle il est véritablement l'auteur du texte emprunté. Le plagiat est donc à distinguer du *stellionat*, parfois appelé aussi *auto-plagiat*, qui consiste à "vendre ou hypothéquer un même bien à plusieurs personnes" (*Lexis*).

On ne parlera pas non plus de plagiat dans le cas de *rencontre involontaire*, de *similitude d'idée*s et de *réminiscence*, phénomènes assez naturels chez des écrivains qui participent d'un même tissu historique ou qui traitent d'un même sujet. Il n'y a pas davantage de plagiat dans le cas d'une *imitation* avouée, quand le modèle est sous les yeux de tous et possède valeur d'archétype. La *citation* n'est pas non plus du plagiat, car l'emprunt est isolé du texte d'accueil par des diacritiques ou des procédés typographiques et est normalement attribué à sa source. Un texte entièrement tissé de citations s'appelait jadis un *centon* et participait d'un genre littéraire tout à fait légitime.

Dans certains cas, des renvois à une œuvre connue peuvent rester implicites et emprunter la forme de l'*allusion*. Ce type de fonctionnement textuel est systématisé dans le *pastiche* et la *parodie*, où le texte emprunté sert de base à des pratiques "*hypertextuelles*" (Gérard Genette) qui peuvent être elles-mêmes affectées d'un coefficient d'originalité plus ou moins élevé. Tout en étant distinct des diverses notions énumérées ci-dessus, le plagiat couvre une aire aux frontières floues. Des critiques moralisateurs ou justiciers peuvent parfois donner au terme une extension maximale, tandis que les écrivains incriminés justifieront leur "emprunt" en l'appelant citation, pastiche, rencontre involontaire ou, aujourd'hui, *intertextualité*. Certains se retourneront même contre leurs modèles en les accusant de "plagiat par anticipation" selon le mot d'Alexis Piron, qui sera repris par l'Oulipo (voir Benbou, p. 21. Les références sont en fin d'article) [11].

1.6.2 Types de plagiat

Le plagiat, ce n'est pas seulement « copier-coller » le travail de quelqu'un d'autre. La traduction et la paraphrase de textes ou l'utilisation de synonymes sont considérées comme étant du plagiat tout autant que la reprise d'une théorie existante avec vos propres mots, sans mentionner la source et l'auteur. Autrement dit, utiliser l'idée de quelqu'un d'autre sans mentionner dûment la personne propriétaire de l'idée, c'est du plagiat.

Le plagiat prend diverses formes ; de la réutilisation d'un document entier à la réécriture d'un seul paragraphe. En fin de compte, tous les types de plagiat se résument à faire passer les idées ou les mots de quelqu'un d'autre pour les vôtres [12].

1.6.2.1 Plagiat par copier-coller

Le plagiat par copier-coller, également appelé plagiat direct, consiste à utiliser un texte provenant d'une autre source sans la citer. Si vous voulez vraiment inclure mot pour mot un passage d'une autre source, vous devez apprendre à le citer.

1.6.2.2 Plagiat en mosaïque

Copier et coller ensemble différents morceaux de texte pour créer une sorte de « mosaïque » ou de « patchwork » des idées d'autres chercheurs est un plagiat. Bien que le résultat soit un morceau de texte complètement nouveau, les mots et les idées ne sont pas nouveaux.

1.6.2.3 Autoplagiat

Lorsque vous utilisez des parties de vos travaux antérieurs (par exemple un article, une analyse documentaire ou un ensemble de données) sans les citer correctement, vous commettez ce que l'on appelle de l'autoplagiat. Bien que cela semble un peu fou d'être pénalisé pour avoir plagié votre propre travail, vous devez savoir que cela se fait parce que cela va à l'encontre des attentes des lecteurs de votre article. Ils s'attendent à ce que l'œuvre soit originale.

1.6.2.4 Acheter des documents

Lorsque vous utilisez le papier de quelqu'un d'autre, vous commettez un plagiat car vous prétendez que les mots et les idées sont les vôtres. Utiliser le travail de quelqu'un d'autre signifie, par exemple, demander à un ami ou à un membre de la famille d'écrire le texte pour vous ou acheter un document en ligne.

1.6.2.5 Traduction des textes

Lorsque vous utiliser un outil pour traduire un paragraphe que vous venez de copier-coller, cela ne signifie pas que vous devenez l'auteur de la traduction. Copier-coller le travail de quelqu'un d'autre en langue étrangère et en faire la traduction sans mentionner la source reste du plagiat.

1.6.2.6 Paraphrase

Paraphraser signifie traduire l'idée d'autrui avec vos propres mots. Si vous ne citez pas la source de l'idée ou du concept paraphrasés, vous vous les approprier. Il s'agit une fois de plus de plagiat [12].

1.6.3 Conséquences du plagiat

Les conséquences du plagiat dépendent du type de plagiat et du fait que vous soyez un étudiant de première année, un universitaire expérimenté ou un professionnel en activité. On peut citer quelques conséquences possibles du plagiat :

- Échec au cours et expulsion ou suspension.
- Violation du droit d'auteur.
- Une réputation ruinée et potentiellement la fin de votre carrière [12].

1.6.4 Comment éviter le plagiat

Pour éviter le plagiat, il suffit de suivre ces deux étapes : Faites une citation ou paraphrasez les mots ou les idées d'autrui et donnez la source originale dans le texte et la bibliographie [12]. Il est important de travailler de manière structurée afin de :

- Ne faites pas de copier-coller intentionnel (logique !).
- Assurez-vous que vous sauvegardez toutes les sources que vous utilisez dès le début (en note de bas de page par exemple).
- Citez et paraphrasez de manière correcte.
- Utilisez toujours le style de citation adéquat quand vous citez le travail de quelqu'un d'autre (APA, Harvard, Chicago…).

1.6.5 Citer ses sources pour éviter le plagiat

Pour citer vos sources, vous pouvez utiliser plusieurs styles de citation, tels que le style APA, le format MLA ou les citations du style Chicago. Les universités et les revues vous indiquent souvent le style de citation à utiliser. Vous devez citer les sources à la fois dans le texte courant avec une citation dans le texte, une note de bas de page ou une note de fin de texte et dans la liste de référence.

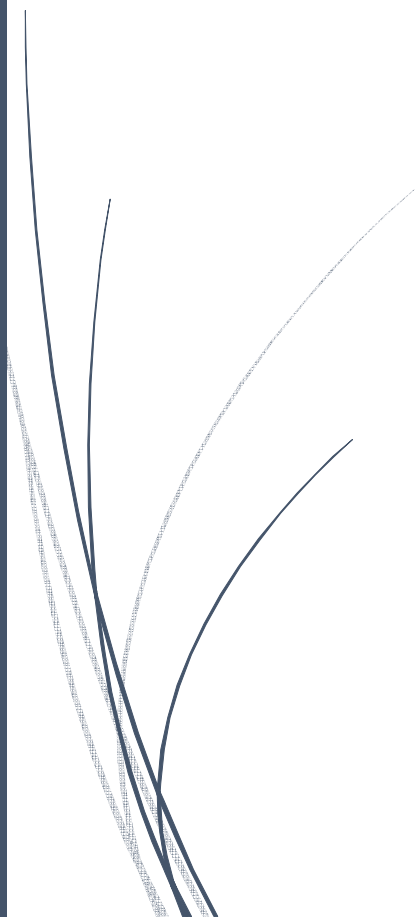
Souvent, la citation dans le texte ne mentionne que le nom de l'auteur ou des auteurs et l'année de publication. La liste de référence contient toutes les informations sur une source, y compris le titre de l'ouvrage et l'URL du site web [12].

1.7 Conclusion

Nous avons présenté quelques généralités sur l'attribution de la paternité (AA) dans ce chapitre, ainsi qu'un bref historique des traits d'auteur et des étapes d'attribution de la paternité. Ensuite, des définitions de la stylométrie sont proposées, ainsi qu'un bref historique de cette dernière. D'autre part, nous avons discuté de la catégorisation automatique des textes avant de passer à la définition et aux types de plagiat. Nous présenterons l'AT et la méthodologie de recherche, ainsi que les approches employées dans cette étude, dans le chapitre suivant.

CHAPITRE-2

FONCTIONNEMENT DU TRANSCRIPTEUR AUTOMATIQUE



CHAPITRE-2

FONCTIONNEMENT D'UN TRANSCRIPTEUR

2.1 Introduction

Dans ce chapitre, nous décrivons la procédure suivie par un transcripteur automatique pendant la conception de la base de donnée textuelle conçue pour la validation des tests de notre système d'identification d'auteurs. Le point central dans cette procédure est l'étape d'alignement entre un corpus audio avec les transcriptions textuelles associées.

En plus de l'application VEED en reconnaissance vocale, l'alignement entre les longs audios et leurs transcriptions correspondantes est utile dans un certain nombre d'applications, les auteurs proposent la synchronisation de livres audio avec des livres électroniques pour l'apprentissage des langues ou l'amélioration de la lecture. La création de corpus alignés est utilisée pour construire des systèmes TTS, l'accent est mis sur l'alignement et la correction automatiques de transcriptions inexactes à travers un processus itératif.

Les données collectées contiennent généralement enregistrements audio longs qui doivent d'abord être alignés avec les transcriptions et ensuite divisés en plus petits morceaux pour que l'entraînement VEED soit réalisable. Les travaux fondamentaux traitant directement de la question de l'alignement comprennent, une sortie VEED imparfaite est alignés au niveau des mots avec la transcription du texte pour trouver des mots d'ancrage reconnus avec une grande confiance.

Le décodage VEED avec des modèles de langage adaptés est ensuite effectué sur chacun de ces morceaux. Ce processus est itéré jusqu'à ce que toutes les données aient été correctement alignées. Cette approche nécessite qu'un système VEED bien entraîné sur la langue cible soit initialement disponible. Alternativement, les auteurs montrent qu'une sorte d'alignement est toujours nécessaire.

L'alignement forcé au niveau de la phrase basée sur les HMM est effectuée utilisant un sous-ensemble de modèles de phonèmes entraînés avec des données plusieurs langues. Enfin, l'alignement au niveau de la phrase en exploitant l'information sur les syllabes dans la parole et les transcriptions.et les transcriptions. D'autres moyens d'entraîner les systèmes de reconnaissance de la parole de reconnaissance de la parole avec peu de ressources incluent l'adaptation d'un système VEED préexistant sur la même langue, mais avec un dialecte différent , le boots trappions inter-langue ou l'entraînement de modèles acoustiques de semences à partir de très peu de données alignées manuellement à partir de très peu de données alignées manuellement .Dans cet article, nous utilisons des outils génériques et des données faciles à obtenir pour entraîner un système VEED viable sur une nouvelle langue. Tout d'abord, nous recueillons données pour la langue cible à partir d'Internet, avec la seule contrainte que certaines transcriptions manuelles (même inexactes) doivent être disponibles.

En particulier, nous avons expérimenté l'utilisation dédies livres audio pour et des discours parlementaires pour le catalan. Ensuite, nous décodons l'audio en une séquence de phonèmes en utilisant une machine de reconnaissance de phonèmes du commerce entraînée sur une autre langue.

Ensuite, nous alignons la transcription du texte sur la séquence de phonèmes en utilisant la programmation dynamique et une matrice de transformation qui fait correspondre les phonèmes graphèmes arabe. Enfin, les transcriptions audio et textuelles sont divisées en petits segments et un VEED est construit à l'aide de la boîte à outils de reconnaissance vocale en utilisant des modèles basés sur les graphèmes (pour éviter d'avoir à entraîner un système de graphème à phonème système). Les résultats obtenus en termes de précision d'alignement et de performance redémontrent la faisabilité de l'approche. Le sitelle même approche peut être utilisée pour n'importe quelle langue à condition que des données audio texte soient disponibles.

2.2 Historique

Se lancer dans l'établissement d'un corpus de parole spontanée sous-entend des possibilités d'enregistrement et de traitement post-enregistrement importantes. Comment en effet envisager d'analyser un objet dont on ne saurait avoir une

quelconque trace ? A cet égard, l'historique des corpus qui nous intéressent est étroitement lié aux avancées technologiques du vingtième siècle. Queneau considérait dans Bâtons, chiffres et lettres que « l'usage du magnétophone a provoqué en linguistique une révolution assez comparable à celle du microscope avec Swammerdam » et, bien que quelques travaux précurseurs sur le sujet n'aient pu bénéficier d'un tel support, il est incontestable que le fait de pouvoir « capturer » l'oral en a radicalement modifié la perception.

Pourtant, avant même que ne naisse cette invention, D'amourette et Pichon [D'amourette et Pichon, 1911-1940], en s'appuyant sur des conversations recueillies auprès d'un médecin, d'une institutrice... avaient dessiné les premiers contours morphosyntaxiques d'une « langue orale » dont la communauté linguistique avait encore pourtant du mal à admettre l'existence.

Puis il y a eu Bally [Bally, 1929] et surtout Frei [Frei, 1929] qui s'est attaché à analyser les lettres non parvenues aux soldats de la grande guerre. Ces lettres étaient rédigées par des familles souvent peu familières de l'écriture, et le style employé était en conséquence très oralisé. Toutefois, l'objectif de Frei était moins l'analyse de ce style oral que la hiérarchisation des « fautes » qui pouvaient s'y trouver. [14] [15]

Mais ce sont véritablement Guggenheim, Sauvageon, Michée et Rivent [Guggenheim étal., 1964] qui, s'appuyant sur 275 enregistrements sonores, ont révélé sans que cela constitue leur objectif premier la véritable nature de l'oral spontané. Ils souhaitaient avant tout proposer un équivalent du Basic English [Ogden, 1932], et ainsi favoriser l'apprentissage du français langue étrangère. Ils ont effectué notamment une étude quantitative du nombre d'occurrences des formes rencontrées. Il apparut alors que des mots comme « on », « hein » ou « ben » étaient parmi les plus utilisées de la langue française parlée, ce qu'aucune grammaire de l'époque ne prenait en compte. Aujourd'hui, certaines d'entre elles continuent de ne pas référencer ces termes. Il faut dire que les intégrer à partir des catégories en usage n'est pas chose facile. S'agit-il d'interjections, d'onomatopées, d'adverbes, de conjonctions, etc. ? Leur mention passe donc par l'invention de nouveaux concepts qui prennent du temps à s'installer.

Cela va des appuis du discours aux conjonctions, en passant par les connecteurs [Roulet et al.,1985]. À l'époque où D'amourette et Pichon entreprirent leurs recherches, les ordinateurs n'existaient évidemment pas, et les machines à écrire en étaient à leurs balbutiements. Les transcriptions étaient donc réalisées « à la volée », ce qui ne permettait pas de faire des études précises : il était par exemple impossible d'enrichir ou de modifier une transcription après-coup, puisque l'objet sonore n'avait pas été capturé. Par la suite, les ordinateurs ont changé l'adonne : le traitement de texte a tout d'abord permis de numériser les transcriptions, et d'en assurer une certaine pérennité. Un grand pas a ensuite été franchi lorsqu'il est devenu possible de numériser également le signal sonore, évitant ainsi l'inévitable dégradation dont souffraient les bandes magnétiques, microfilms, cassettes, etc. Parallèlement à ces avancées, l'arrivée de logiciels permettant d'aligner le signal audio avec le texte de la transcription a été une petite révolution : il est désormais devenu possible, en quelques secondes, d'écouter n'importe quelle partie d'un enregistrement, et de voir apparaître à l'écran la transcription qui en a été faite. Cette synchronisation offre entre autres la possibilité de réécouter très facilement un extrait pour voir si les propos transcrits y correspondent, et ainsi de corriger rapidement une erreur ou une interprétation. Les logiciels d'aide à la transcription qui proposent cette fonctionnalité sont aujourd'hui très répandus

2.3 Présentation des convertisseurs audio

2.3.1 Définition

Comme une définition on peut dit que Le terme de transcription n'est pas habituel. Que veut-il dire exactement ? Est-ce la même chose que retranscription ? Pas facile de s'y retrouver, d'autant que les professionnels du secteur n'ont pas tous le même usage de ce vocabulaire. Voilà qui mérite bien un petit effort de définition ! Occupez-vous du sens et les mots s'occuperont d'eux-mêmes.

Lewis Carroll, Alice au pays des merveilles, 1865Le mot « transcription » qualifie l'opération par laquelle on reproduit un contenu en l'écrivant. Transcrire, c'est donc avant tout réaliser une « copie écrite ». C'est pour cela que sa racine latine, transcrire, contient le mot « écrire » (scribe).

Le terme de transcription concerne d'abord la copie écrite de textes (ex : la copie de certains actes officiels). Il s'applique aussi à la reproduction écrite d'autres

types de contenus (ex : transcrire de la musique, des souvenirs ou des idées, un contenu audio enregistré…). Avec le temps, son application s'est élargie. Aujourd'hui, le terme est même utilisé en biologie pour qualifier le processus par lequel un segment d'ADN (acide désoxyribonucléique contenant notre information génétique) est copié en ARN (acide ribonucléique).

2.3.2 Différence entre Reproduction et transformation

La transcription peut être effectuée avec le même système d'écriture que celui utilisé pour le contenu initial. Sinon, elle peut être réalisée avec un système d'écriture (code ou mode d'enregistrement) différent. Par exemple : utiliser des caractères latins pour écrire un mot chinois, utiliser un code pour transcrire un message secret… C'est aussi le cas lorsqu'on fait la transcription d'un contenu audio : on passe d'un enregistrement sonore à une langue écrite.

La transcription reproduit-elle à l'identique ou transforme-t-elle le contenu transcrit ? Il y a souvent des débats à ce propos… En fait, ce sont deux vérités qui coexistent. En tant que copie, une transcription vise à l'exacte reproduction d'un contenu. Pour autant, cela ne va pas sans risque d'erreurs ou de déformations ! Surtout, à partir du moment où l'on utilise pour la copie un système d'écriture différent de celui du contenu initial, il y a forcément une transformation de ce contenu, avec un risque de perte d'informations ou de modification de sens. Le changement de code d'écriture peut aussi gêner la compréhension du contenu. Par exemple, la langue orale transcrite à l'écrit peut se révéler très pénible à lire… Ce qui peut amener à l'adapter pour la rendre plus lisible (ex : suppression des répétitions, intégration d'une ponctuation…).

Tout cela fait la difficulté de la transcription audio. En effet, il faut trouver comment transcrire des informations sonores, verbales et non verbales, de façon à perdre le moins possible d'information. En même temps, il faut trouver l'équilibre entre la fidélité au contenu initial et la lisibilité du texte transcrit, selon l'usage qu'on en a. [13]

2.3.3 Déférence entre Transcrire et retranscrire

Et la retranscription, alors ? Alors, là, ça se complique ! Le dictionnaire fait pourtant plutôt simple : selon le Petit Robert, comme selon le Larousse, retranscrire,

c'est transcrire de nouveau, recopier. Autrement dit : il s'agit de la même opération, mais elle est répétée dans le cas de la retranscription. Cependant, le terme est rarement utilisé en ce sens dans le langage courant et donne lieu à de nombreuses interprétations.

Ainsi, les professionnels de la transcription ne sont pas tous d'accord sur le sens à donner au terme de retranscription ! En effet, pour certains c'est un synonyme de transcription et ils l'emploient volontiers à sa place pour parler de toute transcription de contenus audio (qu'on appelle aussi audiotypie).

Pour d'autres, il y a une différence entre transcription et retranscription : la première qualifierait l'opération de transcription audio lorsque celle-ci cherche à être la plus fidèle possible au contenu initial (donc sans reformulation ni adaptation, par exemple dans les cas de la transcription intégrale ou épurée) ; tandis que la seconde concernerait les travaux impliquant une correction ou une reformulation, partielle ou totale (ex : compte-rendu intégral, révisé ou synthétique). Dans cette dernière acception, transcrire c'est reproduire un contenu, tandis que retranscrire, c'est le retravailler.

2.3.4 Principe de fonctionnement d'un transcripteur

Dans cette section, nous décrivons comment nous alignons les fichiers audio longs sur les leurs transcriptions textuelles correspondantes dans un contexte de faibles ressources. Les données résultantes sont ensuite utilisées pour entraîner un système VEED sur la langue cible.

L'entrée consiste en un fichier audio et la transcription textuelle associée. Bien que la transcription doive suivre le plus fidèlement possible ce qui est dit dans le fichier audio, nous sommes en mesure de gérer un certain nombre d'erreurs.

Nous sommes en mesure de gérer un certain nombre d'inadéquations, comme du texte supplémentaire ou manquant, ou des sons supplémentaires (par exemple, des applaudissements, des hésitations, du bruit, etc.)

Dans l'approche proposée, nous générons une transcription phonétique à partir de l'audio.

À partir de l'audio, qui est ensuite alignée avec le texte en utilisant la programmation dynamique. C'est l'approche inverse de ce que

La méthode proposée consiste à convertir le texte en forme acoustique et à l'aligner sur l'audio. En forme acoustique, puis nous l'alignons avec l'audio. Ensuite, nous

Ensuite, nous décrivons en détail les quatre étapes principales de ce processus, à savoir : le prétraitement de l'audio et du texte, l'alignement de l'audio sur le texte.et la préparation et l'entraînement des données [13]

2.4 Normalisation du texte

Avant l'étape d'alignement, la transcription du texte d'entrée est normalisée. La sortie de l'étape de normalisation du texte est un ensemble de symboles individuels de type graphème, plus un symbole de silence.

Symboles individuels de type graphème, plus un symbole de silence. Le site choix d'utiliser des graphèmes au lieu d'une transcription de phonèmes inspiré par les résultats de travaux antérieurs dans le domaine de la reconnaissance vocale (voir par exemple), où la reconnaissance vocale en arabe basée sur les graphèmes avérés dégradée, basée sur les graphèmes, n'a pas entraîné une dégradation importante des performances par rapport à un système basé sur les phonèmes. De plus, dans cet article En outre, dans cet article, nous avons voulu épargner les ressources nécessaires à la construction d'un système de conversion pour obtenir des phonèmes.

Un système de conversion pour obtenir des phonèmes. Au lieu de cela, quelques règles simples inspirées de l'acoustique ont été appliquées à certains graphèmes et paires de graphèmes et paires de graphèmes pour obtenir des phonèmes uniques.

Nous avons éliminé tous les diacritiques et le graphème 'h' lorsqu'il n'apparaît pas après 'c', et nous avons converti tous les chiffres dans leur forme textuelle.

Enfin, nous avons remplacé tous les signes de ponctuation par le phonème 'sil' pour indiquer les endroits où il peut y avoir une pause dans l'audio. Bien que spécifiques à la langue, ces conversions peuvent être définies avec une simple connaissance de base de la langue cible. [14] [15]

2.5 Difficultés liminaires

Transcrire de la parole spontanée présente de nombreuses difficultés. La première que nous mentionnerons est également la première à laquelle se trouve

nécessairement confronté tout transcripteur, à savoir la perception auditive. En effet, si un journal d'informations ne nécessite bien souvent qu'une écoute par segment de parole pour pouvoir être transcrit sans erreur, il n'en va pas de même pour un débat où plusieurs invités se coupent la parole, hésitent, bégayent, bafouillent... Il s'agit alors de discerner ce qui est effectivement prononcé, par qui et à quel moment. Les écoutes multiples deviennent donc indispensables pour peu que l'on souhaite une transcription rigoureuse. Toutefois, cela va de pair avec une perte de temps dès lors qu'on se trouve face à des masses de données à transcrire

Outre cette difficulté majeure, la transcription fait rapidement apparaître un autre souci saillant : l'orthographe. En premier lieu, celle de la langue française n'est pas des plus « intuitives », et nombreux sont les cas où même un annotateur aguerri peut être amené à hésiter : problèmes d'homonymies (ces/ses, été/était...), accords de participes passés de verbes pronominaux, pluriels incertains (pas de problème/problèmes), etc.

Même si l'hésitation peut parfois ne durer que quelques secondes, il suffit de ramener ce laps de temps au nombre d'hésitations que le transcripteur peut avoir dans un corpus de cent heures...Qui plus est, et cette fois sans même avoir à parler de spécificités du français, un autre versant de cet aspect orthographique est tout aussi important, sinon plus: les noms propres. Bien que plus fréquents dans un cadre préparé, comme l'une de nos expériences l'montré, ils n'en demeurent pas moins un élément prégnant de certains débat sou interviews, pour peu que le sujet y soit propice.

Une joute verbale politique sera rapidement sujette à l'évocation de présidents, ministres, députés... et parfois dans des pays peu familiers à l'annotateur francophone, ce qui peut poser parfois de vrais problèmes, ne serait-ce que pour identifier clairement la personne dont il est question. S'ajoute en cela les noms de journalistes ou de techniciens mentionnés à la fin d'une émission, et qui demeurent bien souvent assez confidentiels, donc délicats à identifier.

2.6 Les conventions de 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. Le premier aspect à aborder est celui de la

représentation textuelle. Comment représenter par écrit des conversations orales ? La majorité des corpus francophones créés 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. Les exemples les plus courants concernent l'élision de certaines lettres comme les « e » muets, dont nous aurons l'occasion de reparler longuement dans notre cinquième et dernier chapitre. Pour le dire simplement ici, un mot tel que petit sera souvent prononcé [pit] à l'oral : se pose alors la question de la représentation à l'écran. Retranscrire le mot sous sa forme « petit » est la démarche la plus naturelle, et évite les processus artificiels comme l'emploi de l'apostrophe (« p'tit »). Cela étant, l'information concernant l'élision du « e » est perdue, et rien ne permet alors de distinguer à l'écrit les prononciations [pâti] et [pit]. Une des solutions permettant de pallier ce problème est l'emploi de l'alphabet phonétique international (ou de la norme équivalente SAMPA).

Néanmoins, transcrire une masse de données importante en API est une tâche démesurément longue, et qui pose entre autres de nombreux problèmes acoustiques [Durand et Tartrier,2006]. Un compromis intéressant peut-être le suivant : adopter une transcription orthographique « classique », complétée par l'utilisation de l'API pour les phénomènes que l'on souhaite mettre en valeur. C'est, nous l'avons vu dans notre premier chapitre, cette méthode qui a été adoptée lors de la réalisation du corpus EPAC. Bien qu'assez coûteuse en temps si la granularité de l'annotation est fine (réalisation ou non des schwas, liaisons, assimilations, etc.), elle offre des possibilités d'analyse a posteriori très riches. D'autres études proposent même de réaliser deux version d'un même corpus : l'une richement annotée, et l'autre beaucoup plus épurée [Benzitoun et Véronis, 2005].[13]

2.7 Répétitions faux départs troncations

On a d'autres spécificités de la parole spontanée posent régulièrement problème aux systèmes de reconnaissance automatique. Les répétitions, faux départs, troncations ou autres diffluences sont autant d'anomalies langagières qu'ils n'ont pas l'habitude de rencontrer. Pour les premières citées, il est intéressant de constater que

LIUM RT s'est même, en de rares occasions, refusé à proposer deux occurrences consécutives du même mot, bien que la prononciation ne laissait planer aucune ambiguïté : « faut faut faut faut » : faut fois font fois

Les troncations ou les faux départs génèrent inévitablement de nouvelles alternatives homonymiques. Et à nouveau, le système de reconnaissance automatique se retrouve à traiter des suites de sons qu'il va chercher à associer à des mots qui lui sont connus, et jamais à des amorces ou fins de mots, particularités qu'on ne retrouve (presque) que dans la parole spontanée. Ce qui ne manque pas, à nouveau, de créer de nouvelles confusions. [13]

2.8 Fonctionnement du site « VEED »

Le site est disponible gratuitement en ligne et vous permet de convertir n'importe quel clip audio dans n'importe quelle langue. D'abord vous coupez l'audio à convertir en texte a pas plus de 10 minutes parce que le site n'accepte pas les clips de plus de 10 minutes, puis vous ouvrez la page du site et choisissez de saisir le fichier audio à convertir dans la base de données, puis attendez qu'il soit téléchargé après le téléchargement spécifie la langue et l'accent du haut-parleur et attendez un peu que le site produise le texte pour vous

Le site forme le texte sous forme de phrases intermittentes dans des cadres sporadiques pour transmettre le texte complet et harmonieux. Vous devez faire un processus de transcription de collage et le suivre progressivement image par image, puis le coller dans une page de mot vierge en suivant ces étapes et en utilisant un site qui a converti un fichier audio en texte L'interface du site Web est comme indiqué dans la figure [13].

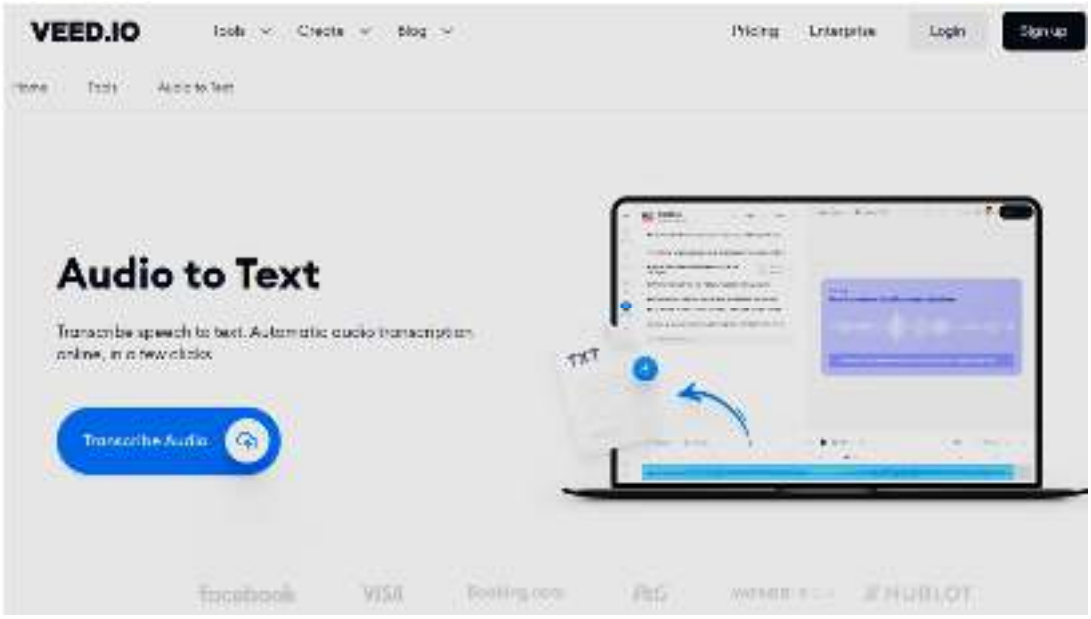


Figure-2.1 : L'interface du VEED

2.8.1 Préparation et entraînement des données VEED

La dernière étape du processus correspond à la préparation des données audio originales et des transcriptions de texte à utiliser comme entrée pour l'entraînement d'un système VEED. Pour ce faire, nous devons segmenter l'audio et le texte d'origine en phrases courtes, chacune d'entre elles avec leur texte correspondant. Nous avons expérimenté deux méthodes. La première méthode divise les phrases à tous les signes de ponctuation, même si ceux-ci ne correspondent pas toujours à la réalité.

La deuxième méthode utilise à nouveau le système de d'irisation LIUM, mais cette fois-ci avec une durée maximale de 10 secondes pour déterminer les points de séparation possibles. Ensuite, la limite de mot la plus proche pour un temps de séparation donné est choisie comme point de séparation. Choisi comme point de séparation. La boîte à outils de reconnaissance vocale est utilisée pour entraîner veed, comme décrit en détail dans la section expérimentale [14] [15].

Nous faisons le processus comme indiqué dans la figure suivante :

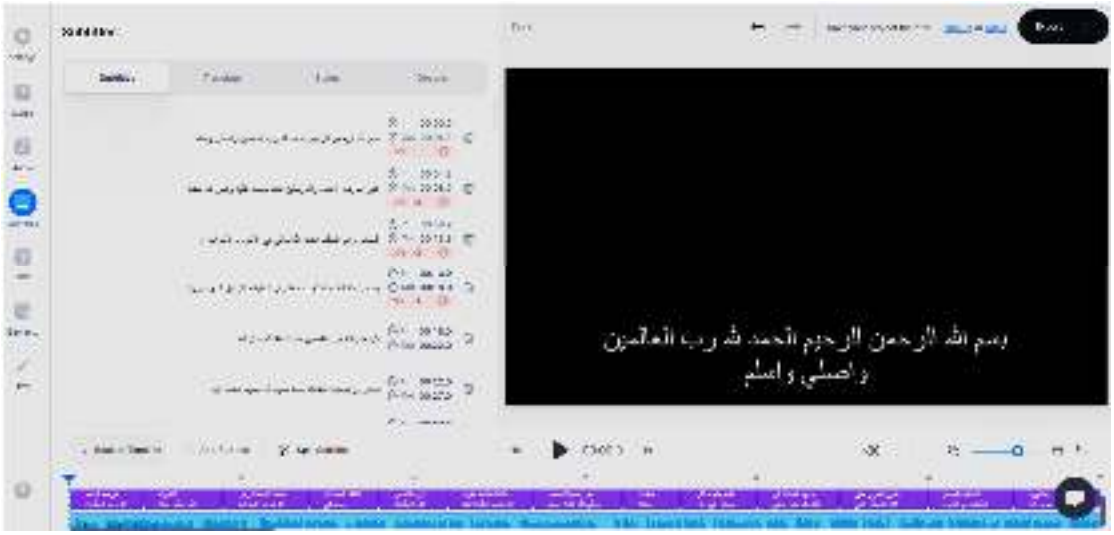


Figure-2.2 : Une illustration du travail du site VEED

2.8.2 Résultats de l'évaluation de la reconnaissance vocale

Dans cette section, nous présentons les résultats de la reconnaissance de la parole par l'entraînement de plusieurs systèmes VEED. Nous avons utilisé la boîte à outils de reconnaissance de la parole. En particulier, nous avons adapté la recette d'entraînement de Switch bocard et entraîné deux systèmes différents par test recette. D'une part, un modèle HMM utilisant des unités tri phoniques avec un arbre de décision de 3200 feuilles et un total de 30000 mélanges gaussiens.

D'autre part, le modèle précédent est réutilisé, mais au lieu Delles mélanges gaussiens, un DNN de taille moyenne a été entraîné avec4 couches cachées et 1024 neurones par couche. Sauf mention contraire Sauf mention contraire, la quantité totale de données disponibles dans chaque cas a été répartie entre la formation (90 %) et l'apprentissage (10 %) a été divisée en données de formation (90%) et de test (10%). Les mots des données de test ont été inclus dans le lexique pour s'assurer qu'il n'y avait pas de mots OOV dans le test mots OOV dans le test.

Les données de formation utilisées par le moteur de reconnaissance proviennent de trois ensembles de données différents. Le premier jeu de données, décrit et appelé Cedex, est composé d'environ 500 phrases uniques phonétiquement équilibrées, prononcées par 167 locuteurs arabes. Il représente environ 4 heures de parole propre. Dans ce cas, un jeu de test avec 1000 phrases provenant de la base de données Satis (également décrite) de la base de données Satis (également décrite) a été utilisé pour éviter la répétition des mêmes tests.

Cette configuration a été utilisée comme base de référence contrôlée. Une deuxième base de données est composée des 20 premiers chapitres de la version parlée du livre El Quioute de Miguel de Crevantes, téléchargée avec le texte associé. L'ensemble de données représente 5,3 heures au total. À troisième ensemble de données contient des discours parlementaires téléchargés depuis le site web du Parlement arabe. Ils correspondent à deux sessions (25/09/2013 et 26/09/2013) où l'état de la région autonome a été discuté.

L'état de la région autonome a été discuté. En particulier, 9 interventions du président Artur Mas ont été collectées, dont 8 ont une durée comprise entre 10 et 45 minutes. D'entre elles avec des durées comprises entre 10 et 45 minutes, et une de1h45m. L'ensemble de données représente 5,4 heures au total. Tous les discours, saule plus long, n'étaient pas scriptés. Le texte associé est une transcription professionnelle proposée sur le site web du Parlement. Le tableau 3 résume les taux d'erreur sur les mots (WER) obtenus sur les ensembles de test pour ces ensembles de données.

Tous les tests, à l'exception du premier test Cèderont été entraînés à l'aide de modèles de graphèmes. Les premier et deuxième Les premier et deuxième tests utilisant la base de données Cedex sont considérés comme la base de référence pour le système, car nous n'avons pas utilisé le modèle proposé. Système, car nous n'avons pas utilisé l'algorithme d'alignement proposé sur les données proposées sur les données (nous avons entraîné le système avec la base de données originale). Comme prévu, les erreurs pour le système basé sur les graphèmes sont juste modérément plus élevées qu'en utilisant les phonèmes [14] [15].

2.9 Techniques Proposées

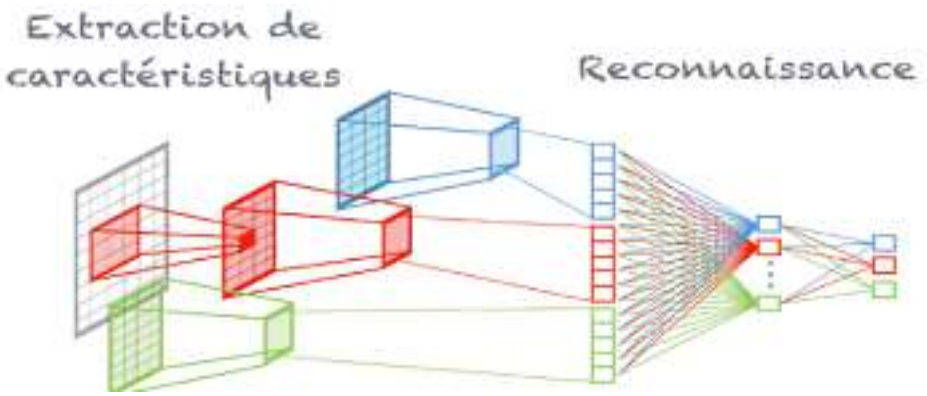


Figure-2.3 : Techniques Proposées

2.9.1 Absoute Centroid driver

Le clustering k-means est un algorithme d'apprentissage automatique non supervisé qui cherche à segmenter un ensemble de données en groupes en fonction de la similarité des points de données. Un modèle non supervisé a des variables indépendantes et aucune variable dépendante.

Supposons que vous disposiez d'un ensemble de données d'attributs scalaires bidimensionnels : $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

Si les points de ce jeu de données appartiennent à des groupes distincts avec des attributs variant de manière significative entre les groupes mais pas au sein, les points doivent former des clusters lorsqu'ils sont tracés [16].

2.9.2 Juola-Wyner Cross Entropy

La recherche présentée ici s'appuie sur nos connaissances actuelles de la structure autorail d'un corpus néerlandais précédemment collecté, et compare les résultats de l'entropie croisée en tant que technique d'inférence de la paternité aux résultats présentés dans l'article cité [17].

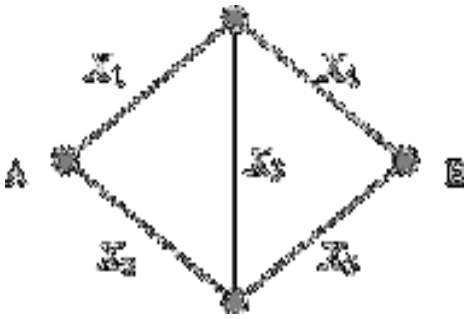


Figure-2.5 : Structure Juola-Wyner Cross Entropy

2.9.3 Centroid driver

Le pilote de Centroïde est une méthode de quantification vectorielle, issue du traitement du signal, qui vise à partitionner n observations en k clusters dans lesquels chaque observation appartient au cluster avec la moyenne la plus proche (centres de cluster ou centroïde de cluster), servant de prototype du cluster. Il en résulte un partitionnement de l'espace de données en cellules de Voronia. Le clustering du pilote centroïde minimise les variances intra-cluster (distances euclidiennes au carré), mais pas les distances euclidiennes régulières, ce qui serait le problème de Weber le plus

difficile : la moyenne optimise les erreurs au carré, alors que seule la médiane géométrique minimise les distances euclidiennes. Par exemple, de meilleures solutions euclidiennes peuvent être trouvées en utilisant les k-médianes et les k-médoïdes [2].

2.10 Extraction des caractéristiques

La notion de N-grammes de caractères a été utilisée de manière fréquente dans l'identification de la langue ou dans l'analyse de corpus oraux. L'utilisation de profils de fréquence N-gramme, qui est une tranche de N caractères d'une chaîne de caractères, est un moyen simple et fiable de classification des documents dans un large éventail de tâches de catégorisation utilisent les N-grammes pour la classification de langues complexes. Les expérimentations fondées sur diverses valeurs de N (de 2 à 7-grammes), Les types de caractéristiques qui ont été proposées et utilisées dans ce travail sont N grammes (avec N= 2, 3, 4,5 ,6,7) comme illustré ci-dessous:

- Caractères bi-grammes (n=2),
- Caractères tri-grammes (n=3),

Pour utiliser ces caractéristiques, une liste de tous les mots est extraite du texte, puis les caractères n-grammes de chaque mot sont pris, ainsi un profil de caractères n-grammes est créé (contenant les caractères n-grammes et leurs fréquences).

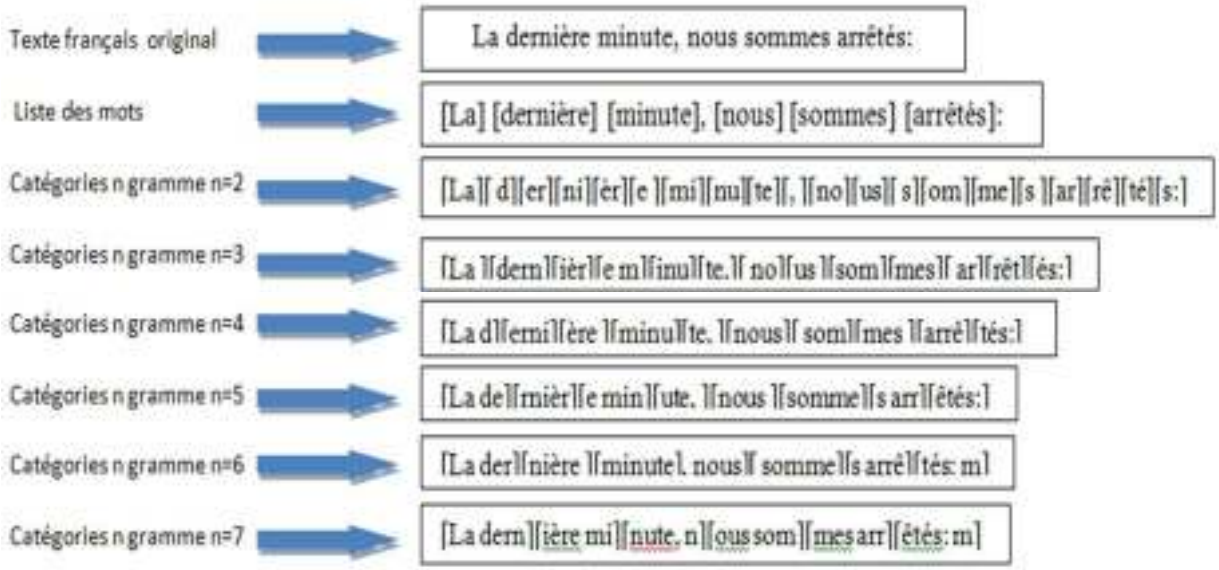
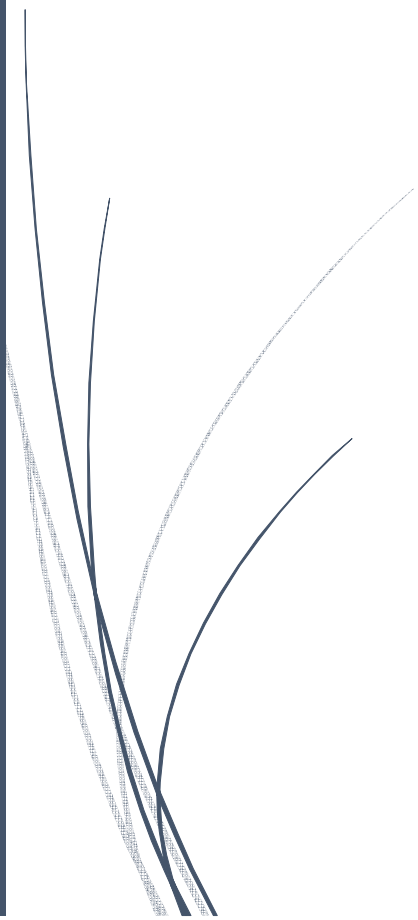


Figure-2.6 : Exemple d'extraction des caractères N-grammes d'un texte.

2.11 Conclusion

Construire un système de reconnaissance vocale pour une nouvelle langue est toujours un processus ardu. Alors que certains travaux antérieurs font l'utilisation de petites quantités de données d'entraînement bien alignées ou l'utilisation de systèmes de RVA existants dans des langues très similaires. Dans ce travail, nous avons utilisé des outils standard disponibles et un peu de connaissance de la langue cible pour construire avec succès un système viable pour l'arabe à partir de données très similaires.

CHAPITRE-3
EXPERIENCES ET RESULTATS



CHAPITRE-3
EXPERIENCES ET RESULTATS

3.1 Introduction

Dans ce chapitre nous allons exposer les séries d'expériences d'attribution d'auteur effectuées sur notre corpus qui est composé de 10 auteurs (5 masculins et 5 féminins) dont chacun a écrit 6 textes d'une longueur moyenne de 800 mots. Ces textes, numéroté de 1 à 6 et qui ont été obtenus après conversion de conférences audio en texte (CAT), Nous avons corrigé les erreurs survenues après l'opération (CAT).

Ces textes ont fait l'objet d'une série d'expériences pour voir l'effet de transcription sur le Taux d'Attribution d'Auteurs (TAA). Par la suite, les résultats obtenus ont été examinés et discutés et des interprétations et des conclusions objectives ont été donnée.

3.2 Corpus d'évaluation

3.2.1 Description du Corpus

L'évaluation expérimentale occupe une place importante dans la classification des textes A l'aide des corpus de tests, nous pouvons voir l'effet Traducteur Automatique (TA) sur l'attribution d'auteurs Cependant, les études en attribution d'auteur des textes obtenus après une opération CAT disposent d'un nombre relativement restreint de corpus.

3.2.2 Constituants du Corpus

Le corpus que nous avons conçu contient 10 écrivains arabes contemporains (5 Féminins et 5 Masculins) qui sont : Muhammad Al Arifi, Nabil Al Awadi, Bader Al Mashari, Jasssem Al Mutawa, Mohammed Rateb Al Nabulsi, Buthaina Ghalbazuri, Hala Samir, Hanan Al Qttan, Lina Al himsi, Nawara Hachem.

Nous préparons des conférences audio, les coupons en 10 minutes, puis les mettons sur le site VEED. Ensuite, nous convertissons l'audio en texte, une fois terminé, nous mettons les textes obtenus dans Word. Pour chaque auteur on sélectionne 6 textes d'une longueur moyenne de 800 mots et on les classe en deux catégories :

- Textes Corrigés : Textes qui ont subi des opérations de correction des erreurs de transcription après leur extraction du site VEED.
- Textes Non-Corrigés : Textes qui n'ont pas subi des opérations de correction (textes brutes tels qu'ils sont extraits du site VEED.

Les textes utilisés pour l'opération d'apprentissage sont les textes numérotés ; 1, 2, 3 et 4 pour chaque auteur en utilisant les deux catégories, tandis que les textes utilisés pour l'opération de test sont les textes numérotés ; 5 et 6 pour chaque auteur. Les textes étudiés sont obtenus à partir des conférences de ces écrivains. Les détails des informations sur les auteurs et les textes de notre collection sont donnés dans les tableaux suivants.

Tableau-3.1 : Récapitulatif du Corpus

Auteur	Pays Natal	Période	Textes		Nombre de mots	Utilisation
Muhammad Alarifi	Arabie saoudite	1970- à ce jour	Corrigé	Alarifi_Corrigé1	801	Apprentissage
				Alarifi_Corrigé2	800	Apprentissage
				Alarifi_Corrigé3	820	Apprentissage
				Alarifi_Corrigé4	810	Apprentissage
				Alarifi_Corrigé5	806	Test
				Alarifi_Corrigé6	813	Test
			Non-Corrigé	Alarifi_Non_Corrigé1	805	Apprentissage
				Alarifi_Non_Corrigé2	807	Apprentissage
				Alarifi_Non_Corrigé3	812	Apprentissage
				Alarifi_Non_Corrigé4	809	Apprentissage
				Alarifi_Non_Corrigé5	808	Test
				Alarifi_Non_Corrigé6	817	Test
Nabil Al Awadi	Koweit	1970- à ce jour	Corrigé	Awadi_Corrigé1	833	Apprentissage
				Awadi_Corrigé2	800	Apprentissage
				Awadi_Corrigé3	813	Apprentissage
				Awadi_Corrigé4	809	Apprentissage
				Awadi_Corrigé5	811	Test
				Awadi_Corrigé6	813	Test
			Non-Corrigé	Awadi_Non Corrigé1	814	Apprentissage
				Awadi_Non Corrigé2	801	Apprentissage
				Awadi_Non Corrigé3	812	Apprentissage
				Awadi_Non Corrigé4	807	Apprentissage
				Awadi_Non Corrigé5	809	Test
				Awadi_Non Corrigé6	811	Test
Bader Al Mashari	Arabie saoudit	1973- à ce jour	Corrigé	Mashari_Corrigé1	807	Apprentissage
				Mashari_Corrigé2	820	Apprentissage
				Mashari_Corrigé3	816	Apprentissage
				Mashari_Corrigé4	813	Apprentissage
				Mashari_Corrigé5	810	Test
				Mashari_Corrigé6	825	Test
			Non-Corrigé	Mashari_Non_Corrigé1	807	Apprentissage
				Mashari non corrigé2	807	Apprentissage
				Mashari non corrigé3	808	Apprentissage
				Mashari non corrigé4	810	Apprentissage
				Mashari non corrigé5	811	Test
				Mashari non corrigé6	812	Test
Jasssem Al Mutawa	Koweit	1965- à ce jour	Corrigé	Motaouia corrigé1	812	Apprentissage
				Motaouia corrigé2	811	Apprentissage
				Motaouia corrigé3	815	Apprentissage
				Motaouia corrigé4	811	Apprentissage
				Motaouia corrigé5	816	Test
				Motaouia corrigé6	821	Test
			Non-Corrigé	Motaouia non corrigé1	812	Apprentissage
				Motaouia non corrigé2	804	Apprentissage
				Motaouia non corrigé3	807	Apprentissage

				Motaouia non corrigé4	810	Apprentissage
				Motaouia non corrigé5	813	Test
				Motaouia non corrigé6	814	Test
Mohammed Rateb Al Nabulsi	Syrie	1938- à ce jour	Corrigé	Nabolsi corrigé1	816	Apprentissage
				Nabolsi corrigé2	805	Apprentissage
				Nabolsi corrigé3	814	Apprentissage
				Nabolsi corrigé4	823	Apprentissage
				Nabolsi corrigé5	813	Test
				Nabolsi corrigé6	813	Test
			Non-Corrigé	Nabolsi non corrigé1	809	Apprentissage
				Nabolsi non corrigé2	805	Apprentissage
				Nabolsi non corrigé3	809	Apprentissage
				Nabolsi non corrigé4	811	Apprentissage
				Nabolsi non corrigé5	808	Test
				Nabolsi non corrigé6	807	Test
Buthaina Ghalbazuri	Maroc	1970 à ce joure	Corrigé	Bothaina corrigé1		Apprentissage
				Bothaina corrigé2		Apprentissage
				Bothaina corrigé3	851	Apprentissage
				Bothaina corrigé4	1513	Apprentissage
				Bothaina corrigé5	1483	Test
				Bothaina corrigé6	1412	Test
			Non-Corrigé	Bothaina non corrigé1	802	Apprentissage
				Bothaina non corrigé2	801	Apprentissage
				Bothaina non corrigé3	814	Apprentissage
				Bothaina non corrigé4	802	Apprentissage
				Bothaina non corrigé5	802	Test
				Bothaina non corrigé6	803	Test
Hala Samir	Egypte	inconnu	corrigé	hala corrigé1	800	Apprentissage
				hala corrigé2	800	Apprentissage
				hala corrigé3	804	Apprentissage
				hala corrigé4	800	Apprentissage
				hala corrigé5	801	Test
				hala corrigé6	800	Test

			Non corrigé	hala non corigé1	800	Apprentissage
				hala non corigé2	800	Apprentissage
				hala non corigé3	800	Apprentissage
				hala non corigé4	800	Apprentissage
				hala non corigé5	800	Test
				hala non corigé6	800	Test
				Hanan corrigé1	800	Apprentissage
				Hanan corrigé2	802	Apprentissage
Hanan Al Qttan	Koweit	inconnu	Corrigé	Hanan corrigé3	800	Apprentissage
				Hanan corrigé4	800	Apprentissage
				Hanan corrigé5	809	Test
				Hanan corrigé6	807	Test
			Non corrigé	Hanan non corrigé1	802	Apprentissage
				Hanan non corrigé2	800	Apprentissage
				Hanan non corrigé3	800	Apprentissage
				Hanan non corrigé4	801	Apprentissage
				Hanan non corrigé5	803	Test
				Hanan non corrigé6	802	Test
Lina Al himsi	Syrie	1965- à ce jour	Corrigé	Lina corrigé1	800	Apprentissage
				Lina corrigé2	800	Apprentissage
				Lina corrigé3	800	Apprentissage
				Lina corrigé4	802	Apprentissage
				Lina corrigé5	801	Test
				Lina corrigé6	801	Test
			Non corrigé	Lina non corrigé1	800	Apprentissage
				Lina non corrigé2	800	Apprentissage
				Lina non corrigé3	800	Apprentissage
				Lina non corrigé4	805	Apprentissage
				Lina non corrigé5	800	Test
				Lina non corrigé6	800	Test
			Corr	Nawara corrigé1	803	Apprentissage
				Nawara corrigé2	804	Apprentissage

				Nawara corrigé3	**803**	**Apprentissage**
				Nawara corrigé4	**801**	**Apprentissage**
				Nawara corrigé5	**816**	**Test**
				Nawara corrigé6	**820**	**Test**
Nawara Hachem	**Syrie**		**Non corrigé**	**Nawara non corrigé1**	**806**	**Apprentissage**
				Nawara non corrigé2	**801**	**Apprentissage**
				Nawara non corrigé3	**801**	**Apprentissage**
				Nawara non corrigé4	**801**	**Apprentissage**
				Nawara non corrigé5	**800**	**Test**
				Nawara non corrigé6	**801**	**Test**

3.3 Préparation des documents du corpus

Les documents du corpus doivent être préparés avant leur utilisation pour l'attribution de leurs véritables auteurs. La phase de préparation se résume en opérations pour préparer ce texte :

- Couper les fichiers audio à 10 minutes ou moins,
- Insertion de fichiers audio dans le site VEED,
- Les fichiers corrigés et non corrigés sont enregistrés dans Word,
- Les documents textes obtenus sont enregistrés sous forme UTF-8 (Encodage basé sur l'Unicode qui peut être codé sur 4 octets),

Remarque : Il est à noter, qu'on a utilisé l'encodage UTF-8 pour encoder tous les textes du corpus, car ce dernier couvre un très grand nombre de caractères, et qui est implicitement capable d'encoder la majorité des langues vu qu'il est encodé sur 4 octets. En revanche, l'utilisation de cet encodage est payée en termes de temps de calcul et en termes de mémoire.

- Par la suite, le corpus est divisé en deux sous-ensembles (apprentissage et Test) selon la règle (2/3 pour l'apprentissage et 1/3 pour le test) appliquée dans les bases de données. L'ensemble d'apprentissage est constitué des textes numérotés ; 1, 2, 3 et 4 pour chaque auteur) et l'ensemble de test est constitué des textes numérotés ; 5 et 6 pour chaque auteur. Au totale, le corpus contient 120 textes : 80 textes corrigés pour l'apprentissage et 40 textes pour le test.

3.4 Exemples de textes obtenus après une opération de transcription

Après le processus de conversion des fichiers audio en texte, les résultats obtenus sont considérés comme des documents modifiables (format Word) pour corriger les erreurs, ajouter ou supprimer tout élément supplémentaire. Ci-dessous, nous passons en revue des exemples de textes que nous avons obtenus après le processus de conversion de fichiers audio en textes.

Figure-3.1: Etapes du processus de transcription

Figure-3.2: Exemple de textes convertis en Word non-corrigés

Figure-3.3: Exemple de textes convertis en Word corrigés

Ces résultats ont été obtenus après le processus de conversion des fichiers audio en texte et de leur correction manuelle.

3.4 Travaux d'expérimentation

3.4.1 Protocole expérimental

Dans ce mémoire, la tâche d'attribution d'auteurs est effectuée en utilisant six caractéristique : [Word, Word Grams (N=2 et N=3), Suffixes N= (3-5), Caractères, Leave kout W.G] et trois classifieurs : Absoulut centroid drive (ACD), Centroid Driver et JWCross Entropy.

On a utilisé ces techniques pour voir si la tâche d'attribution à l'auteur existe toujours en utilisant les textes obtenus par le processus de transcription des fichiers audio en textes avant et après correction. Le Taux d'Attribution d'Auteurs (TAA) est défini par la relation suivante :

$$TAA=\frac{\text{nombre } \textbf{documents} \text{ correctement attribué}}{\text{nombre } \textbf{document} \text{ testé}} \times 100$$

Ce travail expérimental est organisé en quatre séries d'expériences et chaque série comporte plusieurs cas d'applications selon six caractéristique. Dans la première série, nous avons utilisé des textes non corrigés en phase d'apprentissage et en phase de test. Dans la deuxième série, nous mettons les textes non corrigés des auteurs masculins et féminins. Dans la seconde série, les textes utilisés en phase d'apprentissage sont des textes corrigés et les textes utilisés en phase de test sont des textes corrigés.

3.4.2 Séries d'expériences et résultats obtenus

3.4.2.1 Série-I : Utilisation des textes corrigés pour l'apprentissage et pour le test

Le but de cette série d'expériences est de déterminer le type approprié de caractéristiques pour avoir le meilleur taux TAA en utilisant les textes corrigés. On a utilisé 4 textes pour l'apprentissage et 2 textes pour le test. Après investigation, nous avons choisis d'utiliser la méthode d'analyse des Evénements les Plus Courants (en Anglais Most Commun Events « MCE = 100 »). Les résultats obtenus de cette série d'expérience sont présentés dans les tableaux et les figures qui suivent :

A) Auteurs Masculins

Tableau-3.2: TAA pour les textes corrigés des auteurs Masculins

Caractéristique / Classifier	Word	Word Grams		Suffices N=(3-5)	Characters	Leave kout W.G
		N=2	N=3			
ACD	90	90	90	60	80	90
Centroid Driver	80	90	90	80	70	90
JWCross Entropy	80	90	70	90	80	70

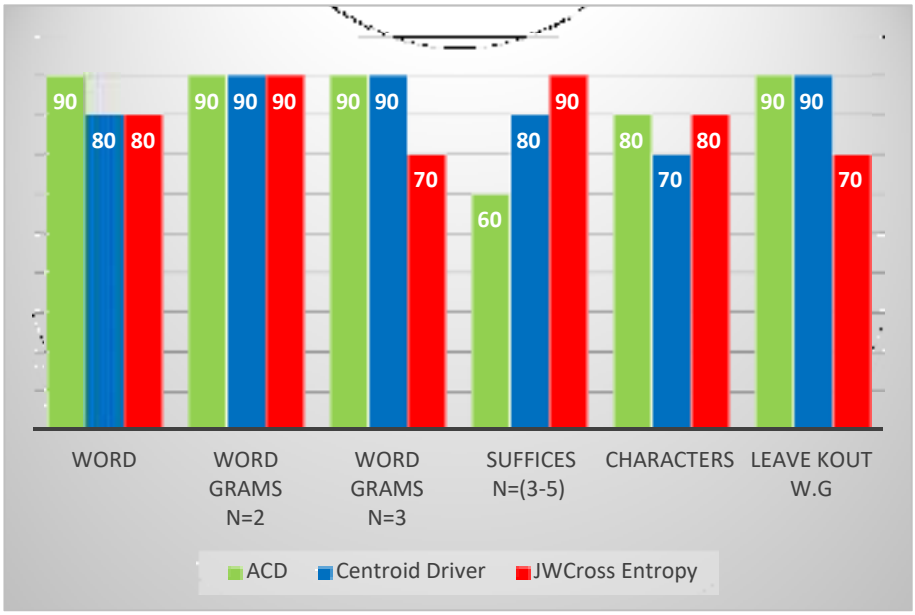


Figure-3.4: TAA pour les textes non corrigés des auteurs Masculins

D'après les résultats obtenus, on peut voir clairement que le taux TAA de cette expérience est entre 60-90% pour ACD 70-90 % pour Centroid Driver et 70-80% pour le JWCross Entropy. Ceci peut être expliqué par la présence des erreurs de conversion dans le texte résultant et aussi à cause de la similitude du style de l'auteur et la difficulté de la langue arabe.

B) Auteurs Féminins

Tableau-3.3: TAA pour les textes corrigés des auteurs Féminins

Caractéristique / Classifier	Word	Word Grams		Suffices N=(3-5)	Characters	Leave kout W.G
		N=2	N=3			
ACD	60	90	90	80	60	90
Centroid Driver	60	90	90	60	70	90
JWCross Entropy	100	90	80	90	80	80

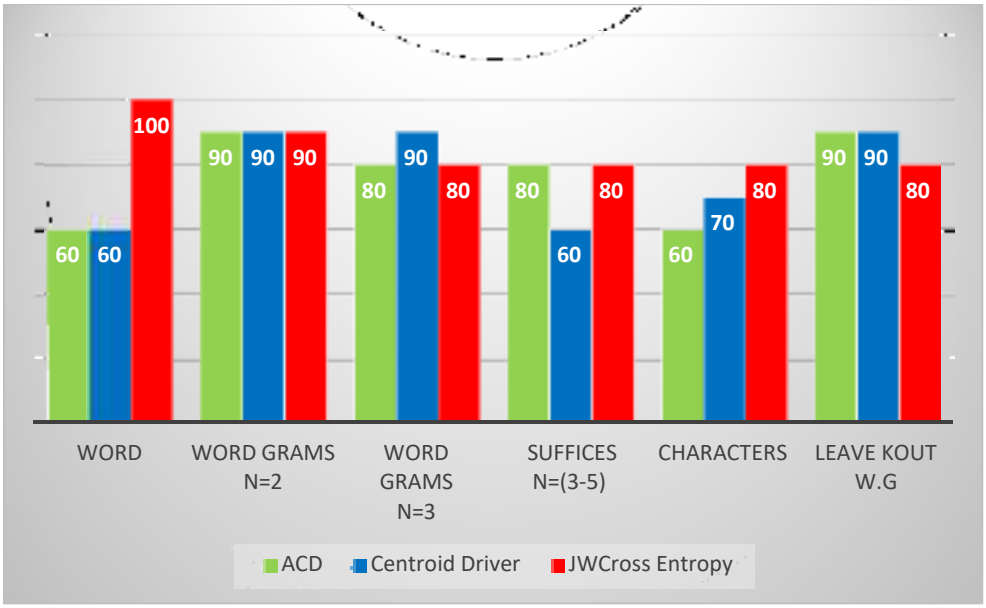


Figure-3.5: TAA pour les textes non corrigés des auteurs Féminins

D'après les résultats obtenus, on peut voir clairement que le taux TAA de cette expérience est entre 60-90% pour ACD 60-90 % pour Centroid Driver et 80-100% pour le JWCross Entropy Ceci peut être expliqué par la présence des erreurs de conversion dans le texte résultantEt aussi à cause de la similitude du style de l'auteur Et la difficulté de la langue arabe.

3.4.2.2 Série-II : Utilisation des textes non corrigés pour l'apprentissage et pour le test

Le but de cette série d'expériences est de déterminer le type approprié de caractéristiques pour avoir le meilleur taux TAA en utilisant les textes corrigés. On utilise 4 textes pour l'apprentissage et 2 textes pour le test. Après investigation, nous avons choisis d'utiliser la méthode d'analyse des Evénements les Plus Courants (en Anglais Most Commun Events « MCE = 100 »). Les résultats obtenus de cette série d'expérience sont présentés dans les tableaux et les figures qui suivent :

A) Auteurs Masculins

Tableau-3.4: TAA pour les textes corrigés des auteurs Masculins

Caractéristique / Classifier	Word	Word Grams		Suffices N=(3-5)	Characters	Leave kout W.G
		N=2	N=3			
ACD	70	80	90	60	60	90
Centroid Driver	70	80	90	60	60	90
JWCross Entropy	70	80	70	80	80	70

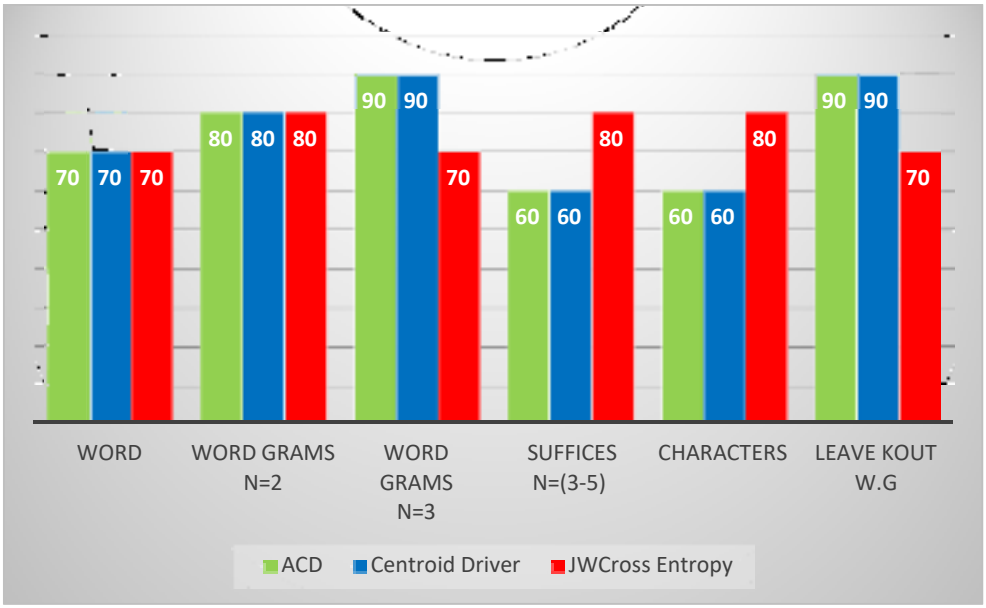


Figure-3.6: TAA pour les textes corrigés des auteurs Masculins

D'après les résultats obtenus, on peut voir clairement que le taux TAA de cette expérience est entre 60-90% pour ACD 60-90 % pour Centroid Driver et 70-80% pour le JWCross Entropy. On constate qu'après la correction, les résultats ont diminué, et cela est dû à la correction des mots-clés dans lesquels les auteurs partagent un caractère religieux, et c'est ce qui rend difficile leur différenciation.

B) Auteurs Féminins

Tableau-3.5: TAA pour les textes non corrigés des auteurs Féminins

Caractéristique / Classifier	Word	Word Grams		Suffices N=(3-5)	Characters	Leave kout W.G
		N=2	N=3			
ACD	60	90	80	60	60	90
Centroid Driver	60	90	80	60	70	90
JWCross Entropy	100	80	80	70	80	80

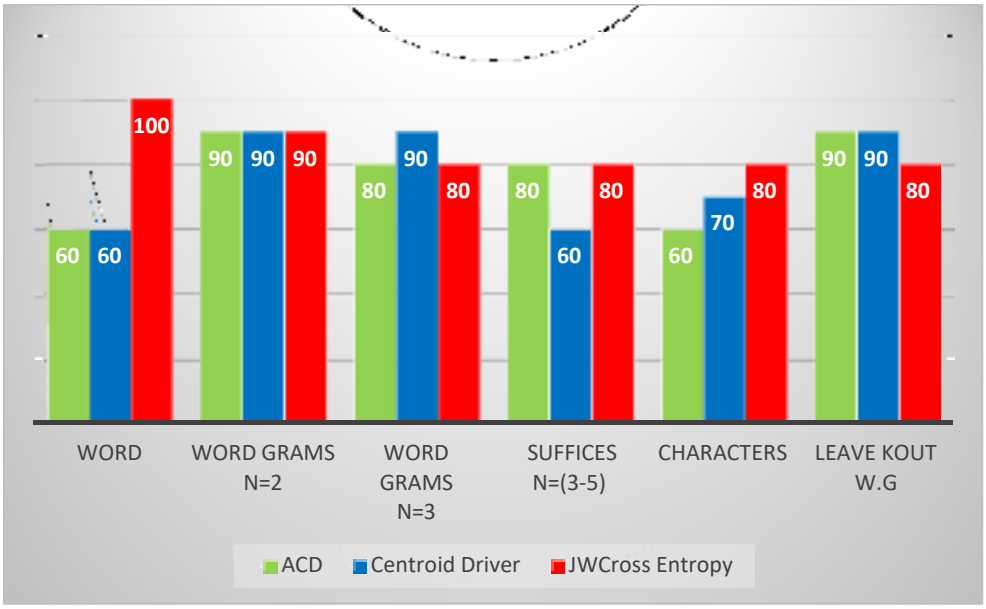


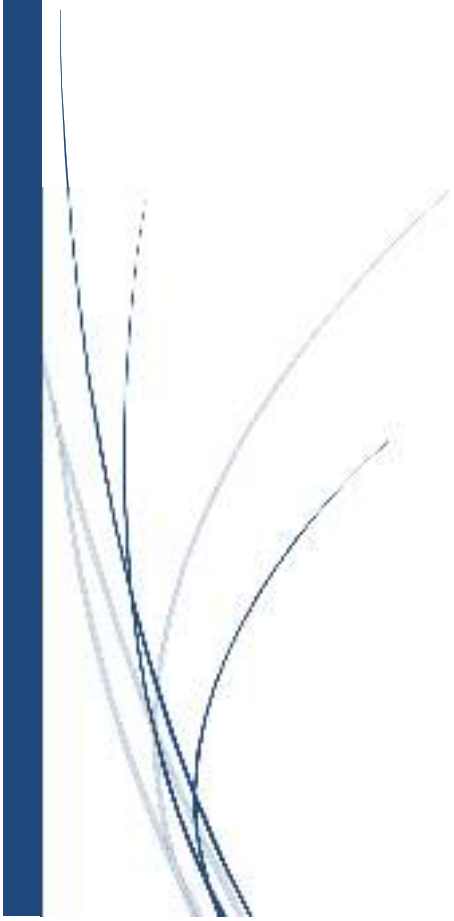
Figure-3.7: TAA pour les textes corrigés des auteurs Féminins

D'après les résultats obtenus, on peut voir clairement que le taux TAA de cette expérience est entre 60-90% pour ACD 60-90 % pour Centroid Driver et 70-100% pour le JWCross Entropy Ceci peut être expliqué par la présence des erreurs de conversion dans le texte résultantEt aussi à cause de la similitude du style de l'auteur Et la difficulté de la langue arabe.

3.5 Conclusion

Dans ce chapitre, nous effectuons des expériences d'attribution d'auteur pour des documents texte dactylographiés et des textes traduits automatiquement obtenus après transcodage de fichiers audio et texte via un site Web VEED. L'évaluation expérimentale a été réalisé en utilisant une base de données qu'on conçu pour cette fin. Les résultats obtenus ont montré que les erreurs dans la conversion des fichiers audio en textes affectent significativement le taux d'attribution du TAA. La méthode utilisée pour l'attribution d'auteurs est basée sur l'utilisation six caractéristique et elle : [Word, Word Grams (N=2 et N=3), Suffices N= (3-5), Characters, Leave kout W.G] et trois classifieurs : Absoulut cntroid driver (ACD), Centroid Driver et JWCross Entropy. Les expériences sont effectuées sur une base de données (Corpus) On peut conclure des expériences que nous avons menées précédemment que la meilleure façon de connaître les textes attribués à l'auteur est d'utiliser caractéristique et elle : [Word Grams (N=2 et N=3)], et classifieur Centroid Driver.

CONCLUSION GENERALE



Conclusion Générale

Le thème que nous avons étudié dans ce mémoire s'intéresse à l'effet de transcription automatique, après acquisition à l'aide d'un transcripteur en ligne, des documents textes sur la tâche d'attribution d'auteurs. Dans ce travail, nous avons abordé l'attribution d'auteurs des textes anonymes, en particulier des textes obtenus après transcription des fichiers audio.

Le corpus que nous avons conçu pour réaliser nos expériences, est construit autour d'une base de données constituée de 10 auteurs dont on a choisi leurs conférences audibles pour chaque auteur on fait l'extraction aléatoire d'un certain nombre de pages contenant (~800) mots, ensuite on sélectionne 6 textes d'une longueur moyenne de 800 et on les classe en deux catégories : textes corrigés et textes non-corrigés. Les textes utilisés pour l'opération d'apprentissage sont les textes numérotés ;1, 2, 3 et 4 pour chaque auteur en utilisant les deux catégories. Cependant, les textes utilisés pour l'opération de test sont numérotés ; 5 et 6 (pour chaque auteurs).

Ce mémoire avait pour ambition d'étudier le style des auteurs afin de trouver le véritable auteur, en appliquant des caractéristiques telles que caractère [Word, Word Grams (N=2 et N=3), Suffices N=(3-5), Characters, Leave kout W.G] et trois classifieurs : Absoulut centroid driver (ACD), Centroid Driver et JWCross Entropy. L'originalité de ce travail de recherche est que la tâche d'Attribution d'Auteurs (AA) a été appliquée aux textes transcrit qui ont été reporté fidèlement en les comparants avec les fichiers et des textes (inconnues) et non pas été écrits directement par les auteurs.

REFERENCES BIBLIOGRAOHIQUES

Références bibliographiques

[1] Pignot, L., & Saez, J. P. Le droit d'auteur sous toutes ses facettes. Lectures, Publications reçues.

[2] https://fr.wikipedia.org/wiki/%C3%89tat_de_l%27art

[3] Brixtel, R., Lecluze, C., & Lejeune, G. (2015, June). Attribution d'Auteur: approche multilingue fondée sur les répétitions maximales. In Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles. Articles longs (pp. 208-219).

[4] Lutoslawski, W. (1898). Principes de stylométrie appliqués à la chronologie des œuvres de Platon. Revue des études grecques, 11(41), 61-81.

[5] MENASRI, R., & YAKOUBI, M. (2020). Etude et analyse des effets d'acquisition optique à l'aide d'un OCR des textes arabes sur l'attribution d'auteurs (Doctoral dissertation, Univ M'sila).

[6] Réhel, S. (2005). Catégorisation automatique de textes et cooccurrence de mots provenant de documents non étiquetés.

[7] Rimouche, N., & Hadjira, H. (2016). Amélioration du produit scalaire via les mesures de similarités sémantiques dans le cadre de la catégorisation des textes. Université abou Beker Belkaid Tlemcen.

[8] OUALI, C. (2014). Classification automatique de textes (Doctoral dissertation, universite mohamed boudiaf m'sila: faculte des mathematiques et de l'informatique: Département d'Informatique).

[9] Vandendorpe, C. (1992). Le plagiat.

[10] https://www.scribbr.fr/category/le-plagiat/

[11] Bazillon, T. (2011). Transcription et traitement manuel de la parole spontanée pour sa reconnaissance automatique (Doctoral dissertation, Université du Maine).

[12] https://www.upf.edu/en/web/universitat/-/departament-de-tecnologies-de-la-informacio-i-les-comunicacions

[13] Anguera, X. (2012). Telefonica Research System for the Query-by-example task at Albayzin 2012. Proceedings of Iber SPEECH, 626-632.

[14] https://towardsdatascience.com/create-your-own-k-means-clustering-algorithm-in-python-d7d4c9077670

[15] Juola, P., & Baayen, R. H. (2005). A controlled-corpus experiment in authorship identification by cross-entropy. Literary and Linguistic Computing, 20(Suppl), 59-67.

ملخص

إن إسناد نص مجهول لمؤلف معين هو واحد من أكثر المشاكل المطروحة منذ القدم... في هذا العمل البحثي، حاولنا القيام بإسناد نصوص محولة بشكل آلي من محاضرات صوتية إلى نصوص مكتوبة. هذه النصوص قد تم الحصول عليها باستعمال موقع خاص يدعى (VEED) والذي يقوم بتحويل المحاضرات الصوتية إلى نصوص مكتوبة.

في هذه الدراسة قمنا بإنشاء قاعدة بيانية جديدة لاستعمالها في تجارب التعرف على كاتب النص، كما قمنا باقتراح خوارزميات إحصائية لحل مشكلة التصنيف الأوتوماتيكي للمؤلفين والتعرف على الكتاب الأصليين..

كلمات مفتاحية: التعرف الآلي على الكاتب، الناسخ الآلي الصوتي، إسناد المؤلف.

Résumé

L'attribution d'un texte inconnu à un auteur précis est l'un des problèmes les plus courants depuis l'Antiquité... Dans ce travail de recherche, nous avons essayé de faire l'attribution de textes automatiquement convertis de cours magistraux en textes écrits. Ces textes ont été obtenus à l'aide d'un site Web spécial appelé VEED qui convertit les conférences audio en textes écrits. Dans cette étude, nous avons créé une nouvelle base de données à utiliser dans des expériences de reconnaissance d'auteurs de texte. Nous avons également proposé des algorithmes statistiques pour résoudre le problème de la classification automatique des auteurs et de la reconnaissance des auteurs originaux.

Mots-clés : Reconnaissance Automatique d'Auteur, Transcripteur Automatique, Attribution d'Auteur.

Abstract

The attribution of an unknown text to its specific author is one of the most common problems since Antiquity... In this research work, we tried to make the attribution of texts automatically converted from lectures into written texts. These texts were obtained using a special website called VEED that converts audio lectures into written texts. In this study, we created a new database to be used in text author recognition experiments. We also proposed statistical algorithms to solve the problem of automatic author classification and recognition of original authors.

Keywords: Automatic Author Recognition, Automatic Transcription, Author Attribution.