Détection de dérivation de texte

Fabien Poulard

▶ **To cite this version:**

Fabien Poulard. Détection de dérivation de texte. Informatique [cs]. Université de Nantes, 2011. Français. NNT : . tel-00590708v2

HAL Id: tel-00590708
https://theses.hal.science/tel-00590708v2

Submitted on 26 May 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Université de Nantes

École Doctorale

« Sciences et Technologies de l'Information et de Mathématiques »

Année : 2011

Thèse

de

Doctorat

de l'Université de Nantes

Spécialité : Informatique

Présentée et soutenue publiquement par

Fabien B. Poulard

le 24 Mars 2011

à l'UFR des Sciences et Techniques de Nantes

Titre

Détection de dérivation de texte

Jury

Présidente :	Josiane Mothe Professeur	IRIT
Rapporteurs :	François Yvon, Professeur	LIMSI/CNRS
	Patrice Bellot, Maître de conférences	LIA
Examinateur :	Claude de Loupy, Dirigeant et co-fondateur	Syllabs

Directrice de thèse : **Béatrice Daille**, Professeur
Laboratoire : LINA – UMR CNRS 6241
Co-encadrant: Nicolas Hernandez, Maître de conférences
Laboratoire : LINA – UMR CNRS 6241
Composante de rattachement du directeur de thèse : UFR Sciences et Techniques – Université de Nantes

N° ED 503-123

Université de Nantes

École Doctorale

« Sciences et Technologies de l'Information et de Mathématiques »

Année : 2011

Thèse

de

Doctorat

de l'Université de Nantes

Spécialité : Informatique

Présentée et soutenue publiquement par

Fabien B. Poulard

le 24 Mars 2011

à l'UFR des Sciences et Techniques de Nantes

Titre

Détection de dérivation de texte

Jury

Présidente :	Josiane Mothe Professeur	IRIT
Rapporteurs :	François Yvon, Professeur	LIMSI/CNRS
	Patrice Bellot, Maître de conférences	LIA
Examinateurs :	Claude de Loupy, Dirigeant et co-fondateur	Syllabs

Directrice de thèse : **Béatrice Daille**, Professeur
Laboratoire : LINA – UMR CNRS 6241
Co-encadrant: Nicolas Hernandez, Maître de conférences
Laboratoire : LINA – UMR CNRS 6241
Composante de rattachement du directeur de thèse : UFR Sciences et Techniques – Université de Nantes

N° ED 503-123

Détection de dérivation de texte

Detecting textual derivatives

Fabien B. Poulard

UNIVERSITÉ DE NANTES

lina

Fabien B. POULARD

Détection de dérivation de texte

xvi+248+XVI p.

Ce document a été préparé avec LaTeX2e, logiciel libre, et la classe `mythesis`. Tous les graphiques ont été créés avec `pgfplots`. Les sources sont disponibles sur simple demande auprès de l'auteur.

Version du 17/05/2011 à 15:24.

Napoléon : Monsieur de Laplace, je ne trouve pas dans votre système mention de Dieu ?
Laplace : Sire, je n'ai pas eu besoin de cette hypothèse. [...] Cette hypothèse, Sire, explique en effet tout, mais ne permet de prédire rien. En tant que savant, je me dois de vous fournir des travaux permettant des prédictions

— Pierre-Simon Laplace à Napoléon I^{er}

Table des matières

Table des figures

Table des exemples

Liste des tableaux

Introduction

Si on commence avec des certitudes, on finit avec des doutes. Si on commence avec des doutes, on finit avec des certitudes.

— Francis Bacon

Toute nouvelle découverte se fonde sur les briques d'une connaissance auparavant acquise. Ptolémée I^{er} l'avait bien compris lorsqu'il fit construire la bibliothèque d'Alexandrie et y accumula des manuscrits de tout le monde antique. Cette bibliothèque — compilant environ 700 000 ouvrages à son apogée — attira les plus grands savants de l'époque : Euclide, Archimède... Les ouvrages s'y empilèrent indépendamment de la volonté de leurs auteurs puisque des scribes recopiaient les manuscrits présents à bord des bateaux entrant au port d'Alexandrie. Dans l'antiquité, la paternité d'une œuvre était secondaire, état de fait qui perdura jusqu'à la fin du Moyen-Âge. L'individualisme, puis l'humanisme, portés par la Renaissance ont changé la donne. Les auteurs ont commencé à chercher une certaine reconnaissance et à signer leurs œuvres. En parallèle, la généralisation de l'imprimerie a entraîné la mise en place de règles concernant la reproduction des œuvres. En France, c'est à Beaumarchais que l'on doit la première société de défense de la reconnaissance de droits au profit des auteurs. La I^{re} République après avoir dans un premier temps supprimé ces privilèges, instaurera les bases du droit d'auteur tel que nous le connaissons aujourd'hui.

Les textes représentent aujourd'hui encore la manière la plus aisée de diffuser de l'information. Les scribes et les presses ont disparu, cédant leurs places aux photocopieurs et à la numérisation tandis que le réseau Internet est en train de sonner le glas du contrôle de la production et de la diffusion de l'information par un nombre restreint d'individus. La copie et la diffusion des œuvres s'effectuent désormais en un click. L'avenir nous dira si cette diffusion incontrôlable est néfaste ou non à la production culturelle. Toutefois, si les outils modernes rendent l'acte de copie aussi aisé, il n'existe pas d'outils d'une puissance équivalente pour l'authentification de la paternité, ni pour la détection de celle-ci dans d'autres œuvres. À l'image du glaive et du bouclier, il est nécessaire de contre-balancer le pouvoir de la numérisation et des réseaux par des outils permettant d'identifier les plumes et les inspirations des œuvres.

Nous présentons la dérivation par le prisme du plagiat qui est à la fois l'appelation grand public, le mot-clé de la communauté et qui reste de plus l'application principale de la détection de dérivation.

1 Le plagiat, nerf de la guerre

Un *plagiarius* désignait, dans la Rome antique, soit un voleur d'esclave soit quelqu'un qui vendait comme esclave une personne libre. Le terme « plagiat » qui en découle est défini dans le TLFI[1] (Trésor de la Langue Française Informatisée) comme

1. `http://atilf.atilf.fr/`

« l'action d'emprunter à un ouvrage original et à son auteur des éléments dont on s'attribue abusivement la paternité en les reproduisant, avec plus ou moins de fidélité, dans une œuvre que l'on présente comme personnelle ». Tout le paradoxe de la notion de plagiat réside dans cette évolution du concept : du vol de propriété physique[2] à celui de la reproduction de l'écrit ou de l'idée d'un autre. Le premier est condamnable car il entraîne la privation du bien pour celui à qui on le dérobe (le maître ou l'homme libre lui-même), le second est condamné par la morale en ce que le plagiaire profite d'une gloire indue qui aurait dû profiter à l'auteur original. Ce qui représentait pour Diderot le « délit le plus grave qui puisse se trouver dans la République des Lettres ».

Le plagiat est un délit moral qui n'a pas forcément de pendant juridique. Les délits assimilables à du plagiat se distribuent entre les atteintes au droit d'auteur (Code de la propriété intellectuelle, Première partie : la propriété littéraire et artistique) et la contrefaçon (Code de la propriété intellectuelle, Deuxième partie : la propriété industrielle) mais une grande partie de ces délits ne sont pas constitués. Les affaires ne sont d'ailleurs portées devant les tribunaux que lorsque le préjudice a des conséquences commerciales (en littérature majoritairement).

Multiplication des accusations de plagiat

Les accusations de violation du droit d'auteur défraient parfois la chronique lorsqu'elles concernent des œuvres à succès[3]. C'est notamment le cas dans le procès opposant *Willy the Wizard* à *Harry Potter*. Les avocats du plaignant estiment les dommages à près d'un milliard de dollars. Comment les accusateurs vont-ils justifier la violation ? Comment les accusés vont-ils se défendre ? Existe-t-il des méthodes juridiquement cautionnées permettant de valider une accusation de violation de droit d'auteur, ou bien au contraire de la lever ?

Les institutions juridiques commencent à s'équiper de moyens leur permettant de prendre des décisions concernant les affaires de « plagiat ». Ainsi, l'*Institute for Linguistic Evidences*[4] financé notamment par le *United States Department of Justice* travaille à la mise au point de méthodes permettant d'obtenir des preuves reconnues par un tribunal. Cet institut travaille notamment sur l'outil ALIAS (*Automated Linguistic Identification and Assessment System*) (Chaski, 2001) qui a partiellement passé les normes de Frye et Daubert[5].

Les établissements scolaires s'équipent également de tels outils afin de déceler les fraudes aux examens. Encore récemment, un étudiant d'une école privée suisse s'est vu déchu de son diplôme pour avoir puisé l'essentiel de son mémoire dans Wikipédia (Pingoud et Fabre, 2010). Les condamnations à l'encontre de Wikipédia sont légions, mais elles relèvent principalement d'incompréhensions à l'égard des licences impliquées. Dans le cas présent, il est tout à fait légal de recopier intégralement le contenu de Wikipédia. La licence le permet sous réserve de citer l'encyclopédie, ce qu'a omis ici de faire l'étudiant et ce que lui reproche l'administration.

Les outils employés dans ces différents cas sont très sensibles aux modifications opérées par les plagiaires, et il est assez aisé de les contourner. De plus, la présence de longs passages retrouvés dans d'autres textes est un peu trop rapidement considéré comme plagiat. Ainsi, l'initiative *PhraseBank*[6] entraîne certainement une augmenta-

2. Aussi triste que ce soit, les esclaves étaient les propriétés de leurs maîtres.

3. Voir notamment `http://britainnews.ru/fr/j45/3773.html`, `http://www.accesshollywood.com/harry-potter-author-hit-with-plagiarism-lawsuit_article_29139` ou encore `http://www.guardian.co.uk/books/2010/feb/18/harry-potter-jk-rowling-willy-wizard`

4. `http://www.linguisticevidence.org/`

5. Les normes de Frye et Daubert définissent l'admissibilité d'un témoignage scientifique lors d'un procès aux États-Unis.

6. `http://www.phrasebank.manchester.ac.uk/index.htm`

tion des faux positifs en sortie de ces logiciels. En effet, ce site compile une collection de phrases outils directement utilisables dans des papiers académiques. Il a pour but d'aider les étudiants non-anglophones dans la rédaction d'articles scientifiques internationaux. Le site informe que les phrases proposées « sont neutres de contenus et génériques en nature ». Il ne s'agit donc pas de vol des idées d'un autre, leur utilisation ne constitue donc pas un plagiat. Les outils automatiques peuvent pourtant s'y laisser prendre.

L'accusation de plagiaire est très grave et ne peut-être décidée que par une cour. Cependant, les milieux académiques le font assez peu ce qui, au final, est dommageable à la présomption d'innocence. Ainsi, l'accusation envers plusieurs scientifiques turcs parue dans la revue Nature en 2007 (Brumfiel, 2007) avait fait l'effet d'une bombe. Ces accusations étaient fondées sur un fort recouvrement de segments de texte entre les articles de ces scientifiques et d'autres articles présents sur le serveur arXiv [7]. Les accusés s'étaient défendus dans une réponse également publiée dans Nature en 2007 (Yilmaz, 2007). Ils réfutaient le plagiat clamant la réutilisation, juste et morale, de phrases en bon anglais tirées desdits articles dans le but d'améliorer la qualité de l'anglais de leurs propres articles.

Au regard des paragraphes précédents, la réutilisation d'expressions vides de contenu, ne serait pas du plagiat, contrairement la réutilisation d'expressions avec du contenu. Pour autant, que ce soit en sciences, en journalisme ou même en littérature, les nouvelles œuvres recyclent majoritairement les idées des œuvres antérieures. La figure 1 illustre ce paradoxe. La similitude entre de nombreux passages de ces thèses soutient la thèse du plagiat, mais les universités concernées ne tranchent pas en ce sens. La présence de citations de la première thèse dans la seconde est peut-être une première piste d'explication.

Le plagiat comme renfort créatif ?

Le plagiat est perçu, à dessein, comme un acte moralement condamnable tandis que la réutilisation de contenu ou d'idées est souvent associée au processus créatif. En réalité, la frontière entre les deux est assez floue.

Pour Giraudoux (1928, Acte I, scène 6), « le plagiat est la base de toutes les littératures, excepté de la première, qui d'ailleurs est inconnue ». Dans le Palimpsestes (Genette, 1982), Genette analyse des enchevêtrements d'œuvres : de la copie à l'allusion lointaine. Il montre qu'un texte peut toujours en cacher un autre, mais qu'il le dissimule rarement tout à fait, prenant pour exemple l'*Ulysse* de Joyce qui est, d'après sa terminologie, l'hypertexte (œuvre dérivée) de l'*Odyssée* d'Homère qui en est l'*hypotexte* (œuvre originale). Les pastiches et les parodies sont autant d'exemples du potentiel créatif de la réutilisation. Ainsi, totalement ancré dans notre époque, le livre Twittérature (eds. Saint-Simon) est une réécriture de 70 romans classiques en 20 phrases de 140 signes chacun. Une œuvre qui n'apporte aucun nouvel élément et que l'on pourrait considéré comme un simple exercice d'écriture mais qui relève de la création.

En réalité, comme tout ce qui relève de la morale, le plagiat est une notion culturelle. Ainsi, bien que le droit d'auteur soit assez uniformisé à l'internationnal grâce au traité de Berne (1886), chaque pays possède son propre système de protection du droit d'auteur, lorsqu'il en possède. Les États-Unis par exemple n'ont signé le traité qu'un siècle plus tard, réticents à la vision franco-allemande centrée sur l'auteur et son droit moral. Cette notion est même encore plus éparse puisqu'elle est perçue différemment au sein des branches d'un même métier. Thomas (2009) montre qu'un processus d'écriture considéré comme du plagiat par la presse écrite et qui entraînerait le renvoi

7. `http://arxiv.org/`

FIGURE 1 – Dans cette image, tirée de (wha, 2009), représentant les pages d'une thèse, les passages qui correspondent mot pour mot à des passages d'une autre thèse sont surlignés en jaune. Le partage de nombreuses sous-chaînes communes est la preuve la plus souvent mise en avant dans les cas de plagiat. On trouve également dans le cas présent des correspondances dans la structuration de la thèse (titres des sections et sous-sections).

du journaliste est très couramment usité à la TV et à la Radio. D'une manière générale, la frontière entre plagiat et réutilisation est très ténue dans la presse. Cela est peut-être dû au fait que les journalistes manipulent des faits qui sont généralement directement accessibles à tous. Il est alors difficile de savoir si l'information a été captée à la source du fait ou au travers d'un autre article de presse. Clough (2003a) consacre le chapitre 3 de sa thèse à la description de la réutilisation dans le journalisme.

Finalement, c'est également un concept qui évolue au travers des pratiques et des générations. Comme le présente Erny-Newton (2010) :

> « Dans la culture dématérialisée des natifs du numérique, l'emprunt n'est pas associé au vol mais à la création communautaire. Pour les membres de la culture du remix, l'emprunt est au cœur de la création, en même temps qu'il représente un hommage à une création antérieure. Ce qui frustre un/e Gen Y [individu de la nouvelle génération, nda], ce n'est pas que quelqu'un puisse réutiliser ses productions sans son consentement, c'est qu'il ne puisse mettre les doigts dans celles des autres –particulièrement celles qui forment le canevas de sa propre culture. La culture du remix crée de nouvelles phrases à partir d'un alphabet social partagé par une génération. »

Nous faisons le choix dans cette thèse de laisser complètement de côté toutes les considérations morales, sociales, juridiques et économiques liées au plagiat et de nous concentrer sur ce que nous appellerons **la dérivation**. Cette notion, détaillée dans le chapitre 1, se focalise sur le processus créatif lié à la réutilisation et non sur ses implications.

Exemple de dérivation

Le GIEC est un groupe d'observation qui compile les résultats des études scientifiques sur le changement climatique. Les rapports qu'il publie à l'attention des dirigeants gouvernementaux dérivent donc des articles scientifiques utilisés comme source. Dans son édition du 17 janvier 2010, *The Sunday Times* titrait sur la méconduite du monde concernant la fonte des glaciers de l'Himalaya pour 2035. L'abattement médiatique qui s'en suivit a permis de connaître avec une bonne certitude le parcours de l'information erronée de son origine à sa publication dans le rapport du GIEC. Nous étudions brièvement par la suite la manière dont l'information *disparition des glaciers de l'Himalaya en 2035* a été relayée de textes en textes.

Le parcours de la mauvaise information a été établie par plusieurs journalistes. L'information originale proviendrait d'une interview téléphonique d'un scientifique nommé Hasnain, source d'un article du *New Scientist* que nous prenons comme point de départ. L'article du *New Scientist* aurait été repris en premier lieu par un rapport du *WWF*, puis ce rapport aurait été lui-même repris par le GIEC. Nous détaillons chacune de ces étapes en partant de l'hypothèse que cette suite de dérivation est la bonne.

La première dérivation a pour source l'article du *New Scientist* et pour dérivé le rapport du WWF (*cf. Exemple 1*). La reprise du WWF est citationnelle, elle se compose donc d'un contexte de citation et du discours repris. La contextualisation par le WWF est intéressante car elle rattache par erreur les propos à un rapport du WGHG (*Working Group on Himalayan Glaciology*) alors que le texte source cite clairement l'étude sur quatre ans de Hasnain. Le WWF a interprété que la prévision faisait partie du rapport de l'ICSI car l'auteur original fait référence à une étude de l'auteur principal d'un rapport du ICSI. Le discours repris est tout aussi intéressant. Le texte source parle d'une disparition possible des glaciers du centre et de l'est de

NewScientist – Flooded out

"All the glaciers in the middle Himalayas are retreating," says Syed Hasnain of Jawaharlal Nehru University in Delhi, the chief author of the ICSI report. A typical example is the Gangorti glacier at the head of the River Ganges, which is retreating at a rate of 30 metres per year. Hasnain's four-year study indicates that all the glaciers in the central and eastern Himalayas could disappear by 2035 at their present rate of decline.

WWF – An Overview of Glaciers, Glacier Retreat, and Subsequent Impacts in Nepal, India and China (p.29)

As discussed in the thematic introduction to this regional status review, there is particular concern at the alarming rate of retreat of Himalayan glaciers. In 1999, a report by the Working Group on Himalayan Glaciology (WGHG) of the International Commission for Snow and Ice (ICSI) stated : "glaciers in the Himalayas are receding faster than in any other part of the world and, if the present rate continues, the livelihood of them disappearing by the year 2035 is very high". Direct observation of a select few snout positions out of the thousands of Himalayan glaciers indicate that they have been in a general state of decline over, at least, the past 150 years.

Exemple 1: Première dérivation : de l'article du *New Scientist* au rapport du WWF

l'Himalaya pour 2035. Le WWF étend la zone à tous les glaciers de l'Himalaya et modifie la modalité, parlant de *très forte probabilité*.

La seconde dérivation a pour source le rapport du WWF et pour dérivé le rapport du GIEC (*cf. Exemple 2*). La reprise du GIEC est une référence, un type de dérivation commun pour les articles scientifiques. Elle se compose d'un passage synthétisant le document auquel il est fait référence ainsi qu'une référence bibliographique à ce document, en l'occurrence le rapport du WWF. Le passage synthétique est une reprise quasi identique à la citation du texte source. Il se différencie de la source par la correction orthographique (*livelihood* corrigé en *likelihood*), ainsi que l'ajout d'une référence à un tableau de données appuyant la première constatation (fonte plus rapide des glaciers de l'Himalaya) et l'ajout d'une condition à la phrase principale « *if the Earth keeps warming at the current rate* ».

L'étude de ces phénomènes de dérivation et la mise au point d'outils permettant de les détecter automatiquement permettrait de retrouver plus facilement les sources des informations, ce qui, comme le montre cet exemple, a un réel intérêt.

2 Problématique et motivations

Problématique

Les auteurs sont conscients des textes qui leur servent de source, des passages ou des idées qu'ils exploitent dans ces textes, et comment ils les intègrent dans leur propre création. Cependant, ces informations sont confinées au processus de création et donc inconnues du lecteur, comme l'illustre la figure 2. En effet, une fois le processus accompli, le texte créé n'est pas différent d'un autre texte écrit à partir de rien, exception faite des cas où l'auteur laisse transparaître la dérivation (citations, références...). L'identification d'une dérivation nécessite de fouiller la seule trace accessible : le texte dérivé.

L'identification de la dérivation se décline elle-même en de nombreuses tâches. S'agit-il de découvrir les liens de dérivation au sein d'une collection de textes ? S'agit-

WWF – An Overview of Glaciers, Glacier Retreat, and Subsequent Impacts in Nepal, India and China (p.29)

As discussed in the thematic introduction to this regional status review, there is particular concern at the alarming rate of retreat of Himalayan glaciers. In 1999, a report by the Working Group on Himalayan Glaciology (WGHG) of the International Commission for Snow and Ice (ICSI) stated : "glaciers in the Himalayas are receding faster than in any other part of the world and, if the present rate continues, the livelihood of them disappearing by the year 2035 is very high". Direct observation of a select few snout positions out of the thousands of Himalayan glaciers indicate that they have been in a general state of decline over, at least, the past 150 years.

IPCC – 4^{th} Assessment Report of IPCC (section 10.6.2)

Glaciers in the Himalaya are receding faster than in any other part of the world (see Table 10.9) and, if the present rate continues, the likelihood of them disappearing by the year 2035 and perhaps sooner is very high if the Earth keeps warming at the current rate. Its total area will likely shrink from the present 500 000 to 100 000 km2 by the year 2035 (WWF, 2005).

Exemple 2: Seconde dérivation : du rapport du WWF à celui du GIEC

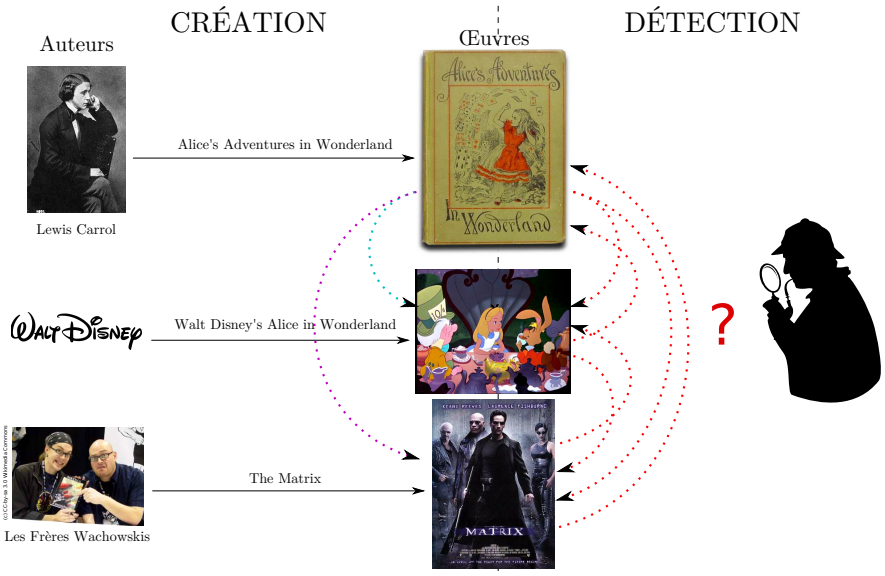


Figure 2 – Les processus de création et de détection de dérivation de texte sont complètement déconnectés l'un de l'autre rendant la tâche particulièrement difficile. Les flèches en pointillé indiquent des relations de dérivation entre les œuvres.

il, étant donné deux textes, de décider si l'un dérive de l'autre ? S'agit-il, étant donné deux textes dérivés l'un de l'autre, de décider lequel des deux est la source ? S'agit-il encore d'identifier les éléments qui ont été prélevés dans le texte source pour produire le dérivé ? Ces différentes tâches peuvent être mises en œuvre afin de répondre à des besoins applicatifs différents : détection du plagiat académique (Lyon et collab., 2004), identification de révision (Hoad et Zobel, 2002), recyclage de texte (Aizawa, 2003) ou filtrage de copies approximatives (Zobel et Bernstein, 2006; Yang, 2006a).

Ces tâches sont d'autant plus difficile que le processus de dérivation peut prendre de nombreuses de formes. Le nombre de sources, les transformations opérées, la quantité des éléments prélevés... sont autant d'éléments qui participent à la configuration du processus de dérivation et qui ajoutent à la difficulté de recomposer ledit processus.

Motivations

La détection automatique de dérivation a de nombreux intérêts applicatifs. L'application la plus couramment évoquée est celle de la protection de la propriété intellectuelle. Les ayants droits y voient un moyen de défendre leurs intérêts, y puisant à la fois un moyen de veille aussi bien que des preuves d'infraction. Le corps enseignant y voit un moyen d'endiguer le plagiat et la collusion parmi les étudiants.

La recherche de la source ou le suivi d'une information dans la presse est une autre application prometteuse. Ainsi, Bendersky et Croft (2009) s'intéressent au suivi du développement d'une information sur le Web. Par extension, il serait possible de construire les filiations des textes concernant l'évolution d'une information, à la façon des arbres phylogénétiques pour la classification des espèces. Les services de communication d'entreprises, d'institutions ou d'individus pourraient également tirer parti d'un tel suivi concernant leurs clients.

Il s'agit également d'une piste d'amélioration pour la Recherche d'Information (RI). Les contenus présents sur le Web sont largement dupliqués et versionnés de sorte qu'il est difficile de sélectionner l'information pertinente parmi toutes les informations redondantes. Fetterly et collab. (2003) rapportent que près de 28 % des pages Web sont des répliques des 72 % de pages restantes, et que parmi celles-ci seules 6 % ne sont pas des copies à l'identique. Ainsi, Bernstein et collab. (2006) s'intéressent à la détection de doublons lors de la fusion des résultats dans les systèmes de RI distribués.

Toujours en RI, le regroupement de textes co-dérivés est une première étape permettant la production de descriptions prototypiques des textes traitant d'un même sujet à partir desquels il serait possible d'identifier les positionnements spécifiques de chaque co-dérivé. Cette approche irait dans le sens de la recherche guidée par l'opinion. La détection de dérivation permettrait également l'émergence d'un nouveau paradigme où la recherche partirait d'un texte et non plus d'une requête. Le moteur de recherche Etblast [8] permet ainsi de rechercher non pas sur la base de mots clés, mais en mesurant la similarité entre des passages de texte.

Le monde de la recherche profiterait également de meilleurs outils de détection de dérivation. Ainsi, le regroupement de documents dérivés multilingues permettrait de construire automatiquement des corpus comparables. De tels outils devraient alors être en mesure de fonctionner sur un corpus ouvert, c-à-d pour lequel la liste des textes évolue au fil du temps. La critique génétique textuelle profite d'ores et déjà de tels algorithmes (Bourdaillet, 2007) ce qui facilite l'étude de l'évolution des différentes versions de l'œuvre d'un auteur. Le contexte de mise en œuvre correspond cependant à une recherche dans un corpus fermé, c-à-d pour lequel la liste des textes est figée et connue.

8. `http://invention.swmed.edu/etblast/etblast.shtml`

Positionnement

Dans le cadre de ces travaux de thèse, nous n'explorons pas toute l'étendue des formes de dérivation. Tout d'abord, nous nous limitons aux dérivations de texte à texte, nous délaissons le problème de la multimodalité. Ensuite, nous nous concentrons sur les dérivations qui reprennent du texte ou au moins des idées du texte source, nous délaissons les dérivations de structure ou de style (*cf. Section 1.4*). Finalement, nous travaillons résolument à l'échelle du document. Nous délaissons les recherches de dérivation à l'échelle de la séquence textuelle telles qu'opérées pour le *textual entailment* (Giampiccolo et collab., 2007) par exemple, ou même du passage. Les méthodes appliquées à l'échelle du document peuvent être appliquées à ces niveaux, même si d'autres méthodes plus coûteuses peuvent également être mises en œuvre (*cf. Section 2.2.4*).

3 Plan de la thèse

Cette thèse comporte cinq chapitres qui discutent l'état de l'art et rapportent nos diverses propositions et contributions.

Chapitre 1 : la notion de dérivation

Le premier chapitre introduit la notion de dérivation de texte et la définit par rapport aux différentes formes abordées dans la littérature. Ainsi, nous présentons un état de l'art de la détection de plagiat et des phénomènes voisins. Nous montrons notamment que les travaux convergent vers la détection de relations de filiation plus larges que la duplication. Nous proposons un cadre théorique pour ces relations de dérivation, ainsi qu'un modèle multidimensionnel permettant de les caractériser.

Chapitre 2 : les méthodes de détection de dérivation

Le second chapitre reprend l'état de l'art en se focalisant sur les méthodes déployées pour détecter automatiquement des relations de dérivation. Nous catégorisons ces méthodes en deux grandes familles : (i) les méthodes intrinsèques qui ne tirent parti que des indices contenus dans le document étudié et (ii) les méthodes extrinsèques qui reposent sur la comparaison de paires de documents. Nous en dégageons deux méthodes que nous mettons en œuvre par la suite : la détection intrinsèque sur la base d'éléments contextuels pour les citations et la détection extrinsèque reposant sur la comparaison d'éléments textuels pour diverses autres formes de dérivation.

Chapitre 3 : détection de reprises contextualisées par apprentissage artificiel

Dans le troisième chapitre nous nous intéressons à un type particulier de dérivation que sont les reprises contextualisées. Il s'agit des dérivations textuellement signalées par l'auteur et dont l'instanciation la plus commune est la citation. Nous rapportons une approche de détection de ces segments dérivés par l'exploitation de divers indices contextuels par des algorithmes d'apprentissage supervisé.

Nous présentons l'évaluation de cette approche sur un corpus d'articles de presse constitué par nos soins. Nous en déduisons notamment que les modèles probabilistiques exploités ne sont pas forcément les mieux indiqués et que le gain que nous en tirons pour la désambiguïsation des segments entre guillemets n'est pas significatif au regard de la complexité de la méthode.

Ce travail a été réalisé dans le cadre du projet PIITHIE [9] soutenu par l'Agence Nationale de la Recherche sous la référence 2006 TLOG 013 03.

Chapitre 4 : ressources et protocoles d'évaluation pour la détection extrinsèque

Dans le quatrième chapitre nous discutons des protocoles et des ressources pour l'évaluation des méthodes de détection extrinsèques de dérivation. Nous présentons les méthodes d'évaluation les plus utilisées, les mesures et les corpus associés. Sur la base de ces connaissances, nous proposons un protocole d'évaluation composé de mesures tirées de la RI, de corpus constitués dans le cadre de cette thèse ainsi que des résultats de référence pour chacune des formes de dérivation étudiées.

Chapitre 5 : détection extrinsèque de dérivations par des signatures de taille réduite

Le cinquième chapitre rapporte l'ensemble de nos propositions concernant la détection de dérivation. Nous y relatons les différentes séries d'expérimentations que nous avons menées afin d'éprouver nos méthodes : filtrage statistique des n-grammes mots, emploi de descripteurs ayant un ancrage linguistique ainsi que la combinaison de ces différentes approches. Nous montrons que la sélection de descripteurs pertinents permet d'obtenir des résultats similaires à l'approche de référence pour un coût de traitement bien inférieur et que la combinaison de ces descripteurs permet d'obtenir de meilleurs résultats que l'approche de référence.

4 Principales contributions

Nous nous intéressons dans cette thèse à la détection de plusieurs formes de dérivation (reprises de presse, révisions et plagiat artificiel), principalement en français. Nous faisons le choix d'aborder la détection automatique de ces formes de dérivation en confrontant des textes, considérés comme sources, à une collection de textes suspects. Nous nous attelons donc à la tâche d'identification automatique, au sein d'une collection fermée de candidats, des textes qui dérivent d'un texte source connu. Nous nous donnons notamment pour objectif de réduire le coût en termes d'espace de stockage et de temps d'exécution des méthodes existantes.

Nos contributions portent sur la définition d'un cadre théorique unifié de la dérivation de texte, la mise à disposition de corpus pour l'observation et l'expérimentation, la proposition d'approches de détection et d'évaluation des systèmes ainsi que la diffusion sous licence libre des outils réalisés :

Définition d'un cadre théorique unifié de la dérivation Notre première contribution consiste en la définition d'un cadre théorique posant les concepts de la dérivation. Ce cadre nous a permis de définir une taxinomie des formes de dérivation. Celle-ci se fonde sur différents traits qui caractérisent les différents liens de dérivation qui composent la globalité de la dérivation à l'échelle du document. Nous avons été en mesure, à l'aide de ce cadre théorique et de cette taxinomie, d'unifier les travaux de la littérature qui traitent majoritairement de formes, de tâches ou encore de configurations particulières liées à la dérivation.

Dérivation en langue française Notre travail de thèse est, à notre connaissance, le premier à s'intéresser au cas de la dérivation dans la langue française. La littérature relate presque exclusivement des travaux sur l'anglais. La conséquence de

9. www.piithie.com

cette singularité fût la nécessité de compiler des actes de dérivation en français, aucun corpus de ce type n'étant disponible dans cette langue.

Corpus Nous avons constitué trois corpus dans le cadre de ce travail qui ont été utilisés pour la validation expérimentale de nos méthodes. Le corpus Piithie, dont la construction est le fait de nos partenaires du projet PIITHIE [10], illustre les dérivations opérées dans la presse française. Le corpus Wikinews illustre les dérivations de type révision (rédaction itérative à plusieurs main), également sur des articles de presse et en français. Enfin, le corpus PANini est une version réduite du corpus anglais de référence PAN. Le corpus Wikinews ainsi que le corpus PANini sont distribués librement de sorte qu'ils puissent être réutilisés par la suite.

Signatures de taille réduite Nous avons étendu l'approche par signature en utilisant des descripteurs linguistiquement motivés. Les approches antérieures emploient majoritairement des n-grammes de mots. Les signatures qui en découlent sont assez volumineuses et donc difficilement opérationnelles. La contrainte linguistique permet d'alléger les signatures tout en maintenant une qualité de prédiction comparable.

Outils Nous avons développé la boîte à outils logicielle TDDTS (*Textual Derivation Detection ToolSet*) [11], reposant sur le cadriciel Apache UIMA [12], qui permet d'expérimenter les différentes approches que nous avons expérimenté dans le cadre de cette thèse. TDDTS est distribué librement sous licence Apache UIMA afin de faciliter les recherches futures.

10. Le projet PIITHIE (`www.piithie.com`) a été financé par l'Agence Nationale française pour la Recherche (ANR, numéro de financement : 2006 TLOG 013 03).
11. `http://www.fabienpoulard.info/download/Recherche/TDDTS/`
12. `http://uima.apache.org`

Chapitre 1

La dérivation de texte

Les bons artistes copient, les grands artistes volent.

— Pablo Picasso

Sommaire

Les travaux sur le rapprochement de textes semblables débutent certainement avec la mise au point des algorithmes de recherche sous-chaînes exactes (Aho et Corasick, 1975; Knuth et collab., 1977; Boyer et Moore, 1977; Karp et Rabin, 1987a) et de sous-chaînes communes acceptant de légères variations (Jaccard, 1912; Dice, 1945; Hamming, 1950; Levenshtein, 1966; Winkler, 1999). Le développement de ces outils destinés aux passages de texte a récemment permis de s'intéresser au rapprochement de textes complets (Manber, 1994; Brin et collab., 1995; Si et collab., 1997; Shivakumar et Garcia-Molina, 1999). Ces derniers travaux ont eux-même inspirés les travaux plus récents sur le problème de la détection de réutilisation partielle de textes réduit au cas du plagiat (Monostori et collab., 2000; Clough, 2003a; Fetterly et collab., 2003). Finalement, les plus récents travaux (Bernstein et Zobel, 2005; Stein, 2005; Seo et Croft, 2008; Bendersky et Croft, 2009) élargissent la problématique à d'autres scénarios : création de textes littéraires, résumés, transpositions de genre, traduction ou encore révision. Tous ces travaux se sont intéressés à des formes particulières de production de textes à partir d'autres. Ils semblent lentement converger vers des phénomènes communs sans jamais les étudier dans leur globalité. Nous souhaitons offrir une vision globale de ce problème commun que nous nommons la *dérivation de texte*.

Le terme de dérivation a été initialement introduit par Bernstein et Zobel (2004) dans le sens de réutilisation [1] pour définir la notion adjacente de co-dérivation. Dans le cadre de cette thèse, nous proposons d'adopter une perspective qui vise à généraliser ces phénomènes. Ainsi nous définissons informellement la notation de *dérivation de texte* comme le processus rédactionnel permettant de produire un nouveau texte (le texte dérivé) à l'aide d'autres textes (les textes sources). Par extension, nous définissons la *co-dérivation de texte* comme un cas particulier où deux textes sont liés par une dérivation mais ni la source ni le dérivé ne sont identifiés.

La dérivation de texte est courante et se retrouve dans beaucoup de méthodes de production de textes. Ainsi, un ou plusieurs textes sources interviennent dans le processus rédactionnel des activités suivantes :

> Un journaliste dérive d'un texte scientifique un article de presse vulgarisé afin de le rendre accessible à ses lecteurs. Ce même journaliste peut également dériver de ce texte scientifique des passages qu'il présentera dans son article sous la forme de citations.
>
> Dans leur communication, ces scientifiques se positionnent par rapport aux travaux antérieurs en résumant le contenu des articles correspondants et en y accolant une référence bibliographique.
>
> Un traducteur traduit le dernier roman d'un auteur : il produit une œuvre dérivée de ce roman écrit dans une langue différente et s'ancrant dans la culture du pays visé par l'ouvrage.

La détection de ces différentes formes de dérivation de texte est au cœur de nombreux enjeux. Elle permettrait de maîtriser la redondance d'information inhérente au Web (Bernstein et Zobel, 2005), d'optimiser le parcours de sites (Manku et collab., 2007), le stockage de fichiers (Kulkarni et collab., 2004) ou encore de protéger la propriété intellectuelle (Özlem Uzuner et Davis, 2003). Les travaux sur la détection de dérivation de texte profiteraient également à des tâches telles que le rapprochement de passages de texte, la recherche de copies à quelques modifications près, l'identification de réutilisations de texte ou le suivi des révisions d'un document ou même d'une information.

Nous nous attachons dans ce chapitre à introduire la notion de dérivation comme un prolongement logique des travaux antérieurs sur les différentes formes de dérivations étudiées. Nous cherchons pour cela à définir formellement cette notion et à

1. « for two documents to be co-derived, some portion of one must be *derived* from the other »

cadrer les différentes formes par lesquelles elle est mise en œuvre. Nous faisons tout d'abord le bilan des différentes formes de dérivation qui ont été étudiées dans la littérature : duplication, version, résumé, plagiat et collusion, citation et référence, transposition de genre et traduction (*cf. Section 1.1*). Nous proposons ensuite un cadre théorique permettant de fixer ce qu'est la dérivation et quelques autres notions connexes (*cf. Section 1.2*). Enfin, à partir des différentes propositions de classification de la littérature (*cf. Section 1.3*), nous proposons une vision multidimensionnelle de la dérivation qui décrit les relations de dérivation à l'origine de la création d'un texte dérivé (*cf. Section 1.4*). Ce modèle permet d'une part d'unifier les différentes propositions de classification existantes et d'autre part de classer et différencier les différentes formes de dérivation présentées auparavant.

1.1 Différentes formes de dérivation

Nous présentons dans ce chapitre différents phénomènes étudiés indépendamment les uns des autres dans la littérature et qui relèvent, selon nous, de la dérivation de texte : duplication, version, résumé, plagiat et collusion, citation et référence, transposition de genre et traduction. La dénomination de ces formes varie selon les travaux. Nous tentons d'utiliser l'appellation la plus consensuelle tout en nous efforçant de présenter les éventuelles variations. Nous cherchons à illustrer, pour chacune de ces formes, les propriétés qui en font une dérivation de texte.

1.1.1 Duplication

Nous proposons le terme « duplication »comme traduction du terme anglais « *duplicate* »et celui de « presque-duplication » pour le terme anglais « *near-duplicate* »ou « *near-replicas* ». Intuitivement, le phénomène de duplication consiste à copier un texte à l'identique, contrairement à la presque-duplication pour laquelle des modifications peuvent être apportées. Nous présentons tout d'abord la notion de duplication et comment le phénomène de la presque-duplication a émergé de la duplication des sites internet. Nous présentons ensuite les différents travaux qui ont cherché à cadrer cette notion de presque-duplication.

DUPLICATION

La détection de *duplication* de documents est le centre d'intérêts de nombreux travaux, que le phénomène porte sur des documents isolés (Manber, 1994; Doermann et collab., 1997) ou des collections de documents (Broder et collab., 1997; Cho et collab., 2000). La duplication est le phénomène le plus trivial de production d'un texte à partir d'un autre puisqu'elle consiste à *recopier* intégralement (dupliquer) un texte ou une collection de textes. Les textes source et dérivé partagent globalement la même forme textuelle. Ils se différencient éventuellement par leur aspect visuel (sauts de ligne...) ou leur stockage informatique (encodage, format...). Pour autant, ces différentes variations graphiques ne suffisent à en faire des textes différents.

La création de miroirs[2] sur Internet a donné un nouvel élan au phénomène de duplication. Cependant, la modification locale des copies de documents au sein de ces miroirs entraîne potentiellement un éloignement des formes textuelles de la source et des dérivés. Ainsi, les textes peuvent diverger à cause d'infimes variations textuelles dûes à l'application (ou la non-application) des règles d'orthographe et de grammaire du langage associé mais aussi par l'ajout de données, le redécoupage des fichiers ou d'autres opérations rendues nécessaires par le maintient du miroir. Les textes produits au sein de ces miroirs ne sont plus exactement des copies à l'identique mais

2. Un site miroir est une copie exacte, à un moment donné ou synchronisé fréquemment, d'un autre site.

PRESQUE-DUPLICATION

des copies approximatives. Il ne s'agit alors plus du phénomène de duplication mais de *presque-duplication*. La littérature anglophone y fait référence sous des dénominations diverses : *near-replicas*, *near-duplicates* ou encore *near-copy* (Heintze, 1996; Shivakumar et Garcia-Molina, 1999).

Si le concept de presque-duplication est légitime, ses frontières avec la copie et d'autres formes de dérivation ne fait pas consensus. Broder et collab. (1997) liste plusieurs cas de documents proches mais pas identiques rencontrés sur le Web : différentes versions du même document, le même document à la mise en forme près, le même document avec des liens ou une personnalisation spécifiques au site, la combinaison d'autre matériel source pour former un document plus important ou encore la séparation en plus petits documents. Il parle de documents « trouvés dans une incarnation identique »sans cadrer plus précisément cette notion similaire à la presque-duplication. Toujours dans le cadre du Web, Shivakumar et Garcia-Molina (1999) introduisent la notion de copie approximative. Leurs *near copies* sont « les vieilles versions de documents populaires, ou dans des formats différents, ou qui peuvent avoir des boutons, liens ou images en plus » [3]. Malheureusement, le format des documents Web mêle contenu et mise en forme au sein de la dimension textuelle du document. Finalement, Cho et collab. (2000); Fetterly et collab. (2003) introduisent la notion de recouvrement de texte pour mesurer la quantité de matière en commun. Le recouvrement de texte désigne les expressions textuelles (séquences de caractères) en commun entre des textes. Fetterly et collab. (2003) introduit le terme de *near-duplicates* pour désigner les documents qui se recouvrent au minimum à 95 %. L'utilisation du recouvrement de texte a depuis été repris par d'autres comme une mesure objective de la distance entre des textes dérivés (Seo et Croft, 2008; Potthast et collab., 2010b).

Au final, les notions de duplication et de presque-duplication sont associées à des textes qui se recouvrent de manière importante, c-à-d qui partagent un nombre improbable [4] de séquences textuelles communes. Cette forme de dérivation désigne donc un texte dérivé très proche dans sa forme textuelle du texte source. Par abus de langage et par souci de simplicité terminologique, nous utiliserons le terme duplication pour désigner indépendamment le phénomène de duplication ou de presque-duplication.

1.1.2 Version

VERSION

Si le stockage numérique des textes facilite leur duplication, leur création au travers de l'outil informatique entraîne la multiplication de leurs éditions et résulte en la cohabitation de plusieurs *versions* pour un même document. Intuitivement, les versions correspondent à différents états transitionnels d'un texte produits par les mêmes auteurs à partir d'une même base d'écriture commune. Nous présentons tout d'abord les travaux de Hoad et Zobel (2002) qui traitent sur le même plan les phénomènes de version et de plagiat, puis les travaux de Bourdaillet (2007) qui perçoit les versions comme des étapes du processus de création. Enfin, nous illustrons par un exemple les opérations d'édition introduites par ce dernier.

Hoad et Zobel (2002) cherchent à déterminer si deux documents sont des versions différentes du même texte ou bien des textes différents. Ils traitent les différentes versions des textes comme des codérivés (des documents dérivant d'un même texte source). Pour les auteurs, révisions et plagiat sont semblables dans le sens où ils peuvent être détectés par les mêmes techniques et ne cherchent pas à caractériser ce

3. « documents may be older versions of some popular documents, or may be in different formats, or may have additionnal buttons, links and inlined images that make them slightly different from other versions of the document »

4. Improbable dans le sens où il est peu probable que ces séquences se retrouvent dans deux textes écrits indépendamment.

qui différencie ces deux formes de dérivation.

Bourdaillet (2007) s'intéresse aux versions dans le cadre de la critique génétique textuelle. Cette discipline, issue de la philologie, étudie la genèse des œuvres littéraires au travers des brouillons d'écrivains. Elle ne reconnaît pas forcément l'existence de versions et lui préfère un continuum textuel établit entre le premier brouillon et la version finale d'un texte. Cependant, dans un but opérationnel, Bourdaillet (2007) fait l'hypothèse d'un ensemble discrétisé de versions. Dans ce cadre, le passage d'une version à une autre s'effectue au moyen de quatres opérations menées par l'écrivain sur les phrases, mots ou caractères : les insertions, les suppressions, les substitutions et les déplacements.

L'exemple 3 illustre les opérations d'insertion, de ~~suppression~~ et de remplacement sur un paragraphe. Cet exemple est tiré de la version francophone de Wikinews [5]. Le contexte du projet Wikinews est particulier puisque l'auteur des opérations menant à la nouvelle version n'est pas nécessairement un des auteurs originaux de l'article. Le projet souligne l'idée qu'une version est un type de dérivation qui ne se caractérise pas par la continuité des auteurs mais plutôt par l'intention qui dirige la production d'un écrit. Nous considérerons que les versions d'un document sont des dérivations menées durant la genèse du texte pour la production d'un nouveau texte.

Version du 22 février 2010 à 20 :17	Version du 23 février 2010 à 00 :32
Un des chefs de la nouvelle junte militaire, le colonel Djibrilla Hima Hamidou, a justifié le coup d'État de jeudi, affirmant que l'armée avait renversé le président Mamadou Tandja pour rétablir la stabilité, celui-ci ayant refusé de quitter le pouvoir à la fin de son mandat, qu'il avait par ailleurs allonger sans demander	L' un des chefs de la nouvelle junte militaire, le colonel Djibrilla Hima Hamidou, a justifié le coup d'État de jeudi, affirmant que l'armée avait renversé le président ~~Mamadou~~ Tandja pour rétablir la stabilité, celui-ci ayant refusé de quitter le pouvoir à la fin de son mandat, qu'il avait par ailleurs prolongé à la suite d'un référendum controversé.

Exemple 3: Deux versions d'un paragraphe d'un article de presse sur le coup d'état au Niger en 2010 (source : Wikinews Fr)

1.1.3 Résumé

Nous présentons dans cette section la forme de dérivation que l'on nomme couramment le *résumé*. Nous présentons les deux niveaux de résumé classiquement reconnus : le résumé indicatif et le résumé informatif. résumé

Le *résumé indicatif* décrit brièvement ce dont parle le texte et constitue un moyen d'appréhender rapidement son contenu. Il s'agit du type de résumé que l'on trouve couramment apposé aux dos des livres ou en en-tête des articles scientifiques. Ils apparaissent également au sein d'articles de magazines ou d'articles scientifiques afin de présenter un livre ou un article. Leur compilation digeste, notamment dans le domaine juridique, est un exemple des plus pertinents de leur utilité.

Le *résumé informatif* est une représentation abrégée du document « qui transpose l'information importante du texte original, et qui n'est pas plus long que la moitié du texte original et souvent bien moins long que cela » (Radev et collab., 2002). Afantenos et collab. (2005) entre autres proposent une typologie reposant sur leur mode de

5. `http://fr.wikinews.org/wiki/Niger_:_la_junte_promet_une_nouvelle_constitution_et_des_Ã©lections`. WikiNews est un des projets de la fondation Wikimedia (`wikimediafoundation.org`, à qui l'on doit également le projet Wikipedia (`fr.wikipedia.org`

construction. Ainsi, les auteurs font une distinction entre les *résumés abstraits* (*abstracts*) et les *résumés extraits* (*extracts*). Les premiers sont une présentation textuelle des concepts les plus saillant du texte source. Les seconds sont une organisation, et éventuellement une réécriture superficielle, de passages verbatim sélectionnés dans le texte source.

L'exemple 4 tiré d'une note méthologique sur la création de résumé [6] illustre cette compression textuelle sans perte informationnelle.

Passage complet

Pékin 2008, un pari risqué
Depuis maintenant vingt ans, la Chine est en pleine mutation. Le Parti communiste sous l'égide de Deng Xiaoping puis de Jiang Zemin a engagé le pays dans une réforme brutale vers l'économie de marché et l'intégration dans l'économie mondiale. Cette "révolution" a permis à une minorité de Chinois d'améliorer leur situation économique et de pouvoir décider plus librement de leur futur.
Cette transformation a eu d'autres conséquences qui font de la Chine un pays à la fois répressif et instable. Des dizaines de millions de Chinois ont été chassés de leur travail et des millions de paysans ont fui leurs campagnes appauvries vers les grandes villes, notamment Pékin. Depuis maintenant quelques années, les révoltes ouvrières et les jacqueries se sont multipliées. A Pékin, des centaines de milliers de chômeurs et de paysans tentent de survivre, exclus des fruits de la croissance économique. Les inégalités se creusent et les autorités sont bien incapables d'assurer un minimum de subsistance à ces dizaines de millions de défavorisés. Les experts chinois qui dénoncent les risques d'explosion sociale sont systématiquement censurés. Ainsi, l'économiste He Qinglian est interdite de publication depuis plus d'un an pour avoir écrit des articles sur les échecs de la politique économique gouvernementale.

Proposition de résumé

Bien que depuis 20 ans leur pays s'ouvre, seule une minorité de Chinois ont vu leur sort s'améliorer. La Chine constitue un pays instable et répressif : instable du fait de diverses révoltes paysannes et ouvrières, répressif pour éviter une explosion sociale que certains experts prédisent.

Exemple 4: Comparaison d'un texte complet et un résumé de ce dernier

Si les différents types de résumé ne traitent habituellement que d'un texte, les orientations récentes en résumé automatique portent sur le résumé multidocuments.

6. `http://aix1.uottawa.ca/~fgingras/cybermetho/modules/resume.html`
Le texte complet est tiré de l'appel « Au nom des droits de l'homme, non à la candidature de Pékin aux J.O en 2008 »par *Reporters sans frontières*, *Solidarité Chine* et le *Comité de soutien au peuple tibétain*.
Le résumé est proposé par l'auteur de la note : *François-Pierre Gingras*, professeur au Département de science politique de l'Université d'Ottawa

1.1.4 Plagiat et collusion

La recherche de plagiat est probablement le cas d'étude ayant trait à la dérivation qui est le plus commun (Clough, 2000, 2003b; Eissen et Stein, 2006; Hannabuss, 2001; Joy et Luck, 1999; Lyon et collab., 2004; Martin, 1994; Si et collab., 1997; Stein et Eissen, 2006; Hoad et Zobel, 2002). Nous ne pouvions donc pas nous permettre de faire l'impasse sur cette notion. Nous prétendons néanmoins qu'il est maladroit d'utiliser le terme de plagiat dans le contexte de la détection automatique. Nous présentons la notion de plagiat comme objet d'étude éthique et son opposition à la notion reprise dans les travaux liés à sa détection, puis nous présentons la notion de collusion.

PLAGIAT

Pour Vandendorpe (1992) « le *plagiat* est un terme à connotation morale et esthétique »qui n'a pas d'existence légale (on parle plutôt de contrefaçon ou d'infraction au code de la propriété intellectuelle). De nombreux chercheurs ont négligé cet aspect de la notion et ont utilisé le terme pour signifier essentiellement la copie plus ou moins exacte d'un texte. Ainsi Joy et Luck (1999) définissent le plagiat comme « la copie non reconnue de documents ou programmes ». Hannabuss (2001) précise qu'il s'agit de « l'utilisation non autorisée ou l'imitation proches des idées et du langage de quelqu'un d'autre et la représentation de ce travail comme le sien ». Il explicite ainsi le « non reconnue » de Joy et Luck (1999) comme la représentation du travail d'un autre ou du travail imitant un autre sans accorder les crédits à ce dernier. Il considère que les idées tout comme leur mise en forme (le langage) peuvent être plagiées. Dans ces deux définitions détachées de la tâche de détection, l'intention de l'auteur transparaît sans être explicitée. Lyon et collab. (2006), dans un contexte académique, confirment le caractère de non reconnaissance de l'auteur source en décrivant le plagiat comme un scénario où « les étudiants prélèvent du matériau à partir du Web et l'utilisent sans le référencer convenablement dans des essais ou des rapports qui sont censés être leur propre travail ». Le plagiat ne relève alors plus de l'intention de l'auteur mais d'une mauvaise maîtrise des techniques bibliographiques. Il s'agit d'un plagiat perçu mais pas intentionnel. Irving (2004) explicite ce caractère intentionnel en définissant le plagiat comme « la soumission d'une partie ou de la totalité du travail d'une autre personne comme si c'était la sienne, sans informer l'auteur, et avec l'intention de tromper » [7]. L'absence de marques citationnelles ne suffit pas à reconnaître le plagiat. De plus, il ne se traduit pas par une forme de transformation textuelle particulière. En revanche il se distingue par l'intention de l'auteur de plagier. C'est d'ailleurs certainement la seule caractéristique invariante de tout plagiat. Ainsi, il ne faut pas faire le racourci entre une reprise verbatim et du plagiat. De la même façon que l'on peut plagier en transformant la forme textuelle, il est possible d'utiliser licitement un passage verbatim, sous la forme d'une citation par exemple.

COLLUSION

La *collusion* se différencie du plagiat en ce que les deux parties se copient l'un et l'autre de manière consentie. Ainsi, Irving (2004) définit la collusion comme « la soumission d'un travail comme le sien alors que ce travail a été partiellement ou totalement fait par une autre personne, et que cette autre personne est partie prenante de la tromperie » [8]. Clough (2000) et Lyon et collab. (2006) s'intéressent à la détection de collusion et proposent une vision plus pragmatique correspondant encore une fois au contexte académique. Ainsi, pour Clough (2000), la collusion correspond aux scénarios où « les étudiants se copient les uns les autres ou travaillent ensemble » [9], ce qu'il résume à une « collaboration inacceptable » (Clough, 2003b). Lyon et collab. (2006)

7. « By plagiarism we mean the submission of part or all of another person's work as if it were ones own, without the knowledge of the author, and with intention to deceive. »

8. « Collusion, on the other hand is the submission of work as ones own when (at least some of) that work has been done partly or wholly by another person, and that other person is party to the deception. »

9. « students copy from each other or work together »

ont une vision très similaire, ils précisent notamment qu'il s'agit du partage de données entre étudiants alors que ces derniers sont censés travailler indépendamment. En résumé, la collusion est une forme de coopération frauduleuse pour le bénéfice mutuel des parties. Tout comme le plagiat, c'est une notion morale qui ne s'instancie pas dans les textes sous une forme particulière et reconnaissable. Ainsi, les différents travaux sur la détection de collusion (Clough, 2003b; Lyon et collab., 2006) distinguent la collusion du plagiat en termes de phénomène et de scénario, mais les mêmes techniques sont utilisées pour leur détection. Bull et collab. (2001) simplifient même la collusion « aux documents qui se recoupent et sont liés les uns avec les autres » [10].

Le plagiat comme la collusion sont des notions morales qui n'ont pas de contraintes ou de caractéristiques particulières de réalisation textuelle. En d'autres termes, seul le processus dérivatif permet de qualifier une dérivation de plagiat ou de collusion. Le texte en lui-même n'en porte aucune marque. C'est d'ailleurs une des raisons pour lesquelles nous n'avons aucun exemple pour illustrer ces notions. Toutefois, ces concepts restent incontournables et fortement présents dans la littérature.

1.1.5 Citation et référence

CITATIONS
RÉFÉRENCES

La dérivation de texte peut également être très ponctuelle, c-à-d l'expression empruntée à la source est précisément localisée et s'intègre localement dans le texte dérivé. Cette forme de dérivation se retrouve sous la forme de *citations* ou de *références bibliographiques*. La citation porte l'attention sur une insertion car elle introduit en contextualisant un passage d'un autre texte. La référence met en avant la mention d'une source identifiée.

Les citations font partie du champs d'étude du discours rapporté (Muñoz et collab., 2004). Elles ont également une place importante dans les travaux sur l'attribution du discours (Jackiewicz, 2006; Bethard et collab., 2004; Choi et collab., 2005). Pour Rosier (1999), le discours rapporté est la mise en rapport de deux discours dont l'un est mis à distance de l'espace énonciatif créé par l'autre. Le discours mis à distance est attribué à une autre source, de manière univoque ou non. Pour Rabatel (2001), les sources sont des énonciateurs différents, renvoyant à des locuteurs différents. La distance provoquée entre les discours par l'un des locuteurs correspond alors à une prise de distance de ce locuteur envers le discours de l'autre. L'exemple 5 illustre cette forme de dérivation : le discours de S. Royal est dérivé sous la forme d'une citation dans l'article du Figaro.

Les références scientifiques sont un tout autre acte langagier qui ne repose pas sur un passage contextualisé. Les éléments rapportés de la source sont beaucoup plus diffus et éventuellement ponctués par un code bibliographique. Ritchie et collab. (2006) se sont notamment intéressés aux références dans les textes scientifiques où les sources sont clairement identifiées. Le texte source est dans ce cas dérivé pour poser un cadre de réflexion permettant à l'auteur de se positionner.

Les citations et les références sont toujours le résultat d'un processus de dérivation car le retrait des textes sources entraînerait la disparition des citations et références des textes dérivés. La modification en conséquence du texte dérivé est la caractéristique fondamentale d'un processus de dérivation.

1.1.6 Transposition de genre

Toujours dans l'objectif de faciliter l'accès à l'information, si le résumé permet de condenser l'information, il est des processus dérivatifs qui permettent de transcrire

10. « The term collusion is used where documents overlap and inter-link with each other to varying extents, indicative of work copied from peers. »

Extrait du discours de S. Royal
(19/02/2007)

Je lance un appel à toutes celles et ceux qui veulent que la France fasse triompher la République du respect, parce que nous savons qu'il n'y a pas de liberté sans justice, qu'il n'y a pas d'efficacité économique sans progrès social.[...] J'appelle ce soir au rassemblement de toutes celles et ceux qui se reconnaissent dans les valeurs du pacte présidentiel, et qui pensent que l'on peut réformer la France sans la brutaliser, [...]

Extrait d'un article du Figaro
(20/02/2007)

Elle a appelé à "faire triompher la République du respect" et au "rassemblement de tous ceux qui se reconnaissent dans les valeurs du pacte présidentiel", et qui "pensent qu'on peut réformer la France sans la brutaliser".

Exemple 5: Exemple d'une dérivation de type citation

l'information d'un lecteur à un autre au travers d'une *transposition du genre*.

TRANSPOSITION DU GENRE

Les genres discursifs catégorisent les actes de langage selon les conventions et normes qui y sont mises en œuvre. Ainsi, à chaque genre est associé un certain agencement de la matière langagière utile à l'auteur et au lecteur (Veron, 1988). Ils fournissent au premier des modèles d'écriture, et au second des horizons d'attente (Todorov, 1987). Finalement, la notion de genre fournit à l'observateur extérieur des points de comparaison entre les textes. Le passage d'un genre à un autre peut être nécessaire lorsque le lectorat visé change. C'est ainsi le cas lorsqu'un journaliste reprend les résultats d'une étude scientifique, ou que cette étude scientifique est reprise dans un discours électoral. Le niveau de langue n'est pas le seul modifié puisque l'article de presse qui découle de la dérivation ne suit plus les règles du genre de l'article scientifique. Ainsi, l'exemple 6 illustre la différence de ton entre une dépêche et un billet de blog reprenant la même information.

Dépêche de presse : AFP

WHISTLER (Canada) (AFP) - Biathlon/Mass-start : Martin Fourcade décroche l'argent, 8^{e}médaille pour la France
Mal parti avec deux pénalités dès son premier passage sur le pas de tir, le Français Martin Fourcade a décroché dimanche à 21 ans une improbable médaille d'argent dans la mass-start des jeux Olympiques 2010 grâce à un superbe final.
Il apporte à la France sa 8^{e}médaille. Le biathlon en a fourni 5 à lui seul.

Blog : buzznews.fr

Martin Fourcade a remporté la médaille d'argent dans l'épreuve de la Mastart en biathlon aux JO de Vancouver. Avec trois pénalités, Fourcade a réussi un superbe parcours sur les skis. Avec un sans-fautes au tir, il remportait la médaille d'or ! :))
Le relais français va être terrible !

Exemple 6: Textes co-dérivés chacun dans un genre différent (dépêche de presse et billet de blog)

1.1.7 Traduction

La traduction permet de transcrire un texte dans une langue source en un nouveau texte écrit dans une langue cible, avec pour objectif que les deux textes signifient la même chose dans les deux langues et que le texte dans la langue cible soit compréhensible pour des personnes ne connaissant pas la langue source. Nous présentons

TRADUCTION

les différentes considérations de ce qu'est une traduction et nous positionnons par rapport à celles-ci.

La traduction ne se limite cependant pas forcément à la transposition du lexique, de la syntaxe et de la sémantique d'un texte dans une langue cible. La pragmatique du texte, c-à-d l'effet du texte sur ses lecteurs, doit également être transposée, ce qui nécessite une adaptation culturelle du message. Certaines études accordent également une dimension sociale à l'acte de traduction, parlant alors de « traductologie » (Ivekovic, 2009). Deguy (2009) présente les différentes facettes de cette généralisation de la traduction, de la transcription d'une séquence verbale de la langue source à la langue cible à la notion plus générale de faire-passer une chose d'un état à l'autre « selon la logique complexe de l'exprimer ».

Chacune de ces visions de la traduction repose sur un processus de dérivation, la connaissance et l'interprétation d'un état des choses initial préemptant une nouvelle expression de cet état des choses. Ces considérations sont cependant bien plus larges que la dimension de la dérivation que nous souhaitons traiter dans cette thèse. Nous nous limiterons donc à la signification généralement acceptée de la transposition d'une expression verbale d'une langue à une autre tel que l'illustre l'exemple 7. Ce dernier présente en parallèle deux articles de presse dérivés d'une même dépêche, mais écrits dans des langues différentes.

Synthèse

Nous avons présenté dans cette section les différents objets d'étude de la littérature autour de la détection de plagiat, de copies, de documents similaires... Nous avons cherché à montrer en quoi ces différents objets d'étude correspondaient à des formes de dérivation, c-à-d des archétypes de textes résultant d'un processus de dérivation. Nous avons ainsi voulu justifier notre choix de tisser des ponts entre tous ces travaux au travers du concept de dérivation que nous définissons formellement à la section suivante.

Le regroupement de tous ces phénomènes autour d'un même concept a pour objectif de partager les approches et les techniques mises au point sur les différentes formes de dérivation. La définition d'un contexte global permet également de clarifier chaque notion en la positionnant par rapport aux autres. Le cadre théorique général permet lui de discuter, comparer et traiter des problèmes qui sont supposés distincts dans la littérature. Cette mise en commun avait déjà été esquissée auparavant. Ainsi, Hoad et Zobel (2002) tentent de détecter des versions et du plagiat à l'aide des mêmes approches. Les auteurs ont ainsi montré que les textes dérivés que sont les versions ou le plagiat ne portent pas dans leur expression textuelle de spécificités permettant de les différencier. Cela nous conforte dans l'idée que les formes de dérivation se définissent en grande partie par le processus de dérivation, et non par le texte dérivé résultat.

1.2 Proposition d'un cadre théorique

Nous proposons dans cette section une définition de la notion de dérivation de texte. Cette notion permet de désigner sous une même appellation les différents phénomènes discutés dans la section précédente de production d'un texte à partir d'autres. Nous tentons de formaliser autant que possible la dérivation afin d'offrir un cadre théorique rigoureux dans lequel nous évoluerons par la suite. Nous définissons tout d'abord le principe de dérivation que nous rattachons ensuite à la notion de codérivation introduite par Bernstein et Zobel (2004). Enfin, nous discutons certaines propriétés de ces relations.

TEXTE ANGLAIS : *The Australian – 18/02/2010*

A NORTH Carolina man who spent nearly 17 years in prison on a first-degree murder conviction has been released with his name cleared.
The courtroom erupted in cheers, while Greg Taylor and his family members wept and hugged each other as the court was adjourned.
After the ruling was read, Wake County District lawyer Colon Willoughby walked over to Taylor, shook his hand and apologised for his conviction.
Taylor, 47, was the first person in the state's history to be exonerated through the North Carolina Innocence Inquiry Commission, the only state-run agency in the United States probing post-conviction claims of innocence.
A three-judge state Superior Court panel reached its unanimous decision to free Taylor after a week-long review of his case prompted by an investigation by the commission, established by state lawmakers in August 2006.
"It is ordered that the release sought by Gregory F. Taylor, the convicted person, is granted, and the charge of first-degree murder against Gregory F. Taylor is dismissed," presiding judge Howard Manning said after reading the ruling of each of the three judges on the review panel.
Taylor was convicted in April 1993 for the murder of Jacquetta Thomas on September 26, 1991.
His lawyers argued that no physical evidence linked Taylor and the victim. They also called into question the credibility of witnesses who had testified against him.
According to the statute that established the innocence commission, a person can have his conviction overturned during the review process only if the three-judge panel reaches a unanimous decision that « clear and convincing evidence » proves the person is innocent.
Taylor had exhausted all avenues of appeals when the commission reviewed his case and decided in September it warranted special review under the statute.
« We have been blessed in the state of North Carolina to make more progress as it relates to the system of justice. This is one of those fantastic days, » Joseph Cheshire V, one of Taylor's lawyers, told reporters after the ruling.
AFP

TEXTE FRANÇAIS : *Libération – 18/02/2010*

Un Américain de Caroline du Nord (sud-est) qui a passé près de 17 ans en prison pour meurtre a été libéré et innocenté mercredi, grâce à une commission exceptionnelle qui enquête sur les cas litigieux dans cet Etat depuis 2006.
Des explosions de joie se sont fait entendre dans la salle du tribunal lorsque Greg Taylor et ses proches se sont serrés dans les bras après la décision de la cour. Le procureur du comté de Wake, Colon Willoughby, a alors traversé la salle d'audience pour serrer la main de Greg Taylor et s'est excusé pour sa condamnation injuste.
Greg Taylor, 47 ans, est le premier homme à être innocenté par la Commission d'enquête sur l'innocence de Caroline du Nord, seul organe géré par les autorités aux Etats-Unis qui enquête sur les proclamations d'innocence après une condamnation.
Un panel de trois juges est parvenu à la décision unanime de libérer Greg Taylor après avoir examiné l'affaire pendant une semaine, suite à une enquête de la Commission créée par les parlementaires de l'Etat en août 2006.
«Ordre est donné que la libération réclamée par Gregory F. Taylor soit accordée et que l'accusation de meurtre à l'encontre de Gregory F. Taylor soit annulée», a dit le juge Howard Manning qui présidait l'audience.
Greg Taylor avait été condamné en avril 1993 pour le meurtre de Jacquetta Thomas le 26 septembre 1991. Ses avocats ont fait valoir qu'aucun élément matériel ne le liait à la victime et ont mis en doute les témoignages qui l'accusaient.
Selon le statut de la Commission, un condamné ne peut être innocenté que si le panel de trois juges parvient à une décision à l'unanimité.
(Source AFP)

EXEMPLE 7: Articles de presse anglais et français dérivés de la même source (AFP)

1.2.1 Dérivation de texte

Nous appelons « dérivation [11] de texte » le processus de production d'un nouveau texte en utilisant un ou plusieurs textes préexistant comme base d'écriture. Nous réunissons sous cette notion différentes formes de productions (duplication, version, résumé, plagiat, collusion, citation, référence, transposition de genre et traduction).

TEXTE SOURCE
TEXTE DÉRIVÉ

Nous considérons qu'un texte dérive d'un autre lorsque la préexistence de l'un, que nous appelons *texte source*, est une condition nécessaire à l'aboutissement de l'écriture de l'autre, que nous appelons *texte dérivé* (*cf. Définition 1*).

Le texte source influence (dans ses idées, dans son expression, dans sa structure...) l'auteur du texte dérivé, de telle façon qu'il a eu un impact dans ses choix d'écriture. Il en découle que si cet auteur est privé de la connaissance d'un texte source alors il produira un texte différent de celui qu'il aurait produit en sa connaissance.

Définition 1 *Soient T_d et T_s des textes, si T_d est dérivé de T_s, le retrait du texte source T_s lors du processus d'écriture du texte dérivé T_d résulte en un texte $T_{d'}$ différent de T_d.*
T_s est appelé le texte source et T_d est appelé le texte dérivé.

1.2.2 Relations de dérivation et de codérivation

RELATION DE DÉRIVATION

Lorsqu'un texte dérive d'un texte source tel que défini dans la définition 1, nous considérons qu'il y a une *relation de dérivation* entre ces deux textes (*cf. Définition 2*). Cette relation binaire entre deux textes est irréflexive et n'est pas symétrique. Elle est irréflexive car un texte ne peut dériver de lui-même, cela reviendrait à préexister à sa propre existence. De plus, elle n'est pas symétrique car si T_s est un texte source de T_d alors T_d ne peut en aucun cas être également texte source de T_s, ce qui violerait ici encore le principe de préexistence de l'un par rapport à l'autre.

Définition 2 *Soit T_d un texte dérivé d'un texte source T_s, alors (T_s, T_d) appartient à la relation de dérivation $\mathbf{R_D}$ notée $T_s \mathbf{R_D} T_d$ (ou $\mathbf{R_D}(T_s, T_d)$).*

La notion de codérivation a été originalement introduite par Bernstein et Zobel (2004) pour qui « deux textes qui dérivent l'un de l'autre ou bien dont certaines portions des deux dérivent d'un même texte tiers », sont des codérivés [12]. Leur utilisation originelle de la notion de dérivation est à rapprocher de celle de réutilisation de texte. Nous reprenons leur définition de co-dérivation en y injectant notre vision de la dérivation. Nous parlons alors de *relation de codérivation* entre deux textes lorsque soit les deux textes sont en relation de dérivation l'un avec l'autre, soit ils sont tous deux en relation de dérivation avec un même texte source tiers :

RELATION DE CODÉRIVATION

Définition 3 *On note $\mathbf{R_C}$ la relation de codérivation et soient T_i et T_j deux textes, alors :*

$$\forall T_i \, \mathbf{R_C} \, T_j \Leftrightarrow \begin{cases} \mathbf{R_D}(T_i, T_j) \vee \mathbf{R_D}(T_j, T_i) \\ \textit{ou} \\ \exists T_s \textit{ tel que } \mathbf{R_D}(T_s, T_i) \wedge \mathbf{R_D}(T_s, T_j) \end{cases}$$

Nous pouvons directement déduire de la définition précédente le lemme 1, à savoir que toute relation de dérivation entre deux textes implique que ces deux textes sont également liés par une relation de codérivation. Ces différents éléments sont illustrés

11. La notion de dérivation ne correspond pas ici à la notion de dérivation morphologique laquelle produit de nouvelles unités lexicales à partir de la racine d'un mot.

12. « for two documents to be co-derived, some portion of one must be derived from the other or some portion of both must be derived from a third document. »

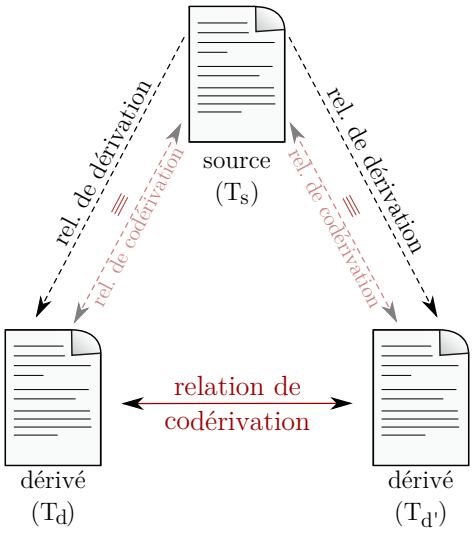


FIGURE 1.1 – Liens de dérivation entre des codérivés

dans la figure 1.1. Cette dernière expose une configuration où deux textes (T_d et $T_{d'}$) sont en relation de dérivation avec le même texte source (T_s). Chacun des textes est alors en relation de codérivation avec les deux autres : les deux textes dérivés sont en relation de codérivation avec leur source car la relation de dérivation est subsumée par la relation de codérivation, et ils sont en relation de codérivation entre eux car ils dérivent d'un ancêtre commun (T_s) :

Propriété 1

$$\forall T_i \, \mathbf{R_D} \, T_j \Rightarrow \left\{ \begin{array}{l} \mathbf{R_C}(T_i, T_j) \\ \mathbf{R_C}(T_j, T_i) \end{array} \right.$$

La relation de codérivation est symétrique puisqu'elle n'identifie ni un texte source ni un texte dérivé (*cf. Propriété 2*). En ce sens, la relation de codérivation est plus souple à manipuler que la relation de dérivation.

Propriété 2 *La relation de codérivation ($\mathbf{R_C}$) est une relation symétrique.*

Synthèse

Nous avons défini dans cette section la notion de dérivation qui repose sur la nécessaire préexistence du texte source à l'écriture du texte dérivé. Nous avons également introduit les notions de relation de dérivation et de co-dérivation qui permettent de tisser des liens entre les différents textes (sources et dérivés) impliqués dans une dérivation.

Nous avons cherché autant que possible à conserver les propositions existantes (notamment Bernstein et Zobel (2004)) ou du moins à rester cohérents avec les travaux antérieurs les plus directement liés à la notion de dérivation (Seo et Croft, 2008; Clough et Gaizauskas, 2008; Bendersky et Croft, 2009).

L'introduction des relations de dérivation et de co-dérivation permet de découper le processus de dérivation menant à la production du texte dérivé en autant de liens que nécessaires avec les textes sources. Ce découpage est une première étape nécessaire à la mise en place de méthodes de détection automatique. Nous pensons toutefois

que les liens qui unissent les textes sources et les textes dérivés dans le processus de dérivation ne sont pas tous comparables. Nous explorons dans les sections suivantes les traits qu'ils ont en commun et sur lesquels ils diffèrent. Cette exploration nous mènera à la proposition d'un modèle multidimensionnel en section 1.4.

1.3 Différentes classifications des formes de dérivation

Les différentes formes de dérivation présentées dans la section 1.1 font l'objet d'étude de plusieurs travaux mais aucune ne les a appréhendées dans leur ensemble. Nous présentons les différentes tentatives d'organisation, souvent partielle ou du moins parcellaire, de ces formes de dérivation. Nous considérons notamment trois grandes familles de taxinomies : celles consacrées au plagiat, celles décrivant les relations textuelles entre les documents et finalement celles qui ont émergé du domaine de la détection de réutilisation de texte.

1.3.1 Autour du plagiat

L'étude du plagiat d'un point de vue éthique plus qu'informatique a permis de dresser des catégorisations du phénomène détachées de toute considération opérationnelle. Nous rapportons notamment la taxinomie de Martin (1994) qui est la pierre angulaire de plusieurs travaux à visée opérationnelle, Kleppe et collab. (2005) notamment.

Martin (1994), cité par Clough (2003b) notamment, propose une synthèse des différentes formes de plagiats relevées dans la littérature. Premièrement le *plagiat mot-à-mot* qui consiste à recopier des passages de textes à partir d'un texte publié sans utiliser de guillemets ou sans citer la source. Le *plagiat de paraphrase* en est une extension, la différence étant que certains mots des passages recopiés sont modifiés. Deuxièmement, le *plagiat de sources secondaires* est une forme de plagiat plus subtile. Il correspond à la citation de sources inconnues de l'auteur mais obtenues d'une source secondaire elle-même non citée. Troisièmement, la réutilisation de la structuration des arguments d'une source, sans employer les mêmes mots, ce que l'auteur appelle *plagiat de la forme de la source*. De manière analogue, le *plagiat d'idée* est la réutilisation d'une idée sans employer les mêmes mots, ni la même structure. Finalement, il nomme *plagiat de paternité* le fait de marquer son nom sur le travail d'un autre. Ces différentes formes de plagiat sont détachées de toute considération opérationnelle ce qui leur permet d'embrasser une large vision des formes de plagiat. Nous y distinguons ainsi des dérivations de forme (plagiat de paternité, plagiat mot-à-mot, plagiat de paraphrase ainsi que plagiat de sources secondaires) où l'on va globalement retrouver les mots employés dans la source ; et des dérivations plus structurelles où ce qui est repris de la source ne réside pas dans sa dimension textuelle, mais dans sa dimension discursive (plagiat de la forme de la source) ou bien dans sa dimension sémantique voire pragmatique (plagiat d'idée).

La taxinomie du plagiat proposée par Kleppe et collab. (2005) (*cf. Figure 1.2*) reprend globalement les formes de plagiat proposées par Martin (1994) et que nous avons classées comme dérivations de formes ou de fond. Ils explicitent les cas où seule une partie du document est plagiée. Il est intéressant de noter qu'ils introduisent également la traduction comme une forme potentielle de plagiat. Si cette taxinomie est globalement cohérente avec les catégories proposées par Martin (1994), elle manque à notre avis de cohésion. Le placement au même niveau du plagiat d'idée, de paragraphe, de phrase/segment ou de synonymie est maladroit dans le sens où l'idée relève de la

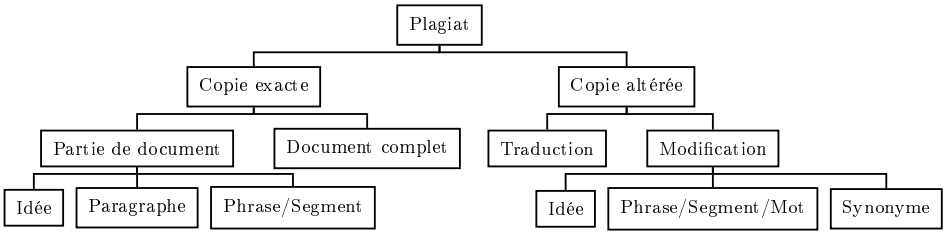


FIGURE 1.2 – Taxinomie des formes de plagiat proposée par Kleppe et collab. (2005). Eissen et Stein (2006) proposent une classification similaire que nous reprenons dans la figure 1.5

dimension sémantique ou pragmatique du texte tandis que les autres appartiennent à la dimension lexicale ou syntaxique.

Nous retenons de ces deux contributions majeures sur la catégorisation des formes de plagiat, l'ancrage de la forme de plagiat dans la dimension textuelle, structurelle ou sémantique du texte source. Nous notons également les différents niveaux de modification proposés : du simple copier-coller à la retranscription d'idée en passant par le paraphrasage et la synonymie ou encore la traduction.

1.3.2 Autour de relations textuelles

D'autres propositions de classifications et de catégories ont été faites, non plus pour caractériser le plagiat, mais les relations entre les textes (*textual relationships*). Nous en relevons au moins deux importantes, celle de Shivakumar et Garcia-Molina (1995) et celle de Heintze (1996).

Ainsi, Shivakumar et Garcia-Molina (1995) considèrent quatre tests de relation entre un document candidat et un document cible (*Document Target Test*). Le test de *plagiat* (*plagiarized*), déconnecté de la notion morale de plagiat telle que présentée précédemment, est satisfait si un document contient des parties d'un autre. Le test de *sous-ensemble* (*subset*) est satisfait si un document est complètement contenu dans un autre à une marge de recouvrement près. Le test de copie (*copies*) est satisfait si un document apparaît être l'exacte copie d'un autre. Les tests de plagiat, de sous-ensemble et de copie correspondent chacun à des niveaux de recouvrement de texte définis. Ces différents tests ne sont pas exclusifs, comme le précisent les auteurs, une paire de documents peut ainsi valider plusieurs de ces tests. La validation du test de copie notamment semble impliquer la validation de tous les autres tests. Finalement, dans le cas où ces tests échouent, les textes semblent relever du test de *liaison* (*related*).

Heintze (1996) propose une classification des relations textuelles en six catégories : documents identiques, documents résultant de modifications ou de corrections mineures d'autres documents, réorganisations d'autres documents, révisions, versions condensées ou étendues, documents intégrant des portions (plusieurs centaines de mots) d'autres documents. L'auteur s'est concentré sur les types de relations les plus importantes, il ne considère donc pas sa proposition comme exhaustive. Nous relevons ici que ces catégories suivent une progression du recouvrement de textes telle que décrite par Shivakumar et Garcia-Molina (1995).

Ces deux classifications proposées sont ancrées sur la forme textuelle des documents. Les dimensions structurelles et sémantiques que l'on retrouvait dans les taxinomies des formes de plagiat ont disparu. De plus, les nuances des différentes catégories sont liées à la proportion de texte commun entre les documents impliqués : le recouvrement de texte. Cette proportion de texte peut être calculée sur la base des

deux documents ou bien d'un seul. Cette dernière approche permet d'identifier les cas où un document constitue une petite portion d'un autre.

1.3.3 Autour de la réutilisation de texte

Nous présentons les trois propositions principales autour de la classification des réutilisations de texte : celle à l'échelle du document de Piao et McEnery (2003) et Clough (2003a), celle de Seo et Croft (2008) qui différencie l'important de la dérivation du point de vue de la source et du dérivé, et celle de Metzler et collab. (2005) qui intègre une dimension sémantique en plus de la seule dimension textuelle. Globalement, elles reposent toutes les trois sur une catégorisation en termes de recouvrement et se distinguent par la portée du recouvrement en termes de matériel (dimension textuelle ou sémantique) ou de niveau d'application (texte dérivé uniquement ou couple source/dérivé). De plus elles font toutes l'hypothèse de l'implication d'un seul texte source. Les catégories de ces taxinomies sont exclusives et construites sur une similarité de contenu (Metzler et collab., 2005), la granularité des passages dérivés (Seo et Croft, 2008) ou un mélange des deux (Piao et McEnery, 2003; Clough, 2003a).

Piao et McEnery (2003) et Clough (2003a) proposent une classification des relations de dérivation spécifique à la réutilisation de texte et qui se positionne à l'échelle du document. Cette classification repose sur une discrétisation du recouvrement de texte entre source et dérivé sur trois niveau et du point de vue du dérivé : totalement dérivé, partiellement dérivé et non-dérivé. La catégorie *complètement dérivé* (*wholly-derived*) correspond au scénario où le texte constituant le dérivé provient intégralement d'un unique document source. La catégorie *partiellement dérivé* (*partially-derived*) est un assouplissement de la catégorie précédente puisque le texte du document dérivé provient en parti seulement du document source. Dans ce cas, plusieurs autres sources peuvent être impliquées. Finalement la catégorie *non-dérivé* (*non-derived*) concerne les cas où il n'y a pas de réutilisation de texte. Cette proposition permet de classer sans trop de difficultés les relations de dérivation et a donc un réel intérêt opérationnel. Le corpus METER (Gaizauskas et collab., 2001) a notamment été annoté à partir de cette classification. Cependant, elle ne rend pas compte des subtilités entre les différentes formes de dérivation : la catégorie « partiellement dérivé » englobe les résumés, les traduction et le plagiat alors que la catégorie « complètement dérivé » se résume aux seules duplications.

Seo et Croft (2008) proposent une classification des dérivations également à l'échelle du document mais fondée non plus sur le contenu partagé mais sur la quantité d'éléments partagés entre le texte dérivé et sa source. Ils proposent ainsi de considérer séparément la proportion d'éléments tirés de la source et la proportion d'éléments qu'ils représentent dans le texte dérivé. Ils définissent ainsi trois proportions : une majorité (*most*), une partie importante (*considerable*) et une petite partie (*partial*). Les catégories qu'ils proposent sont donc constituées d'une proportion pour la source associée à une proportion pour le texte dérivé, soient six catégories :

- *Majorité-Majorité* Une majorité des éléments de la source sont repris et représentent une majorité des éléments du dérivé. Cette configuration correspond à la conception classique des duplications où les deux textes sont presque identiques.
- *Majorité-Important* Une majorité des éléments de la source représente une grande partie du texte dérivé. D'après les auteurs, cette configuration est typique de l'ajout d'un court passage à un autre texte, phénomène observé notamment dans les blogs.
- *Majorité-Partie* Une majorité des éléments de la source représente une petite partie du texte dérivé. Dans cette configuration, le texte source entier constitue une partie du texte dérivé : un article de presse constitué de plusieurs dépêches

par exemple.
- *Important-Important* Une grande partie du texte source constitue une grande partie du dérivé.
- *Important-Partie* Une grande partie du texte source constitue une petite partie du texte dérivé.
- *Partie-Partie* Une petite partie du texte source constitue une petite partie du dérivé. Il s'agit de la configuration des citations ou références.

Les auteurs font le parallèle entre les catégories *Majorité-Majorité*, *Majorité-Important* et *Important-Important* et les dérivations de type duplication largement traités dans la littérature. En contrepartie, ils proposent le concept de *réutilisations locales de texte* pour les autres catégories. Leur classification se différencie ainsi des précédentes en ce qu'ils délaissent la similarité de contenu afin de se concentrer sur la distribution des portions dérivées entre les différents documents. Ils confirment que la dérivation d'un texte doit être étudiée en tenant compte du texte source associé. Nous notons toutefois que leur classification ne s'applique pas directement aux cas de dérivation impliquant plusieurs sources.

Dans la lignée des propositions de classification selon un niveau de recouvrement de texte, Metzler et collab. (2005) proposent de considérer la similarité entre des documents comme une graduation linéaire continue. La similarité porte sur le recouvrement des textes sources et dérivé à la fois selon leur dimension textuelle, comme les taxinomies précédentes, mais également sémantique. Les auteurs proposent à des fins de classification d'identifier certains points de ce spectre. Ils proposent ainsi les catégories *sans lien* (*unrelated*), même thématique générale (*on the general topic*), même thématique spécifique (*on the specific topic*), mêmes faits (*same facts*) et copies (*copied*). La graduation linéaire varie d'une similarité thématique large d'un côté, que l'on peut partiellement rapprocher d'une similarité d'idée telle qu'introduite par Martin (1994), aux documents identiques. Par document identique, les auteurs entendent identiques jusque dans leur forme textuelle, soit la forme la plus naïve correspondant à la catégorie complètement dérivé de Clough (2003a). Les auteurs proposent plusieurs exemples de relations entre documents pour lesquels cette classification s'appliquerait : les documents qui en résument ou paraphrasent d'autres, les documents codérivés (selon la définition 3 de la section 1.2) ou encore les documents qui partagent la même structure ou les mêmes faits (choix des arguments). Cette proposition de taxinomie est bien plus précise que celle de Clough (2003a), tout en se focalisant exclusivement sur le contenu des documents. Ils délaissent notamment la différence entre un document dérivant de plusieurs sources ou d'une seule, concept qui avait été rapidement introduit dans la proposition de Clough (2003a). Ce gain dans la précision des catégories est contrebalancé par une mise en œuvre plus complexe. Ainsi, la différence entre une même thématique générale et une même thématique spécifique reste très subtile.

En résumé, les taxinomies proposées dans le cadre de la détection de réutilisation de texte s'inscrivent uniquement dans des configurations où un texte ne dérive que d'une seule source.

Synthèse

La taxinomie de Martin (1994), même si elle a été conçue pour le plagiat, offre une vision haut niveau des différentes formes de dérivation, détachée des considérations opérationnelles. Elle a l'avantage de s'affranchir de la dimension textuelle des documents, contrairement à Shivakumar et Garcia-Molina (1995) et Heintze (1996). Kleppe et collab. (2005) ont proposé une structuration hiérarchique qui reprend partiellement la taxinomie de Martin (1994). Malheureusement cette structuration pré-

sente quelques failles : elle place notamment la notion de dérivation d'idée au même niveau que les dérivations purement textuelles et elle évince la dérivation de structure.

Les taxinomies proposées dans le cadre de la détection de réutilisation de texte ont enrichi les nuances portant sur le contenu des textes (Metzler et collab., 2005). Elles ont également introduit la notion de granularité des éléments mis en relation entre la source et le texte dérivé (Seo et Croft, 2008), permettant de détacher la notion de recouvrement, fondamentale pour Shivakumar et Garcia-Molina (1995) et Heintze (1996), de la dimension textuelle des documents.

Toutes ces taxinomies ne se positionnent que dans le contexte d'une dérivation n'impliquant qu'une seule source. Si cette restriction est pertinente dans un certain nombre de cas, elle est trop limitative pour la notion de dérivation dans son intégralité.

Chacune des taxinomies présentées modélise une facette particulière de la dérivation. Il nous semble nécessaire de pouvoir caractériser les formes de dérivation sur plusieurs dimensions afin d'y positionner les différentes formes décrites dans la section 1.1 et offrir une vision globale de la dérivation. Nous proposons une telle vision dans la section suivante.

1.4 Proposition d'une classification multidimensionnelle de la dérivation

Nous avons présenté dans la section précédente différentes taxinomies qui décrivent chacune une portion des éléments caractérisables de la dérivation. Nous composons dans cette section une vision multidimensionnelle de la dérivation qui repose en partie sur ces taxinomies.

Notre proposition de classification s'articule autour de trois principes. Premièrement, le processus de la dérivation n'est perçu qu'à travers le texte dérivé. Notre classification se positionne donc du point de vue du texte dérivé et repose sur les éléments communs entre le texte dérivé et ses différents textes sources.

Deuxièmement, nous envisageons l'implication de plusieurs textes sources dans le processus de dérivation décrit. En conséquence, nous choisissons d'appréhender le processus de dérivation comme une collection de relations de dérivation entre un texte source et un texte dérivé. Plus précisément, nous considérons qu'il est envisageable d'avoir une relation de dérivation caractérisable par collection homogène d'éléments d'un texte source reprise dans le texte dérivé. En pratique, cela revient à potentiellement une relation de dérivation par nature des éléments dérivés (*cf. Section 1.4.2*) et par source.

Enfin, troisièmement, notre classification se veut multidimensionnelle sans dépendance hiérarchique. Il ne s'agit donc pas d'une taxinomie. Chaque dimension du modèle correspond à un axe de la classification *a priori* indépendant des autres. Nous en proposons huit : l'arité, la nature, la granularité côté source et la granularité côté dérivé, la paternité, l'intention, la similarité et l'intégration. Pour chacune de ces dimensions nous proposons un ensemble de valeurs permettant de caractériser la dérivation. La combinaison de toutes les valeurs de chaque dimension couvre, selon nous, l'ensemble des variations caractérisables d'un processus de dérivation.

La figure 1.3 illustre le type de classification que peut fournir notre modèle. L'icône au centre représente le texte dérivé tandis que les textes sources sont placés en périphérie. Chaque source est reliée au texte dérivé au travers d'une relation de dérivation. Les dimensions que nous proposons caractérisent soit le texte dérivé dans son intégralité (arité), soit une relation de dérivation dans laquelle il est impliqué en tant que dérivé et qui porte sur certains éléments d'une source. Ainsi, l'ensemble du processus

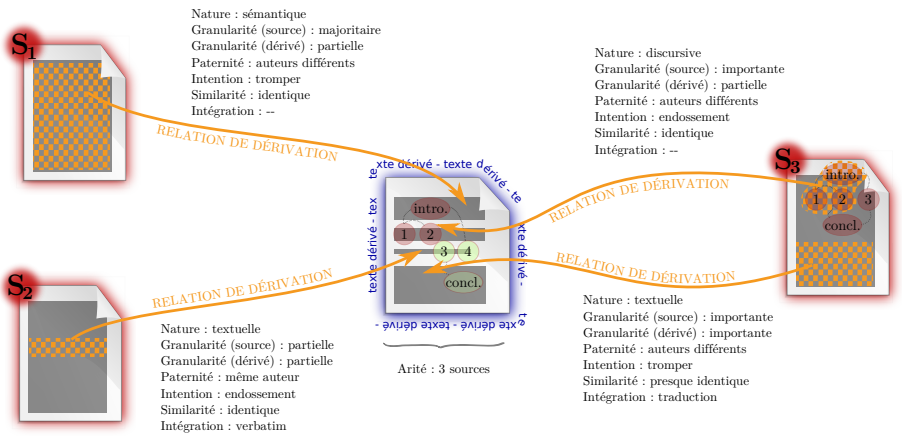


FIGURE 1.3 – Caractérisation multidimensionnelle d'une dérivation

de dérivation est décrit en caractérisant chacune de ces relations de dérivation.

Nous décrivons par la suite en détail chacune des dimensions proposées : l'arité, la nature des éléments dérivés, la granularité du passage dérivé pour la source et pour le dérivé, la paternité des textes, l'intention de l'auteur du dérivé, la similarité entre les éléments dérivés et l'intégration. Nous détaillons en parallèle la figure 1.3. Enfin, nous appliquons notre modèle à la citation qui fait aussi l'objet du chapitre 3.

1.4.1 L'*arité* : nombre de sources impliquées dans la dérivation

La première dimension que nous présentons est la seule qui ne caractérise pas une relation de dérivation mais seulement le texte dérivé. L'*arité* décrit le nombre de textes source qui sont en relation de dérivation avec le texte dérivé en focus. ARITÉ

Les différentes taxinomies proposées dans la littérature se confinent aux configurations n'impliquant qu'un seul texte source. Ces configurations sont trop réductrices dans le cas de la dérivation puisqu'elles ne permettent pas de prendre en compte les dérivations tels que les résumés ou encore les articles journalistiques croisant plusieurs sources. D'une manière générale, si l'on exclut les duplications et les traductions qui sont par essence presque toujours monosource, la plupart des dérivations impliquent plus qu'un seul texte source. L'*arité* palie ce manque. ARITÉ

1.4.2 La *nature* des éléments dérivés depuis la source

Comme l'avait introduit Martin (1994), puis l'avait repris Kleppe et collab. (2005), il n'y a pas que le niveau textuel qui puisse être dérivé. Lors du processus de dérivation, l'auteur va sélectionner des éléments dans le texte source. Il peut s'agir de séquence textuelle, mais également d'une tournure de phrase, d'une figure de style ou même d'une idée. La *nature* de ces éléments sélectionnés est selon nous une dimension caractéristique de la dérivation. Nous présentons tout d'abord les différentes natures que nous considérons dans notre modèle, puis nous les illustrons au travers de la figure 1.3. Enfin, nous discutons des natures acceptables pour différentes formes de dérivation. NATURE

Nous associons quatre valeurs à la nature : textuelle, sémantique, discursive et stylistique. La *nature textuelle* correspond au prélèvement de séquences textuelles. La réutilisation de texte est certainement la dérivation la plus facilement identifiable. En effet, le processus de dérivation ne nous est accessible qu'au travers du texte dérivé.

Ainsi, la réutilisation d'une séquence textuelle de la source est presque directement accessible dans le texte dérivé. Pour autant c'est une dérivation également complexe à appréhender car le texte emprunté, s'il est suffisamment long, embarque également des éléments de contenu et de style, relevant de deux autres types de nature. La *nature sémantique* correspond au prélèvement d'idées, d'éléments de contenu. Les éléments de cette nature peuvent s'appuyer sur diverses réalités textuelles et peuvent être difficiles à délimiter dans le texte source. La *nature discursive* correspond au prélèvement d'éléments liés à l'organisation des idées. Elle correspond à la copie de structure introduite par Martin (1994). Le découpage en section, en paragraphe ou même en phrases relèvent potentiellement de la nature discursive, tout comme la mise en opposition ou la coordination d'idées, à l'aide de connecteurs discursifs par exemple. Finalement, la *nature stylistique* correspond au prélèvement d'éléments de style. L'idée d'une dérivation du style est plus simple à comprendre pour les arts plastiques, la peinture notamment. Nous pensons que ceci est également possible dans le cas des œuvres littéraires. Pour Ducrot et Schaeffer (1995) le style relève des choix de l'auteur parmi « les disponibilités contenues dans la langue ». Les registres de langue (courant, soutenu, familier...) sont une de ces disponibilités, tout comme les idiosyncrasies. Dans la même veine, Özlem Uzuner et Davis (2003) différencient le style de l'expression : « le style fait référence aux éléments linguistiques qui, indépendamment du contenu, persistent dans les œuvres d'un auteur », « l'expression implique des éléments linguistiques qui se rapportent à la façon dont l'auteur exprime un contenu particulier » [13].

La figure 1.3 expose des dérivations de plusieurs natures. Ainsi, la relation de dérivation entre S_1 et le texte dérivé est de nature sémantique et celle impliquant S_2 est de nature textuelle. S_3 est impliqué dans deux relations de dérivation : une de nature textuelle et une autre de nature discursive. Pour cette dernière, plutôt qu'une représentation abstraite des éléments par une zone grisée, nous avons préféré représenter la structure du texte par un arbre dont les branches correspondraient à des sections. Dans les autres cas, la zone grisée représente indifféremment des éléments textuels ou sémantiques.

Les duplications empruntent principalement des éléments de nature textuelle. La forme textuelle de ces dérivés sont en effet très proches de la forme de leurs sources. Les approches de détection de dérivation travaillant sur les duplications (Manber, 1994; Conrad, 2003; Yang et Callan, 2005; Manku et collab., 2007; Potthast et Stein, 2007; Kołcz et Chowdhury, 2008) tirent parti de cette caractéristique en calculant le taux de recouvrement de textes entre documents. Les citations, les versions et les traductions empruntent également des éléments de cette nature, mais pas seulement. Ainsi pour la traduction, il peut être nécessaire de considérer le contenu pour mieux restituer le message, ainsi que des éléments de style tels que la construction des phrases. Les résumés par abstraction sont certainement la forme de dérivation qui réutilise le moins les expressions et dérive plutôt du contenu et de la structure.

1.4.3 La *granularité* des éléments dérivés du texte source

GRANULARITÉ

La notion de *granularité* a été introduite par Seo et Croft (2008). Ils proposent de décrire une dérivation à partir de la proportion d'éléments empruntés au texte source par rapport à la proportion qu'ils représentent dans le texte dérivé. Nous reprenons cette notion afin de caractériser une relation de dérivation. Tout comme Seo et Croft, nous différencions la granularité côté source de celle côté dérivé en les considérant dans

13. « Style refers to the linguistic elements that, independently of content, persist over the works of an author and has been widely studied in author- ship attribution. Expression involves the linguistic elements that relate to how an author phrases par- ticular content and can be used to identify potential copyright infringement or plagiarism. »(Uzuner et collab., 2005)

deux dimensions différentes. La dimension équivalente côté dérivé est présentée dans la section suivante (*cf. Section 1.4.4*). Par contre, nous considérons que cette granularité ne s'applique pas uniquement à des dérivations de nature textuelle mais également à celles de nature sémantique ou discursive. Elle est plus difficile à concevoir pour les dérivations de nature stylistique. Nous traitons dans cette section de la granularité du passage dérivé par rapport au texte source. Nous présentons la dimension, puis les valeurs que nous lui affectons et enfin nous donnons quelques exemples.

La granularité des éléments dérivés du texte source correspond à la proportion que représentent ces éléments par rapport à l'intégralité des éléments de même nature au sein du texte source. Par exemple, pour une dérivation de nature textuelle, supposons que l'on dérive un paragraphe d'un texte source qui en contient deux (de tailles équivalentes). La proportion est de 50 % : les éléments dérivés du texte source représentent 50 % de tous les éléments de même nature (en l'occurrence le texte) du texte source. Dans la figure 1.3, nous avons symbolisé les éléments du texte source par un fond gris et les éléments dérivés par un motif en damier orange. Ainsi, tous les éléments de S_1 sont dérivés tandis que seul un sous-ensemble des éléments de S_2 le sont. Nous avons également impliqué ce code visuel pour la dérivation de nature discursive impliquant S_3 : seule la partie contenant *intro*, *1* et *2* de la structure de S_3 est dérivé, raison pour laquelle seule cette partie est recouverte par le damier.

La granularité du passage dérivé du texte source évolue le long d'un intervalle avec à une extrémité aucun élément repris et de l'autre la réutilisation de l'intégralité des éléments (pour une nature donnée) du document source. Nous proposons toutefois, à l'instar de Seo et Croft (2008), de jalonner cet intervalle par trois proportions :

- *majoritaire* (*most*) lorsque la proportion d'éléments partagés représente pratiquement la totalité des éléments de la source ;
- *importante* (*considerable*) lorsque la proportion d'éléments partagés représente une grande partie des éléments de la source ;
- *partielle* (*partial*) pour les autres proportions.

Seo et Croft (2008) associent à chacune de ces valeurs un seuil indicatif (respectivement 80 %, 50 % et 10 %) pour les dérivations de nature textuelle. Le pourcentage correspond alors au nombre de mots repris par rapport aux mots composants le texte source. Ces seuils sont moins intuitifs lorsque l'on traite de dérivation d'autres natures. Nous utiliserons donc les jalons intuitifs (majoritaire, importante et partielle) plutôt que ces valeurs arbitraires.

Les duplications reprennent par définition une majorité des éléments de leur source. Par conséquent, pour les relations de dérivation de type duplication, la granularité des éléments repris de la source est majoritaire. À l'opposé, dans le cas des citations seul une petite partie de la source est habituellement dérivée. La granularité de la source est partielle. Comme l'illustre la figure 1.3 (page 31), et plus particulièrement la première relation de dérivation impliquant S_3, la granularité a également du sens pour les dérivation de nature discursive. Dans le cas présent, une partie importante de la structure source est dérivée, la granularité est importante.

1.4.4 La *granularité* des éléments dérivés du texte dérivé

La dimension de granularité des éléments dérivés du texte dérivé est très similaire à celle présentée dans la section précédente si ce n'est qu'elle s'applique du point de vue du texte dérivé et non du texte source. Nous présentons la dimension et quelques exemples. Les valeurs de la dimension granularité pour le dérivé sont les mêmes que celles pour la source présentées dans la section précédente (*cf. Section 1.4.3*). Nous ne les représentons pas dans cette section.

La granularité des éléments dérivés du texte dérivé correspond, pour une relation

de dérivation donnée, à la proportion que représentent ces éléments par rapport à l'intégralité des éléments de même nature au sein du texte dérivé. Par exemple, pour une dérivation de nature textuelle, supposons qu'une relation de dérivation considérée résulte en un paragraphe dans le texte dérivé qui en contient deux (de tailles équivalentes). La proportion est de 50 % : les éléments dérivés de la relation de dérivation en focus représentent 50 % de tous les éléments de même nature (en l'occurrence le texte) du texte dérivé. Dans la figure 1.3, les éléments résultant d'une relation de dérivation sont pointés par la flêche matérialisant la relation de dérivation (zones grises ou bulles rouges).

En ce qui concerne les duplications, les textes dérivés sont habituellement constitués exclusivement des éléments dérivés du texte source. La granularité des éléments dérivés dans le texte dérivé est alors majoritaire. Il en est de même pour les résumés monodocument qui sont constitués uniquement d'éléments (de nature textuelle ou sémantique) du texte qu'ils résument. Le cas des citations est différent. Celles-ci ne représentent habituellement qu'une toute petite partie du texte dérivé. La granularité est partielle.

1.4.5 La *paternité* des textes

PATERNITÉ

La *paternité* des textes explicite le lien entre les auteurs du texte source et ceux du texte dérivé pour les parties de ces textes impliquées dans la relation de dérivation en focus.

Nous ne proposons que deux valeurs pour cette dimension : *même auteur* et *auteur différent*. Nous aurions pu rajouter une troisième valeur : *un des auteurs originaux*. En effet, il se peut que le texte source ait été écrit par plusieurs auteurs et que le texte dérivé ne soit l'œuvre que d'un seul. Dans ce cas il ne s'agit ni tout à fait du même auteur, ni tout à fait d'un auteur différent. Nous avons considéré que la distinction complexifiait inutilement la classification. Elle posait alors la question de l'introduction d'autres valeurs : deux des auteurs originaux, la moitié des auteurs originaux, un des auteurs originaux et un autre qui n'en est pas... Nous considérons que lorsque les auteurs sont identiques entre la source et le dérivé nous emploierons *même auteur* et nous emploierons *auteur différent* dans tous les autres cas.

Les versions de documents sont par définition écrites par les mêmes auteurs. Leur paternité est donc de valeur *même auteur*. À l'opposé, le plagiat est une forme de dérivation où l'on cherche à s'attribuer la paternité d'une œuvre qui n'est pas sienne. Leur paternité serait donc de valeur *auteur différent*. La figure 1.3 illustre des dérivations de différentes paternités. Ainsi, pour les sources S_1 et S_3 les auteurs sont différents de ceux du texte dérivé, tandis que la source S_2 est du même auteur. Si l'on faisait abstraction de la dimension de paternité pour S_2 nous pourrions penser à du plagiat puisque la dérivation ressemble à du copier-coller. Toutefois la dimension paternité nous indique qu'il s'agit plus probablement d'une dérivation de type révision ou bien, étant donné la présence d'autres sources, de ce que l'on pourrait qualifier de « recyclage d'écrits antérieurs », pratique assez courante dans le milieu scientifique notamment.

1.4.6 L'*intention* de l'auteur du texte dérivé

Dans la section 1.1.4, nous avons présenté le plagiat et la collusion, deux notions morales. Nous avons conclu que seul le processus dérivatif permettait de qualifier ces formes, le texte lui-même n'en portant aucun indice. En d'autres termes, le plagiat ne se distingue pas des autres formes de dérivation discutées dans la section 1.1 autrement que par l'intention de plagier. Nous avons donc choisi d'introduire la dimension

d'*intention* qui permet de caractériser la raison du choix d'une dérivation pour l'auteur. Si cette dimension nous semble nécessaire pour caractériser complètement une dérivation, elle n'est pas forcément accessible pour autant. Seul l'auteur du texte dérivé connaît la véritable valeur de cette dimension. INTENTION

L'intention de l'auteur de la dérivation définit sa relation par rapport au texte source et à son auteur. L'auteur du texte dérivé peut avoir l'*intention de tromper* en dissimulant l'existence ou la paternité du texte source. Cette configuration est spécifique au plagiat et à la collusion. À l'opposé, il peut avoir l'*intention de reconnaître* la source d'une information. Il peut le manifester en contextualisant ses reprises au sein du texte dérivé mais pas forcément : par maladresse ou du fait d'une mauvaise maîtrise de la technique de citation (Martin, 1994). Nous proposons deux valeurs qui correspondent à chacune de ces configurations : tromper pour la première et reconnaître pour la seconde.

L'intention de tromper s'accorde avec toutes les dérivations qui peuvent être qualifiées de plagiat, voir notamment Vandendorpe (1992); Martin (1994); Hannabuss (2001); Maurer et collab. (2006); Duggan (2006). L'intention de reconnaître caractérise les autres.

1.4.7 La *similarité* entre les éléments source et dérivé

La *similarité* entre les éléments prélevés dans le texte source et leurs équivalents dans le texte dérivé sont une autre dimension qui caractérise les relations de dérivation. Le calcul de la similarité estime la proximité entre des éléments distincts, que cette similarité soit textuelle (Friburger et Maurel, 2002; Stein et Eissen, 2006; Bao et collab., 2007), sémantique (Pirró, 2009), structurelle (Jeh, 2002; Pereira et Ziviani, 2003), cognitive (Tversky et collab., 1977; Hahn et collab., 2003; Lee et collab., 2005), ou bien un mélange de ces niveaux d'analyse (Martins, 2004; Metzler et collab., 2005). Nous avons choisi d'ajouter la dimension de similarité pour porter la notion d'éloignement ou de proximité des éléments dérivés par rapport aux éléments issus de la source. SIMILARITÉ

Metzler et collab. (2005) proposent un spectre de similarité du contenu discrétisé par trois jalons : *sujet général*, *sujet spécifique* et *faits identiques*. Nous souhaitons élargir la perception de la similarité à toutes les dérivations et pas seulement celles qui portent sur le contenu. Si nous pensons également que la similarité est continue, nous considérons tout autant qu'il est préférable de la discrétiser. Ainsi, la similarité prend la valeur *identique* lorsque les éléments du texte dérivé correspondent exactement à ceux correspondant dans le texte source, *presque-identique* lorsqu'ils admettent quelques modifications et *différent* dans les autres cas. Nous avons fait le choix de ne pas proposer de valeur de similarité nulle. En effet, étant donné qu'il y a une relation de dérivation, nous pensons qu'il y a une similarité minimale entre les éléments de la source et ceux du dérivé.

La figure 1.3 (page 31) illustre des relations de dérivations dont la dimension similarité prend les valeurs identique ou presque-identique. Nous pouvons rencontrer des similarités plus faibles (valeur différent) dans les articles journalistiques qui reprennent des éléments d'une conférence de presse et les repositionnent dans un contexte plus global, ou encore dans les travaux de synthèse où de la même façon les éléments dérivés sont repositionnés dans un contexte plus global ou bien sont adaptés pour mieux s'intégrer au développement de l'argumentation ou de la narration.

1.4.8 L'*intégration* des séquences textuelles

Les dérivations de nature textuelle, notamment lorsqu'elles sont de granularité partielle, reviennent à injecter des séquences textuelles dans un texte en construction (le texte dérivé). L'auteur de la dérivation doit adapter ces séquences textuelles afin de s'assurer que le texte produit conserve ses propriétés de cohésion et de cohérence (Ducrot et Schaeffer, 1995). Pour le maintien de la cohérence, l'auteur doit manuellement tisser les liens entre les concepts introduits tirés du texte source et ceux qui ne le sont pas, ou bien tisser de nouveaux liens permettant de se passer de ceux ayant disparu dans le processus. De la même façon pour la cohésion, il doit s'assurer de rétablir les relations mutuelles entre syntagmes ou entre phrases, par le maintien ou la correction des liens anaphoriques par exemple. Il doit également introduire de nouvelles relations pour cimenter les différents emprunts. Nous caractérisons ce travail d'adaptation des séquences textuelles du texte source pour les injecter dans le texte dérivé par la dimension d'*intégration*.

INTÉGRATION

Les valeurs qui peuvent être associées à l'intégration d'une relation de dérivation esquissent les opérations nécessaires à l'auteur pour intégrer les éléments prélevés dans le texte source : verbatim, syntaxique, paraphrase, contraction/dilatation, adaptation de genre et traduction. La valeur verbatim correspond à une intégration directe de l'élément du texte source, sans aucune modification. La valeur syntaxique implique une intégration nécessitant des corrections syntaxiques telles que la mise en œuvre de la concordance des temps, des accords... La valeur paraphrase correspond à des modifications locales minimes mais qui ne rentrent pas dans les deux catégories précédentes. La valeur contraction/dilatation correspond à une réduction ou une augmentation de la matière composant les éléments tirés du texte source. La valeur adaptation de genre implique des modifications nécessaires à une transposition du genre (p. ex. de scientifique à vulgarisé) ou du registre (p. ex. de formel à familier). Et enfin, la valeur traduction correspond à un changement de langue entre le texte source et le texte dérivé.

La citation est certainement l'exemple le plus accessible en ce qui concerne l'intégration. Les journalistes utilisent les citations pour, par exemple, gagner en objectivité. Ils tentent de reprendre tels quels les écrits ou les propos afin de ne pas y intégrer de biais. Cependant, ils leur arrivent de devoir modifier quelque peu la forme de ces citations afin, par exemple, d'intégrer des propos portant sur des éléments temporels ou spatiaux et où le référentiel de l'énonciateur est différent de celui du lecteur. Ils leur arrivent également de devoir expliciter le contexte ou détailler des termes (dilatation). La plupart du temps ils utilisent l'acronyme *ndlr* ou *ndla*[14] pour indiquer que ces extensions sont de leur fait. Enfin, ils peuvent également être amenés à modifier le registre de langue afin de le faire correspondre aux attentes de leur lectorat ou de leur rédaction.

1.4.9 Un exemple de projection : la citation

L'objectif de notre modèle est de pouvoir décrire des instances particulières de dérivation. Dans cette section, nous l'utilisons pour caractériser non pas une instance particulière, mais une forme de dérivation : la citation (*cf. Section 1.1.5*). La citation est un cas particulier de dérivation où l'intégration — détaillée par la suite — intègre des informations de contexte qui ne relèvent pas directement du texte mais de la situation d'énonciation. Dans ce cadre, nous n'utilisons pas notre modèle de classification pour décrire précisément une citation mais pour préciser l'espace multidimensionnel (sous-ensemble des valeurs pour chaque dimension) occupé par les citations.

14. Respectivement « note de la rédaction »et « note de l'auteur ».

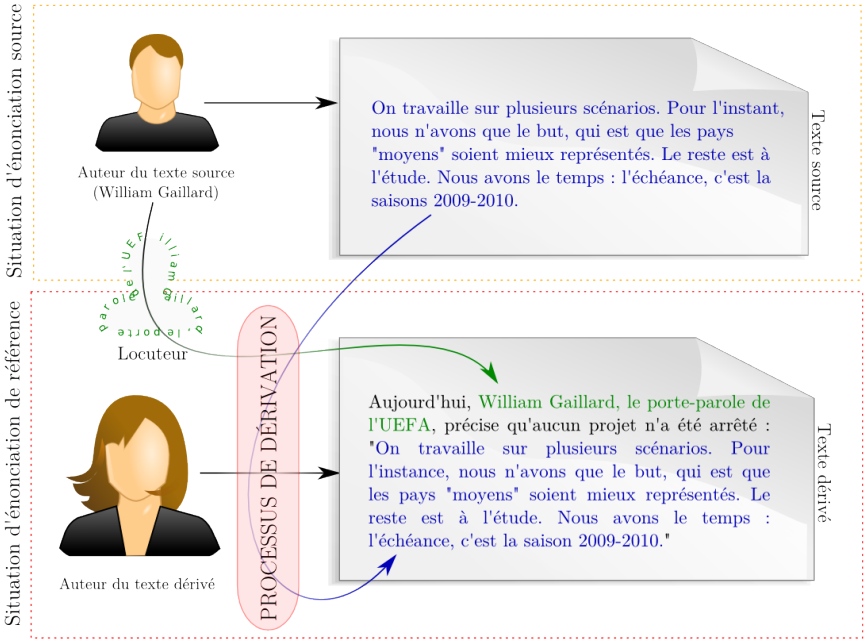


FIGURE 1.4 – Particularités du processus dérivationnel de la citation

La figure 1.4 illustre le processus de citation. Dans un premier temps, un texte est produit dans une situation d'énonciation source. L'auteur (William Gaillard) est l'énonciateur (ou auteur) source, et le texte produit est le texte source. Dans un second temps, un nouvel auteur va rapporter ce texte source au sein d'une nouvelle situation d'énonciation que nous appelerons situation d'énonciation de référence. Dans notre exemple, l'auteur source est introduit dans le texte dérivé comme locuteur par l'expression « William Gaillard, le porte-parole de l'UEFA », tandis que le texte source est retranscrit tel quel dans le texte dérivé.

Projection du modèle

Premièrement, la nature des éléments prélevés dans le texte source est la forme textuelle exclusivement. Toutefois, l'auteur complète ces éléments textuels prélevés dans le texte source par des informations repositionnant ceux-ci dans le contexte de la situation d'énonciation source : les éléments de contextualisation. C'est le cas dans notre exemple (*cf. Figure 1.4*) où sont précisés le temps de l'énonciation (*Aujourd'hui*) ainsi que le nom et le titre de l'auteur source (*William Gaillard, le porte-parole de l'UEFA*).

Deuxièmement, nous pouvons vraisemblablement dire que l'intention de l'auteur d'une citation n'est pas de plagier mais de citer dans le but de légitimer l'argument qu'il rapporte. Le processus citationnel permet à l'auteur d'être transparent par rapport à l'origine des propos rapportés.

Troisièmement, les éléments prélevés pour constituer une citation sont des passages du texte source et représentent également des passages dans le texte dérivé. La granularité de cette forme de dérivation est résolument partielle. La citation d'un document complet est peu probable (ou bien il s'agira d'une référence).

Quatrièmement, la dimension intégration dépend du style de discours utilisé par l'auteur pour rapporter la citation. Les travaux sur le discours rapporté font princi-

palement état de quatre styles de discours : le discours direct, le discours indirect, le discours indirect avec îlots textuels et le discours indirect quasi-textuel. Le style le plus commun est le *discours direct* (DD) : l'auteur opère une transcription littérale du passage sélectionné dans le texte source et marque cette retranscription dans le texte dérivé à l'aide d'éléments typographiques. Ce dernier s'intègre alors directement dans le texte dérivé sans modification ou bien sous couvert de transformations syntaxiques ou de contractions mineures. L'exemple 8 montre un texte source et sa citation dans laquelle des transformations de compression ont été opérées afin de rendre le texte source au discours direct. Le *discours indirect* (DI) est le pendant du discours direct. Contrairement au premier, les éléments dérivés ne sont pas marqués et les transformations qui y sont appliquées peuvent être plus importantes, comme l'illustre l'exemple 9. Cependant, le discours indirect est assez peu utilisé dans les articles de presse où il est avantageusement remplacé par une forme hybride nommée *DI quasi-textuel* qui mêle du DI, profitant ainsi de sa souplesse d'intégration, et du DD pour rapporter des segments *verbatim* marqués typographiquement (*cf. exemple 10*).

Discours direct

Discours indirect

DI quasi-textuel

Discours de Ségolène Royal	Reprise du Figaro
Et je tends la main à toutes celle et ceux qui pensent comme moi qu'il est non seulement possible mais urgent de quitter un système qui ne marche plus.	« Je tends la main à tous ceux qui pensent, comme moi, possible et urgent de quitter un système qui ne marche plus », a-t-elle ajouté

Exemple 8: Exemple de discours rapporté au style direct

Les gouvernements britannique et allemand ont, quant à eux, déjà déclaré que tout contrat militaire avec EADS serait succeptible d'être revu en cas de fermetures de sites.

©*Journal* Le Monde

Exemple 9: Exemple de discours rapporté au style indirect

Discours de Ségolène Royal	Reprise du Figaro au discours indirect quasi-textuel
J'appelle ce soir au rassemblement de toutes celles et ceux qui se reconnaissent dans les valeurs du pacte présidentiel, et qui pensent que l'on peut réformer la France sans la brutaliser, qui veulent faire triompher toujours les valeurs humaines sur les valeurs boursières, qui veulent mettre fin aux insécurités et aux précarités qui se sont douloureusement creusées au cours de ces dernières années, qui veulent faire reculer toutes les formes de violence grâce a un ordre juste et à de nouvelles sécurités durables.	Elle a appelé à «faire triompher la République du respect» et au «rassemblement de tous ceux qui se reconnaissent dans les valeurs du pacte présidentiel», et qui « pensent qu'on peut réformer la France sans la brutaliser ».

Exemple 10: Exemple de discours rapporté au style indirect quasi-textuel

Finalement, les dimensions paternité et similarité peuvent prendre pratiquement n'importe quelle valeur. La citation n'a pas réellement de spécificités pour ces traits ci.

Synthèse

Nous avons dans cette section proposé notre vision multidimensionnelle de la dérivation. Celle-ci tente de caractériser complètement ce qui constitue le processus dérivationnel menant au texte dérivé, en nous positionnant du point de vue du texte dérivé. Ainsi, contrairement à ce que l'on peut trouver dans la littérature (*cf. Section 1.3*), nous avons pris en compte l'existence de multiples sources en caractérisant séparément les différentes relations de dérivation qui les lient au texte dérivé.

Notre proposition se compose de sept dimensions : l'arité, la nature, la granularité (du texte source ou dérivé), la paternité, l'intention, la granularité, la similarité et l'intégration. Pour chacune de ces dimensions, nous avons défini et présenté les différentes valeurs que l'on pouvait y associer. Elles sont synthétiquement reprises dans le tableau 1.1.

Ce modèle multidimensionnel a pour but premier de caractériser une instance particulière d'un processus de dérivation. Toutefois, comme nous l'avons illustré dans la section 1.4.9, il permet également de cadrer les valeurs que peuvent prendre les différentes formes de dérivation qui ont été traitées dans la littérature (*cf. Section 1.1*). Il définit pour chacune des dimensions les valeurs que pourraient prendre des instances de la forme en focus. Cela revient à définir une sorte d'espace vectoriel dans lequel seraient cloisonnées chacune des formes de dérivation.

1.5 Conclusion

Nous avons discuté dans ce chapitre des différents travaux de la littérature ayant trait à ce que nous avons nommé la dérivation de texte, ce terme étant lui-même tiré des travaux de Bernstein et Zobel (2004).

Nous avons pu noter que ces différents travaux portaient sur des objets d'étude partageant un même socle commun : la préexistance nécessaire de textes sources à leur propre production. Ce socle commun est également la fondation de notre proposition de définition de ce qu'est la dérivation de texte et du cadrage théorique qui en découle.

Nous avons également présenté l'ensemble des formes de documents qui émergeaient de la littérature. Nous avons proposé, à partir des classifications existantes, un ensemble de dimensions et leurs valeurs associées permettant de caractériser, et donc différencier, ces archétypes.

Notre effort d'unification des travaux antérieurs nous a permis de poser clairement ce qui constituait notre sujet d'étude : les dérivations de textes. Ce cadre théorique unifié reste cohérent avec les travaux existant tout en nous permettant de définir le processus de dérivation et de proposer notre classification multidimensionnelle. Eissen et Stein (2006) ont proposé une taxonomie partielle des méthodes de dérivation décrites dans la littérature qui reposent largement sur les propositions de taxinomies antérieures (*cf. Section 1.3*). Nous l'avons retranscrite dans la figure 1.5 en l'ajustant [15] pour faire le parallèle avec notre classification multidimensionnelle (*cf. Section 1.5*). Les traits proposés par Eissen et Stein correspondent aux dimensions de similarité, d'intégration et de granularité. Nous avons précisé pour chaque valeur de trait des auteurs, les valeurs correspondantes dans notre modèle. Nous présentons les méthodes qui apparaissent, en gris, à l'extrême droite de la figure dans le chapitre suivant.

15. Il est intéressant de noter que ledit schéma est une forme de dérivation graphique.

Dimension	Description	Valeurs
Arité	le nombre de textes sources	nombre entier
Nature	la nature des éléments prélevés du texte source	textuelle sémantique discursive stylistique
Granularité (source)	la proportion d'éléments de la source partagés avec le dérivé par rapport à la totalité des éléments contenus dans la source	majoritaire importante partielle
Granularité (dérivé)	la proportion d'éléments du dérivé partagés avec la source par rapport à la totalité des éléments contenus dans le dérivé	majoritaire importante partielle
Paternité	liens entre les auteurs de la source et du dérivé	même auteur auteurs différents
Intention	raisons du choix de la dérivation pour l'auteur	tromper reconnaître
Similarité	notion d'éloignement ou de proximité des éléments dérivés par rapport aux éléments issus de la source	identique presque-identique différent
Intégration	esquisse les opérations nécessaires à l'auteur pour intégrer les éléments prélevés dans le texte source	verbatim syntaxique paraphrase contraction/dilatation adaptation de genre traduction

TABLE 1.1 – Synthèse des dimensions de notre proposition et de leurs possibles valeurs

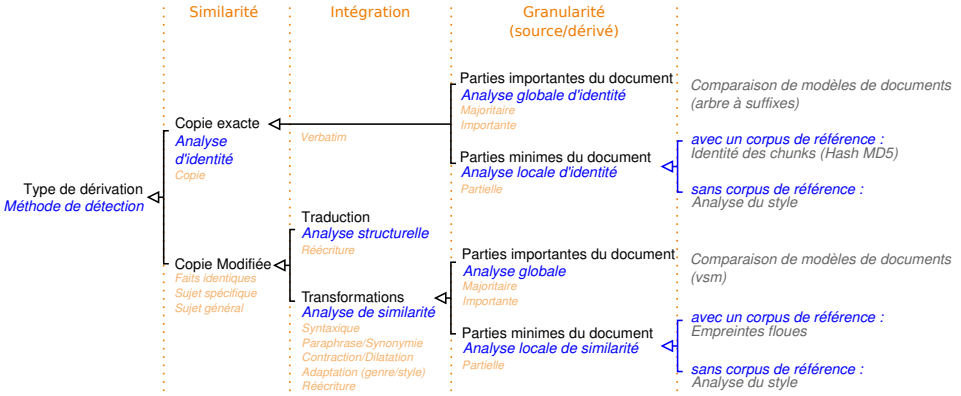


FIGURE 1.5 – Taxinomie des types de dérivation (en noir) et des méthodes de détection associées (en bleu) proposée par Eissen et Stein (2006, Fig. 1). Projection du schéma original dans notre proposition de vision multi-dimensionnelle (*cf. Tableau 1.1, page 40*) en terme de dimension et de valeur (en orange).

Chapitre 2

La détection de dérivation de texte

Rien ne se perd, rien ne se crée, tout se transforme

— Antoine Lavoisier

Sommaire

Nous avons introduit dans le chapitre précédent le concept de dérivation et présenté différents phénomènes (duplication, version, résumé, plagiat et collusion, citation et référence, transposition de genre, traduction) qui ont été autant d'objets d'études de travaux antérieurs comme des formes particulières de dérivation. Nous avons notamment proposé un cadre théorique permettant de définir ce qu'est une dérivation et un modèle de classification permettant de caractériser les dérivations du point de vue du texte dérivé résultat. Nous présentons dans ce chapitre les différentes méthodes de détection automatique des formes de dérivation. Nous cherchons autant que possible à les repositionner par rapport à notre cadre théorique et à notre taxinomie.

La détection de dérivation telle que traitée dans la littérature gravite autour de quatre grandes tâches applicatives : le test de dérivation, la recherche de dérivés (ou de sources) à partir d'un texte requête, le regroupement de co-dérivés et l'alignement de textes ou de passages dérivés.

TEST DE DÉRIVATION

Le *test de dérivation* consiste à savoir si un texte donné est un texte dérivé ou non. Le texte en focus est considéré isolément. Le système décide s'il s'agit d'un texte dérivé ou non et identifie éventuellement les passages incriminés.

RECHERCHE DE DÉRIVÉS

La *recherche de dérivés (ou de source)* à partir d'un texte requête consiste à identifier dans une collection de textes ceux qui dérivent totalement ou partiellement du texte requête si celui-ci est une source ou bien les textes sources s'il s'agit d'un dérivé. Le système prend en entrée un texte et retourne une collection de textes (ordonnés par pertinence ou non) qui répondent à la requête de l'utilisateur, ce qui n'est pas sans rappeler les systèmes de Recherche d'Information (RI). Par exemple, Bendersky et Croft (2009) s'intéressent à la détection de réutilisation de texte dans le contexte d'une recherche Web. Leur objectif est d'identifier, sur le Web, la source originale d'une information en tirant parti des dates de publications et des hyperliens entre documents.

REGROUPEMENT DE CO-DÉRIVÉS

Le *regroupement de co-dérivés* — ou plus précisément le regroupement de textes selon leurs liens de dérivation — consiste à rapprocher des textes d'une collection selon la proportion d'éléments dérivés qu'ils partagent. Contrairement à la tâche précédente, aucun texte n'est pointé comme cible. Plus le lien de dérivation entre deux documents est fort, plus les documents sont proches. Inversement, plus le lien de dérivation entre deux documents est faible, plus les documents sont éloignés. Un tel système prend en entrée une collection de documents et retourne un regroupement de ces documents. Ainsi, Yang (2006b) regroupe des lettres de citoyens destinées à l'administration fédérale américaine qui dérivent souvent de lettres-types écrites par des organismes tiers. Ce regroupement permet à l'administration de traiter les lettres par lot tout en identifiant les variations qui peuvent exister entre la lettre type source et les lettres dérivées par les citoyens.

MISE EN CORRESPONDANCE DE DÉRIVÉS

L'*alignement de textes ou de passages dérivés* consiste à lier des textes ou des passages selon leur implication dans une dérivation. Il est préférable d'identifier une dérivation à l'échelle du document avant de tenter un alignement. La critique génétique textuelle (Bourdaillet, 2007) réalise ce genre d'opérations sur des textes très proches (versions d'un même texte par le même auteur). Un système dédié à cette tâche prend en entrée une collection de documents dont un est considéré comme central et retourne une mise en correspondance entre des passages de ce document et des passages des documents du reste de la collection. Ainsi, Hose (2003) s'est intéressé aux liens qui pouvaient exister entre les manuscrits de la mer Morte et la prose biblique. Il a mis en œuvre des techniques de détection de dérivation pour rapprocher des phrases tirées de ces manuscrits de la Bible hébraïque.

Nous situons ce travail de thèse dans le cadre de la tâche de détection de dérivés à partir d'un texte source. Si nous considérons toutes les granularités de dérivation

(*cf. Section 1.4.3*), nous travaillons résolument à l'échelle du document. Les méthodes de recherche de dérivés (ou de sources) se répartissent en deux groupes selon le matériel en présence : la détection intrinsèque et la détection extrinsèque. La détection intrinsèque correspond à la tâche du test de dérivation. La détection extrinsèque est la tâche d'identification des textes sources à partir d'un texte dérivé, ou l'inverse. Les méthodes de détection de dérivation de la littérature se répartissent dans ces deux groupes : la détection intrinsèque ou extrinsèque. Soit elles utilisent uniquement le texte suspect et recherchent en son sein des marques de dérivation (*cf. Section 2.1*), soit elles tentent de mettre en correspondance des passages dudit texte avec d'autres (*cf. Section 2.2*).

2.1 Techniques de détection intrinsèque

DÉTECTION INTRINSÈQUE

Les méthodes de *détection intrinsèque* reposent sur l'utilisation exclusive du texte en focus. L'approche est similaire à celle, certainement inconsciente, des enseignants qui doutent de l'originalité d'un travail d'étudiant car celui-ci est d'un niveau différent de son niveau habituel : emploi d'un vocabulaire riche, de tournures de phrases élégantes... Le style de l'énonciation est en rupture avec celui communément rencontré.

Nous décrivons tout d'abord le principe général des méthodes de détection intrinsèque (*cf. Section 2.1.1*). Nous détaillons ensuite les deux types de marques qui sont exploitées par ces méthodes : les marques explicites qui font références à des éléments de contexte de la source (*cf. Section 2.1.2*) et les marques inhérentes aux éléments dérivés (*cf. Section 2.1.3*). Ces deux types de marques ne sont pas exclusives et peuvent être communément présentes dans des textes dérivés.

2.1.1 Approche générale

Le principe de fonctionnement d'une méthode intrinsèque repose sur la recherche dans le texte en focus de marques de la mise en œuvre d'un processus de dérivation. L'objectif d'une méthode de détection intrinsèque est de décider si le texte est dérivé et éventuellement d'identifier les passages qui le sont.

Certaines marques sont introduites volontairement par l'auteur, ce sont les *éléments de contextualisation de la dérivation*. Les dérivations dans cette configuration sont à rapprocher de la notion de citation (Muñoz et collab., 2004). L'auteur marque le passage emprunté à la source par des éléments typographiques (guillemets, deux points...), de mise en forme (bloc en retrait...) ou une contruction lexico-syntaxique particulière (« il dit que », « a-t-il dit »...). Il peut également compléter par des informations caractérisant la source : nom des auteurs cités, éléments spatio-temporels... L'identification de ces passages dérivés passe par la recherche de ces marques (*cf. Section 2.1.2*). Leur présence nous renseigne sur l'*intention* (*cf. Section 1.4.6*) qui caractérise la relation de dérivation. Étant donné que ces marques sont du fait de l'auteur, il reconnaît le processus de dérivation devant son lecteur ou du moins ne le dissimule pas. La valeur d'intention caractérisant cette relation de dérivation serait plutôt la reconnaissance que la tromperie.

D'autres marques sont inhérentes aux éléments dérivés. Lors de la dérivation, l'auteur doit intégrer des éléments d'un texte source dans un nouveau texte. L'intégration de ces éléments implique une nécessaire adaptation du texte qui les recueille afin de garantir à la fois sa cohérence (« configuration des concepts et des liens entre concepts »(Ducrot et Schaeffer, 1995, p.604)) et sa cohésion (« relations mutuelles entre syntagmes intraphrastiques ou entre phrases »(Ducrot et Schaeffer, 1995, p.603)). Des marques de dérivation se nichent en partie dans les inconsistences

provoquées par le maintien de ces propriétés et se manifestent dans les choix stylistiques [1], c-à-d le choix des mots et des constructions syntaxiques qui donnent une touche particulière au texte (Glover et Hirst, 1996) (*cf. Section 2.1.3*).

2.1.2 Exploitation des marques de contextualisation

L'auteur peut introduire volontairement des indices dans le but de contextualiser la dérivation. Il marque le passage emprunté et le repositionne dans son contexte d'énonciation original (auteur, lieu, date...). Les dérivations opérées dans cette configuration rappellent clairement les notions de citations et de références et s'ancrent dans la théorie du discours rapporté (Muñoz et collab., 2004). Nous introduisons tout d'abord la notion de reprise contextualisée qui explicite la configuration particulière de ce type de dérivation, puis nous décrivons rapidement les méthodes de détection reposant sur l'exploration contextuelle.

2.1.2.1 Reprises contextualisées

Nous nommons *reprise* un segment de texte du dérivé qui résulte de l'intégration d'un élément ayant une réalité textuelle dans la source. Il s'agit, en d'autres termes, des dérivations de nature textuelle. Les citations sont des reprises de ce type. Les références, telles qu'elles sont instanciées dans les articles scientifiques notamment, sont plutôt des dérivations de nature contenu (*cf. Tableau 1.1, page 40*). L'auteur emploie, le plus souvent aux frontières de la reprise, des segments textuels de contextualisation qui permettent au lecteur d'identifier (i) la présence d'une reprise et éventuellement, (ii) le texte source duquel le matériel textuel provient. Nous parlons alors de *reprise contextualisée*.

REPRISE

REPRISE CONTEXTUALISÉE

> **Le quotidien économique souligne** : " Si le rapport ne veut pas associer ces montants à l'idée d'une nouvelle 'cagnotte' budgétaire, ni au débat électoral sur le niveau de prélèvements obligatoires, le montant est équivalent au déficit budgétaire de l'État, à savoir 36,5 milliards d'euros l'an dernier."

EXEMPLE 11: Citation extraite d'un article de presse. L'auteur contextualise le texte rapporté en indiquant sa source « le quotidien économique ».

Les citations sont la forme la plus commune de reprise contextualisée. Elles sont particulièrement présentes dans les articles de presse. L'illustration textuelle 11 est une citation extraite d'un article de presse en ligne. Le texte en gras est ajouté par l'auteur pour contextualiser la reprise. Cette dernière est elle-même délimitée à l'aide d'éléments ponctuatifs et typographiques (*: "..."*). Ces différents éléments sont autant d'indices pour la détection intrinsèque de dérivation.

Les références bibliographiques, plus communes dans les articles scientifiques, peuvent être considérées comme une autre forme de dérivation accompagnée d'éléments de contextualisation. Ainsi, dans l'illustration textuelle 12, le segment « *[4, 12]* » est un élément de contextualisation. Toutefois, contrairement aux citations, la dépendance entre cet élément de contextualisation et le texte dérivé est floue. L'auteur ne délimite pas les éléments repris.

En résumé, certaines dérivations peuvent être accompagnées d'éléments textuels renseignant l'auteur sur le contexte de l'énoncé original. Ces éléments contextuels sont les plus présents dans les citations, nous parlons alors de reprise contextualisée car

1. Proposition de traduction du terme anglophone *writing style*.

> Instead of using k-grams, the strings to fingerprint can be chosen by looking for sentences or paragraphs, or by choosing fixed-length strings that begin with "anchor" words **[4, 12]**.

EXEMPLE 12: Référence bibliographique extraite de (Schleimer, 2003).

le texte dérivé a une réalité textuelle dans la source et s'accompagne dans sa forme dérivée desdits éléments de contextualisation.

2.1.2.2 Détection par exploration contextuelle

Les informations de contextualisation peuvent être exploitées pour détecter automatiquement des dérivations, et ce de manière intrinsèque. Les travaux ayant exploité cette idée sont uniquement dédiés à la détection de citation.

Mourad et Minel (2000); Mourad (2001); Mourad et Desclés (2002) exploitent l'exploration contextuelle en corpus pour extraire des indices et des marqueurs pour la détection de citation. L'exploration contextuelle consiste en l'acquisition de régularités lexicales caractéristiques d'une forme linguistique d'intérêt. Ces régularités lexicales sont appelées *indices* ou *embrayeurs*. Lorsque l'indice est réellement spécifique à l'objet linguistique étudié et qu'il marque avec une bonne probabilité la présence du dit objet, les auteurs le nomment *marqueur*. Cette méthode a été mise en œuvre sur des corpus d'articles journalistiques, des textes scientifiques, techniques et des œuvres littéraires, pour lesquels les citations avaient été annotées. Les indices extraits sont décrits dans Mourad et Desclés (2004), il s'agit d'introducteurs verbaux, typographiques, prépositionnels et expressionnels.

Giguet et Lucas (2004) ont également exploré l'utilisation d'informations contextuelles pour la détection de citations. Toutefois, leur approche se distingue de celle de Mourad (2001) sur deux points. Premièrement, leur étude a porté uniquement sur des textes journalistiques. La restriction à ce genre leur a permis de définir un patron invariant constitué de trois constantes citationnelles (*cf. Section 3.1*). Les différents indices qu'ils utilisent permettent d'identifier ces différentes constantes citationnelles, le patron leur permettant de reconstituer la citation. Deuxièmement, contrairement au travail de Mourad (2001) qui tire uniquement parti de ressources lexicales, Giguet et Lucas s'imposent une utilisation parcimonieuse de ressources. Ils tirent ainsi parti uniquement d'indices typographiques (*ponctuation, casse...*), d'indices morpho-syntaxiques (*morphèmes grammaticaux comme « que », suffixes « ent »...*) et d'indices positionnels (*début, fin d'unités textuelles...*), soient une liste de mots outils, des séquences de ponctuation et une liste de suffixes. La présence de citation est déduite de la présence combinée de ces indices.

Notons également les travaux de Siddharthan et Teufel (2007) sur la détection de références dans les articles scientifiques. Celle-ci repose sur le repérage de motifs alphanumériques couramment utilisés (crochets, parenthèses...), complété par la section bibliographique du papier ainsi que par des bases de données de publications.

En résumé, les travaux de détection de dérivation qui tirent parti des indices contextuels se limitent à la détection de citations et de références bibliographiques. Ces méthodes reposent sur l'identification d'ancres textuelles à l'aide d'indices divers (typographiques, lexicaux, syntaxiques...), et l'extraction de l'objet linguistique selon la cooccurrence de ces ancres. Deux approches s'opposent alors : (i) l'exploitation de ressources externes importantes (Mourad, 2001; Siddharthan et Teufel, 2007) ou (ii) l'utilisation restreinte d'indices de surface (Giguet et Lucas, 2004).

2.1.3 Exploitation des irrégularités stylistiques

Lorsque l'auteur ne contextualise pas sa dérivation, il est possible d'étudier les variations stylistiques de son écrit. Nous présentons tout d'abord les travaux autour de la caractérisation stylistique des écrits et des auteurs, puis nous discutons de leur adaptation à la détection de dérivation intrinsèque telle que proposée originalement par Eissen et Stein (2006).

2.1.3.1 Dérivations non contextualisées

Le style est défini par Ducrot et Schaeffer (1995) comme « résultant de la combinaison du choix que tout discours doit opérer parmi un certain nombre de disponibilités contenues dans la langue et des variations qu'il introduit par rapport à ces disponibilités ». Cette définition laisse penser que la langue est complètement accessible aux auteurs, et qu'ils y sélectionnent les variations [2] propres à leur style. En d'autres termes, selon eux, le style réside dans les spécificités de l'instanciation d'un acte langagier particulier par rapport à l'ensemble des actes langagiers qui auraient la même valeur sémantique, voire pragmatique.

Les travaux autour de l'attribution d'auteur, van Halteren et collab. (2005) notamment, ont une vision légèrement différente du style, même si elle rejoint celle de Ducrot et Schaeffer (1995) dans le fond. Ils considèrent que le style est propre à un individu et qu'il se définit par certains traits propres à son langage. Il est en effet généralement accepté aujourd'hui que notre connaissance de la langue relève d'un apprentissage [3]. Cet apprentissage se fait, selon les théories constructivistes, sur la base d'exemples. La langue est alors reconstruite par l'individu à partir d'actes langagiers qu'il peut observer. Ce corpus d'apprentissage, s'il est majoritairement commun aux autres individus parlant la même langue, contient des exemples propres qui entraînent des divergences minimes dans la connaissance de ladite langue. Chaque individu utilise alors un langage propre qui est une instanciation particulière de la langue et dont les spécificités font le style. Les stylostatisticiens parlent d'un ensemble de motifs mesurables : les *marqueurs de style*.

STYLE D'AUTEUR

MARQUEURS DE STYLE

À l'instar des peintres, les écrivains ont un style, une patte, une touche, une signature. Des études littéraires ont très certainement été menées sur la plupart des œuvres classiques afin de décrire le style de leurs auteurs. L'avènement de l'informatique, et sa combinaison avec la *stylométrie* [4] ont permis de supporter leur existence et parfois les décrire. Ainsi la collection d'essais *The Federalist Papers* a servi de corpus d'entraînement à plusieurs travaux (Mosteller et Wallace, 1964; Levitan et Argamon, 2006), tout comme certaines œuvres des auteurs classiques français (Labbé et Labbé, 2006). Toutefois, les auteurs « établis »ont chacun développé un style personnel de par leur statut et leur usage prolifique de la langue, tout en cherchant peut-être à se différencier de leurs prédécesseurs et contemporains. L'existence d'un style propre à chacun n'est pas évident. Il n'est notamment pas démontré que l'on puisse identifier un nombre suffisant de traits mesurables permettant de distinguer un auteur d'un autre. Les résultats des travaux de van Halteren et collab. (2005) sont en faveur de l'existence d'un tel *stylome* [5]. Toutefois, de nombreux traits restent à découvrir afin

STYLOME

2. Les auteurs laissent entendre que ces variations sont définies selon un continuum avec pour extrêmes les *registres collectifs* (p. ex. registres de la langue) à un pôle et les idiosyncrasies à l'autre.

3. Ceci ne remet pas forcément en cause les théories selon lesquelles nous héritons, à travers notre génome, de capacités de développement d'une langue naturelle (grammaire universelle). Voir notamment les études autour des aires de Broca et de Wernicke.

4. Le domaine de caractérisation statistique du style peut-être désigné de plusieurs façons : stylométrie, stylostatistique, ou stylistique computationnelle.

5. Un stylome est l'ensemble des traits de langage permettant de différencier les écrits d'un auteur des écrits d'un autre.

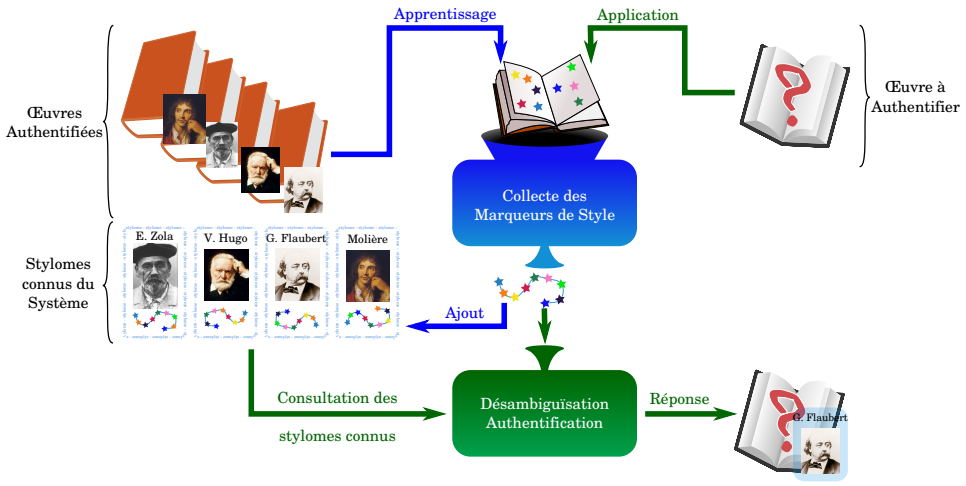


FIGURE 2.1 – Fonctionnement général d'un système d'attribution d'auteur

de le caractériser. De plus, la méthodologie utilisée est critiquable en ce qu'elle ne considère qu'un nombre réduit d'individu (8 auteurs, un texte par auteur) et en ce que chaque auteur a produit un texte sur un sujet différent. Nous pouvons nous interroger sur le fait que les traits retenus (majoritairement lexicaux) ne caractérisent pas les sujets des textes plutôt que le style de leurs auteurs.

Attribution d'auteur Le calcul du stylome d'un auteur est une tâche impossible. En effet, un stylome n'est connu que par son exemplification au travers d'actes de langages (textes, interventions orales...), l'auteur lui-même n'en est pas conscient. De plus, ce stylome peut évoluer au fur et à mesure de la pratique de la langue par l'auteur. Les recherches en attribution d'auteur ont cependant montré qu'une description partielle peut suffire à identifier l'auteur d'un texte.

La tâche d'attribution d'auteur consiste à authentifier automatiquement un texte et l'associer à son auteur, sans autre connaissance que le texte lui-même et le stylome partiel du dit auteur. Dans les faits cette tâche revient soit à attribuer un texte à un auteur parmi une palette limitée d'auteurs connus du système, soit vérifier si un texte a été écrit par un auteur donné. Les systèmes d'attribution d'auteur fonctionnent habituellement en deux étapes :

1. Le système collecte les marqueurs de styles présents dans un texte ;
2. Une méthode de désambiguïsation permet de classer, à partir de cette collection de marqueurs de style, le texte dans une catégorie prédéfinie.

La classification repose sur un apprentissage supervisé qui permet de caractériser les catégories prédéfinies. Cet apprentissage nécessite par conséquence une phrase d'entraînement réalisée sur des œuvres dont les auteurs sont identifiés, comme l'illustre la figure 2.1. Le système ne peut donc authentifier que les œuvres dont les auteurs sont connus au moment de l'entraînement.

Grieve (2007) a mené une évaluation quantitative des différentes techniques utilisées pour l'attribution d'auteur. Il a expérimenté la capacité de certains marqueurs stylistiques de surface (longueur des mots et phrases, richesse du vocabulaire, distribution des caractères, distribution des mots et leurs cooccurrences) à rapprocher un texte anonyme d'un auteur. Les tests du χ^2, opérés pour mesurer l'adéquation des traits aux textes des auteurs correspondant, montrent la validité globale des ap-

proches. Les meilleurs traits permettent de choisir le bon auteur entre deux dans 95 % des cas, et le bon auteur entre quarante auteurs dans 63 % des cas.

Stamatatos (2009b) propose une description exhaustive des marqueurs stylistiques employés :

Lexicaux longueur des mots et des phrases, richesse du vocabulaire, fréquence des mots, des n-grammes mots, idiosyncrasies orthographiques ;

Caractères types de caractères (lettres, nombres...), n-grammes caractères de taille fixée ou variable, méthodes de compression ;

Syntaxiques rôles grammaticaux (*Part-of-speech*), syntagmes, structure des groupes de mots et des phrases, fréquences des réécritures, idiosyncrasies syntaxiques ;

Sémantiques synonymes, dépendances sémantiques, rôle fonctionnel des mots outils ;

Spécifiques structure du document (HTML, RTF...), contenu, éléments propres à une langue.

Détection des variations stylistiques Les travaux en attribution d'auteur montrent qu'il est possible de mesurer et comparer les styles. Si la stylométrie peut-être utilisée pour authentifier les auteurs d'un texte, tout comme nous le faisons pour les tableaux de maître, nous devrions pouvoir également en tirer profit pour identifier les variations stylistiques au sein même des textes. Ces détections de variation nous permettraient, à l'instar des enseignants, de détecter les ruptures stylistiques suspectes au sein des textes et lever de potentielles dérivations.

Certains travaux exploitent déjà avec succès les ruptures stylistiques pour d'autres tâches que la détection de dérivation. Chase et Argamon (2006) s'inscrivent dans la problématique de résumés de texte. Ils cherchent à compléter les méthodes classiques d'extraction de phrases pertinentes reposant sur le contenu par une segmentation des textes en passages sur la base de traits stylistiques. Glover et Hirst (1996) utilisent les techniques d'attribution d'auteur pour la détection d'inconsistences stylistiques dans les textes écrits de manière collaborative, et ce afin d'homogénéiser le style d'écriture.

2.1.3.2 Détection de dérivation par identification des ruptures stylistiques

Eissen et Stein (2006) proposent de mettre en œuvre une identification automatique des passages suspects au sein d'un document sans le besoin d'une collection à laquelle se comparer. Les travaux antérieurs autour de la caractérisation du style sont la principale ressource où puiser des techniques pour la mise en œuvre de cette approche, à la différence que la détection des inconsistences stylistiques n'est plus une fin (comme c'était le cas pour Glover et Hirst (1996) par exemple), mais un moyen. Le schéma des approches que l'on retrouve dans la littérature est alors le suivant :

1. *Découpage en passages* le texte est découpé en segments définis selon une taille en mots ou en caractères, ou bien linguistiquement motivés. Le plus souvent, il s'agit de fenêtres glissantes de taille fixe (environ 250 mots).
2. *Caractérisation* le passage considéré est caractérisé à l'aide de traits stylométriques (graphiques, lexicaux, ponctuatifs, syntaxiques, structurels...) ;
3. *Détection* la phase finale consiste à différencier, par le calcul d'une déviation par rapport à une caractérisation de référence, les passages stylistiquement concordant (*target*) de ceux qui ne le sont pas (*outlier*), et potentiellement généraliser à l'ensemble du document.

L'approche d'Eissen et Stein (2006) [6] est très similaire à celle de Chase et Argamon

6. Les travaux présentés dans Eissen et Stein (2006) ont fait l'objet d'une description plus détaillée dans zu Eissen et collab. (2007)

(2006). Le document est tout d'abord découpé en parties, reflétant plus ou moins un découpage « naturel » ou du moins linguistiquement motivé. Le style de chaque partie est décrit à l'aide de traits stylométriques. Les auteurs ont utilisé des traits stylométriques classiques tels que la taille moyenne des phrases, le nombre moyen d'instances de certaines catégories grammaticales (*part-of-speech*) et le nombre moyen de mots outils. Ils ont également utilisé la classe de fréquence moyenne des mots (*cf. Annexe I*) afin de caractériser la complexité du style de l'auteur et la richesse de son vocabulaire.

Les auteurs ont évalué leur approche sur un corpus de 450 documents construit manuellement à partir d'articles scientifiques tirés de la bibliothèque ACM. Chaque document contient entre 3 et 6 passages copiés, exactement ou avec des reformulations, de longueurs variables. Ils ont été en mesure d'identifier les passages copiés avec une précision d'environ 70 % et un rappel variant de 80 % à 95 % selon la proportion de texte plagié (Eissen et Stein, 2006, Figure 2). Les traits les plus discriminants dans le contexte de leur corpus sont la classe de fréquence moyenne des mots, le nombre moyen de prépositions et la taille moyenne de phrases.

Stein et collab. (2010) reprennent la caractérisation stylistique des passages de texte de Eissen et Stein (2006) en complétant les traits stylométriques employés (notamment par des mesures de complexité du style), et en utilisant une classification naïve bayesienne pour classer les traits stylométriques des passages comme concordant avec le reste du document (*target group*) ou non (*outlier group*). Ils ont également expérimenté trois stratégies de classification pour décider si les documents complets sont des dérivés ou non d'après la sortie du réseau bayésien : le risque minimum, le vote et le démasquage. La stratégie du risque minimum consiste à classer un document comme dérivé si au moins un trait stylométrique est classé comme *outlier*, l'heuristique de vote décide par rapport à un seuil concernant le nombre de traits classés comme *outlier* et la stratégie de démasquage utilise un seuil haut et un seuil bas combiné à un meta-learning (Stein et zu Eissen, 2007).

Les auteurs ont évalué leur approche sur le corpus PAN'09 (*cf. Section 4.2*). Les traits stylométriques les plus efficaces sont le score de lecture de Flesch (Flesch, 1948), le nombre moyen de syllabes par mots, la fréquence du mot « of »et les trigrammes *Nom-Verbe-Nom* et *Nom-Nom-Verbe*. En ce qui concerne les stratégies de classification, la stratégie de démasquage obtient la meilleure précision, même si les trois stratégies expérimentées sont comparables en termes de F-score.

La campagne d'évaluation PAN'09 (Potthast et collab., 2009) a permis de comparer quatre méthodes de détection intrinsèque sur le même corpus et avec le même protocole d'évaluation. Le tableau 2.1, extrait de Potthast et collab. (2009) synthétise les performances des différents systèmes. L'approche de référence consiste à considérer tous les documents comme non plagiés. Stamatatos (2009a) caractérise une fenêtre glissante de 1 000 caractères par la fréquence de ses trigrammes caractères. Le document est considéré comme plagié si la différence entre les profils de deux fenêtres adjacentes est supérieure à un seuil prédéterminé. Zechner et collab. (2009) travaillent à l'échelle de la phrase et emploient un modèle vectoriel compilant la classe de fréquence des mots et la distribution de la ponctuation, des rôles grammaticaux, de certains pronoms et de certains mots outils. Les passages *outlier* sont identifiés par le calcul de la dérivation standard par rapport à un passage moyen. Finalement, Seaward et Matwin (2009) reprennent une approche similaire à celle de Eissen et Stein (2006) en introduisant une mesure de complexité de la distribution des classes de mots (complexité de Kolmogorov) approximée par un algorithme de compression sans perte (Lempel-Ziv).

Rang	F-score	Précision	Rappel	Participants
1	0,308 6	0,232 1	0,460 7	(Stamatatos, 2009a)
2	0,195 6	0,109 1	0,943 7	Approche de référence
3	0,228 6	0,196 8	0,272 4	(Zechner et collab., 2009)
4	0,175 0	0,103 6	0,563 0	(Seaward et Matwin, 2009)

TABLE 2.1 – Résultats de l'évaluation de détection sans point de comparaison réalisée dans le cadre de PAN'09. Le protocole d'évaluation est décrit dans la section 4.1.1

Synthèse

La détection de dérivation intrinsèque est une considération récente bien qu'elle repose sur des traits stylométriques expérimentés depuis plusieurs décennies pour la tâche d'attribution d'auteur : les premières recherches sur la caractérisation du style datent du milieu du XIX^{e} siècle. Les quelques travaux qui s'y emploient décrivent des segments de texte à l'aide de traits stylométriques classiques et recherchent une rupture parmi ces traits entre deux passages. Cette rupture est interprétée comme le commencement (ou la fin) d'un passage dérivé. Les performances de ces approches si elles sont encourageantes, sont bien inférieures à celles des approches avec point de comparaison : au challenge PAN'09 (Potthast et collab., 2009) les cinq meilleures approches avec point de comparaison obtiennent une F-mesure comprise entre 0,46 et 0,69 tandis que la meilleure approche sans point de comparaison obtient 0,30. Elles ont toutefois le très grand avantage de ne pas nécessiter de collection de référence, ce qui borne leur complexité à la taille du texte.

En général les traits stylométriques sont considérés indépendamment les uns des autres lors de la recherche de rupture, ce qui ne correspond pas exactement à la définition de stylome donnée par van Halteren et collab. (2005). De notre point de vue, la dérivation est un phénomène multidimensionnel et les approches devraient le considérer comme tel lors de la détection. Dans le cas présent il serait intéressant de combiner les différents traits stylométriques. Il faudrait également s'assurer que ceux-ci soient indépendant du sujet et du genre du texte.

Une approche originale de détection intrinsèque de dérivation, et plus particulièrement de plagiat, consiste à interroger l'auteur du texte sur ses choix lexicaux. C'est ce que réalise le programme *Glatt Screening* (gla, 1999) qui invite l'auteur d'un texte à remplir des blancs dans son propre texte. L'hypothèse est que l'auteur original du texte doit réutiliser globalement les mêmes mots que dans la version d'origine. Si les propositions ne correspondent pas au texte original, c'est qu'il s'agit d'une dérivation.

Au final, la détection de dérivation sans point de comparaison est une piste très récente qui a été assez peu étudiée. La majorité des travaux se concentre sur la détection avec point de comparaison, c-à-d avec un corpus de référence auquel confronter le texte considéré. Nous décrivons les différentes méthodes qui vont dans ce sens dans la section suivante.

2.2 Techniques de détection extrinsèque

Les méthodes de *détection extrinsèque* reposent sur la confrontation entre le texte en focus et une collection de textes suspects. Elles s'opposent en ce sens aux méthodes de détection intrinsèque qui exploitent exclusivement les données du texte en focus. Les méthodes de détection extrinsèques de dérivation ont reçu plus d'attention que les méthodes intrinsèques alors que leur mise en œuvre nécessite des ressources plus importantes (collection des textes suspects). Les causes de cet intérêt particulier

DÉTECTION EXTRINSÈQUE

sont peut être à chercher du côté de la similarité de ces méthodes avec des méthodes classiques de RI. Ce type d'approche est encore une fois similaire avec l'approche des enseignants pour détecter un plagiat. Le travail soumis à l'enseignant peut lui rappeler un travail antérieur et exposer différentes similitudes avec celui-ci : instanciation textuelle, plan, organisation des idées, figures ou encore certaines idiosyncrasies — fautes d'orthographe ou de syntaxe en particulier.

Nous décrivons tout d'abord le principe général des méthodes de détection extrinsèque (*cf. Section 2.2.1*). Nous détaillons ensuite les trois grandes approches mises en œuvre : les approches par calcul de couverture (*cf. Section 2.2.2*), les approches directement inspirées de la RI et qui utilisent des modèles vectoriels (*cf. Section 2.2.3*), et finalement les approches par alignement (*cf. Section 2.2.4*). Les différentes approches présentées ont été évaluées individuellement avec des protocoles d'évaluation différents et sur des corpus distincts. Par conséquent, nous ne rapportons pas leurs performances respectives car elles ne pourraient être comparées.

2.2.1 Approche générale

Le principe de fonctionnement d'une méthode extrinsèque repose sur la confrontation du texte en focus avec des textes suspects. La confrontation se fait par paire : des éléments communs sont recherchés entre le texte en focus et chacun des textes suspects. La révélation d'un nombre suffisant de tels éléments est considérée comme une preuve de l'existence d'une relation de dérivation entre les deux textes. L'objectif d'une méthode de détection extrinsèque est d'identifier les relations de dérivation impliquant le texte en focus et d'éventuels textes dérivés (respectivement sources) issus de la collection de textes suspects.

Le fait de considérer le texte en focus comme une source ou un dérivé oriente la recherche vers des dérivés ou bien des sources. Il est tout aussi envisageable de s'intéresser uniquement à la découverte de liens de dérivation sans chercher à en nommer la source et le dérivé (lien de co-dérivation). Pour alléger les explications, nous nous concentrons par la suite sur la configuration consistant à rechercher des textes dérivés d'un texte source en focus. Les explications s'appliquent tout aussi bien aux autres configurations à une réattribution des statuts des textes près.

L'idée partagée par toutes les méthodes de la littérature revient à collecter des traces d'un document source dans un document dérivé. Ces approches peuvent également être caractérisées par les trois mots clés du principe d'échange de Locard (1940) [7] : le *transfert* d'éléments du document source dans le dérivé, la *persistance* de ces éléments dans la version observée du dérivé et leur *pertinence* comme éléments ayant un réel rôle dans la construction discursive du texte dérivé.

Les éléments repris du texte source dans le texte dérivé peuvent être conservés dans leur forme originale ou bien être modifiés (*cf. Section 1.4.8*). Dans le premier cas une recherche exacte permettra de les identifier. Dans le second cas la recherche devra être approximative, ce qui est plus coûteux d'un point de vue algorithmique. L'utilisation de mesures de similarité est la technique privilégiée d'approximation.

La figure 2.2, dérivée de Potthast et collab. (2009, Fig. 1), schématise le fonctionnement global d'un système de détection extrinsèque de dérivation en trois phases : filtrage des candidats parmi la collection de référence, comparaison exhaustive du

7. Dans son *Traité de criminalistique*, Locard (1940) pose l'hypothèse qu'un malfaiteur laisse sur les lieux de l'infraction des traces de son passage et emporte avec lui des éléments qui détermineront sa présence et son action sur la scène du crime. Cette hypothèse fondatrice de l'enquête criminelle est depuis appelée « principe de Locard » ou « principe d'échange de Locard »et mise en application lorsque le criminel a été au contact de la scène de crime. D'après ce principe il y apporte quelque chose et repart avec autre chose. Ce principe se résume en trois mots clés : transfert, persistance et pertinence.

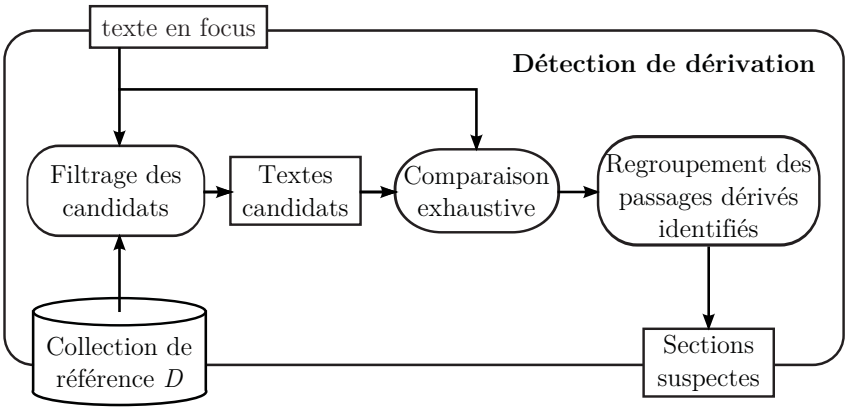


FIGURE 2.2 – Vision générale d'un système de détection extrinsèque. Figure dérivée de Potthast et collab. (2009, Fig. 1)

texte en focus (soumis au système) avec chacun des textes candidats et regroupement des passages dérivés identifiés. La première et la dernière étape sont optionnelles. La première phase consiste à filtrer une collection de référence D afin de constituer une collection de candidats pertinents : même thématique, distribution similaire des mots... Cette étape optionnelle a pour but de réduire l'espace de recherche et doit être la plus efficace possible. Les critères de sélection reposent le plus souvent sur des principes classiques de RI. Les textes sont représentés sous la forme de vecteurs de mots. Une mesure de similarité est appliquée entre le texte en focus et chaque texte de D. Un score supérieur à un seuil empirique conduit à conserver le texte de D comme candidat.

La second phase consiste à comparer de manière exhaustive cette fois, le texte d_q avec chaque texte candidat. Selon Clough et Gaizauskas (2008, p. 6), les techniques de comparaison nécessitent également trois étapes : (i) le pré-traitement linguistique du texte et la génération d'une représentation intermédiaire adaptée à la comparaison ; (ii) la comparaison de ces représentations intermédiaires ; (iii) le calcul d'une mesure de similarité ou la visualisation de la sortie de la comparaison. Stein et zu Eissen (2005) catégorisent les techniques de comparaison exhaustive en trois grandes classes associées à des systèmes opérationnels :

- Les approches par **couverture de texte**, ou **analyse d'empreintes** (COPS (Brin et collab., 1995), KOALA (Heintze, 1996), SCAM (Shivakumar et Garcia-Molina, 1995) et DSC (Broder et collab., 1997)) reposent sur l'hypothèse que plus des documents partagent un nombre important de séquences textuelles, plus il est probable que ces documents soient dérivés l'un de l'autre. Ces systèmes modélisent les documents comme des collections de sous-séquences textuelles et en maintiennent un index inversé. La probabilité de dérivation est alors proportionnelle au nombre de collisions au sein de cet index.
- Les approches par **similarité de mots-clés** (CHECK (Si et collab., 1997) et I-MATCH (Chowdhury et collab., 2002)) reposent sur l'extraction et la pondération des mots-clés censés représenter la thématique du texte. Les documents sont comparés sur la base de ces mots-clés pondérés. Si deux documents sont globalement similaires, c-à-d s'ils partagent à l'échelle du texte un nombre important de mots-clés, le système itère récursivement à travers la structure du document afin d'isoler les parties similaires.

– Les approches par **alignement de sous-chaînes** (MatchDetectReveal (Monostori et collab., 2001), TESAS (Piao, 2001) et l'approche de Bourdaillet (2007)) identifient des paires de chaînes de caractères entre les documents. Ils utilisent dans ce but des algorithmes de recherche de plus longues sous-chaînes communes (Aho et Corasick, 1975; Knuth et collab., 1977; Boyer et Moore, 1977; Karp et Rabin, 1987a) [8].

La dernière phase d'un système de détection extrinsèque de dérivation est optionnelle et a pour objet de regrouper les passages dérivés proches les uns des autres. Elle peut tirer partie de connaissances liées au genre, à la structure du document...

Dans la suite de cette section nous présentons en détail le fonctionnement de ces différentes approches ainsi que les travaux de la littérature qui s'en inspirent.

2.2.2 Approches par couverture de texte

Les approches par couverture de texte consistent à découper les textes comparés en tuiles et à compter le nombre de tuiles communes. Nous en présentons tout d'abord l'approche générale commune aux systèmes COPS (Brin et collab., 1995), KOALA (Heintze, 1996) et plus particulièrement DSC (Broder et collab., 1997) ; puis nous détaillons le cœur des méthodes qui repose sur les fonctions de sélection des sous-séquences textuelles (Π) et les mesures de similarité (φ). Nous détaillons ensuite la signature complète et ses variations, considérées comme approche de référence du domaine. Enfin, nous discutons les qualités et défauts d'une telle approche.

2.2.2.1 Approche générale

Les approches par mesure du recouvrement de texte ont été introduites par les premiers systèmes dédiés à la détection de plagiat dans le contexte de bibliothèques numériques : COPS (Brin et collab., 1995), KOALA (Heintze, 1996) et SCAM (Shivakumar et Garcia-Molina, 1995). Ces approches reposent sur l'hypothèse que la probabilité que deux textes soient dérivés est corrélée au nombre de sous-séquences textuelles que ces textes ont en commun. Ce que nous exprimons plus formellement :

$\forall d_1, d_2 \in \mathcal{D}$ deux documents,

$\mathcal{D}$ l'ensemble des documents,

$\mathcal{S}$ l'ensemble des séquences textuelles de $\mathcal{D}$,

$$\begin{cases} p((d_1 \,\mathbf{R_D}\, d_2) \vee (d_2 \,\mathbf{R_D}\, d_1)) \propto \varphi(\Pi(d_1), \Pi(d_2)) \\ \quad \text{avec } \Pi : \mathcal{D} \mapsto \mathcal{P}(\mathcal{S}), \mathcal{P}(\mathcal{S}) \text{ ensemble des parties de } \mathcal{S} \\ \quad \text{et } \varphi : \mathcal{S}^n \times \mathcal{S}^m \mapsto [0..1], \text{ mesure de similarité} \end{cases}$$

Ces approches varient dans les faits selon deux éléments : (i) la fonction de projection Π qui sélectionne les séquences textuelles des documents et (ii) la mesure de similarité φ permettant de comparer deux résultats de la projection de Π.

Ce type d'approche permet d'identifier des dérivations à toutes les granularités du texte selon le type de filtrage des séquences mis en place. Ainsi, une dérivation de granularité partielle sera rendue visible par la concentration de séquences textuelles communes dans une partie localisée du texte. Toutefois, l'utilisation de correspondances exactes entre ces séquences par les mesures de similarités rend ces approches sensibles aux intégrations autres que verbatim. Une expression du texte source réécrite est perçue par le système comme une séquence différente, la dérivation n'est

8. Voir Charras et Lecroq (2004) pour un état de l'art complet sur la recherche efficace de sous-chaînes.

alors pas détectée. Ces propriétés en font des systèmes adaptés aux tâches de détection de duplication ou presque-duplication.

2.2.2.2 Fonctions de sélection des sous-séquences textuelles (Π)

La fonction de sélection des sous-séquences textuelles décrit comment sont prélevés les éléments des textes qui seront comparés et comment ils sont réunis pour représenter les textes. Nous introduisons tout d'abord les deux niveaux d'unité sur lesquels reposent ces fonctions. Nous décrivons ensuite leurs stratégies de construction puis leur assemblage.

Les premiers travaux expérimentant des approches par couverture de texte (Brin et collab., 1995; Heintze, 1996; Broder et collab., 1997) s'accordent sur le besoin de deux niveaux d'unité nécessaires au découpage du texte en sous-séquences (*chunking*). La première unité définit le niveau indivisible du texte, ce que Brin et collab. (1995) nomment *unit*, Heintze (1996) la nomme *unit of chunking* et Shivakumar et Garcia-Molina (1996) *primitive unit*. Nous la nommerons *unité primitive*.La seconde unité est l'unité de la comparaison et correspond à un regroupement séquentiel d'instances de la première unité. Broder (1997) la définit comme une « sous-séquence contiguë contenue »dans le document, et la nomme *w-tuile* (que l'on pourrait traduire par « tuile de taille w ») où w indique sa taille en unités primitives. Le choix de cette taille décide de la granularité à laquelle la dérivation pourra être détectée. Brin et collab. (1995) et Shivakumar et Garcia-Molina (1996) y font référence par le terme *chunk*, et Hoad et Zobel (2002) nomment *minutia* la version hachée de ce segment. Nous choisissons de conserver le terme *w-tuile* (tuile de texte composée de w unités primitives), le terme *chunk* étant trop ambigu.

CHUNKING

UNIT

UNITÉ PRIMITIVE

w-tuile

D'après Brin et collab. (1995, p.9), les *w-tuiles* peuvent être construites selon quatre stratégies : (i) une seule unité primitive (Brin et collab. (1995) utilisent les phrases) ; (ii) plusieurs unités primitives qui ne se recouvrent pas (Manber (1994) utilise des chaînes de 50 octets, Wise (1996) utilise des mots en filtrant les noms propres et les mots outils) ; (iii) plusieurs unités qui se recouvrent [9] (Lyon et collab. (2001) utilisent des trigrammes de mots) ; (iv) plusieurs unités primitives sans recouvrement de taille variable [10]. La seule contrainte est de pouvoir définir une relation d'égalité entre les *w-tuiles* afin de mettre en œuvre la mesure de similarité.

La définition de ces unités est la première étape nécessaire à la définition de la fonction de sélection des sous-séquences textuelles Π. Il est également nécessaire de définir la façon dont les sous-séquences ainsi construites vont être assemblées afin de refléter l'unité du document. Cet assemblage constitue la *signature* du document, et le nombre de *w-tuile* qui la compose en est la *résolution* (Hoad et Zobel, 2002). Deux structures mathématiques sont proposées. Brin et collab. (1995); Heintze (1996); Broder (1997) proposent d'utiliser un ensemble ou un multiensemble. Ainsi Broder (1997) définit la *w-couverture* comme le multiensemble [11] des *w-tuiles* d'un document. Shivakumar et Garcia-Molina (1995) proposent d'utiliser un vecteur de fréquence. Ce choix de structure, selon le type de *w-tuile* choisi, est à la limite entre les approches

SIGNATURE

RÉSOLUTION

9. Brin et collab. précisent que chaque *w-tuile* partage k-1 unités primitives avec ses voisins, ce qui n'empêche pas d'imaginer d'autres proportions de recouvrement.

10. Brin et collab. proposent d'agglomérer les unités primitives jusqu'à ce que leur hash soit congruent à une certaine valeur modulo k. Ils parlent de hash avec paramètre secret. Cette approche a été expérimentée par Shivakumar et Garcia-Molina (1996) sous l'appelation *hashed breakpoint chunking*

11. En mathématiques, un multiensemble (*multiset* ou *bag* en anglais) est une généralisation de la notion d'ensemble où la contrainte d'appartenance unique est levée. En d'autres termes, alors que dans un ensemble un élément ne peut apparaître au maximum qu'une seule fois, dans un multiset il peut apparaître plusieurs fois.

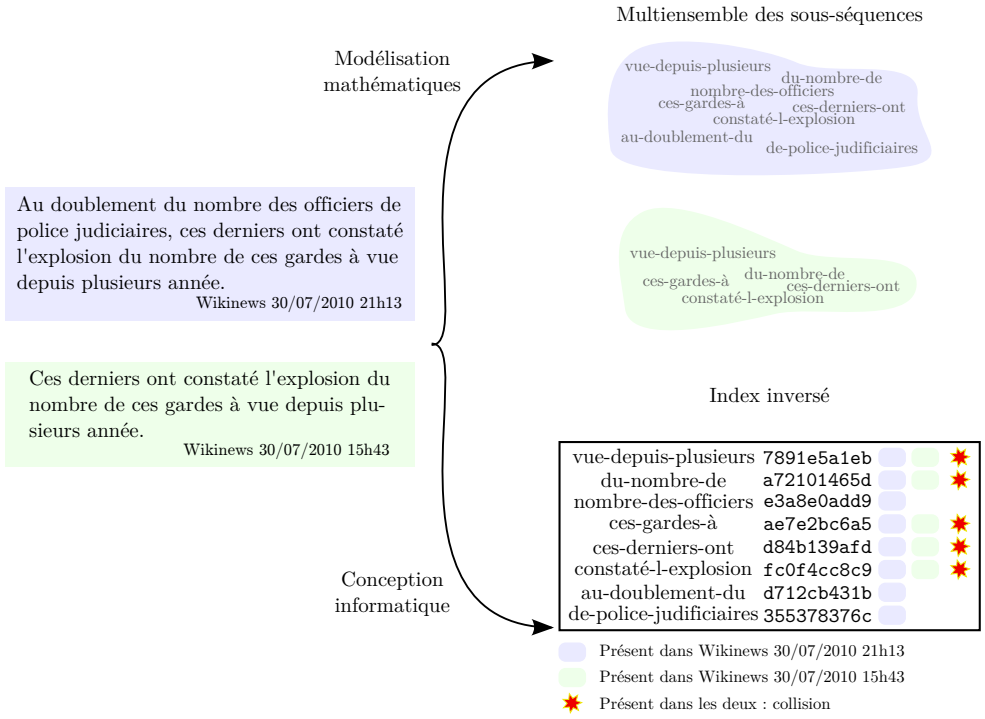


FIGURE 2.3 – Modélisation de deux documents (*sic*) dans le cadre d'une approche par recouvrement de texte et son éventuelle implémentation à l'aide d'un index inversé.

par couverture de texte et par similarité de mots clés (*cf. Section 2.2.3*). Nous choisissons de le classer dans cette dernière catégorie et nous nous concentrons par la suite uniquement sur la modélisation par (multi)ensemble.

La figure 2.3 (page 58) illustre les différentes étapes de modélisation pour les approches par recouvrement de texte. Dans cet exemple, nous avons choisi le mot comme unité primitive, les *w-tuile* consistent en un regroupement sans recouvrement de trois unités primitives, ce qui revient à modéliser les textes par l'ensemble des trigrammes de mots qui les composent.

2.2.2.3 Mesures de similarité (φ)

Les mesures de similarité utilisées pour ces approches reposent largement sur des opérations ensemblistes de par l'utilisation des ensembles et multiensembles pour représenter les documents. Deux mesures ont été particulièrement employées : la *resemblance* et le *containment*. Elles ont toutes les deux été définies par (Broder, 1997).

RESEMBLANCE

La *resemblance* mesure le nombre d'éléments en commun entre deux documents par rapport à la totalité des éléments des deux documents réunis (*cf. Équation 2.1*). La formule n'est pas sans rappeler le *coefficient de Jaccard* (Jaccard, 1912). Elle capture ainsi l'idée que les deux documents sont globalement similaires et a été utilisée à cette fin par Lyon et collab. (2004) notamment. Monostori et collab. (2002) y font également référence sous l'appelation similarité symétrique, tandis que Stein et Eissen (2006) parlent de *mesure de similarité locale* ou *similarité de recoupement*.

$$r(\Pi(d_1), \Pi(d_2)) = \frac{|\Pi(d_1) \cap \Pi(d_2)|}{|\Pi(d_1) \cup \Pi(d_2)|} \tag{2.1}$$

Le *containment* mesure le nombre d'éléments en commun entre deux documents par rapport aux éléments présents dans l'un des deux (*cf. Équation 2.2*). Cette mesure capture ainsi l'idée que l'un des documents est globalement contenu dans l'autre et a été utilisée à cette fin par Yang et Callan (2005) notamment. Monostori et collab. (2002) y font également référence sous l'appelation *similarité asymétrique*. La non-symétrie de cette mesure la rend particulièrement adaptée aux configurations où la dérivation depuis d_1 ne constitue qu'une petite partie de d_2. Elle est une solution pour la détection des cas où la composante source a une granularité importante tandis que la composante dérivée a une granularité partielle. CONTAINMENT

$$c(\Pi(d_1), \Pi(d_2)) = \frac{|\Pi(d_1) \cap \Pi(d_2)|}{|\Pi(d_1)|} \tag{2.2}$$

Par exemple, pour les textes de la figure 2.3, nous aurions :
- $\Pi(d_1) = \{d84b139afd, fc0f4cc8c9, a72101465d, ae7e2bc6a5, 7891e5a1eb\}$
- $\Pi(d_2) = \{d712cb431b, e3a8e0add9, 355378376c, d84b139afd, fc0f4cc8c9, a72101465d, ae7e2bc6a5, 7891e5a1eb\}$

alors :

$$\begin{aligned} r(\Pi(d_1), \Pi(d_2)) &= \frac{|\Pi(d_1) \cap \Pi(d_2)|}{|\Pi(d_1) \cup \Pi(d_2)|} \\ &= \frac{|\Pi(d_1)|}{|\Pi(d_2)|} \\ &= \frac{5}{8} \end{aligned}$$

et :

$$\begin{aligned} c(\Pi(d_1), \Pi(d_2)) &= \frac{|\Pi(d_1) \cap \Pi(d_2)|}{|\Pi(d_1)|} \\ &= \frac{|\Pi(d_1)|}{|\Pi(d_1)|} \\ &= 1 \end{aligned}$$

Ces mesures, étant donné Π, montrent que d_1 est moyennement similaire à d_2, mais qu'il est complètement contenu dans d_2.

2.2.2.4 La signature complète

La signature complète (*full fingerprinting* ou *k-gram*) repose sur le principe de la couverture de texte. Elle en est une mise en œuvre simple et efficace, et constitue l'approche de référence du domaine. Elle consiste à utiliser une fonction de sélection des sous-séquences textuelles Π telle que toute sous-chaîne de d soit contenue dans les éléments de $\Pi(d)$. FULL FINGERPRINTING

Les signatures complètes utilisent des fonctions Π qui ne filtrent, par définition, aucune sous-chaîne. Ces fonctions Π se définissent donc uniquement comme des méthodes de découpage du texte par la définition d'une unité primitive (caractère, mots, phrases...), d'une taille de *w-tuile* et de contraintes de recouvrement ou non des *w-tuiles* . Étant données ces propriétés, la littérature fait également référence à ces méthodes sous l'appellation recouvrement de *n*-grammes (*n-gram overlap*).

Cette méthode a fait ses preuves pour la détection de dérivation de nature textuelle, sous réserve de choisir une taille de sous-chaîne adaptée. Elle a notamment été mise en œuvre avec succès pour la détection de presque-duplication (Broder et collab., 1997; Cho et collab., 2000; Yang et Callan, 2005; Seo et Croft, 2008), de réutilisation de texte (Clough, 2003a), et de plagiat (Lyon et collab., 2004).

Dans plusieurs de ces mises en œuvre les segments sont des trigrammes mots, c-à-d des séquences contiguës de trois mots. Cette taille de sous-chaîne est suffisamment grande pour filtrer les mots communs sans pour autant être trop grande et n'extraire que des sous-chaînes singulières. En effet, Lyon et collab. (2004) ont observé que des textes écrits indépendamment ont un taux de recouvrement assez bas en termes de trigrammes mots : de l'ordre de 8 %. Ils ont également observé sur des corpus de genres différents (TV News, Federalist Paper, Wall Street Journal) qu'en règle générale 80 % des trigrammes mots rencontrés dans les textes sont uniques. Monostori et collab. (2002) ont comparé plusieurs autres méthodes de découpage des textes et de sélection des *w-tuiles* .

2.2.2.5 Améliorations de la signature complète

La signature complète est difficilement opérationnelle puisqu'elle revient, en termes de coût, à comparer les textes complets les uns aux autres. Plusieurs heuristiques ont été proposées pour palier cela : filtrages positionnels ou structurels, selon la fréquence des *w-tuile* ou en combinant des fonctions de hachage. Potthast et Stein (2007) ont mené une évaluation comparative de ces différentes méthodes (voir notamment le tableau 2 et la figure 3 de leur article). Ils rapportent que ces méthodes donnent des résultats comparables, à l'exception du découpage selon la valeur numérique d'un hachage [12].

Les filtrages positionnels ou structurels consistent à retenir ou filtrer un élément de la signature selon sa position dans le texte ou bien selon son contenu. Heintze (1996) fait le choix d'une signature de taille fixe (*fixed size selective fingerprinting*) constituée des *w-tuiles* dont les hachages ont la plus petite valeure numérique. L'algorithme *winnowing* (Schleimer, 2003) opère de façon similaire mais sur des fenêtres de taille fixe, s'assurant que le texte est représenté de façon homogène dans la signature. Broder et collab. (1997) font le choix d'une signature proportionnelle à la taille du texte en ne conservant que les *w-tuiles* dont la position est un multiple de n. D'autres heuristiques ne permettent pas de contrôler la taille des signatures. Ainsi, l'heuristique « modulo » consiste à ne conserver un w-shingle que si son hash est congruent à a modulo b ($hash(w) \mod b \equiv a$) (Brin et collab., 1995; Fetterly et collab., 2003; Bernstein et Zobel, 2004). Manber (1994) sélectionne les débuts des *w-tuiles* selon des ancres textuelles (séquence de caractères). Wise (1996) filtre quant à lui les noms propres et les mots outils. La sélection aléatoire a également été expérimentée (Heintze, 1996; Kołcz et Chowdhury, 2008) mais sans résultat probant.

Les filtrages par fréquence consistent à juger de la pertinence d'un élément selon sa distribution dans le document ou dans une collection. La plupart de ces stratégies filtrent les mots les plus fréquents ou les moins fréquents. Ainsi, Chowdhury et collab. (2002) utilisent l'idf (*inverse document frequency*) pour retirer les mots les plus et les moins fréquents. Bernstein et Zobel (2004); Kołcz et Chowdhury (2008) conservent uniquement les *w-tuiles* qui apparaissent dans plusieurs documents de la collection. Tandis que Heintze (1996) opère une racinisation grossière en ne retenant que les *w-tuiles* dont les cinq lettres de tête sont peu fréquentes dans le document.

Finalement, plusieurs travaux ont tenté de combiner des fonctions de sélection des *w-tuiles* , par combinaison de fonctions de hachage. DSC-SS (Broder et collab., 1997) introduit la notion de super-tuile comme une *w-tuile* de *w-tuiles* , les premières *w-tuiles* sont alors nommées *pré-images*. Cette approche a été reprise par Fetterly et collab. (2003); Bernstein et collab. (2006). Ceux-ci tirent également parti des dis-

12. Une fonction de hachage calcule une empreinte à partir d'une donnée fournie en entrée permettant d'identifier rapidement, bien qu'incomplètement, la donnée initiale. Nous nommons ici *hachage* la donnée incomplète en sortie d'une telle fonction.

tributions supposées uniformes des fonctions de hachage pour ne conserver qu'une partie du hash (jusqu'à un seul bit).

2.2.2.6 Discussion

Les approches par couverture de texte reposent sur l'idée que des documents dérivés partagent un nombre important de séquences textuelles communes. Ces approches ont notamment été développées dans le cadre de la détection de duplication et de presque-duplication, c-à-d des documents très similaires dans leur forme textuelle. Ces approches sont particulièrement efficaces pour des granularités majoritaires ou importantes, mais il est possible, par des manipulations au niveau des signatures et des mesures de similarité, de repérer des dérivations de granularité plus fine.

L'hypothèse de recouvrement important de séquences textuelles entre les textes dérivés est critiquable. Ainsi, le recouvrement de texte de deux documents dérivés dont les tailles respectives sont très différentes (granularité partielle) peut être minime de par la dissolution des séquences textuelles communes. Cela est également accentué par des intégrations impliquant des modifications importantes (paraphrases, réécriture...), et le bruit résiduel dû à la présence d'éléments communs malgré l'absence de dérivation. La pertinence du choix des éléments de la signature joue alors un rôle prépondérant dans l'efficacité de l'approche.

Les stratégies de sélection des éléments de la signature permettent de réduire la taille et le coût de calcul des méthodes mais sont déconnectées de la dimension linguistique des textes. Ainsi, elles intègrent peu de justification linguistique à l'exception de quelques traitements morphologiques. De plus, le choix d'un hachage numérique réduit la comparaison entre les *w-tuiles* à des correspondances exactes de leur forme textuelle. Kołcz (2004) propose de palier ces défauts en créant plusieurs signatures pour un même document par l'introduction d'une dimension aléatoire dans la sélection. De cette façon, l'introduction de légères modifications dans un document dérivé, plutôt qu'impacter fortement l'unique signature, devrait impacter peu les différentes signatures.

Nous pensons qu'une sélection efficace des éléments doit tenir compte de la pertinence de l'élément au regard de la tâche et du texte. Nous nous interrogeons notamment sur l'intérêt d'une représentation plus linguistique.

2.2.3 Approches par similarité de mots-clés

Les approches par similarité de mots-clés consistent à comparer les textes sur la base de leurs distributions de mots. Les modélisations employées ne reflètent plus la structure linéaire des textes contrairement à l'approche par couverture de texte : elles mesurent des similarités globales (à l'échelle du texte complet) entre les textes, là où la couverture de texte mesure des similarités locales (morceaux identifiés dans le texte). Pour cette raison, nous pourrions également parler d'approche par modélisation globale des textes. Nous introduisons tout d'abord le modèle vectoriel sur lequel reposent ces approches ainsi que les mesures de similarité associées. Nous présentons ensuite deux mises en œuvre particulières pour la détection de dérivation : le modèle à fréquence relative et la signature floue. Enfin, nous discutons les qualités et défauts d'une telle approche.

2.2.3.1 Le modèle vectoriel

La popularité du modèle vectoriel réside dans la simplicité de sa compréhension et dans la robustesse des théories mathématiques sur lesquelles il repose (algèbre

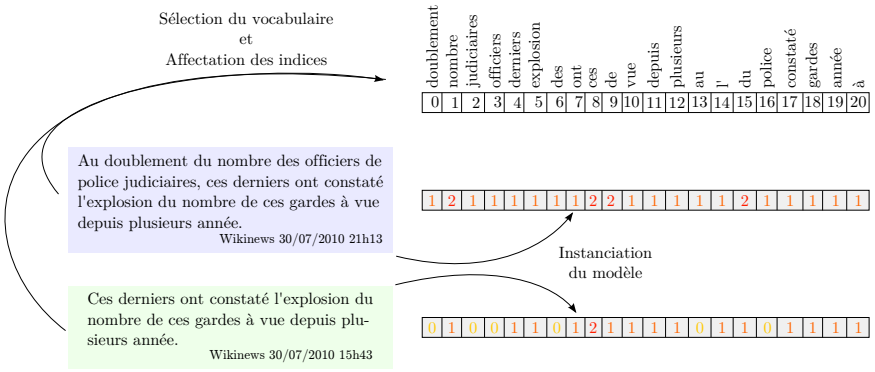


FIGURE 2.4 – Instance pour un document d d'un modèle vectoriel dont le vocabulaire est l'ensemble des mots du corpus et où le poids correspond au nombre d'occurrences du mot dans le document modélisé.

linéaire). Il est couramment mis en œuvre en RI, notamment pour la recherche documentaire ou la catégorisation de documents (Salton et McGill, 1986). Nous présentons tout d'abord le principe du modèle vectoriel. Nous discutons ensuite de sa mise en œuvre pour la détection de dérivation en exploitant les mots ou les éléments stylistiques.

Le modèle vectoriel (ou *Vector Space Model* (VSM)) est une projection algèbrique du contenu d'un document sous la forme d'un vecteur v. Chaque composante i du vecteur correspond à un élément du vocabulaire, cette association est arbitraire et n'impacte pas le modèle mais doit correspondre à une application bijective (tous les indices du vecteur doivent être liés à un et un seul élément du vocabulaire). La taille du vecteur correspond au nombre d'éléments dans le vocabulaire du modèle. Le vocabulaire n'est pas ici synonyme de lexique, ces éléments peuvent être des mots, des groupes de mots ou n'importe quelle entité linguistique. Chaque valeur v_i correspond à un poids affecté à l'élément du vocabulaire associé à l'index i. Un exemple de poids est le nombre d'occurrences de l'élément du vocabulaire dans un document comme l'illustre la figure 2.4, d'autres pondérations permettent de normaliser par rapport à la taille du document (tf...), la diffusion dans le corpus (idf...), la représentativité de l'élément (tf · idf, BM25...) ou à d'autres considérations selon la tâche.

L'efficacité de la mise en œuvre d'un modèle vectoriel repose en grande partie sur le choix du vocabulaire utilisé et sur la qualité de la pondération. Dans le cadre de la détection de dérivation, les mots et les éléments de styles ont été utilisés comme vocabulaire. Nous les détaillons dans les paragraphes suivants. En ce qui concerne la pondération, la fonction tf · idf (*cf. Équation 5.4*) est majoritairement utilisée (Clough, 2003a; Özlem Uzuner et Davis, 2003) ainsi que BM25 dans une moindre mesure (Robertson et collab., 1999). D'après Hose (2003), le choix de l'une ou l'autre mesure a peu d'impact sur les résultats.

Les mots sont les éléments de vocabulaire les plus communément utilisés pour la détection de dérivation par modèle vectoriel. Tous les mots du texte considéré peuvent être retenus (Clough, 2003a, p.152) ou bien être soumis à un filtrage des mots outils (Hose, 2003; Hoad et Zobel, 2002) ou une autre forme de sélection (p. ex. en utilisant idf (*cf. Équation 5.2*) (Chowdhury et collab., 2002)). Outre une normalisation de la casse (*case-folding*), d'autres normalisations peuvent être apportées aux formes textuelles utilisées dans la modélisation. Ainsi, Si et collab. (1997) et Hoad et Zobel (2002) ont expérimenté la racinisation et observé que celle-ci entraînait en général une dégradation des performances sur l'anglais. Özlem Uzuner et Davis (2003) a proposé

une normalisation sémantique en regroupant les mots en classes sémantiques et en utilisant cette dernière comme dimensions du modèle vectoriel.

Les éléments stylistiques sont plus rarement utilisés comme éléments de vocabulaire. Özlem Uzuner et Davis (2003); Uzuner et collab. (2005) ont expérimenté l'utilisation de plusieurs éléments stylistiques : distributions statistiques (tailles des mots et phrases, rôles grammaticaux), constructions syntaxiques utilisées, formes des modaux et des négations ou encore classes sémantiques et syntaxiques des verbes.

Si la modélisation ensembliste est la principale modélisation pour les approches par couverture de texte, le modèle vectoriel est son pendant pour les modélisations globales. La divergence du modèle entraîne l'utilisation d'autres mesures de similarité adaptées à la comparaison de vecteurs.

2.2.3.2 Mesures de similarité entre vecteurs

De manière analogue aux approches par couverture de texte, la comparaison des modélisations est assurée par des mesures de similarité. Elles associent un score de similarité, généralement compris entre 0 et 1, à une paire de vecteurs. Le cosinus est la mesure la plus utilisée mais d'autres mesures algébriques ainsi que des mesures probabilistes ont également été expérimentées.

La mesure la plus commune pour le modèle vectoriel est le cosinus (*cf. Équation 2.5*), rapport du produit scalaire des vecteurs (*cf. Équation 2.3*) et du produit de leur distance Euclidienne (ou l^2-norme) (*cf. Équation 2.4*). Elle a notamment été utilisée par Shivakumar et Garcia-Molina (1995); Hose (2003); Clough (2003a); Stein (2005).

$$a \cdot b = \sum_{i=0}^{|a|} a_i b_i \tag{2.3}$$

$$||a|| = \sqrt{\sum_{i=0}^{|a|} a_i^2} \tag{2.4}$$

$$\cos(a,b) = \frac{a \cdot b}{||a||||b||} \tag{2.5}$$

Un certain nombre d'autres mesures algébriques ont également été expérimentées telles que le simple produit scalaire ou le produit scalaire normalisé (Hoad et Zobel, 2002). Shivakumar et Garcia-Molina (1995) ont également introduit le cosinus asymétrique (*cf. Équation 2.6*) qui, de manière analogique à ce que fait le containment (*cf. Équation 2.2*), pondère le produit scalaire par la norme d'un seul vecteur. Selon les besoins, cette mesure asymétrique peut être transformée en une mesure symétrique (*cf. Équation 2.7*).

$$\text{asymcos}(a,b) = \frac{a \cdot b}{||a||} \tag{2.6}$$

$$\text{sim}_{\text{asymcos}}(a,b) = \max(\text{asymcos}(a,b), \text{asymcos}(b,a)) \tag{2.7}$$

Enfin, certains travaux ont expérimenté des mesures probabilistes telles que le log-likelihood (Metzler et collab., 2005; Yang, 2006b) ou la *KL divergence* (Yang, 2006a), mais également des algorithmes d'apprentissage tels que les réseaux bayésiens (Clough, 2003a) et les arbres de décision (Özlem Uzuner et Davis, 2003; Uzuner et collab., 2005).

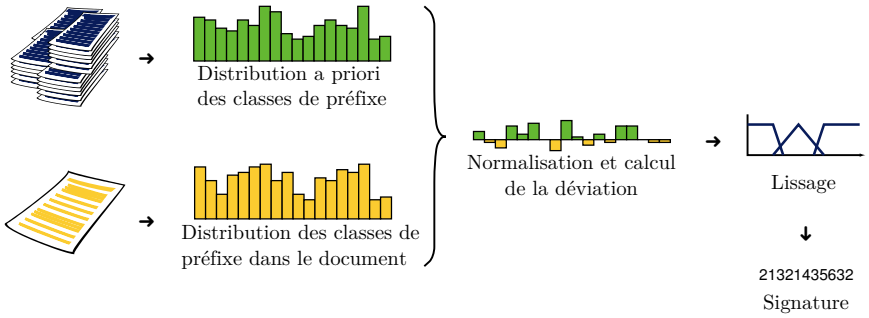


FIGURE 2.5 – Processus de création d'une signature floue d'un document étant donné une collection de référence. Schéma dérivé de Stein (2005, Fig.2).

2.2.3.3 Modèle à fréquences relatives

Le modèle à fréquences relatives (*Relative Frequency Model* (RFM)) a été originalement proposé par Shivakumar et Garcia-Molina (1995, p.5), puis repris par Hoad et Zobel (2002) et Metzler et collab. (2005). Les auteurs l'introduisent comme une alternative au modèle vectoriel couplé à la mesure de *cosinus* qui a des propriétés incompatibles avec la détection de copie : (i) l'indépendance du modèle par rapport au nombre réel d'occurrences d'un mot dans le document, et (ii) son incapacité à détecter le recouvrement de parties du texte.

Pour résoudre le premier problème, seuls les mots qui apparaissent un nombre comparable de fois dans les deux documents sont retenus pour la comparaison. Ce *closeness set* (cs) de mots est défini par :

$$cs(d_1, d_2) = \begin{cases} w_i \in \Pi(d_1) \cap \Pi(d_2), \Pi \text{ projection de l'ensemble des termes} \\ \epsilon - (\frac{\text{occ}(w_i,d_1)}{\text{occ}(w_i,d_2)} + \frac{\text{occ}(w_i,d_2)}{\text{occ}(w_i,d_1)}) > 0 \\ \epsilon \in]2; +\infty[, \text{ défini par l'utilisateur} \\ \text{occ}(w, d) \text{ le nombre d'occurrences du mot } w \text{ dans } d \end{cases} \tag{2.8}$$

Ensuite, le second problème est partiellement résolu par l'utilisation du *cosinus asymétrique*, similaire au containment (*cf. Équation 2.6*).

2.2.3.4 Signature floue

FUZZY-FINGERPRINT

La signature floue, ou *fuzzy-fingerprint* est une évolution plus récente de l'utilisation du modèle vectoriel pour la détection de dérivation introduit par Stein (2005) pour la recherche de proches-copies dans une collection. Il s'agit d'un autre contournement du problème soulevé par Shivakumar et Garcia-Molina (1995) de l'inaptitude du modèle vectoriel classique à comparer des éléments du vocabulaire qui n'apparaissent pas un nombre comparable de fois. Toutefois, plutôt qu'une normalisation à l'échelle de la paire de documents, Stein propose une normalisation à l'échelle de la collection de documents.

Le processus de création d'une signature floue étant donné un document et une collection de référence est décrit par la figure 2.5. La taille du vocabulaire est drastiquement réduite en ne considérant pas l'ensemble des termes, mais des classes de préfixe [13]. L'instance du modèle pour un document particulier est produit en trois

13. Les auteurs proposent ici une catégorisation triviale basée sur la présence d'un caractère alphabétique particulier, soit entre 10 et 30 classes de préfixes. Ils considèrent dans Stein et Eissen (2006)

phases : (i) construction d'un vecteur de fréquence pour les classes de préfixe ; (ii) calcul de la déviation de ce vecteur de fréquence par rapport à un vecteur calculé sur une collection de référence et (iii) lissage (*fuzzyfication*) par discrétisation des variations. La discrétisation des variations est laissée libre à l'utilisateur, l'idée étant d'associer un nombre fini de valeurs entières à des niveaux de variation. La signature floue consiste alors en ce vecteur de valeurs discrètes.

2.2.3.5 Discussion

En résumé, les approches par modélisation vectorielle se différencient les unes des autres par (i) le vocabulaire employé, (ii) la pondération de ce vocabulaire et (iii) la mesure de similarité utilisée pour comparer les vecteurs. Le modèle tel qu'employé en recherche d'information semble fonctionner pour la détection de proches-copies (Clough, 2003a). La détection de dérivation à différentes granularités nécessite toutefois la mise en place de systèmes de catégorisation des mots et des mesures de similarité asymétriques (Shivakumar et Garcia-Molina, 1995; Stein, 2005). Les dérivations de nature stylistique peuvent également être détectées en employant un vocabulaire décrivant le style (Özlem Uzuner et Davis, 2003) à la manière des méthodes de détection intrinsèque.

Les approches par modélisation vectorielle sont des approches de *mesure globale de la similarité* en ce que les modèles abstraient complètement le document et cassent notamment la structure linéaire et séquentielle du texte. Ainsi, contrairement aux approches par couverture de texte, la découverte de similarité ne permet pas d'isoler les éléments similaires du texte. Il est nécessaire de découper les documents en segments et d'appliquer ces modèles sur ces segments pour s'approcher de mesures locales, ce que propose plus ou moins Stein (2005). Le système CHECK (Si et collab., 1997) propose une vision intermédiaire en utilisant une structure arborescente —chaque nœud correspond à une partie du texte— de signatures composées de mots clés racinisés.

SIMILARITÉ GLOBALE

Nous préférons pouvoir revenir au texte et notamment à l'endroit des *lieux de collision* afin d'avoir des analyses plus fines et pouvoir détecter des formes de dérivation à des granularités plus précises, ce que ne permettent pas ces approches. Toutefois, elles peuvent être utilisées en amont de la chaîne pour effectuer un premier filtrage[14] ou bien détecter des dérivations de contenu impliquant de nombreuses réécritures et une quasi-disparition des éléments de formes du document source.

2.2.4 Approches par alignement de sous-chaînes

La dernière famille d'approches utilisée pour la détection de dérivations entre des textes repose sur l'alignement de sous-chaînes. L'objectif de ces approches est de rechercher une correspondance entre des passages de façon à aligner les deux textes. Les méthodes peuvent aller jusqu'à calculer les opérations nécessaires pour obtenir un texte à partir d'un autre, ce qu'on appelle communément la distance d'édition (Crochemore et collab., 2007). Nous rapportons les différentes travaux applicables à la détection de dérivation : distances d'édition, algorithmes de recherche de sous-chaînes communes et méthodes d'alignement de passages. Finalement, nous discutons des avantages et inconvénients de ces approches.

---

que ces classes de préfixe peuvent être constituées par n'importe quelle fonction Π de sélection de segments de texte.

14. L'utilisation de ces modèles peut toutefois entraîner une augmentation du silence pour les dérivations impliquant de la réécriture du fait d'un filtrage trop approximatif.

2.2.4.1 Distances d'édition

Une méthode aussi simple que naïve a été mise en œuvre par Yang (2006a) pour filtrer les copies exactes. Elle consiste à normaliser la casse du texte, à supprimer tous les espaces et à appliquer une fonction de hachage (SHA1). Les textes obtenant la même valeur sont des répliques totalement exactes. Cette méthode a l'avantage de fonctionner directement sur les textes bruts sans nécessiter aucun traitement ou modélisation. Son champs d'action se limite toutefois à l'identification du cas spécifique des duplications exactes.

D'autres méthodes exploitent les distances d'édition. La distance d'édition entre deux textes représente le nombre d'opérations nécessaires pour transformer le premier en le second. La distance de Hamming (Hamming, 1950) mesure, pour deux chaînes de longueur identique, le nombre minimum de remplacements nécessaires pour transformer la première chaîne en la seconde. La distance de Levenshtein (Levenshtein, 1966) en est une extension puisqu'elle considère non seulement les substitutions, mais également les opérations d'insertion et de suppression de caractère. D'autres, telle que la mesure de similarité de Jaro-Winkler (Winkler, 1999), mise au point pour les courtes séquences, peuvent être utilisées par les algorithmes d'alignement de séquences. Crochemore et collab. (2007) présente un état de l'art complet de ces distances.

L'utilisation de ces distances d'édition est limitée à des pré-traitements, des textes très proches ou des dérivations très localisées (groupes nominaux, entités nommées...) de par leur complexité et leur inaptitude à capturer des modifications plus en profondeur.

2.2.4.2 Recherche de sous-chaînes communes

La recherche de sous-chaînes communes est un problème algorithmique largement traité dans la littérature pour ses propriétés de complexité, mais également pour son utilité applicative. Nous pouvons citer les algorithmes Aho-Corasick (Aho et Corasick, 1975), Knuth-Morris-Pratt (Knuth et collab., 1977), Boyer-Moore (Boyer et Moore, 1977) ou encore Rabin-Karp (Karp et Rabin, 1987b).

La méthode *Greedy-String-Tiling* (Wise, 1996; Clough, 2003a) tire partie de ces algorithmes. Elle consiste à identifier les sous-chaînes communes à une paire de textes : les tuiles (*tiles*). En règle générale, les textes dérivés partagent des sous-chaînes plus longues que les textes non-dérivés (Clough, 2003b). L'efficacité de cette approche est assez aléatoire que ce soit par sa sensibilité aux réécritures ou pour le risque d'explosion exponentielle.

2.2.4.3 Alignement de passages

La dernière série de méthodes exploitées pour la détection de dérivation est celle de l'alignement de passages de texte. Nous présentons tout d'abord les algorithmes génériques d'alignement, puis nous détaillons ceux mis en œuvre pour la détection de dérivation. Nous différencions ceux qui alignent des séquences de texte en correspondance exacte de ceux qui alignent des passages qui ont la même signification.

Les méthodes d'alignement identifient des passages locaux identiques et les mettent en relation (*mapping*). L'étude des relations entre ces segments textuels alignés permet de prendre une décision concernant les paires de textes impliquées. Il existe des algorithmes génériques pour l'alignement de séquence utilisés en traduction. Nous pouvons citer l'algorithme de Needleman–Wunsch (Needleman et Wunsch, 1970) ou celui de Smith-Waterman (Smith et Waterman, 1981), qui fonctionnent par maximisation de la similarité entre des sous-chaînes. Smith-Waterman a été expérimenté dans le cadre de la détection de dérivation par Irving (2004). L'algorithme TESAS

utilisé par Piao et Mcenery (2004); Clough et collab. (2002) fonctionne de manière similaire mais considère l'alignement au niveau de la phrase. Il a également été mis en œuvre par Hose (2003) pour recroiser la Bible avec les manuscrits de Qumrân [15].

Les approches spécifiques à la détection de dérivation modélisent les textes par des arbres à suffixes et emploient des mesures tirées de la théorie des graphes. L'objectif est d'isoler les sous-arbres en correspondance qui représentent des passages identiques dans les documents d'origine. Ainsi, Monostori et collab. (2000) construisent [16] un arbre des suffixes mots [17] pour chaque document à comparer. Ils calculent les probabilités de correspondance à l'aide de l'algorithme de Chang et Lawler (1994). Cet algorithme parcourt un des arbres à partir d'un premier nœud et calcule la probabilité de correspondance en traversant tout le second arbre. Les calculs de probabilité de correspondance à partir d'autres positions dans le premier arbre sont effectués par *backtracking* et parcours descendant. Les sous-arbres qui ont les plus hautes probabilités de correspondance sont retenus. Cette approche est implémentée dans le système *MatchDetectReveal* (Monostori et collab., 2001).

Les approches par alignement présentées précédemment ne permettent d'aligner que des segments en correspondance exacte ou très proches. Toutefois, plusieurs mesures de similarité sémantique pourraient être utilisées pour aligner des candidats prenant des formes différentes tout en utilisant les algorithmes d'alignement classiques (voir Metzler et collab. (2007) et Pirró (2009) pour des états de l'art). Le tout est de définir ce qui doit être considéré comme similaire. Pour Pirró (2009), la notion de similarité est plus subtile qu'il n'y paraît. Il différencie notamment ce qui est semblable (p. ex. un vélo et une voiture) de ce qui est lié (p. ex. une voiture et de l'essence). Pour Hatzivassiloglou et collab. (1999), deux unités textuelles sont similaires si « elles s'intéressent à un concept, agent, objet ou une action commune »et que « cet agent ou objet commun doit être impliqué dans ou sujet à la même action, ou être le sujet de la même description ». Les auteurs proposent, compte tenu de cette définition, de caractériser les textes à l'aide des mots simples et des groupes nominaux non récursifs mis en correspondance selon leurs cooccurrences, leurs têtes (pour les groupes nominaux), les liens de synonymie dans WordNet, les classes sémantiques de verbes et le partage des noms propres. Toutes ces approches n'ont pas été expérimentées en condition réelle pour la détection de dérivation, mais uniquement sur des textes courts à cause du coût prohibitif de leur mise en œuvre. Elles restent toutefois intéressantes d'un point de vue théorique.

2.2.4.4 Discussion

Les approches par alignement consistent à rechercher les passages identiques ou suffisamment similaires entre deux textes. Lesdits passages sont alors mis en correspondance et concrétisent une relation de dérivation entre les textes. Les approches par alignement ont certainement été les moins expérimentées et leur mise en œuvre ne s'inscrit pas dans un cadre théorique homogène comme c'est le cas pour les approches par couverture de texte ou par similarité de mots-clés.

Le principe des approches par alignement est que les textes dérivés partagent de longues séquences textuelles. Si cette hypothèse est indéniablement vérifiée pour les cas particuliers de la duplication et éventuellement de la presque-duplication, elle

15. Plus connus sous l'appelation manuscrits de la mer Morte : `http://en.wikipedia.org/wiki/Dead_Sea_scrolls`

16. Les auteurs utilisent l'algorithme d'Ukkonen qui, outre sa simplicité, conserve les liens entre les suffixes.

17. Par « suffixe mot », nous entendons que l'unité indivisible du suffixe n'est pas le caractère mais le mot. Les étiquettes des arcs sont donc au minimum des mots et non des caractères. Les auteurs justifient ce choix par des besoins de performance.

est plus difficile à défendre pour d'autres formes de dérivation pour lesquelles les réécritures et autres réarrangements structurels remodèlent la forme textuelle. De plus, les algorithmes d'alignement nécessaires à la mise en œuvre de ces approches sont inadaptés ou coûteux. Les algorithmes efficaces permettent uniquement l'alignement de séquences identifiques ou extrêmement proches, ce qui est inadapté aux formes de dérivation impliquant des réécritures importantes. Les algorithmes capables de capturer des modifications plus importantes sont très coûteux et sont par conséquent inadaptés au traitement de masses importantes de données qui est le cadre applicatif principal de la détection de dérivation. Ils pourront toutefois être mis à profit dans des cas particuliers tels que l'identification de collusion au sein d'un groupe d'étudiants.

2.3 Conclusion

Nous avons présenté dans ce chapitre les différentes méthodes expérimentées dans la littérature pour des tâches assimilables à la détection de dérivation. Deux grandes approches se dessinent. La détection intrinsèque tout d'abord où les documents sont considérés indépendamment les uns des autres et la détection extrinsèque ensuite où les documents sont considérés par paires afin de chercher dans l'un des traces de l'autre. Les méthodes de détection intrinsèque reposent sur la découverte d'indices contextuels (citations) ou de ruptures (caractérisation du style dans une fenêtre glissante). Les méthodes de détection extrinsèque se divisent en trois grandes familles : le calcul de recouvrement de texte (découpage des textes en *w-tuiles* et recherche des *w-tuiles* communs aux deux textes), la modélisation vectorielle largement inspirée de la RI (comparaison de vecteurs de termes ou autres éléments du texte) et les méthodes d'alignements (alignement générique ou calcul de recouvrement d'arbres à suffixes).

Toutes ces approches partagent à notre avis un biais commun ainsi qu'une limitation systémique. Le biais de toutes ces approches est de ne considérer qu'une seule dimension du texte dans la recherche des dérivations. Les méthodes par couverture de texte et par alignement ne considèrent que la forme (exception faite des quelques initiatives impliquant du filtrage, de la normalisation ou tout autre rapprochement sémantique) tandis que les méthodes par modélisation vectorielle ne représentent que le contenu général des textes. Or, nous avons montré dans notre proposition de modélisation (*cf. Section 1.4*) que la dérivation pouvait porter sur des éléments de diverses natures (forme, contenu, structure, style). Il nous semble alors pertinent de prendre en compte ces différentes dimensions du texte afin de capturer plus justement les différentes manifestations des formes de dérivation.

La limitation systémique concerne la prise de décision finale : y-a-t-il dérivation ? quels éléments/passages sont impliqués ? Les réponses à ces deux questions reposent sur des scores de similarité, des déviations ou des probabilités. Lee et collab. (2005) ont évalué la corrélation entre les scores de plusieurs systèmes et la similarité jugée par des humains. Ils ont montré que cette corrélation était faible et que par conséquent le score mesuré ne correspondait pas réellement à un jugement de similarité tel que nous le produisons. Nous ne prétendons pas apporter de réponse tranchée aux questions susmentionnées, mais les résultats de l'étude de Lee et collab. (2005) montrent qu'elles méritent clairement d'être prises en considération.

Dans le cadre de cette thèse, nous explorons les deux méthodes de détection : la détection intrinsèque pour les citations et la détection extrinsèque pour diverses autres formes de dérivation. En ce qui concerne la détection intrinsèque, nous expérimentons dans le chapitre 3 l'identification de passages citationnels en exploitant les indices contextuels. Dans ce cadre, nos travaux reposent en grande partie sur des travaux existants.

En ce qui concerne la détection extrinsèque, nous explorons plus en profondeur les approches par calcul de la couverture de texte par le biais de modélisations ensemblistes que nous nommons *signatures*. Toutefois, si nous nous cantonons résolument à une comparaison des textes sur la base d'éléments textuels (à une normalisation morphologique près), nous cherchons également à capter des éléments qui décrivent le contenu afin de couvrir plusieurs dimensions du texte. Nous souhaitons identifier des familles d'éléments, ce que nous nommons les *descripteurs*, qui soient très représentatives des textes tout en étant préservées par les processus de dérivation. En effet, nous pensons que l'efficacité d'une méthode de détection par calcul de couverture de texte repose sur la modélisation des textes. Nous pensons qu'une bonne modélisation est :

SIGNATURE

DESCRIPTEUR

- une modélisation représentative du texte : deux textes non-dérivés doivent avoir des modélisations différentes ;
- une modélisation stable malgré la mise en œuvre d'un processus de dérivation : deux textes dérivés doivent avoir des modélisations identiques ou du moins très proches.

Enfin, nous considérons que le problème de détection de dérivation doit être traité par un système d'aide à la décision et non un système de prise de décision. Un tel système doit selon nous proposer à l'utilisateur une liste, éventuellement ordonnée, de suspects ainsi que des éléments justifiant leur état de suspect.

Chapitre 3

Une approche intrinsèque pour la détection des citations

Une citation est une figure particulière, inaccomplie mais indispensable. Elle ouvre un écart entre ce qu'on vient de dire et ce qu'on va dire. [...] Elle suggère une incomplétude tant il lui manque de contexte, et elle complète le sens au sein duquel elle s'inscrit.

— Pierre-Yves Bourdil, Faire la philosophie (1996)

Sommaire

Nous avons posé les bases de la notion de dérivation (*cf. Chapitre 1*) et nous avons présenté les différentes méthodes utilisées pour détecter automatiquement ce phénomène (*cf. Chapitre 2*). Nous nous intéressons dans ce chapitre à la citation qui est une forme de dérivation par laquelle l'auteur place dans le texte produit des éléments permettant au lecteur de retrouver ses sources. Les citations sont des reprises contextualisées pour lesquelles les techniques de détection intrinsèque sont particulièrement adaptées.

Plusieurs motivations dirigent les travaux sur la détection de citation. Pour certains, il s'agit de mesurer l'impact du travail d'un chercheur (mesures bibliométriques) (Hirsch, 2005), pour d'autres d'améliorer les systèmes de recherche d'information (indexation des documents par les citations) (Ritchie et collab., 2006), de question-réponse (« Que pense X au sujet de Y ? ») (Somasundaran et collab., 2007). Finalement, pour d'autres encore, il s'agit de collecter des opinions sur différents sujets (Wilson et collab., 2005; Kim et collab., 2006), potentiellement à des fins commerciales (suivi de l'impact d'un produit) (Dave et collab., 2003; Beauchene et collab., 2002). Ces travaux couvrent plusieurs tâches : le repérage de la citation, son étendue dans le texte ou encore son analyse en contexte (subjectivité, polarité, ...). Néanmoins, les tâches couvertes varient selon la langue, le genre de texte et la motivation applicative.

Dans le cadre de cette thèse, nous nous sommes concentrés sur le genre de l'article de presse en français. Nous avons travaillé sur trois tâches :

1. l'identification des segments porteurs de discours direct ;
2. l'identification des expressions textuelles faisant référence aux sources citées ;
3. le regroupement de ces deux composants afin de capturer ce que l'on nomme une citation.

Nous nous tournons vers une identification indépendante des composants citationnels par la mise en œuvre de méthodes d'apprentissage automatique, puis leur regroupement, à l'aide d'une heuristique, en ce que nous appelons un segment citationnel.

Dans un premier temps, nous détaillons les approches existantes qui reposent sur une exploration contextuelle ou bien sur la reconnaissance de motifs citationnels avant de discuter leurs limites (*cf. Section 3.1*). Dans un deuxième temps nous présentons notre approche basée sur l'identification indépendante par apprentissage des objets citationnels que sont les segments dérivés et les expressions locutrices. Nous y détaillons notamment la construction des corpus nécessaires à l'apprentissage et à l'évaluation (*cf. Section 3.2*). Dans un troisième temps, nous rapportons les résultats expérimentaux de notre approche pour chacune des tâches auxquelles nous nous sommes intéressés (*cf. Section 3.3*). Finalement, nous concluons sur notre approche en proposant plusieurs pistes d'améliorations (*cf. Section 3.4*).

3.1 Approches de détection pour le français

La détection automatique de citations couvre deux tâches : le repérage des passages du texte contenant une ou plusieurs citations et l'identification des éléments constituant la citation (discours, éléments de contexte tels que l'auteur, le lieu, la date...). Cette dernière tâche permet de répondre aux questions : Qu'est-ce qui est dit ? Par qui ?

Deux approches principales ont été développées pour le français (*cf. Section 2.1.2*). La première se base sur l'exploration contextuelle de marques de la citation à l'aide de dictionnaires (Mourad et Desclés, 2004). La seconde définit des motifs de citations constitués de composants citationnels que sont la source, le discours rapporté et le

relateur les liant (Giguet et Lucas, 2004). D'autres travaux se focalisent sur les citations au sein des articles scientifiques et utilisent des éléments extérieurs comme la bibliographie ou une base de données externe pour repérer les références au sein des textes (Teufel et collab., 2006). Nous nous intéressons aux deux premières approches dans le domaine journalistique.

3.1.1 Exploration contextuelle à l'aide de dictionnaires

Mourad et Desclés (2004) décrit l'exploration contextuelle comme une méthode qui consiste à caractériser un objet linguistique à l'aide d'éléments lexicaux appelés *indices* ou encore *embrayeurs* en vue de son repérage automatique. Lorsque l'indice est considéré comme fiable, il change de statut et devient un *marqueur*.

Mourad (2001) a utilisé cette approche pour décrire l'objet citation. Le corpus utilisé est constitué d'articles journalistiques mais également de textes scientifiques, techniques et littéraires. Son étude a permis de recenser « 3 000 formes verbales et introducteurs spécifiques » de la citation en français, classées en deux catégories : les marqueurs typographiques et linguistiques (Mourad et Desclés, 2002).

Les *marqueurs typographiques* relèvent principalement de la ponctuation. La séquence ponctuative *: «* (*deux-points ouverture des guillemets*) prédomine naturellement mais d'autres marques ont été relevées : les incises elliptiques entre guillemets et les points d'exclamation ou d'interrogation aux abords de guillemets fermants.

Les *marqueurs linguistiques* se déclinent en trois sous-catégories. Premièrement, les syntagmes prépositionnels (*selon X, pour X, d'après X, aux yeux de X...*) (*cf. Exemple 13*) et les introducteurs spécifiques (*l'observation de X, les termes de X...*) permettent d'introduire le locuteur. Le *X* est substitué dans les textes par une expression faisant référence au locuteur. Deuxièmement, les indices de référence discursifs (*il, son, elle, etc.*) et la complétive *que* sont des indices linguistiques complémentaires (*cf. Exemple 14*) qui marquent la présence de discours repris. Troisièmement, les verbes de communication introduisent le discours repris. Mourad (2001) regroupe presque 800 verbes d'introduction et en propose une classification selon le degré d'engagement de l'auteur.

> D'après sa mère, Julien faisait une sieste dans l'appartement familial lorsqu'il a disparu.
>
> © *Le Figaro – 20 Février 2007*

EXEMPLE 13: La présence d'un syntagme prépositionnel (*D'après sa mère,*) est également une marque de la présence d'une citation.

> "En 2003, explique-t-il , j' ai fait effectuer douze tests sur des vols en France, dans onze cas sur douze, des armes et explosifs ont pu être introduits dans les avions".
>
> © *Le Figaro – 20 Février 2007*

EXEMPLE 14: La présence d'incises elliptiques (*, explique-t-il,*) et de pronoms personnels (*il*, *j'*) au sein d'un couple apparié de guillemets est un marqueur de citation.

Les auteurs ne donnent pas d'informations sur les performances des algorithmes élaborés à partir de leur travail. Nous pouvons supposer que le grand nombre de marques introduit du bruit du fait de leur ambiguïté. De plus, l'approche reste dépendante de la langue étudiée de par les nombreuses ressources linguistiques sur lesquelles elle repose.

3.1.2 Reconnaissance des composants constitutifs de la citation

Giguet et Lucas (2004) proposent de modéliser la citation selon trois *composants constitutifs* que sont *la source*, *le discours rapporté* et *le relateur*. Ils ramènent le problème d'identification d'une citation au repérage et au raprochement de ces trois composants. Leur méthode s'articule autour des composants constitutifs de la citation, leur ordre et les critères d'identification de ces composants.

COMPOSANTS CONSTITUTIFS DE LA CITATION

La *source* est le « nom propre du locuteur et éventuellement ses qualifiants ». Cette expression fait référence de manière non ambiguë au locuteur qui prend en charge le discours rapporté. Le *discours rapporté* correspond aux expressions textuelles reprises de la source. Enfin, le *relateur* est un « segment établissant la relation entre la source et le discours rapporté ». Il peut être « verbal, conjonctif ou prépositionnel ». L'exemple 15 ci-dessous montre la place de ces composants dans une citation et leur potentielle réalisation textuelle.

> Mais [Brice Hortefeux]$_{source}$ [précise que]$_{relateur}$ « [le seul objectif de cette fusion, c'est de parvenir à une plus grande synergie et à une plus grande coordination de ceux qui s'expriment au nom du candidat]$_{disc.rapp.}$ ».
>
> © *Le Figaro*

EXEMPLE 15: Les différents composants citationnels.

En plus des composants constitutifs, la méthode spécifie deux ordres d'apparition de ces composants pour le français : (i) l'ordre normal (source + relateur + discours rapporté) et (ii) l'ordre inversé (discours rapporté + relateur + source). D'après les auteurs, ces deux ordonnancements des composants sont les seuls qui correspondent à une citation.

Finalement, le repérage des citations est simplifié étant donné que leurs composants constitutifs et l'ordre d'apparition de ces derniers sont connus. L'identification des différents composants profite de la présence à proximité de composants identifiés (indice positionnel). Par exemple, si une source et un discours rapporté sont identifiés au sein de la même phrase, la recherche du relateur est dirigé par la forme de la citation (ordre normal ou inversé). Cette matrice de la citation permet de limiter les besoins en ressources linguistiques importantes au profit de marques de surface. Ces marques de la présence d'un composant reposent sur la co-présence de différents indices au sein d'une même phrase : les indices typographiques (*ponctuation, casse...*), les indices morpho-syntaxiques (*morphèmes grammaticaux comme « que », suffixes « ent »...*) et les indices positionnels (*début, fin d'unités textuelles...*).

Les auteurs ne donnent pas d'informations sur les performances de leur méthode. Elle semble plus facilement adaptable à d'autres langues sous réserve de la stabilité des composants constitutifs et de leur ordonnancement.

3.1.3 Synthèse

Deux des méthodes proposées dans la littérature reposent sur l'identification de marques de natures diverses (textuelles, typographiques, lexicales, syntaxiques...). La première approche (Mourad et Desclés, 2004) à base de lexiques permet d'identifier ce qui est dit, la seconde (Giguet et Lucas, 2004) permet également d'identifier qui le dit. Ces approches se différencient par le choix des éléments qui caractérisent la citation. La première préfère utiliser des éléments lexicaux précis et assez spécifiques de la citation. La seconde approche repose sur un découpage de la citation en composants

constitutifs identifiables à l'aide de marques de surface. Ces deux approches sont liées à la langue même si la seconde semble plus facilement transposable dans une autre langue que la première.

L'autre problème soulevé par les approches proposées est l'absence d'évaluation. Comment se comportent-elles face à des énoncés ambigus ? Quelle est la couverture réelle de ces approches ? Ainsi, aussi précises et spécifiques de la citation qu'elles soient, les marques restent ambiguës. Une simple recherche lexicale n'est donc pas suffisante et peut induire des erreurs. Par exemple, l'introducteur spécifique « l'observation de » considéré par Mourad (2001) comme un marqueur n'en est pas un dans l'expression « l'observation de la nature ». De plus, la liste des motifs citationnels proposée pour le français par Giguet et Lucas (2004) nous semble incomplète (*cf. Annexe C*). Par conséquent, certaines formes de citations pourraient être passées sous silence.

De plus, certains flous subsistent comme la façon dont les marqueurs doivent être combinés pour délimiter les frontières des citations ou comment les indices permettent de délimiter les segments de texte puisque certains indices marquent à la fois la présence d'un des composants, et le début ou la fin d'un autre. Par exemple, le « que » est un indice marquant la présence potentielle d'un relateur aussi bien que le début d'un potentiel discours rapporté.

En résumé, nous proposons de combiner les composants constitutifs de Giguet et Lucas (2004) et les ressources linguistiques disponibles de Mourad et Desclés (2004), et d'adopter une approche par apprentissage supervisé. L'apprentissage permet de simplifier la portabilité de la méthode à une autre langue en extrayant automatiquement les règles pertinentes sous réserve de la constitution d'un corpus où les citations sont annotées pour la langue cible et de ressources linguistiques minimales. Nous constituons nous-même un corpus pour le français et pour l'anglais afin de mettre en œuvre notre approche et de l'évaluer.

3.2 Détection probabiliste des composants citationnels

Nous proposons dans cette section une approche de détection automatique des citations qui tire profit des travaux de l'état de l'art. Nous identifions la citation en détectant ses composants constitutifs à l'aide des différents indices décrits dans l'état de l'art. Toutefois, plutôt que d'élaborer manuellement des règles de combinaison de ces derniers, nous nous tournons vers un apprentissage automatique. Le rôle des algorithmes d'apprentissage est double : identifier les indices et marqueurs les plus discriminants et extraire des règles de combinaison des indices (construction des classifieurs). Le choix de l'apprentissage a pour but de faciliter la portabilité de notre approche vers d'autres langues et d'autres domaines. Nous l'avons d'ailleurs mis en œuvre à la fois pour le français et pour l'anglais (*cf. Section 3.3*).

Nous formalisons tout d'abord notre approche non plus comme un problème de détection mais comme un problème de catégorisation afin de pouvoir exploiter les techniques d'apprentissage (*cf. Section 3.2.1*). Nous discutons ensuite de la constitution des ressources nécessaires à la fois à la mise en œuvre des algorithmes d'apprentissage et de l'évaluation (*cf. Section 3.2.2*).

3.2.1 Catégorisation des sources et discours rapporté

Le repérage des citations dans un texte est abordé par les méthodes de l'état de l'art comme un problème de détection. En faisant le choix d'utiliser des techniques

d'apprentissage, nous changeons de paradigme et considérons la tâche de repérage de citations comme un problème de catégorisation. La catégorisation consiste à classer des objets dans des catégories (ou classes) prédéfinies (Manning et Schütze, 1999). Nous divisons tout d'abord notre problème en deux sous-problèmes de catégorisation : l'un pour l'identification des composants source et l'autre pour les composants de discours rapporté. Puis, nous décrivons pour chacun de ces sous-problèmes les objets à classer et les catégories associées.

3.2.1.1 Division en deux sous-problèmes de catégorisation indépendants

La mise en œuvre d'un problème de catégorisation nécessite d'identifier les objets à classer. Cependant, la notion de citation en tant qu'objet linguistique est difficile à appréhender car difficile à isoler dans le texte. Cette observation est particulièrement vraie pour le genre de l'article de presse. En effet, les journalistes opèrent des transformations afin de raccourcir les discours trop volumineux à rapporter tout en conservant des passages verbatim. Les segments textuels dérivés sont alors constitués d'une juxtaposition de segments reformulés et de passages *verbatim*. Nous devons introduire un objet textuel identifiable de manière opérationnelle pour la mise en œuvre de notre approche à base d'apprentissage.

Nous introduisons un objet textuel à la fois identifiable de manière opérationnelle dans le cadre de notre problème de catégorisation et également suffisamment proche de la notion de citation : le *segment citationnel*. Nous tirons directement profit du schéma générique de la citation constitué autour des composants constitutifs que sont la source, le relateur et le discours rapporté. En effet, nous définissons le segment citationnel comme la portion contiguë de matériel textuel qui englobe au mieux tous ces différents composants.

SEGMENT CITATIONNEL

Objets citationnels textuellement ancrés	Segments citationnels associés
Autre motif de satisfaction pour M. Bayrou : selon *[source/ le sondage IFOP]*, *[discours rapporté/83% des Français seraient "favorables à un gouvernement d'union nationale composé de personnalités politiques de divers camps"]*. *[source/Le candidat de l'UDF]* prétend *[discours rapporté/transcender le clivage droite-gauche qu'il juge "préhistorique"]*. Dimanche, *[source/il]* a déclaré qu'*[discours rapporté/il pourrait nommer un premier ministre de gauche, s'il était élu président de la République]*.	Autre motif de satisfaction pour M. Bayrou : selon *[Segment citationnel/le sondage IFOP, 83% des Français seraient "favorables à un gouvernement d'union nationale composé de personnalités politiques de divers camps"]*. *[Segment citationnel/Le candidat de l'UDF prétend transcender le clivage droite-gauche qu'il juge "préhistorique"]*. Dimanche, *[Segment citationnel/il a déclaré qu'il pourrait nommer un premier ministre de gauche, s'il était élu président de la République]*.

EXEMPLE 16: Annotation sur un texte exemple des composants source et discours rapporté et les segments citationnels correspondants.

Dans les faits, nous ne nous intéressons qu'aux composants source et discours rapporté qui forment le cœur de la citation. Nous délaissons le composant relateur qui a une réalisation textuelle trop variable (*cf. Annexe B*). Le problème de catégorisation du segment citationnel se divise alors en deux sous-problèmes : la catégorisation des composants source et discours rapporté. La reconstruction du segment citationnel à partir de ces deux composants relève d'un simple calcul de couverture de texte. L'exemple 16 illustre le lien entre ces deux composants et le segment citationnel sur un cas pratique. Nous détaillons par la suite les problèmes de classification des

composants source et discours rapporté et notre proposition de résolution.

3.2.1.2 Catégorisation des composants source

La source est l'un des trois composants constitutifs de la citation, elle correspond au « nom propre du locuteur et éventuellement ses qualifiants »(Giguet et Lucas, 2004). Le sous-problème de catégorisation des composants source consiste à identifier les objets qui peuvent correspondre à des composants source, les candidats, et à les catégoriser à l'aide de critères opérationnels.

Choix des candidats En premier lieu, nous devons identifier les objets qui seront passés en entrée du classifieur. Ces objets sont soumis à deux contraintes : (i) ils doivent pouvoir être extraits automatiquement et (ii) ils doivent couvrir un maximum des composants sources si ce n'est tous. Nous avons choisi les entités nommées comme candidats à la catégorisation. Ces objets linguistiques respectent les deux contraintes auparavant énoncées.

Premièrement, l'extraction d'entités nommées est une technologie mature qui obtient des niveaux de performance de l'ordre de 90 % en précision et en rappel sur les articles de presse en anglais (MUC, 1993). Nous utilisons l'extracteur d'entités nommées Némesis (Fourour, 2004). Bien que celui-ci ait été initialement développé pour le français nos résultats montrent qu'il peut aussi s'avérer efficace pour l'anglais.

Deuxièmement, nos observations en corpus ont montré que les composants source prennent un nombre limité de formes. L'observation en corpus (*cf. Annexe C.2*) montre qu'il s'agit majoritairement de composés nominaux ou des noms propres capitalisés qui correspondent à des entités nommées. Plus précisément, la première introduction des sources se fait au travers d'une forme explicite et précise : des entités nommées. Cette forme est rappelée par la suite par des références anaphoriques pronominales ou nominales, ou bien des formes contractées de la forme précise originale (abbréviations...). Ainsi, pour les corpus français et anglais que nous avons étudié, les composants sources sont composés d'entités nommées dans respectivement 55 % et 67 % des cas.

Catégorisation En second lieu, nous devons catégoriser les candidats ci-avant présentés (les entités nommées). L'objectif de la catégorisation est de différencier les candidats qui correspondent à des composants source de ceux qui n'en sont pas. La tâche correspondante est une classification non hiérarchique[1] de telle sorte que les candidats ne soient catégorisés que dans une seule classe et qu'il n'y ait donc pas de recouvrement des classes. Nous proposons deux classes : les *expressions locutrices* et les *non locutrices*.

EXPRESSION LOCUTRICE

Les *expressions locutrices* sont les entités nommées utilisées par un auteur pour désigner un locuteur (*Brigitte Liberman* dans l'exemple 17). Les *expressions non locutrices* sont des entités nommées qui ne prennent pas ce rôle de locuteur et qui ne sont donc pas impliqués dans une citation (*La Roche-Posay* ou *Jean-Paul* dans l'exemple 17). Nous nous refusons à utiliser une classe *composant source* pour deux raisons. Tout d'abord, nous souhaitons conserver la différenciation entre la sortie du classifieur et les composants constitutifs de la citation. Ensuite, les composants source, s'ils contiennent une entité nommée, ne s'y limitent pas forcément. Ainsi, dans l'exemple 17, *Brigitte Liberman* n'est qu'une partie du composant source qui inclue également son titre (*la directrice générale*...). L'expression locutrice est donc une approximation opérationnelle du composant source et pas forcément un composant source en tant que tel.

1. Une telle classification est également appelée un partitionnement, ou *clustering* en anglais.

Entités nommées	Catégorisation
[EN/Brigitte Liberman], la directrice générale de *[EN/Cosmétique Active]* (*[EN/La Roche-Posay]*, *[EN/Vichy]*), rétorque : « Les grandes innovations ne peuvent pas naître chaque année. »[...] Mais elle admet que « *[EN/Jean-Paul]* nous poussant, nous pouvons faire encore mieux ».	*[LOC/Brigitte Liberman]*, la directrice générale de *[NLOC/Cosmétique Active]* (*[NLOC/La Roche-Posay]*, *[NLOC/Vichy]*), rétorque : « Les grandes innovations ne peuvent pas naître chaque année. »[...] Mais elle admet que « *[NLOC/Jean-Paul]* nous poussant, nous pouvons faire encore mieux ».

Exemple 17: Exemples d'entités nommées candidats et leur catégorisation correspondante (LOC pour expression locutrice et NLOC pour expression non locutrice).

Exemple	Caractérisation	
Brigitte Liberman , la directrice générale de Cosmétique Active (La Roche-Posay, Vichy), rétorque : « Les grandes innovations ne peuvent pas naître chaque année. »	Contexte phrastique	
	Dist. verbe d'énonciation	9
	Dist. préposition	∞
	Dans juxtaposée	Oui
	Entité Nommée	
	Type EN	Anthroponyme
	Taille en mots	2

Table 3.1 – Exemple de caractérisation d'une entité nommée pour sa catégorisation en expression locutrice ou non locutrice.

Critères opérationnels de catégorisation Nous proposons un certain nombre de critères opérationnels pour différencier les entités nommées appartenant à chacune des deux catégories. Nous parlons de critères opérationnels car ceux-ci doivent pouvoir être identifiés automatiquement. Nous nous concentrons sur des marques de surface ainsi que des marques lexicales et syntaxiques. Le tableau 3.1 illustre, sur un exemple, les différentes marques utilisées.

Nous distinguons deux types de marques : celles internes à l'entité nommée et celles qui apparaissent dans son contexte phrastique. En ce qui concerne les marques intrinsèques à l'entité nommée, nous considérons la taille en mots de l'entité nommée et son type. Nous distinguons cinq types d'entités nommées calqués sur ceux retournés par Némesis : les anthroponymes, les toponymes, les ergonymes, les praxonymes, les phénonymes ou encore les zoonymes (Daille et collab., 2000). Les anthroponymes correspondent à la notion prototypique de nom propre, c-à-d un nom de personne composé d'un prénom, d'un nom de famille ou encore d'un pseudonyme (*Léonard*, *Lincoln*, *Napoléon*...) ou bien un nom d'un groupe de personnes (*les Beatles*, *les Lumières*, *l'ONU*...). Les toponymes sont les noms des lieux (*France*, *Seine*, *Europe*). Les ergonymes sont les noms donnés aux objets et produits manufacturés, incluant les marques et les entreprises (*Da Vinci Code*, *la Joconde*, *iPod*...). Les praxonymes (ou pragmonymes) sont les noms des faits historiques, des évènements culturels mais encore des noms de maladies ou de périodes historiques (*la Révolution Française*, *le Paléolithique*...). Les phénonymes sont les noms donnés aux phénomènes naturels (*ouragan Katrina*, *comète de Halley*, *le soleil*...). Enfin les *zoonymes* sont les noms donnés aux animaux de compagnie.

En ce qui concerne les marques qui apparaissent dans le contexte phrastique de l'entité nommée, nous considérons les prépositions, les verbes de communication et

la structure syntaxique. Nous avons ainsi considéré la présence de syntagmes prépositionnels introducteurs d'espace énonciatif tels que *Pour X*, *D'après X*... Nous tirons également profit des verbes de communication dont une liste est proposée par Mourad et Desclés (2004). Si elle est une ressource linguistique lourde qui peut limiter la portabilité de notre méthode, nous en avons obtenu une traduction à moindre coût [2] du français vers l'anglais qui permet d'obtenir des résultats satisfaisants. De plus, des travaux récents proposent des méthodes semi-automatique d'extraction de ces verbes (Sagot et collab., 2010). Enfin, nous tenons compte du fait que l'entité nommée s'inscrive dans une proposition délimitée par des virgules à la façon d'une incise.

L'ensemble de ces marques sont communes au traitement du français et de l'anglais. Pour ce dernier nous avons également considéré la proximité d'un sous-ensemble des verbes de communication composé de *say* et *tell*.

3.2.1.3 Catégorisation des composants discours rapporté

Le discours rapporté est également l'un des trois composants constitutifs de la citation, il correspond aux expressions textuelles reprises de la source (Giguet et Lucas, 2004). Le sous-problème de catégorisation des composants discours rapporté consiste à identifier les candidats et à les catégoriser à l'aide de critères opérationnels.

Choix des candidats En premier lieu, comme pour les composants source, nous devons identifier les objets qui seront passés en entrée du classifieur. Ces objets sont soumis à deux contraintes : (i) ils doivent pouvoir être extraits automatiquement et (ii) ils doivent couvrir un maximum des composants discours rapporté si ce n'est tous. Nous avons choisi les passages entre guillemets comme candidats à la catégorisation qui respectent ces deux contraintes.

Premièrement, ce type de motif est assez simple à repérer automatiquement en considérant les guillemets comme délimiteur. La seule difficulté réside dans la gestion correcte de l'appariemment des guillemets. Nous n'avons pas traité les cas complexes d'enchassement, peu présents, où les guillemets ouvrant et fermant sont identiques [3].

Deuxièmement, nos observations en corpus ont montré que les composants discours rapporté prennent un nombre limité de formes, la majeure partie incluant des passages entre guillemets, du moins pour le français (*cf. Annexe C.1*). En effet, nous notons que dans notre corpus français 80 % des citations sont composées de segments entre guillemets. Dans notre corpus anglais, la proportion est légèrement plus faible (75 %) mais reste tout de même la plus courante [4].

Catégorisation En second lieu, nous devons catégoriser les candidats ci-avant présentés (les passages entre guillemets). L'objectif de la catégorisation est de différencier les candidats qui correspondent à des composants discours rapporté de ceux qui n'en sont pas. Tout comme pour les composants source, il s'agit d'une classification non hiérarchique sans recouvrement entre les classes. Nous proposons deux classes : les *segments dérivés* et les *non dérivés*.

SEGMENT DÉRIVÉ

Les *segments dérivés* désignent les zones de texte qui sont dérivées du texte source (*En 2003... dans les avions.* dans l'exemple 18). Les *segments non dérivés* sont des

2. Nous avons utilisé le système de traduction en ligne fourni par Google.

3. De tels guillemets utilisent le même signe typographique pour le guillemet ouvrant et fermant. Ainsi les guillemets en double chevrons (à la française) différencient le guillemet ouvrant («) du guillemet fermant (»), ce n'est pas le cas du guillemet double (") qui emploie le même symbole dans les deux cas.

4. Nous pensons que cette différence s'explique du fait que les styles direct et indirect quasi-textuel sont moins utilisés en anglais.

passages entre guillemets qui ne correspondent pas à une citation (*« ultimes »* dans l'exemple 18). Il s'agit dans la plupart des cas d'emphases. De la même façon que pour les composants source et pour les mêmes raisons, nous nous refusons à utiliser une classe *composant discours rapporté*. Nous différencions ainsi la sortie du classifieur et les composants constitutifs de la citation. De plus, les composants discours rapporté ne se limitent pas forcément au passage entre guillemets, notamment dans les cas des citations au style quasi-textuel comme l'atteste l'exemple 18 : *Elle était bien l'« organisatrice »... « pouvoir de représentation » de la ville.* Ces segments dérivés sont des instances potentiellement parcellaires du composant discours rapporté.

Passages entre guillemets	Catégorisation
Washington avance une estimation des réserves mondiales *[« ultimes »]* de pétrole à 2 275 milliards de barils.	Washington avance une estimation des réserves mondiales *[NDERIV/« ultimes »]* de pétrole à 2 275 milliards de barils.
Elle était bien l'*[« organisatrice »]* du concert. Ce concert était une activité de *[« service public »]*. Les agents qui ont commis des fautes disposaient d'un *[« pouvoir de représentation »]* de la ville.	Elle était bien l'*[DERIV/« organisatrice »]* du concert. Ce concert était une activité de *[DERIV/« service public »]*. Les agents qui ont commis des fautes disposaient d'un *[DERIV/« pouvoir de représentation »]* de la ville.
["En 2003, explique -t-il, j'ai fait effectuer douze tests sur des vols en France, dans onze cas sur douze, des armes et explosifs ont pu être introduits dans les avions"].	*[DERIV/"En 2003, explique -t-il, j'ai fait effectuer douze tests sur des vols en France, dans onze cas sur douze, des armes et explosifs ont pu être introduits dans les avions"]*.

Exemple 18: Exemples de passages entre guillemets candidats et leur catégorisation correspondante (DERIV pour les segments dérivés et NDERIV pour les non dérivés).

Critères opérationnels de catégorisation Nous proposons un certain nombre de critères opérationnels pour différencier les passages entre guillemets appartenant à chacune des deux catégories. Nous nous concentrons sur des marques de surface ainsi que la présence de marques discursives. Le tableau 3.1 illustre sur un exemple les différentes marques utilisées.

Comme cela apparaît dans le tableau 3.1, nous différencions les marques présentes au sein des guillemets de celles en-dehors des guillemets mais restreintes à leur contexte phrastique. En ce qui concerne les marques au sein des guillemets, nous considérons principalement des marques discursives qui caractérisent l'énonciation. Nous retenons ainsi la présence de guillemets au sein du passage entre guillemets, en excluant bien entendu les guillemets encadrant le passage. Nous recherchons également la présence de verbes conjugués, de pronoms personnels et d'adjectifs possessifs. De plus, nous tenons compte d'éléments provoquant des ruptures dans l'énonciation tels que les incises (*[...]* ou *(...)*) qui pourraient embarquer un locuteur. Des éléments de surface tels que la présence de parenthèses ou la taille en mots du passage complètent cette caractérisation.

En ce qui concerne les marques dans le contexte phrastique, passage entre guillemets exclu, nous considérons principalement les marques d'une expression locutrice pronominale ou des marques d'introduction de l'énonciation. Nous différencions ainsi les pronoms et adjectifs aux premières et deuxièmes personnes qui peuvent être locutrices de celles à la troisième personne. Nous avons également retenu la présence des constructions *verbe + que* et la présence de verbes d'énonciation.

L'ensemble de ces marques sont communes au traitement du français et de l'anglais

Exemple	Caractérisation	
Brigitte Liberman, la directrice générale de Cosmétique Active (La Roche-Posay, Vichy), rétorque :« Les grandes innovations ne peuvent pas naître chaque année . »	Contexte entre guillemets	
	Présence de guillemets	Non
	Verbe conjugué	Oui
	Pronom/Adjectif 1ere/2e pers.	Non
	Pronom/Adjectif 3e pers.	Non
	Présence d'incise	Non
	Présence de parenthèses	Non
	Taille en mots	9
	Contexte phrastique	
	Verbe d'énonciation	Oui
	Verbe + que	Non
	Pronom/Adjectif 1ere/2e pers.	Non
	Pronom/Adjectif 3e pers.	Non

FIGURE 3.1 – Caractérisation d'un segment dérivé candidat

moyennant la traduction de nos listes de verbes d'énonciation [5]. Nous avons également ajouté un trait pour le cas spécifique des verbes *say* ou *tell* dans le contexte phrastique du candidat. En effet, nos observations en corpus ont montré qu'il y avait peu de variation dans le choix des verbes d'énonciation, en anglais ces deux verbes étant très majoritairement utilisés.

3.2.2 Constitution de ressources pour l'apprentissage et l'évaluation

L'absence de corpus de presse, en français ou en anglais, enrichit de l'annotation des composants citationnels, nous a conduit à construire de tels corpus. Nous avons choisi de travailler sur des articles de presse généraliste car ceux-ci sont très riches en citations. Les articles y sont très factuels et reposent plus sur la présentation des évènements que sur leur analyse. L'unité minimale sur laquelle nous travaillons est le document de sorte que les citations soient accompagnées de leur contexte. Nous justifions cette configuration plutôt que des citations isolées extraites de leur texte d'origine par la difficulté à identifier les frontières textuelles des citations d'une part, et le fait que les indices ne se trouvent pas forcément à proximité immédiate de la citation.

Nous présentons tout d'abord la compilation du corpus en français que nous avons enrichi par l'annotation des composants citationnels. Ensuite, nous présentons ce même travail d'annotation sur le corpus MPQA en anglais (Wiebe et collab., 2005).

3.2.2.1 Constitution et annotation d'un corpus français

Nous avons compilé un corpus en français constitué d'articles de presse généraliste. Les articles sont tirés de plusieurs quotidiens francophones publiés en ligne (*cf. Tableau 3.2, page 83*) : quatre journaux français (LE MONDE, LE FIGARO, CHALLENGES et LIBÉRATION) et un journal Belge (LE SOIR). La prise en compte de plusieurs journaux a pour objectif d'obtenir un corpus représentatif des différentes formes de citations existantes. Ainsi, une dizaine d'articles a été prélevée en février 2007 dans chaque *Une* des éditions en ligne de ces journaux. Nous avons nettoyé ces documents

5. La traduction des verbes d'énonciation a été faite automatiquement à l'aide des outils linguistiques de Google.

	Ventes moyennes par numéro en 2006	Visites mensuelles du site
Le Monde	312 265	38 262 937
Le Figaro	322 497	24 448 823
Challenges	256 730	7 051 790
Libération	127 687	9 189 585
Le Soir	nc	nc

Table 3.2 – Statistiques des ventes des versions papiers et des visites du site pour les différents journaux sélectionnés (source : www.ojd.com).

```
<citation:locuteur id="3">
Le quotidien économique
</citation:locuteur>
souligne : "
<citation:discoursrepris source="3" type="direct">
Si le rapport ne veut pas associer ces montants à l'idée d'
    une nouvelle 'cagnotte' budgétaire, ni au débat électoral
    sur le niveau des prélèvements obligatoires, le montant
    est équivalent au déficit budgétaire de l'Etat, à savoir
    36,5 milliards d'euros l'an dernier.
</citation:discoursrepris>"
```

Figure 3.2 – Extrait annoté du corpus de citations qui illustre comment les attributs *id* et *source* permettent de connecter les composants.

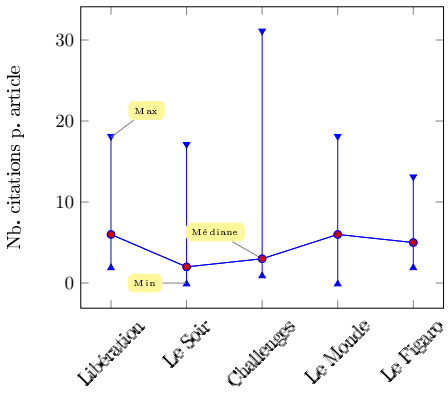
des données ne relevant pas du contenu (publicités, menus...) afin d'obtenir une version texte épurée, simplement formatée en titres et paragraphes. Chacun de ces documents épurés a été stocké en XML avec ses informations de publication (nom du journal, url, date de publication...) afin de le rendre « autonomisable »(Habert, 2000).

Nous avons annoté les composants source et discours rapporté à l'aide de balises XML `<cite:locuteur/>` pour les composants source et `<cite:discoursrepris />` pour les composants discours rapportés. Pour ces derniers, un attribut `type` renseigne le style de discours utilisé (direct, indirect ou indirect quasi-textuel). Nous avons également associé un identifiant à chaque composant source (attribut *id*) afin de le rattacher au composant discours qui lui est associé (attribut *source*). Le détail du schéma d'annotation est présenté dans l'annexe A. La figure 3.2 illustre la mise en œuvre de ce schéma d'annotation sur un court extrait du corpus. Des extraits annotés plus longs sont présentés dans l'annexe B.

Au final, le corpus totalise 53 articles pour environ 36 000 mots. Nous y avons annoté près de 800 objets citationnels correspondant à environ 400 citations. La figure 3.3 illustre la répartition des documents et des objets citationnels du corpus par titre. Nous pouvons observer que les citations sont inégalement réparties au sein du corpus. Le taux de citations n'est apparemment pas propre à un journal car il varie grandement entre les articles.

Nous avons complété notre corpus français par un corpus construit dans les mêmes conditions au sein du LIA [6] (Poulard et collab., 2008). Le corpus résultant est com-

6. Laboratoire Informatique d'Avignon : `www.lia.univ-avignon.fr`



	Nb. articles	Nb. citations
Le Monde	12	82
Le Figaro	10	70
Challenges	11	87
Libération	10	81
Le Soir	10	44
Totaux	53	364

FIGURE 3.3 – Répartition des citations par journaux. Les points de la courbe représentent le nombre médian de citations par article, tandis que les segments verticaux en représentent l'amplitude (les triangles marquent le minimum et maximum par article).

posé de 70 000 mots. Malheureusement la négociation des droits de diffusion avec les différents organes de presse concernés ne nous a pas permis de rendre le corpus constitué librement disponible à la communauté.

3.2.2.2 Annotation d'un corpus anglais

À l'instar du corpus français, le corpus en anglais est constitué d'articles de presse généraliste. Nous avons puisé les articles dans le corpus MPQA (Wiebe et collab., 2005) qui avait déjà été utilisé auparavant dans un contexte semblable par Choi et collab. (2005) pour l'identification des sources d'opinions. Nous avons ainsi pu profiter d'un corpus de qualité déjà nettoyé.

Nous n'avons pas été en mesure d'utiliser directement les annotations sur les sources d'opinions du corpus MPQA car elles ne correspondent pas réellement à des expressions faisant référence à la source d'une citation. Nous avons donc annoté manuellement le corpus en reprenant le format et le protocole utilisé pour le corpus en français. L'annotation a été réalisée par deux personnes au sein de l'outil GATE (Cunningham et collab., 2002) qui nous a permis d'évaluer l'accord inter-annotateur. Le tableau 3.3 détaille cet accord en termes de précision et rappel. Le tableau détaille également le nombre d'annotations en correspondance exacte (*Correct*), qui se recouvrent partiellement (*Partiel*), qui ne sont posées que par un des deux annotateurs (*Manquant*) et qui se recouvrent (*Superposé*). L'accord pour l'annotation des composants de discours rapporté est bon avec une précision de 0,84 et un rappel de 0,75. Les désaccords portent principalement sur les frontières. L'accord pour l'annotation des composants source est également bon avec une précision de 0,90 et un rappel de 0,83. Les expressions locutrices sont plus simples à délimiter car elles se présentent majoritairement sous la forme d'une entité nommée ou d'un syntagme nominal.

Au final, le corpus totalise 42 articles pour environ 17 000 mots. Nous y avons annoté 550 objets citationnels : 256 expressions locutrices et 301 segments dérivés. Le corpus anglais est donc légèrement moins fourni que le français. La licence du corpus MPQA n'est pas explicitée, mais il est librement téléchargeable depuis la page qui lui est dédiée [7]. Nos annotations sont librement disponibles sur demande.

7. `http://nrrc.mitre.org/NRRC/02_results/mpqa.html`

Annotation	Précision	Rappel	Correct	Partiel	Manquant	Superposé
Composant discours rapporté	0,839	0,754	243	28	70	35
Composant source	0,896	0,827	233	7	46	24

TABLE 3.3 – Évaluation de l'accord entre annotateurs pour le corpus anglais.

3.3 Résultats expérimentaux

Nous présentons dans cette section la mise en œuvre des tâches de catégorisation présentées en section 3.2.1 ainsi que la façon dont nous évaluons les classifieurs ainsi créés. Nous rapportons et discutons ensuite les performances des classifieurs obtenus pour les composants source puis pour les composants discours rapporté. Finalement, nous évaluons la combinaison des composants ainsi repérés en segments citationnels.

3.3.1 Protocole d'évaluation et d'expérimentation

Notre méthode consiste, pour chacune des tâches de catégorisation, à extraire les candidats des textes puis les caractériser à l'aide des traits discutés en section 3.2.1 afin de les soumettre au classifieur. Nous appelons l'ensemble des valeurs des traits associés à un candidat une *instance*. Pour l'identification des segments dérivés, nous avons collecté 1 004 instances pour le français et 145 pour l'anglais. Pour l'identification des expressions locutrices, nous en avons collecté 1 703 pour le français et 452 pour l'anglais.

INSTANCE

Nous cherchons à produire un modèle statistique qui devra décider si les instances ainsi compilées correspondent à des segments dérivés (respectivement des expressions locutrices) ou non. Nos annotations sur le corpus permettent, pour l'apprentissage, de savoir si chaque instance correspond réellement à un segment dérivé (respectivement une expression locutrice), auquel cas nous parlons d'un positif (respectivement de négatif). Cet étiquetage des instances selon les annotations manuelles du corpus nous sert de référence pour l'évaluation. Nous parlons alors de *référentiel*.

POSITIF
NÉGATIF
RÉFÉRENTIEL

3.3.1.1 Protocole d'évaluation

Notre méthode s'apparente à un système d'annotation [8], son évaluation porte donc sur la pertinence des traits, l'entraînement et son application globale. L'évaluation a pour but de mesurer la distance séparant les résultats obtenus (en sortie de notre système) de la capacité linguistique idéalement visée (nos annotations manuelles).

L'évaluation de notre système doit être réalisée sur un matériau différent de celui qui a servi à l'entraînement du modèle statistique évalué. Dans la mesure où nos corpus sont de petite taille, nous optons pour une validation croisée plutôt qu'une division du corpus en données d'entraînement et données de tests. La validation croisée consiste à diviser les corpus en partitions, puis entraîner un modèle sur toutes ces partitions sauf une qui sert à l'évaluation. L'opération est répétée en prenant chaque partition comme donnée de tests. Le résultat final correspond à la moyenne des évaluations de chaque partition. Dans le cadre de nos expérimentations, nous avons divisé aléatoirement notre corpus en dix partitions.

8. Soit un système de *type A* selon la classification des systèmes d'évaluation de Popescu-Belis (2008)

Les instances en sortie du classifieur sont étiquetées, par le système, comme positif ou négatif. Pour l'évaluation, cet étiquetage est comparé au référentiel de sortes à classer les résultats en sortie du système en quatre catégories. Les instances classées par le système comme positifs sont des vrais positifs (VP) s'ils sont également positifs dans le référentiel, sinon il s'agit de faux positifs (FP). Les résultats négatifs sont classés de la même façon comme vrais négatifs (VN) ou faux négatifs (FN). D'après Sebastiani (2002), les tâches de catégorisation binaires telles que celles auxquelles nous sommes confrontés doivent être évaluées à l'aide des mesures de précision et de rappel (Manning et Schütze, 1999, p.268) qui permettent de tenir compte de la distribution, potentiellement déséquilibrée, des instances entre les différentes classes. Nous les complétons par la F-mesure qui offre une vision combinée de ces deux mesures. Nous détaillons leurs calculs dans l'annexe D.

3.3.1.2 Algorithmes d'apprentissage

Sebastiani (2002) présente les algorithmes de type *boosting* et *SVM* (*Support Vector Machines*) les plus adaptés aux tâches de classification. Tout d'abord, ces algorithmes donnent les meilleurs résultats lors des évaluations comparatives. Ensuite, leurs fondements théoriques sont les mieux appuyés sur la théorie de l'apprentissage automatique. Nous avons sélectionné *AD Tree* comme algorithme de type *boosting* et *C-SVC* comme algorithme de type *SVM*. Nous avons expérimenté l'utilisation de ces deux algorithmes au sein de la plateforme Weka (Witten et Frank, 2005).

AD Tree (Sattath et Tversky, 1977) est un algorithme de type arbre de décision intégrant un principe d'optimisation appelé *boosting*. Le *boosting* est une méthode de production de prédictions très précises par combinaison successive de prédictions approximatives. Nous faisons varier le nombre d'itérations sur les instances, qui correspond au nombre de combinaison, en parcourant tous les chemins possibles pour la recherche de solutions (algorithme de recherche *Expand all paths*).

C-SVC, intégré à Weka au travers de WLSVM développé par EL-Manzalawy et Honavar (2005), est un algorithme de classification à larges marges. Nous faisons varier le paramètre de coût de cet algorithme sur les instances. Ce paramètre agit sur la souplesse de la marge séparant les positifs des négatifs, plus il est élevé, plus l'algorithme autorise des erreurs de marge. Cela se traduit principalement par la classification correcte de bonnes instances qui auraient été écartées par la maximisation de la marge, c-à-d l'augmentation du rappel au dépens de la précision. Nous avons conservé les valeurs par défaut [9] des autres paramètres concernant notamment le noyau ou la tolérance du critère de terminaison.

3.3.2 Catégorisation des composants source

Cette section discute et analyse les résultats obtenus pour la catégorisation des composants source. Pour rappel, nous catégorisons les entités nommées, qui ont le rôle de candidat, dans les classes expressions locutrices et non locutrices. La première correspond à un sous-ensemble des composants source qui nous intéresse. Nous évaluons dans un premier temps la pertinence des traits choisis pour caractériser nos candidats, puis nous discutons des performances du classifieur obtenu.

3.3.2.1 Évaluation de la pertinence des traits d'apprentissage

Nous avons utilisé l'algorithme *CfsSubsetEval* de Weka (Witten et Frank, 2005, p.232–238) pour déterminer les traits d'apprentissage les plus pertinents d'après les

9. `weka.classifiers.functions.LibSVM -S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.0010 -P 0.1 -B`

Méthode	Précision		Rappel		F-mesure	
	Fr	En	Fr	En	Fr	En
Approche naïve	0,27	0,26	1	1	0,42	0,41
AD Tree	0,53	0,62	**0,31**	**0,50**	0,39	0,55
C-SVC	**0,63**	**0,70**	0,18	0,44	0,28	0,54

TABLE 3.4 – Résultats des classifications étant donné les différents algorithmes testés avec le paramétrage le plus performant.

instances récoltées dans nos corpus. Cet algorithme sélectionne les traits ayant la meilleure capacité de prédiction en se basant sur une évaluation de la corrélation de leurs valeurs avec les données du référentiel. Nous l'avons combiné avec l'algorithme de recherche *ExhaustiveSearch*[10] qui effectue un parcours exhaustif des combinaisons des traits afin de trouver une solution exacte.

Pour le français, l'algorithme ne fait pas de différence entre les traits et les évalue tous au même niveau de pertinence. Il faut noter que le score de confiance pour cette sélection est très faible : 0,085. Cela laisse penser qu'il manque un certain nombre de traits pour pouvoir distinguer les entités nommées qui sont des expressions locutrices de celles qui ne le sont pas.

Les traits les plus pertinents pour l'anglais sont le type d'entité nommée (anthroponyme...), et la distance des verbes *say* et *tell*. Ce dernier trait, spécifique à l'anglais, suffit à faire bondir le score de confiance. Ces résultats doivent cependant être modérés du fait qu'il y a 3,5 fois moins d'instances pour l'anglais que pour le français.

En résumé, aucun trait n'est particulièrement pertinent pour caractériser les expressions locutrices en français alors que la proximité des verbes *say* et *tell* est pertinente pour l'anglais. Cela peut s'expliquer par une plus grande régularité dans la structure des citations en anglais.

3.3.2.2 Évaluation des classifieurs

Nous avons évalué trois classifieurs pour chaque langue : un classifieur pour chaque algorithme expérimenté (ADTree et C-SVC) et un classifieur naïf. L'annexe E détaille le calibrage utilisé pour les différents algorithmes d'apprentissage. Le classifieur naïf consiste à considérer toutes les entités nommées comme des expressions locutrices. Nous l'utilisons à des fins de comparaison, faute de résultats d'évaluation pour les approches existantes. Les performances de ces différents classifieurs sont présentées dans le tableau 3.4 pour le français et l'anglais.

L'approche naïve n'offre pas de bons résultats : le rappel est logiquement maximal puisque tous les candidats sont retenus, mais la précision est à 0,27 pour le français et 0,26 pour l'anglais. Tous nos modèles prédictifs obtiennent une meilleur précision, mais également un rappel plutôt mauvais puisqu'au mieux nous détectons une expression locutrice sur deux (pour l'anglais avec *AD Tree*).

Si l'approche naïve offre des performances équivalentes pour l'anglais et le français, nos classifieurs ne sont pas comparables pour les deux langues. En effet, les résultats sur l'anglais sont meilleurs que ceux sur le français. De plus, en ce qui concerne l'anglais, si les résultats des deux classifieurs sont comparables en terme de F-mesure, il y a tout de même une différence de 10 points pour la précision et le rappel entre les

10. `java -cp weka.jar weka.attributeSelection.CfsSubsetEval -M -s "weka.attributeSelection.ExhaustiveSearch" -i path/to/my/arff`

deux scores. La meilleure précision (0,70) est obtenue avec l'algorithme *C-SVC* tandis que le meilleur rappel est obtenu avec l'algorithme *AD Tree* mais n'est que de 0,5. Il n'y a pas réellement de meilleur modèle pour l'anglais, le choix dépend du compromis souhaité entre précision et rappel.

En ce qui concerne le français, la meilleure performance en terme de F-mesure est obtenue avec le classifieur généré par l'algorithme *ADTree*, c'est également l'algorithme qui donne le meilleur rappel avec une expression locutrice identifiée sur trois. La meilleure précision (0,63) est quant à elle obtenue avec le classifieur généré par *C-SVC*. La disparité des résultats sur le français est assez importante, l'amplitude pour la précision est de 10 points et elle est de 13 pour le rappel. Cette importante volatilité reflète un manque de données ou un manque de traits.

Nous observons également que les résultats pour le français sont bien moins bons que ceux pour l'anglais avec une différence de 16 points entre les meilleures F-mesure de chaque. Les expressions locutrices en anglais semblent clairement plus faciles à identifier que celles du français. Nous pouvons l'expliquer en partie par la régularité des constructions en anglais, la faible diversité des verbes d'énonciation et la présence de l'expression locutrice dans le contexte immédiat de ces verbes.

Finalement, il est nécessaire de relativiser ces résultats. Les modèles ne classent que les candidats et ces derniers ne recouvrent pas l'intégralité des expressions locutrices que nous avons annoté dans nos corpus. Ainsi, étant donné que 56 % des expressions locutrices contiennent des entités nommées pour le français et 67 % pour l'anglais, cela reviendrait pour le modèle *ADTree* à un rappel d'environ 17 % pour le français et 33 % pour l'anglais. Il reste donc une marge de progression importante pour le repérage des expressions locutrices.

3.3.3 Catégorisation des composants discours rapporté

Cette section discute et analyse les résultats obtenus pour la catégorisation des composants discours rapporté. Pour rappel, nous catégorisons les segments entre guillemets, qui ont le rôle de candidat, dans les classes segments dérivés et non dérivés. La première correspond à un sous-ensemble des composants discours rapporté qui nous intéresse. Nous évaluons dans un premier temps la pertinence des traits choisis pour caractériser nos candidats, puis nous discutons des performances du classifieur obtenu.

3.3.3.1 Évaluation de la pertinence des traits d'apprentissage

Les traits sélectionnés pour la caractérisation des candidats sont évalués de la même manière que pour les composants source à l'aide de l'algorithme *CfsSubsetEval*.

En français, les traits considérés comme les plus pertinents pour classer nos segments dérivés candidats sont au nombre de quatre. Le premier trait correspond à la présence d'un verbe d'énonciation dans le contexte phrastique du candidat, ce qui confirme l'importance de ces derniers dans le processus citationnel (Giguet et Lucas, 2004; Mourad et Desclés, 2004; Jackiewicz, 2006). Les autres traits correspondent respectivement à la taille en mots du candidat, la présence d'autres segments entre guillemets dans la même phrase et la présence d'un verbe conjugué au sein du candidat. Ces derniers traits semblent pertinents pour différencier un segment entre guillemets correspondant à un segment dérivé de celui correspondant à une emphase.

En anglais, les traits les plus pertinents pour les segments dérivés sont au nombre de deux. Ces derniers correspondent à la distance en mots entre les frontières du candidat et les verbes *say* et *tell*, respectivement à l'intérieur et à l'extérieur du

segment entre guillemets. Ces marques linguistiques ont été ajoutées spécifiquement pour l'anglais d'après nos observations en corpus (*cf. Annexe C*).

3.3.3.2 Évaluation des classifieurs

Nous avons évalué trois classifieurs pour chaque langue : un classifieur pour chaque algorithme expérimenté (ADTree et C-SVC) et un classifieur naïf. L'annexe E détaille le calibrage utilisé pour les différents algorithmes d'apprentissage. Le classifieur naïf consiste à considérer tous les segments entre guillemets comme des segments dérivés. Nous l'utilisons à des fins de comparaison, faute de résultats d'évaluation pour les approches existantes. Les performances de ces différents classifieurs sont présentées dans le tableau 3.5 pour le français et l'anglais.

Méthode	Précision		Rappel		F-mesure	
	Fr	En	Fr	En	Fr	En
Approche naïve	0,82	0,75	1	1	0,90	0,86
AD Tree	**0,9**	**0,85**	0,94	0,89	0,92	0,87
C-SVC	0,86	0,80	**0,98**	**0,98**	0,92	0,88

TABLE 3.5 – Résultats des classifications étant donné les différents algorithmes testés avec le paramétrage le plus performant.

L'approche naïve offre de bons résultats : le rappel est logiquement à 100 % puisque tous les candidats sont retenus, mais la précision reste élevée avec 82 % pour le français et 75 % pour l'anglais. L'algorithme *AD Tree* génère le modèle qui obtient les meilleures performances en termes de précision à la fois pour le français et l'anglais avec respectivement 90 % et 85 %, soit une augmentation de respectivement 8 et 10 points par rapport à l'approche naïve. Cette performance s'effectue naturellement aux dépens du rappel qui chute respectivement de 6 et 11 points. Au final, la F-mesure est tout juste supérieure à celle de l'approche naïve. L'algorithme *C-SVC* génère quant à lui le modèle avec le meilleur rappel, et une précision de respectivement 4 et 5 points supérieurs à celle de l'approche naïve, ce qui au final lui donne la meilleure F-mesure pour les deux langues.

Il est important de noter que le rappel correspond ici à la proportion de segments dérivés placés entre guillemets (le critère de sélection de nos candidats), et non la totalité des segments que nous avons pu annoter. Ainsi, étant donné que 80 % des citations sont composées de segments entre guillemets pour le français et seulement 43 % pour l'anglais, cela reviendrait pour le modèle C-SVC à un rappel de respectivement 78 % et 42 %.

3.3.4 Identification des segments citationnels

Le problème de l'attribution d'un segment dérivé à une expression locutrice est ici très rapidement abordé et le problème reste certainement entier. Les travaux de Prasad et collab. (2006) et Choi et collab. (2005) proposent des solutions, mais le rapprochement des propos et des expressions locutrices aux énonciateurs nécessite la mise en place de techniques comme la résolution d'anaphores. Les segments citationnels sont une proposition d'unité proche de la citation basée sur l'exploitation des expressions locutrices et des segments dérivés repérés. Nous mettons en œuvre une heuristique qui considère toute phrase contenant un des composants comme un segment citationnel. Nous modérons cette attribution par un degré de confiance : élevé

Langue	Précision	Rappel	F-mesure
Français	0,59	0,67	0,63
Anglais	1	0,56	0,72

TABLE 3.6 – Évaluation de l'heuristique d'identification des segments citationnels sur les composant repérés automatiquement

si la phrase contient à la fois un segment dérivé et une expression locutrice, modéré si elle ne contient que le premier et faible si elle ne contient que le second.

Nous avons évalué manuellement la mise en œuvre de cette heuristique. Le tableau 3.6 présente les résultats obtenus en ne considérant que les segments dont la confiance est modérée ou élevée. La précision est excellente pour l'anglais et est seulement de 0,59 pour le français tandis que le rappel est plutôt mauvais avec 0,67 pour le français et 0,56 pour l'anglais. Le fait de ne retenir que les segments citationnels obtenant une confiance élevé fait bondir la précision pour le français à 0,98 au détriment du rappel qui chute à 0,15.

Les faux négatifs proviennent principalement des discours repris indirects pour lesquels, dans plus de la moitié des cas, aucun candidat d'expression locutrice n'est proposé. La différence entre le rappel pour le français et pour l'anglais s'explique alors par la présence plus marquée du style indirect en anglais. Les faux positifs résultent d'une propagation d'erreurs d'identification de nos modèles. Quelques éléments citationnels sur lesquels nous avions hésité lors de l'annotation se retrouvent également parmi les faux positifs.

3.4 Conclusion

Nous nous sommes intéressé à la détection des citations car nous pensons que celles-ci ont un rôle à jouer dans des problématiques telles que le suivi d'impact ou encore la recherche d'information en temps réel. Les citations sont une forme de dérivation particulière, notamment de par le nombre de marques laissées par l'auteur dans le texte dérivé pour les identifier. Cette particularité permet de mettre en œuvre une détection intrinsèque qui ne repose pas sur des ruptures de style contrairement à ce qui est fait habituellement. Une approche extrinsèque nécessiterait de pouvoir accéder au texte source. Or ce dernier n'est pas toujours disponible, notamment lorsque la citation reprend un discours prononcé et non écrit.

Nous avons mis en place une détection intrinsèque des citations pour le genre des articles de presse en ligne, à la fois pour le français et pour l'anglais. Nous nous sommes orienté vers une approche à base d'apprentissage à laquelle nous déléguons l'extraction des règles de combinaison des indices pour chaque langue. Ce changement de paradigme par rapport aux approches antérieures a nécessité de formaliser le problème de détection des citations comme une tâche de catégorisation (*cf. Section 3.2.1*). Nous nous sommes intéressé à la catégorisation de deux composants constitutifs de la citation : la source et le discours rapporté.

Nous nous sommes concentré sur les composants source contenant une entité nommée et les composants discours rapporté contenant un passage entre guillemets. Nos classifieurs permettent d'identifier, au sein des articles de presse, le premier composant avec une précision de 60 % pour le français et 70 % pour l'anglais, et le second avec une précision de 90 % pour le français et de 85 % pour l'anglais. Nos classifieurs sont meilleurs que l'approche naïve consistant à considérer tous les candidats (tous les segments entre guillemets ou toutes les entités nommées) comme des composants. Le

rappel est dégradé par rapport à cette approche de référence laissant la porte ouverte à de possibles améliorations. En particulier, l'évaluation des traits les plus pertinents nous montre que ceux que nous avons choisis ne sont pas assez discriminants pour notre tâche : soit les candidats que nous avons choisis ne sont pas adaptés, soit ils ne peuvent être discriminés uniquement sur la base de traits lexicaux et contextuels.

Nous maintenons que notre hypothèse qu'une méthode de détection des citations par apprentissage est envisageable et serait portable à plusieurs langues. Toutefois, les gains obtenus par notre approche restent faibles et ne justifient peut-être pas le temps nécessaire à la construction du corpus pour la mise au point des classifieurs. Ainsi, la principale perspective à ce travail est la remise en cause de notre formalisation du problème qui consiste en la catégorisation de certains objets linguistiques (entités nommées et passages entre guillemets) comme des composants constitutifs de la citation (source et discours rapporté respectivement). Étant donné que les traits utilisés sont en grande partie communs aux deux composants, une approche qui consisterait à identifier tout d'abord des segments citationnels sur la base de ces traits, puis en leur sein les différents composants, serait potentiellement plus performante.

Chapitre 4

Évaluer les méthodes de détection extrinsèque de dérivation

« When I use a word, » Humpty Dumpty said, in rather a scornful tone, « it means just what I choose it to mean—neather more nor less. »
« The question is, » said Alice, « whether you can make words mean so many different things. »

— Lewis Carrol, *Through the Looking-Glass*

Sommaire

La détection de dérivation est une tâche complexe à facettes multiples pour laquelle il n'existe ni protocole d'évaluation standard, ni corpus de référence. En effet, l'évaluation des sytèmes peut porter sur différentes capacités : distinguer les textes dérivés et les textes non-dérivés, identifier les types de dérivation (plagiat, version...), isoler dans ces textes les passages dérivés, obtenir des mesures de similarités concordantes avec les jugements humains... Nous proposons un protocole d'évaluation adapté à nos objectifs : identifier les liens de dérivation entre des textes à l'échelle du document. Nous avons notamment déplacé la prise de décision (catégorisation) hors du processus d'évaluation puisque nous considérons que celle-ci est spécifique à la mise en œuvre applicative des méthodes.

L'évaluation d'un système de détection de dérivation repose selon nous sur trois éléments : un protocole d'évaluation associé à un certain nombre de mesures, un (ou plusieurs) corpus et des résultats de référence fournis par une méthode état de l'art. Dans ce chapitre, nous rappelons tout d'abord les différentes méthodes d'évaluation qui ont été mises en œuvre pour les tâches associées à la détection de dérivation (*cf. Section 4.1*) ainsi que les corpus qui ont été utilisés pour ces évaluations qui reprennent le format proposé par les récentes campagnes d'évaluation PAN (Potthast et collab., 2009, 2010b) (*cf. Section 4.2*). Nous présentons alors notre méthode d'évaluation, inspirée de la RI et tirant parti de deux mesures : la MAP et la Sép. Q, et nos corpus français et anglais (*cf. Section 4.3*). Finalement, nous mettons en œuvre cette méthode sur nos corpus pour l'approche par signature complète, qui est considérée comme l'approche de référence, afin d'obtenir des résultats de référence auxquels nous nous comparerons (*cf. Section 4.4*).

4.1 Principales approches d'évaluation

La création récente du challenge PAN (*Uncovering Plagiarism, Authorship, and Social Software Misuse*, (Potthast et collab., 2009, 2010b)) a permis de faire avancer l'idée d'un protocole d'évaluation et d'un corpus de référence. Le protocole proposé évalue le problème comme une tâche de classification entre textes (ou passages) dérivés et non-dérivés. Dans cette section, nous présentons ce type de protocole (*cf. Section 4.1.1*), le plus communément utilisé, ainsi que ceux issus de la RI (*cf. Section 4.1.2*) et enfin une approche qui évalue la concordance entre les scores de similarités des méthodes automatiques et celles de jugements humains (*cf. Section 4.1.3*). Cette présentation des différentes approches d'évaluation a pour objectif de nous aider dans le choix de notre propre méthode d'évaluation.

4.1.1 Évaluation comme une tâche de classification

L'approche pour l'évaluation des systèmes de détection de dérivation la plus courrament mise en œuvre consiste à considérer la détection de dérivation comme une tâche de classification « dérivé vs. non-dérivé ». Cette classification s'effectue soit pour le texte complet, soit en tenant compte de la segmentation en passages.

4.1.1.1 À l'échelle des textes complets

Les approches par classification considèrent des couples de textes dont l'un des deux est potentiellement identifié comme la source. Un couple de textes est classé sur la base du score de similarité obtenu : soit au travers d'un modèle probabiliste (Clough, 2003a), soit en partionnant l'espace des valeurs des scores de similarité. Nous détaillons ce dernier procédé illustré par la figure 4.1.

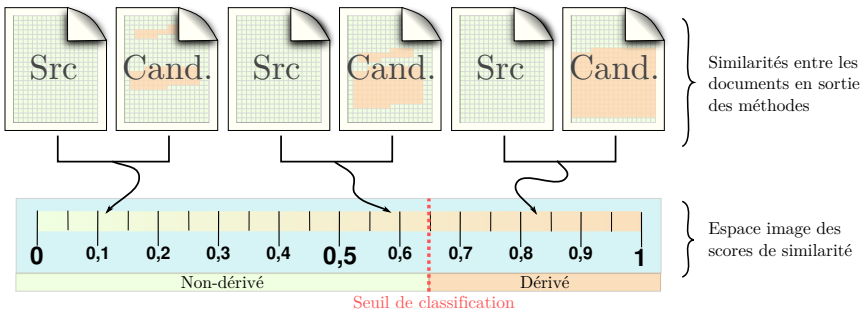


FIGURE 4.1 – Les évaluations par classification définissent des zones dans l'espace image des similarités correspondant aux classes.

L'espace image des mesures de similarité est divisé en zones continues à chacune desquelles est associée une classe. Cette classe caractérise la relation de dérivation entre les textes comparés. Le plus souvent l'espace image est défini entre 0 et 1 et se répartie entre une zone, autour de 1, correspondant à la classe « dérivé » et une autre « non-dérivé » autour de 0 (Si et collab., 1997; Lyon et collab., 2001; Bernstein et collab., 2006). Les deux zones ont une frontière commune définie par un score seuil. Parfois, l'espace est divisé en un plus grand nombre de classes qui reflétent plus finement le lien de dérivation. Leur pertinence dépend de l'application visée. Ainsi Shivakumar et Garcia-Molina (1995) définissent les espaces [0 % ; 33 %], [34 % ; 66 %], [67 % ; 90 %] et [91 % ; 100 %] pour les classes respectives : aucun, peu, beaucoup et total.

Le choix d'un score seuil, s'il a pu être arbitraire dans les travaux pionniers (Si et collab., 1997), s'appuie désormais sur une phase d'apprentissage. Bao et collab. (2007) par exemple, sélectionnent le seuil qui sépare les dérivés des non-dérivés de façon optimale sur un corpus d'apprentissage. Cette phase d'apprentissage doit tenir compte de la distribution non homogène entre les classes : soit les disparités sont conservées (on parle alors de stratification) et doivent être considérées lors de l'analyse des résultats, soit le nombre d'instances est équilibré entre toutes les classes (on parle d'égalisation).

Les métriques utilisées pour mesurer la performance des méthodes sont les classiques précision (*cf. Équation D.1*), rappel (*cf. Équation D.2*) et f-score (*cf. Équation D.5*) (Manning et Schütze, 1999, p.268-269). Elles sont présentées en détail dans l'annexe D.

4.1.1.2 À l'échelle des passages

Il est nécessaire d'adapter le protocole précédent lorsque l'identification des relations de dérivation ne porte plus sur les textes complets mais s'applique à une granularité plus fine. De tels systèmes identifient et catégorisent les relations entre des passages de texte. Les mesures classiques de précision et de rappel ne sont plus appropriées car elles ne rendent pas compte de la fluctuation des frontières des passages incriminés. Potthast et collab. (2010b) considèrent en conséquence l'évaluation au niveau des passages (*psg*) plutôt que des textes complets. Ils introduisent les mesures de précision$_{psg}$ et de rappel$_{psg}$ dérivées des mesures classique de précision et de rappel afin de tenir compte de ce changement, ainsi que la mesure de granularité. Dans le cadre de cette thèse, nous nous limitons à la détection de dérivation à l'échelle du document. Nous présentons ce protocole dans un but informatif.

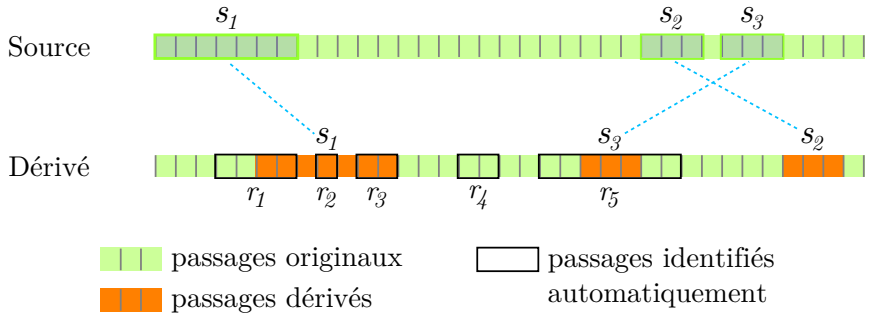


FIGURE 4.2 – Configurations de détection lors d'une évaluation prenant en compte la délimitation des passages. Schéma dérivé de Potthast et collab. (2010b)

Les sytèmes de détection à l'échelle du passage identifient des passages de texte dans le texte candidat ($r_1, \ldots, r_j$) et les rapprochent de passages du texte source ($s_1, \ldots, s_i$). L'ensemble des passages identifiés dans le candidat est noté R et celui du texte source S. L'évaluation doit mesurer la concordance des passages identifiés avec les passages réellements liés par une relation de dérivation. La figure 4.2 illustre différentes configurations de détection de ces passages : recouvrements partiels plus large ou plus court (r_1, r_2, r_3, r_5), faux positifs (r_4) et silences (s_2).

Le rappel$_{psg}$ est défini comme la moyenne du recouvrement des passages identifiés ($r \in R$) par rapport au passage dérivé correspondant ($s \in S$), le tout ramené au nombre de passages à identifier ($|S|$) :

$$\text{rappel}_{psg}(S,R) = \frac{1}{S}\sum_{s\in S}\frac{|s \sqcap \bigcup_{r\in R} r|}{|s|} \tag{4.1}$$

avec $s \sqcap r$ les caractères de r recouvrant correctement s

La précision classique imposerait qu'un seul passage détecté soit rattaché à un passage dérivé. Ce n'est pas forcément le cas comme le montrent les passages r_1, r_2 et r_3 de la figure 4.2. L'idée est de dénombrer la proportion des passages détectés qui sont des dérivés, ce qui revient à calculer le rappel$_{psg}$ de R par rapport à S :

$$\text{précision}_{psg}(S,R) = \frac{1}{R}\sum_{r\in R}\frac{|r \sqcap \bigcup_{s\in S} s|}{|r|} \tag{4.2}$$

La formule du f-mesure est la même que pour les définitions classiques de précision et de rappel (*cf. Équation D.5*).

Ces métriques ne rendent pas compte du nombre de passages détectés qui correspondent à un unique passage dérivé, ce que les auteurs nomment la *granularité* de la détection. Cette métrique est définie comme la moyenne du nombre de passages détectés (r) couvrant un passage dérivé (s) pour chaque passage dérivé détecté :

$$\begin{aligned} \text{granularité}(S,R) &= \frac{1}{|S_R|}\sum_{s\in S_R}|C_s| \\ \text{avec } S_R &= \{s|s\in S \wedge \exists r\in R : s\sqcap r\neq\emptyset\} \\ &\quad \text{soit le nombre de } s \text{ recouverts par au moins un } r \\ \text{avec } C_s &= \{r|r\in R \wedge \bigcup_{s\in S} s\sqcap r\neq\emptyset\} \\ &\quad \text{soit le nombre de } r \text{ qui recouvrent au moins un } s \end{aligned} \tag{4.3}$$

Dans le meilleur des cas, il n'y a qu'un seul passage détecté (r) par passage dérivé (s) auquel cas la granularité$(S,R) = 1$, et dans le pire des cas tous les passages détectés couvrent le même passage dérivé : granularité$(S,R) = |R|$.

Finalement, les auteurs proposent une combinaison de l'ensemble de ces métriques permettant de rendre compte globalement de la qualité de la détection :

$$\frac{\text{f-mesure}_{psg}(S,R)}{\log_2(1+\text{granularité}(S,R))} \tag{4.4}$$

Ces métriques permettent une évaluation élégante des systèmes de détection de dérivation à l'échelle des passages. Elles rendent compte à la fois du classement correct des passages dérivés et non-dérivés, et de la justesse de leurs frontières textuelles.

En conclusion, nous ne retenons aucune de ces méthodes. La phase de création du classifieur dans une approche par classification manque de normalisation et de déterminisme. Cette étape nécessaire relève selon nous de la mise en œuvre applicative et est un frein à la comparabilité des résultats. Par extension nous ne retenons pas non plus l'évaluation par classification à l'échelle des passages, d'autant plus que nous ne nous intéressons pas à la détection de dérivation à cette échelle.

4.1.2 Évaluation comme une tâche de recherche d'information

La recherche d'information profite, de par son ancienneté, de méthodes d'évaluation rigoureuses et largement éprouvées. Le protocole d'évaluation tel qu'opéré en RI se compose de deux étapes : la soumission de requêtes au système évalué et l'application de mesures d'évaluation sur un sous-ensemble ordonné des résultats.

La requête soumise à un système de RI est comparée à un index et donne lieu à un résultat, habituellement une liste de réponses ordonnée. Dans le cadre de la détection de dérivation, les requêtes sont des textes ou des passages de texte, l'index est construit à partir d'une collection de textes candidats et le résultat est un sous-ensemble des textes candidats ordonnés selon un score de similarité avec la requête, comme l'illustre la figure 4.3. Hose (2003) soumet comme requêtes des phrases des parchemins de la mer morte. Metzler et collab. (2005) construisent leurs requêtes à partir du matériel d'évaluation de la campagne TREC. Hoad et Zobel (2002) soumettent le texte source dans son intégralité, ce qui semble le plus cohérent avec la tâche à évaluer.

Les mesures employées évaluent la distribution des réponses du système en les comparant à une distribution idéale. Dans le cas de la dérivation, nous souhaitons que les textes dérivés apparaissent en haut du classement (scores de similarité les plus élevés). Il est possible dans ce but d'utiliser les mesures classiques de précision et de rappel (*cf. Annexe D*) en les appliquant à un sous-ensemble du classement résultat (du rang 1 à 20 (Hoad et Zobel, 2002), 25 (Metzler et collab., 2005) voir même 100 (Hose, 2003)). Ces approches sont également nommées top 10, top 20... La R-précision est un cas particulier qui consiste à calculer la précision sur tous les rangs jusqu'à celui

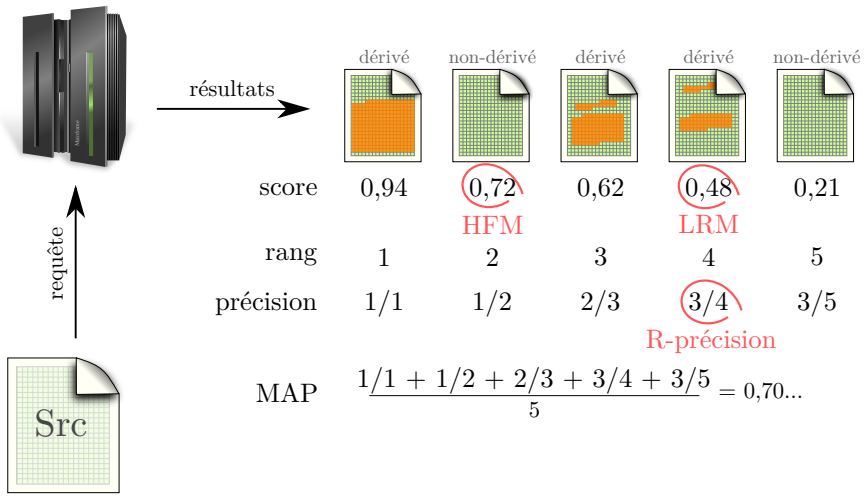


FIGURE 4.3 – Classement des résultats du système et affectation d'un rang tel qu'opéré pour les évaluations en RI.

du dernier dérivé (Hoad et Zobel, 2002). Elle ne peut bien entendu être mise en œuvre que si tous les documents sont connus. Dans l'exemple de la figure 4.3 elle correspond à la précision du rang 1 à 4. Toutefois, la MAP (*Mean Average Precision*) (*cf. Équation 4.5*) est mieux adaptée en ce qu'elle rend directement compte de cette variation du rang (Hoad et Zobel, 2002; Metzler et collab., 2005). La figure 4.3 montre un exemple du calcul de la MAP à partir d'un classement par scores de similarité.

$$\text{MAP} = \frac{\sum_{r=1}^{N}(P(r)\cdot rel(r))}{N} \tag{4.5}$$

avec r le rang
N le nombre de rangs retenus pour l'évaluation
$rel(r)$ une fonction binaire de la pertinence d'un rang, le plus souvent on pose $rel : r \mapsto 1$
$P(r)$ la précision calculée sur les résultats de rang 1 à r

Une approche complémentaire mise en œuvre par Hoad et Zobel (2002) utilise le *highest false match* (HFM) et le *lowest correct result* (LCR) qui correspondent respectivement au plus haut score de similarité obtenu par un document non-dérivé et au plus bas score de similarité obtenu par un document dérivé (*voir la figure 4.3*). L'écart entre ces deux valeurs se nomme la séparation (*cf. Équation 4.6*) et sa combinaison conjointe au HFM (*cf. Équation 4.7*) permet d'évaluer la capacité du système à distinguer les dérivés des non-dérivés (Hoad et Zobel, 2002).

$$\text{séparation} = \text{LCR} - \text{HFM} \tag{4.6}$$

$$\text{discrimination} = \frac{\text{séparation}}{\text{HFM}} \tag{4.7}$$

Les méthodes d'évaluation tirées de la RI ont été très peu utilisées pour la détection de dérivation. Par conséquent, les possibilités de comparaison avec des évaluations antérieures sont limitées.

4.1.3 Corrélation à des jugements humains

La majorité des méthodes de détection de dérivation repose sur la mesure de similarité entre des textes, qui est un problème également étudié d'un point de vue psychologique. Lee et collab. (2005) ont mis en place un protocole d'évaluation qui repose sur la comparaison entre les scores de similarité des méthodes classiques (mots, n-grammes et analyse sémantique latente (LSA)) et des jugements de similarité. Une jugement de similarité est une évaluation sur une intervalle discrète donnée de la similarité entre deux textes telle que perçue par un humain. Uzuner et collab. (2004) ont également mené une évaluation similaire basée sur des évaluations humaines de la similarité des textes afin de les comparer aux résultats de leur méthode.

Les objectifs de l'évaluation sont de vérifier la concordance entre les scores de similarités produits par les méthodes et les jugements humains émis sur les mêmes textes. Cette évaluation repose sur l'hypothèse que ces méthodes reproduisent le processus psycho-cognitif humain.

Dans ce but, les auteurs ont constitué un corpus de courts textes en anglais (articles de journaux) représentant 1 225 liens suspects, de dérivation ou non. Les 83 annotateurs humains, étudiants d'université, devaient noter la similarité (1 pour très peu, et 5 pour très similaires) des paires présentées aléatoirement. La qualité des scores des méthodes est évaluée par rapport à leur corrélation avec les notes des annotateurs.

La démarche est intéressante de par son approche de la problématique de l'évaluation au travers du prisme de la psychologie-cognitive. L'évaluation ne porte pas sur la visée applicative mais sur l'alignement avec la perception humaine. De plus, elle permet de conserver les différents scores de similarité et ne nécessite pas une classification. Elle souffre toutefois d'imprécisions. Tout d'abord, elle réduit l'intégralité de la détection de dérivation à une mesure de similarité entre documents. S'il s'agit effectivement de l'approche de la majorité des méthodes, il semble épistémologiquement discutable de mesurer la performance d'une méthode à une tâche en se basant sur l'approche de cette méthode plutôt que sur l'objectif de la tâche. Enfin, la corrélation est calculée entre les résultats des méthodes définies sur l'espace image continu $[0; 1]$, tandis que les jugements humains sont définis sur un espace discontinu (valeurs discrètes).

En résumé, le protocole bien qu'intéressant nous paraît difficile à mettre en œuvre sur des textes plus longs tels qu'on les trouve dans nos corpus.

Nous discutons dans la section suivante des différentes ressources utilisées pour l'évaluation. Nous revenons dans la section 4.3 sur nos choix d'évaluation au regard des méthodes proposées dans cette section.

4.2 Corpus pour l'évaluation

Quel que soit le protocole d'évaluation choisi, tous impliquent la disponibilité d'un corpus dans lequel les liens de dérivation entre textes source et dérivé ont été annotés manuellement, souvent par des experts, à l'échelle du document ou du passage.

La complexité de la détection de dérivation varie avec la complexité du processus de dérivation comme l'illustre l'exemple 19, ce qui rend difficilement comparables les résultats obtenus sur des corpus différents. Plusieurs échelles de difficulté ont été proposées dans la littérature sans qu'aucune ne s'impose : Fullam et Park (2000) se focalisent sur la granularité [1], Shivakumar et Garcia-Molina (1996, p.5) utilise le

1. Ils proposent les différents échelons, par ordre décroissant de difficulté : copie exacte, copie de paragraphes, copie de phrases, modification de mots, modification de la structure des phrases.

Texte original
The quick brown fox jumps over the lazy dog.
Dérivations manuelles opérées par un humain
Over the dog which is lazy jumps quickly the fox which is brown.
Dogs are lazy which is why brown foxes quickly jump over them.
A fast auburn vulpine hops over an idle canine.

EXEMPLE 19: Comparaison de plusieurs dérivations de texte opérées par des humains lors de la constitution du corpus PAN-PC-2010 (Potthast et collab., 2010b)

taux subjectif de recouvrement des textes [2] et Potthast et collab. (2010b) utilisent la similarité moyenne entre dérivés et sources [3]. Nous nous contenterons de comparer les corpus selon les critères qualitatifs et quantitatifs utilisés pour PAN'10 (*cf. Tableau 4.2, page 104*).

Nous présentons par la suite les corpus majeurs pour l'anglais, à savoir METER (Gaizauskas et collab., 2001) et les corpus PAN-PC-09 et PAN-PC-10 construits pour les campagnes d'évaluation éponymes de la détection de plagiat (Potthast et collab., 2010b).

4.2.1 METER

Le corpus METER (Gaizauskas et collab., 2001) a été créé pour étudier la réutilisation de textes dans le journalisme et évaluer des approches automatiques de mesure de dérivation entre des dépêches d'agence et des articles de journaux (Clough, 2003a). Il s'agit d'un des tous premiers corpus dédié à l'étude de la détection de dérivation et plus précisément ici de la réutilisation de texte (*text reuse*). Ce corpus a été utilisé par la suite par Clough et collab. (2002), Piao et Mcenery (2004) et Uzuner et collab. (2004).

Le corpus est constitué de deux classes de documents : les dépêches d'agence et les articles de journaux. Tous ces documents ont été collectés pendant la même période de temps (entre mars 1999 et juin 2000) et dans les catégories justice et monde du spectacle. Chaque document collecté a été classé selon son origine (dépêche ou article de presse), sa catégorie, sa date de publication, son sujet et apparaît sous cet ordre au sein de la structure arborescente du corpus (*cf. Figure 4.4*). Les dépêches d'agence, tirées du flux de dépêches de la *Press Association*, constituent les sources. Les articles de presse constituent les candidats. Les journaux desquels sont extraits les candidats sont classés par les auteurs selon trois niveaux de « qualité éditoriale » : *tabloid*, *middle-range tabloid* et *broadsheet*.

La réutilisation de texte a été annotée au niveau du document et de la séquence de mots. Ces annotations ont été effectuées par une seule personne : un journaliste. Au niveau du document, il s'agit de définir le taux de dérivation entre la source et le candidat. Les auteurs proposent trois catégories : pas de dérivation (*not derived* : ND), dérivation partielle (*partially derived* : PD) et dérivation totale (*wholly derived* : WD). Les nuances du taux de dérivation sont fonction des évènements communément discutés : si les deux textes abordent les mêmes évènements ils sont totalement dérivés, si certains évènements traités par l'un ne sont pas repris par l'autre ils sont partiellement dérivés (Clough, 2003a, p.82). Au niveau de la séquence de mots, il s'agit de

2. Ils proposent les échelons : copies à l'identique (mot à mot), recouvrement important (*high overlap replies*) et recouvrement partiel.

3. Ils proposent de mesurer la difficulté comme le cosinus moyen entre sources et dérivés modélisés par un modèle vectoriel des n-grammes mots avec racinisation, filtrage des mots fonctionnels et pondération par tf.

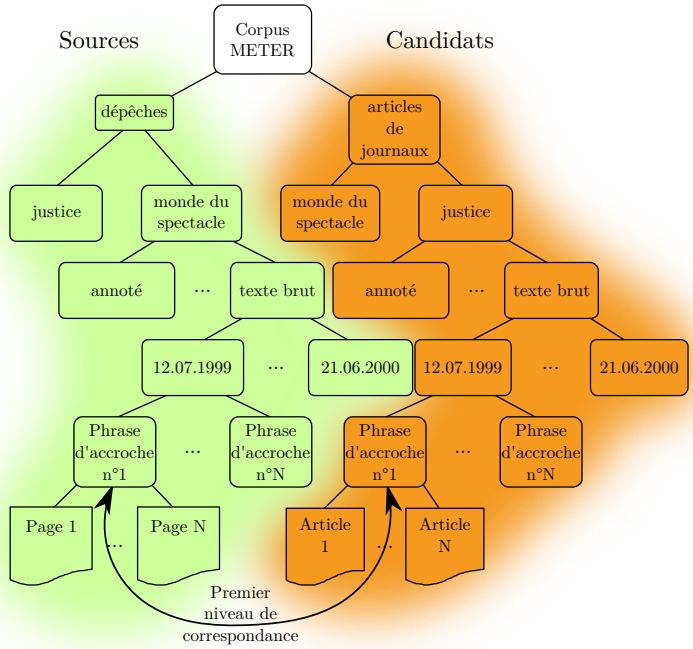


FIGURE 4.4 – Organisation arborescente du corpus METER. Schéma dérivé de Gaizauskas et collab. (2001)

définir les liens entre les passages de texte de la source et du candidat. Les auteurs proposent également trois catégories : verbatim, réécrit (paraphrase) et nouveau. Seul un sous-ensemble du corpus a été annoté de cette manière, et il n'existe pas à notre connaissance de travaux en ayant tiré parti.

Les liens de dérivation entre candidats et sources sont exprimés dans les annotations au niveau du document. Ainsi la combinaison des traits « classification » et « catchline » permet de rapprocher le dérivé de son éventuelle source.

Le tableau 4.1 présente la composition du corpus de manière synthétique. Les articles de journaux candidats sont au nombre de 770 pour la catégorie « justice » et couvrent une période de 24 jours, ceux de la catégorie « monde du spectacle » sont au nombre de 176 et couvrent une période de 13 jours.

4.2.2 PAN-PC-09 et PAN-PC-10

Les corpus PAN-PC-09 et PAN-PC-10[4] (Potthast et collab., 2010b) ont été développés dans le cadre de la campagne d'évaluation PAN (*Uncovering Plagiarism, Authorship, and Social Software Misuse*), le corpus PAN-PC-10 est une extension du corpus PAN-PC-09. Lorsque nous parlerons du corpus PAN, nous ferons référence à PAN-PC-10 mais la plupart des données statistiques sont extrapolées des données connues de PAN-PC-09.

Le corpus PAN ayant été développé pour une campagne d'évaluation, il se divise en une partie de développement (IPAT-DC) et une partie dédiée à l'évaluation pour la compétition (IPAT-CC). Le corpus de développement de PAN'10 correspond au corpus complet de PAN'09, la partie d'évaluation constituant la nouvelle contribution. Le

4. Au moment de l'écriture de ce manuscrit, seule une partie du corpus PAN-PC-10 a été diffusé.

Composition du corpus METER		
Genre des textes	articles de presse en ligne	
Langue	anglais	
Documents sources	152	14 %
Documents suspects	945	86 %
... dérivés	643	59 %
... non-dérivés	302	27 %
Caractéristiques des documents		
Taux de dérivation		
5 %–20 %	460	60 %
20 %–50 %	250	32 %
50 %–80 %	58	8 %
>80 %	0	0 %
Taille des documents		
0–2 500 mots	1 081	98 %
2 500–25 000 mots	16	2 %
25 000–250 000 mots	0	0 %

TABLE 4.1 – Composition du corpus METER. Le taux de dérivation est donné à titre de comparaison, il correspond au taux de recouvrement des mots entre la source et le candidat.

corpus PAN-PC-09 est également subdivisé entre les tâches de détection extrinsèque et intrinsèque. Ces deux tâches ont été regroupées en une seule pour PAN'10, mais la structure du corpus reflète toujours cette division. Les documents sources sont séparés des documents suspects qui regroupent à la fois les dérivés et des documents tiers. La figure 4.5 détaille cette structuration.

Le corpus est construit à partir de 40 000 textes du projet Gutenberg[5]. Il s'agit d'œuvres littéraires tombées dans le domaine public aux États-Unis ou bien disponibles sous une licence permissive. Les textes retenus couvrent de nombreux genres et thématiques. Ils sont de taille variable (de 3 000 à 2,5 millions de caractères). La langue principale est l'anglais mais quelques documents sources sont en allemand ou en espagnol, par contre les textes dérivés correspondant sont des traductions anglaises. Chaque document est représenté dans le corpus par un fichier de contenu (le document au format texte brut) et un fichier d'annotations en XML qui décrit les passages dérivés. Ces passages sont définis comme des segments de texte (index de début et de fin) et lient, pour la tâche de détection extrinsèque uniquement, le passage dérivé au passage source correspondant.

La construction du corpus PAN tire parti de deux techniques différentes : la génération automatique de plagiat et la dérivation manuelle. Le tableau 4.2, largement inspiré de Potthast et collab. (2010b), propose une synthèse de la composition du corpus.

La génération automatique se décompose en deux tâches. La première consiste à prélever des passages de différents textes sources et les insérer dans un texte source tiers afin d'obtenir un texte dérivé. La seconde, l'obufscation, consiste à brouiller certains de ces passages de manière à ce qu'ils ne correspondent pas à des reprises exactes du texte source. Cette technique a permis de générer artificiellement 94 202 passages plagiés et produire ainsi un corpus très diversifié en contrôlant plusieurs

5. `http://www.gutenberg.org`

Composition PAN'10		
Genre des textes	Textes du projet Gutenberg : ouvrages littéraires principalement	
Langue	90 % anglais, 10 % allemand ou espagnol	
Corpus d'entraînement		
Documents sources	14 429	50 %
Documents suspects	14 428 (+6 183 tâche intrinsèque)	50 %
... dérivés	7 214	25 %
... non-dérivés	7 214	25 %
Corpus d'évaluation		
Documents sources	11 148	41 %
Documents suspects	15 925	59 %
Caractéristiques des documents		
Taux de dérivation (sur 18 268 documents dérivés estimés)		
5 %–20 %	env. 8 220	45 %
20 %–50 %	env. 2 740	15 %
50 %–80 %	env. 4 567	25 %
>80 %	env. 2 740	15 %
Taille des documents (sur 62 113 documents estimés)		
0–2 500 mots	env. 31 056	50 %
2 500–25 000 mots	env. 21 739	35 %
25 000–250 000 mots	env. 9 316	15 %
Caractéristiques des dérivations		
Dérivation intra-thématique	env. 9 134	50 %
Dérivation inter-thématique	env. 9 134	50 %
Obfuscation (sur 100 214 cas de dérivation estimés)		
aucune	env. 40 085	40 %
artificielle légère	env. 20 042	20 %
artificielle importante	env. 20 042	20 %
humaine	env. 6 012	6 %
langue différente	env. 14 029	14 %
Taille des passages dérivés		
50 à 150 mots	env. 34 072	34 %
300 à 500 mots	env. 33 070	33 %
3 000 à 5 000 mots	env. 33 070	33 %

TABLE 4.2 – Composition du corpus PAN. Les nombres de documents approximatifs sont estimés en fonction de la taille totale du corpus et des pourcentages publiés dans Potthast et collab. (2010b). La taille des documents en nombre de mots est obtenue en considérant l'étalon de 250 mots par page généralement accepté par les éditeurs.

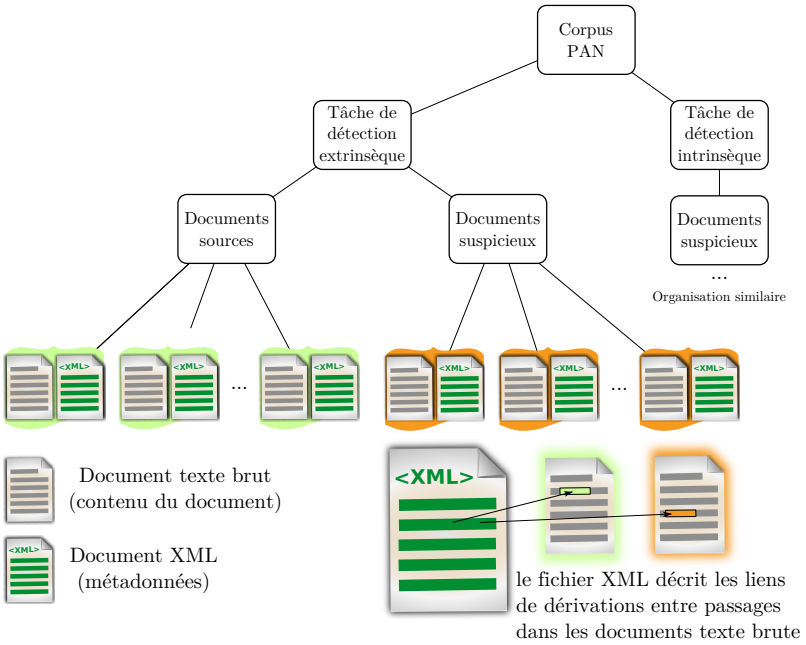


FIGURE 4.5 – Organisation du corpus PAN.

paramètres de la génération automatique :

- la taille des passages sélectionnés dans les textes sources varie uniformément entre 50 et 5 000 mots ;
- les passages sélectionnés proviennent de documents portant sur la même thématique que le document source ou alors une thématique différente. Les auteurs parlent respectivement de plagiats intra-thématiques et inter-thématiques ;
- la proportion de passages provenant de textes sources tiers par document suspect varie de 0 % à 100 %, avec 50 % de ces documents ne contenant aucun passage plagié ;
- plusieurs techniques d'obfuscation des passages sélectionnés ont été mises en œuvre pour simuler une réécriture.

Les différentes techniques d'obfuscation se veulent représentatives des formes de réécriture utilisées par les plagiaires. Cependant, elles ne garantissent pas que le résultat final soit syntaxiquement et stylistiquement correct. Ces obfuscations se basent sur des opérations d'édition (ajout, suppression, remplacement de mots) effectuées aléatoirement, des variations sémantiques (utilisation de synonymes, hyponymes, hyperonymes et antonymes) et le mélange aléatoire des mots de la phrase préservant la séquence originale des rôles grammaticaux. Les exemples 20 et 21 sont des passages tirés du corpus PAN correspondant respectivement à une obfuscation légère et importante.

La dérivation manuelle a été utilisée pour construire la partie évaluation de PAN-PC-10. Les auteurs ont employé la plateforme de *crowdsourcing* Mechanical Turk d'Amazon[6]. Les participants avaient pour mission de réécrire le texte qui leur était présenté de façon à ce que la version réécrite ait la même signification que l'originale mais utilise des tournures différentes (*cf. Exemple 19*). Les auteurs ont ainsi été en mesure de compiler plus de 6 000 dérivations manuelles. Malheureusement au moment

6. `http://www.mturk.com/`

SOURCE (*source-document03240.txt*)

It is to be observed, however, that the male plumage is nuptial merely, and is retained for a very short time; the rest of the year both sexes are plain alike. It is probable, therefore, that the domed nest is for the protection of these delicate little birds against the rain, and that there is some unknown cause which has led to the development of colour in the males only. There is one other case which at first sight looks like an exception, but which is far from being one in reality, and deserves to be

DÉRIVÉ (*suspicious-document00884.txt*)

indignation or reproach upon his good-humored but haggard features. state. It, therefore for, that other the, however the in world to be is happening which first far to be, cabin and deserves against is bridal disguise very short time; the continue for a and plain took disturbers alike color in there an n't, is has led to the development! of year both bondage are the domed beehive is expression like only. There one for and of these delicate asking looking little the, and merely exclusively is being one unknown

EXEMPLE 20: Passage source et la version fortement obfusquée correspondante.

SOURCE (*source-document03240.txt*)

Now there is a difficulty in this view of the origin of the structure of Orchids which the Duke does not allude to. The majority of flowering plants are fertilized, either without the agency of insects or, when insects are required, without any very important modification of the structure of the flower. It is evident, therefore, that flowers might have been formed as varied, fantastic, and beautiful as the Orchids, and yet have been fertilized without more complexity of structure than is found in Violets, or Clover, or Primroses, or a thousand other flowers. The strange springs and traps and pitfalls found in the flowers of Orchids cannot be necessary per se, since exactly the same end is gained in ten thousand other flowers which do not possess them. Is it not then an extraordinary idea, to imagine the Creator of the Universe contriving the various complicated parts of these flowers, as a mechanic might contrive an ingenious toy or a difficult puzzle? Is it not a more worthy conception that they are some of the results of those general laws which were so co-ordinated at the first introduction of life upon the earth as to result necessarily in the utmost possible development of varied forms?

DÉRIVÉ (*suspicious-document00884.txt*)

Is it not then an extraordinary idea, to imagine the Creator of the Universe contriving the various complicated parts of would each confer on the other an advantage in the battle of. life. This would tend to their respective perpetuation and to the constant lengthening of nectaries and probosces. Now, that the Creator of the Universe, a direct act of his Will, so disposed the natural forces influencing the growth of this one species of plant as to cause its nectary to to this enormous length; and at the same time, by an equally special act, determined the flow of nourishment in the organization of the moth, so as to cause its proboscis to increase in exactly the same proportion, having previously so constructed the Angræcum that it could only be maintained in existence by the agency of moth. But what proof is given or suggested that this was the mode by which the adjustment took place?

None whatever, except a, feeling that there is an adjustment of a delicate kind, and inability to see how known causes could have produced such an adjustment. I believe I with have shown, however, that such an adjustment is not only possible but inevitable, unless at some! point or other we deny the action of those simple laws which we have already And admitted to be but the leer of be necessary per se, since get exactly the same end gained in ten thousand other let it be remembered, that what we have to account a foot long.

EXEMPLE 21: Passage source et la version légèrement obfusquée correspondante.

de l'écriture de cette thèse, cette partie du corpus n'a pas encore été publiée.

4.2.3 Corpus secondaires

Tout comme la tâche de détection de dérivation, la constitution de corpus de référence pour l'évaluation est assez récente. De nombreux autres corpus ont été construits mais n'ont été utilisés que par leurs seuls auteurs. Nous les décrivons brièvement par la suite, regroupés selon le genre des documents qui les composent.

Presse Au-delà de METER, au moins trois autres corpus issus de la presse ont été construits. Premièrement, Lyon et collab. (2001) ont collecté 335 transcriptions de journaux télévisés en anglais, soit près d'un million de mots. Deuxièmement, Bao et collab. (2007) ont exploité un corpus d'articles de presse du Financial Times construit dans le cadre de la campagne TREC. Il se décompose en collections de 500 à 1 500 textes qui correspondent à une thématique particulière, décrite par un fichier d'amorce. La similarité thématique des textes de chaque collection avec le fichier d'amorce a été évaluée manuellement. Finalement, le corpus de Lee et collab. (2005) est particulier puisqu'il a été développé dans le cadre d'une évaluation cognitive de la similarité. Les articles tirés de l'*Australian Broadcasting Corporation* sont de petite taille (entre 51 et 126 mots), peu nombreux (une cinquantaine) et couvrent des thématiques différentes.

Technique Plusieurs évaluations ont été menées sur des corpus constitués de textes destinés à des spécialistes : RFC [7], articles scientifiques et requêtes administratives. Ainsi, la collection RFC a été exploitée par Finkel et collab. (2002) et Stein et Eissen (2006). Les premiers se sont intéressés au regroupement par familles à l'aide d'une technique de couverture de texte. Les seconds ont cherché à détecter des révisions de document en exploitant le versionnement de ces RFC. Sorokina et collab. (2006) ont recherché des plagiats dans le site arXiv [8] qui regroupe plus de 375 000 articles scientifiques en physique, informatique et biologie. Finalement, Yang (2006a) a constitué des corpus (NTF1 et NTF2) regroupant chacun un millier de courriers électroniques à destination de l'administration américaine sur la régulation de l'utilisation des gaz toxiques. Ces lettres dérivent fréquemment de lettres types écrites par des ONG. Les liens de dérivation ont été annotés manuellement jusqu'à obtenir un accord inter-annotateur de 90 %.

Version Le recours à des versions de documents est moins fréquent. Stein et Eissen (2006) ont tiré parti du versionnement des RFC. Potthast et Stein (2007) ont été plus loin en exploitant les 6 millions d'articles de Wikipédia et ses 80 millions de révisions. L'intérêt énoncé par les auteurs est que ces révisions correspondent majoritairement à des réécritures et des corrections. Nous pouvons également citer le corpus en français de Bourdaillet (2007) constitué de versions générées artificiellement.

Multilingue La dimension multilingue de la dérivation a reçu assez peu d'attention pour le moment. Outre le corpus PAN-PC-10 qui offre 10 % de dérivations multilingue (allemand et espagnol vers anglais), Özlem Uzuner et Davis (2003) ont utilisé des corpus parallèles pour étudier la similarité de contenu et d'expression pour des dérivations

7. Les *requests for comments* (RFC) sont une série de documents officiels décrivant les aspects techniques d'Internet. Ils sont comparables dans leur contenu aux textes décrivant les normes et standards.

8. `http://arxiv.org/`

d'une langue vers une autre. Les textes utilisés sont majoritairement des œuvres littéraires (20 000 lieux sous les mers de J. Verne, Madame Bovary de G. Flaubert, et Kreutzer Sonata de Tolstoy).

Au final, une grande partie des corpus exploités a été construite par des méthodes automatiques de synthèse des dérivations similaires à celles mises en œuvre pour PAN. Ces corpus sont construits à partir de divers matériaux : pages Web (Fetterly, 2005), travaux d'étudiants (Bao et collab., 2006), articles scientifiques (Eissen et Stein, 2006) ou encore des définitions du dictionnaire Collins Cobuild (O'Shea et collab., 2008).

Discussion

Au final, deux approches se détachent pour la création de corpus dédiés à la dérivation de texte. La première approche consiste à générer artificiellement des dérivations à l'aide d'algorithmes simulant des éditions. PAN en est la meilleure illustration. Elle permet de contrôler précisément le volume du corpus, la répartition entre sources et dérivés et la difficulté des dérivations. Toutefois, de par leur création artificielle, les dérivations correspondent à l'idée que s'en font les auteurs dans la limite des techniques disponibles, notamment en termes de justesse de la syntaxe et du style. La seconde approche consiste à collecter manuellement des documents sources et dérivés. METER en est la meilleure illustration. Elle permet d'obtenir des cas réels de dérivations. Toutefois, la démarche est très lourde et il n'est pas possible de savoir si les dérivés collectés dérivent réellement de la source associé ou non. L'humain en charge de la construction en est le seul juge.

À notre connaissance, les corpus METER et PAN sont les deux seuls à avoir été utilisés dans plusieurs travaux d'auteurs distincts. Le corpus PAN notamment est une initiative heureuse qui permet de mettre à disposition des collections textuelles de référence et qui couvre plusieurs formes de dérivations (variation de la granularité et de l'intégration). La partie obtenue par génération artificielle pose les questions évoqués précédemment, mais la partie évaluation simulée par des humains a un réel intérêt et devrait rapidement s'affirmer comme un corpus de référence du domaine pour l'anglais.

En conclusion, nous notons deux obstacles en l'état à nos travaux : aucun corpus n'est disponible en français et les corpus existant mélangent indifféremment les formes de dérivation.

4.3 Notre méthode d'évaluation inspirée de la RI

La section 4.1 a présenté les protocoles d'évaluation existants. Sur ces protocoles présentés nous avons choisi de retenir celui de la RI même s'il est très peu utilisé dans la littérature liée à la détection de dérivation et permet donc difficilement de se positionner.

Nous présentons dans cette section les objectifs de l'évaluation que nous souhaitons mener ainsi que nos choix méthodologiques et les ressources constituées.

4.3.1 Objectifs de l'évaluation

L'objectif global de l'évaluation est de mesurer la qualité de la détection des liens de dérivation entre documents sources et dérivés pour les différentes méthodes mises en œuvre, afin d'obtenir des arguments permettant de conforter ou réfuter les choix sous-tendant ces méthodes. Pour ce faire et en accord avec une démarche scientifique, l'évaluation doit pouvoir être reproduite et les résultats obtenus doivent rester stables lors des reproductions.

Rang	Classe	similarité	Rang	Classe	similarité
1	dérivé	1	10	non-dérivé	0,3
2	dérivé	1	11	non-dérivé	0,2
3	non-dérivé	0,9	12	non-dérivé	0,15
4	dérivé	0,9	13	non-dérivé	0,1
5	dérivé	0,85	14	non-dérivé	0
6	dérivé	0,8	15	dérivé	0
7	dérivé	0,7	16	non-dérivé	0
8	dérivé	0,7	17	non-dérivé	0
9	dérivé	0,4	18	non-dérivé	0

TABLE 4.3 – Scores de similarité exemples pour illustrer la méthode d'évaluation.

Nous nous concentrons sur les liens de dérivation à l'échelle du document. En d'autres termes nous cherchons à pouvoir répondre à la question « ce document suspect dérive-t-il de ce document source? » L'évaluation doit refléter la capacité des méthodes à identifier correctement les documents dérivés, étant donné un document source. Par extension, l'évaluation doit également mesurer la capacité des méthodes à exclure les documents qui ne sont pas dérivés. Les performances en termes de temps de calcul et d'espace de stockage nécessaires à la mise en œuvre de la méthode font également partie prenante de l'évaluation.

4.3.2 Méthodologie et mesures

Les méthodes que nous expérimentons calculent une mesure de similarité entre deux documents. Les scores de similarité en sortie sont, si la méthode utilisée est correcte, corrélés à la probabilité d'existence du lien de dérivation testé. L'évaluation que nous avons choisie de mener porte sur différentes formes de dérivation (*cf. la section 4.3.3 présentant les corpus*) et repose sur trois axes : la qualité de l'identification des liens de dérivation, la capacité de discrimination entre les liens avérés et non-avérés, et les performances des méthodes en termes de temps de calcul et d'espace de stockage.

Afin d'illustrer le calcul des différentes mesures utilisées, nous utiliserons en exemple la distribution des scores de similarité du tableau 4.3 correspondant à des comparaisons entre sources et suspects fictifs.

4.3.2.1 Qualité de l'identification des liens de dérivation

La qualité de l'identification des liens de dérivation est la propriété du système qui nous paraît la plus pertinente et la plus importante à évaluer.

Dans une évaluation comme une tâche de RI, les paires de textes sont classées par score de similarité décroissant [9] (*cf. Tableau 4.3, page 109*).

Nous évaluons la qualité de l'identification des liens de dérivation en utilisant la métrique MAP (*cf. Équation 4.5*) sur le classement des paires. Cette mesure calcule la précision (*cf. Équation D.1*) en considérant itérativement les rangs de 1 à n et en en faisant la moyenne. Nous choisissons pour n le rang de la dernière paire du classement correspondant à un lien de dérivation avéré. Dans notre exemple, le dernier lien de

MAP

9. Les scores de similarité en sortie de nos systèmes permettent de définir un ordre partiel (plusieurs liens peuvent obtenir le même score), qui peut être simplement étendu en un ordre strict en positionnant aléatoirement les uns par rapport aux autres les paires obtenant le même score.

rang	VP	FP	$\mathcal{P}(rang)$
1	1	0	1
2	2	0	1
3	2	1	$\frac{2}{3}$
4	3	1	0,75
5	4	1	0,8
6	5	1	$\frac{5}{6}$
7	6	1	$\frac{6}{7}$
8	7	1	0,875
9	8	1	$\frac{8}{9}$
10	8	2	0,8
11	8	3	$\frac{8}{11}$
12	8	4	$\frac{2}{3}$
13	8	5	$\frac{8}{13}$
14	8	6	$\frac{4}{7}$
15	9	6	0,6

$$\begin{aligned} \text{MAP} &= \frac{1}{15}\sum_{i=1}^{15}\mathcal{P}(i) \\ &\approx 0,776 \end{aligned}$$

FIGURE 4.6 – Mise en œuvre du calcul de la MAP sur notre exemple. Les colonnes VP et FP correspondent respectivement aux liens de dérivation cumulés et aux non liens de dérivation cumulés. La colonne $\mathcal{P}(rang)$ est la précision du sous-ensemble jusqu'au rang *rang*.

dérivation est situé au rang 15. La figure 4.6 rapporte le détail du calcul de la MAP pour cet exemple.

La MAP rend compte à la fois des notions de précision et de rappel et permet donc d'évaluer la qualité de la classification des paires.

4.3.2.2 Capacité de discrimination

La capacité de discrimination de la méthode consiste à évaluer l'espace tampon qui sépare les scores de similarité des documents dérivés de ceux des documents non-dérivés. Plus cet espace est important, plus les résultats en sortie des méthodes sont différents entre les documents dérivés et non-dérivés et donc moins la marge d'erreur est problématique.

Nous cherchons notamment à ce que les scores de similarité des non-dérivés se concentrent autour de 0, tandis que ceux des dérivés se concentrent autour de 1. Il faut éviter autant que possible que les scores des deux classes se croisent dans l'espace image, et plutôt privilégier une zone tampon la plus large possible dans laquelle aucune des deux classes n'a de valeur comme l'illustre la figure 4.7.

L'utilisation de la métrique de séparation (*cf. Équation 4.6*) nous paraît la plus appropriée pour rendre compte de cette capacité de discrimination. Dans le contexte de notre exemple, le plus bas dérivé (LCR) a un score de 0 et le plus haut non-dérivé (HFM) 0,9, soit une séparation de $-0,9$. Toutefois, les minimums et maximums ne sont pas forcément représentatifs et sont difficiles à interpréter, notamment lorsque ces individus sont éloignés du reste du groupe. Nous introduisons donc une nouvelle mesure reprenant l'idée de la séparation mais utilisant les quartiles de chaque population. Nous la nommons *séparation des quartiles* (Sép. Q). Il s'agit de la différence entre le premier quartile (score de similarité qui sépare les 25 % inférieurs) des documents dérivés et le troisième quartile (score de similarité qui sépare les 75 % supérieurs) des

Sép. Q

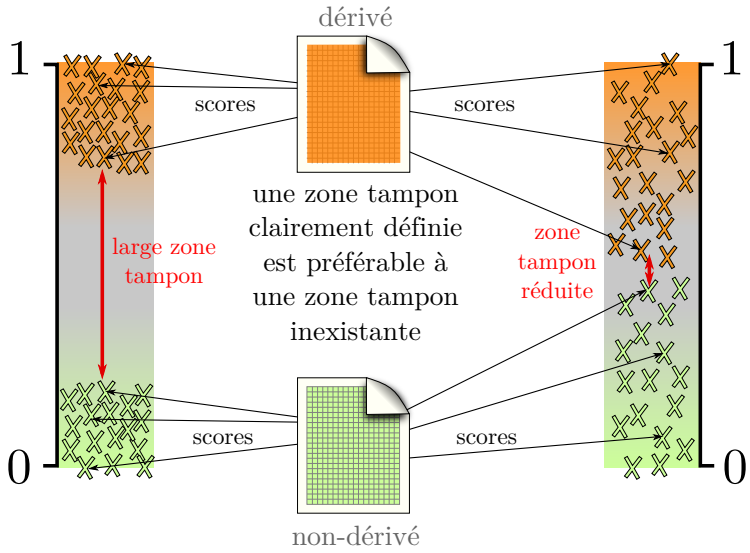


FIGURE 4.7 – Il est préférable que les scores de similarité obtenus pour les dérivés soient différents de ceux obtenus par les non-dérivés, créant ainsi une zone tampon qui sépare les premiers des seconds.

documents non-dérivés (*cf. Équation 4.8*).

$$\text{Sép. Q} = 1^{er}\text{quartile dérivés} - 3^{e}\text{quartile non-dérivés} \tag{4.8}$$

Dans notre exemple, le premier quartile des dérivés vaut 0,4, contre 0,2 pour le troisième quartile des non-dérivés, ce qui donne :

$$\begin{aligned} \text{Sép. Q} &= 1^{er}\text{quartile dérivés} - 3^{e}\text{quartile non-dérivés} \\ &= 0,4 - 0,2 \\ &= 0,2 \end{aligned}$$

4.3.2.3 Performances en temps et en espace

Le temps et l'espace disque requis nécessaires à l'exécution des méthodes sont deux propriétés potentiellement aussi importantes que la qualité de l'identification des liens de dérivation lors de la mise en production.

La complexité spatiale et temporelle la plus précise possible serait le meilleur indice. Cependant, elle est difficile à évaluer, notamment à cause des outils de prétraitement dont la complexité dépend des données en entrée ainsi que de l'état courant du modèle.

Nous faisons le choix d'évaluer cette complexité expérimentalement en calculant, en fonction de la taille de chaque document, le temps d'exécution nécessaire à la modélisation et la taille de ladite modélisation. Les résultats sont bien entendus tributaires de la machine sur laquelle ils sont calculés. Par conséquent, toutes les expérimentations devront être évaluées sur la même machine et dans les mêmes conditions. Nous complétons ces mesures par le temps nécessaire à la comparaison des modélisations.

```
<?xml version="1.0" encoding="UTF-8"?>
<document
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:noNamespaceSchemaLocation="http://www.uni-weimar.de/
      medien/webis/research/corpora/pan-pc-09/document.xsd"
  reference="suspicious-document12001.txt">
  <feature
    name="project-gutenberg" etext_number="19028"
    url="http://www.gutenberg.org/files/19028/19028-8.txt"/>
  <feature name="language" value="en" />
  <!-- Use tags like the one below to annotate plagiarism you
      detected. -->
  <!-- <feature name="detected-plagiarism"
        this_offset="50" this_length="1000"
        source_reference="source-documentX.txt"
        source_offset="750" source_length="1700"/> -->
</document>
```

Listing 4.1 – Exemple de fichier XML de description d'un document du corpus.

4.3.3 Corpus PIITHIE, Wikinews et PANini

Nous souhaitons évaluer la qualité de nos méthodes sur différents types de dérivation. Nous avons pour cela constitué plusieurs corpus. Le corpus Piithie est similaire au corpus METER (Clough et collab., 2002) en ce qu'il représente les dérivations rencontrées dans la presse, mais en français. Le corpus de révision Wikinews représente les dérivations de type révision, toujours en français. Finalement, le corpus PANini est un sous-ensemble du corpus PAN d'une taille abordable pour nos expérimentations.

4.3.3.1 Caractéristiques communes

Nos corpus partagent plusieurs caractéristiques communes, majoritairement structurelles.

DOCUMENTS SOURCES
DOCUMENTS DÉRIVÉS
DOCUMENTS NON-DÉRIVÉS

Tout d'abord, nous distinguons pour chacun les *documents sources*, les *documents dérivés* de ces documents sources et les documents qui ne sont ni l'un, ni l'autre et que nous nommerons *non-dérivés* (ou *tiers*). Les documents sources sont isolés car leur rôle est indépendant du contexte. À l'opposé les documents dérivés et non-dérivés sont mélangés car leur rôle dépend du document source considéré. Par exemple si l'on considère les documents sources s_1 et s_2 tels que $s_1 \mathbf{R_D} d_1$ et $s_2 \mathbf{R_D} d_2$, alors lorsque l'on travaillera sur s_1, d_1 aura un rôle de dérivé et d_2 de non-dérivé. Inversement lorsque l'on travaillera sur s_2, d_2 aura un rôle de dérivé et d_1 de non-dérivé.

Ensuite, le mode d'organisation des corpus est copié sur celui du corpus PAN. Chaque document du corpus est représenté par deux fichiers : le contenu textuel du document et des méta-données décrivant l'origine du document ainsi que les éventuels liens de dérivation dans lesquels il est impliqué. Ce dernier fichier est au format XML, le listing 4.1 donne un aperçu de la syntaxe utilisée.

Enfin, nous faisons le choix d'équilibrer le nombre de dérivés et de non-dérivés comparés à chaque source. Pour ce faire, nous programmons pour chaque source une comparaison avec chaque document dérivé connu ainsi qu'avec autant de documents non-dérivés prélevés aléatoirement dans le corpus, exception faite des sus-mentionnés documents. De plus, à des fins de comparabilité, cette sélection n'est effectuée qu'une

seule fois en amont de toutes nos expérimentations de sorte que cette liste soit figée et partagée entre toutes nos expérimentations.

4.3.3.2 Corpus Piithie

Le corpus Piithie est largement inspiré du corpus METER, notamment dans sa construction.

Dans un premier temps nous avons ciblé les agences de presse telles que l'AFP (Agence France-Presse) et REUTERS qui sont habituellement les fournisseurs de contenus originaux pour la presse en ligne. Nous nous sommes concentré sur des nouvelles décrivant des évènements politiques ou sociaux durant la période de novembre à décembre 2008. Cette sélection constitue les documents sources.

Dans un second temps, nous avons tenté de rassembler, à l'aide de moteurs de recherche, des articles de presse traitant de ces mêmes évènements et qui auraient ainsi pu être dérivés des dépêches d'agence. Nous nous sommes limités à certains sites d'actualités afin d'assurer une homogénéité de genre. Les annotateurs linguistes ont ensuite extrait les documents dérivés et non-dérivés dans les documents résultats. Puis, ils les ont manuellement étiquetés afin de lier les documents dérivés à leur source, en s'assurant entre autre chose de la postériorité de la date de publication.

L'annotation a permis d'identifier des relations de dérivation au niveau du document et au niveau du paragraphe. Le premier niveau permet ainsi de rattacher le document à sa source. De par la méthode employée, chaque document annoté est en relation de dérivation avec une seule source maximum. Contrairement à PAN, Piithie ne contient donc pas de dérivations multisources. Le second niveau tisse les liens de dérivation entre les paragraphes de la source et du dérivé, en précisant pour chaque lien si l'intégration est verbatim ou bien s'il y a eu réécriture.

La forme textuelle des dérivés annotés étant extrêmement proche de celle de leurs sources nous avons décidé de compléter le corpus par des dérivations manuelles de façon similaire à ce qui a été fait pour PAN'10. L'objectif était d'augmenter le nombre de documents dérivés d'une part, et d'inclure des dérivations dont la forme textuelle est plus éloignée de celle de la source d'autre part. Les annotateurs avaient cinq minutes pour opérer chaque dérivation. Ces dérivations « simulées » sont identifiées et représentent 26 % de tous les cas de dérivations.

Au final, nous avons retenu 85 textes sources qui sont impliqués dans des relations de dérivation avec 717 documents dérivés. À ceux-ci s'ajoutent 308 documents tiers qui ne sont en relation de dérivation avec aucun document source, mais qui traitent de sujets similaires. Le tableau 4.4 présente la composition du corpus de manière synthétique.

4.3.3.3 Corpus de révisions Wikinews

Un autre type de dérivation de texte qui a un réel intérêt applicatif tout en étant très différent des dérivations entre agences de presse et journaux est la révision. Une révision est une modification mineure apportée par un contributeur à l'œuvre afin de corriger ou compléter l'existant. La critique génétique textuelle (Bourdaillet, 2007) s'y intéresse pour les œuvres littéraires.

Potthast et Stein (2007) ont utilisé la Wikipédia anglophone comme ressource pour une telle étude. Nous souhaitons toutefois conserver le genre de l'article de presse afin de pouvoir comparer les résultats obtenus avec ceux du corpus Piithie. Nous nous sommes donc tourné vers un autre projet de la fondation Wikimédia : Wikinews.

Wikinews est un site collaboratif dans la lignée de Wikipédia, qui publie des articles de presse sur l'actualité. Les faits sont relatés directement par ceux qui les vivent ou

Composition Piithie		
Genre des textes	articles de presse en ligne	
Langue	français	
Documents sources	85	8 %
Documents suspects	1 025	92 %
... dérivés	717	65 %
... non-dérivés	308	27 %
Caractéristiques des documents		
Taux de dérivation (sur 717 documents dérivés)		
5 %–20 %	2	<1 %
20 %–50 %	5	<1 %
50 %–80 %	6	<1 %
>80 %	704	98 %
Taille des documents (sur 1 110 documents)		
0–2 500 mots	1 104	99 %
2 500–25 000 mots	6	1 %
25 000–250 000 mots	0	0 %
Caractéristiques des dérivations		
Obfuscation (sur 6 585 cas de dérivation)		
aucune	3 656	55 %
réécriture journalistique	1 229	19 %
simulée	1 700	26 %
Taille des passages dérivés		
moins de 50 mots	807	12 %
50 à 150 mots	1 578	24 %
150 à 300 mots	2 055	31 %
300 à 500 mots	1 489	23 %
500 à 3 000 mots	656	10 %

TABLE 4.4 – Composition du corpus Piithie

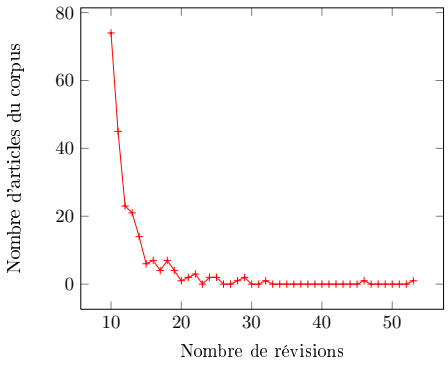


FIGURE 4.8 – Distribution du nombre de révisions par pages pour notre corpus Wikinews.

bien sont des synthèses de sources tierces. Les articles sont publiés sous licence *Creative Commons Attribution 2.5*[10] ce qui nous permet de les utiliser et les redistribuer librement.

Nous avons construit notre corpus à partir de l'export de la version française de Wikinews[11] en date du 13 novembre 2009 disponible sur `http://download.wikimedia.org`. Nous avons utilisé les composants que nous avons développés autour de Wikipédia[12] pour sélectionner automatiquement dans cette archive les articles possédant au moins 10 révisions. Le graphique de la figure 4.8 montre la répartition du nombre de révisions par article ajouté à notre corpus, et le tableau 4.5 expose les articles contenant le plus de révisions. Nous avons ensuite exporté les différentes révisions de ces articles, puis utilisé notre parseur MediaWiki pour supprimer les balises de mise en forme et obtenir ainsi les révisions au format texte brut. Nous avons uniquement supprimé la page d'accueil qui comptabilisait beaucoup plus de révisions que les autres et qui ne correspondait pas à un article de presse, ainsi que l'article 12 412 écrit en langue anglaise.

Nous avons également repris pour ce corpus le modèle source/dérivé que nous avons utilisé pour Piithie. La notion est moins évidente dans le cadre des révisions puisque chaque version du document est dérivée de la version précédente et source de la version suivante. Nous avons fait le choix de considérer la version la plus ancienne comme la source et toutes les versions postérieures comme ses dérivés. Nous avons annoté ces liens de dérivation dans les fichiers de méta-données comme l'illustre le listing 4.2 :

- une annotation au niveau du document qui lie le dérivé à sa source tout en indiquant la profondeur de la version dans l'arbre de révision (la profondeur 0 correspond à la source) ;
- une annotation au niveau du passage, principalement pour des raisons de compatibilité avec PAN, mais qui couvre l'intégralité du texte source et du texte dérivé. Cette annotation pourrait être affinée manuellement mais nous avons manqué de ressources pour le faire.

Au final, le corpus se compose de 221 textes sources qui sont impliqués dans des relations de dérivation avec 2 670 documents dérivés. Le corpus ne contient aucun texte non-dérivé, les dérivés des autres sources jouant le rôle de non-dérivé pour

10. `http://creativecommons.org/licenses/by/2.5/`
11. `http://fr.wikinews.org`
12. Le code de ces composants est disponible sous licence Apache à l'adresse : `http://code.google.com/p/uima-mediawiki-engine/`

Id. page	Date	Phrase d'accroche	Nb. révisions
15 337	11 janvier 2008	L'alpiniste néo-zélandais Sir Edmund Hillary est décédé à 9 heures, heure locale, des suites d'un infarctus du myocarde.	53
20 208	15 janvier 2009	Le vol 1549 US Airways est un vol de la compagnie aérienne US Airways qui a subi un accident aérien le 15 janvier 2009 à New York, aux États-Unis.	46
2 950	7 juillet 2005	Les explosions ont frappé des bus et le métro londonien.	32
8 384	24 novembre 2006	Responsable de Parti National (English : National Party) Néo-Zélandais, Don Brash, a résigné jeudi son emploi comme responsable d'opposition parlementaire.	29
22 755	1 juin 2009	Le vol AF 447 de la compagnie Air France a disparu.	29
3 386	14 août 2005	Un boeing 737 avec le numéro de vol HCW 502 de la compagnie chypriote Ilios, avec 121 passagers à bord dont 56 enfants et 6 membres d'équipage en provenance de Larnaca en Chypre s'est écrasé non pas sur une montagne de la presqu'île d'Eubée, comme la tour de contrôle de l'aéroport d'Athènes l'avait indiqué dans un premier temps mais sur une zone non habitée à moins d'un kilomètre de Varnava, à 40km d'Athènes, ce matin peu avant 11h30.	28
4 419	12 décembre 2005	Une série d'explosions importantes a eu lieu à proximité de Hempstead.	25
4 395	8 décembre 2005	Le Mouvement international de la Croix-Rouge et du Croissant-Rouge se dote d'un nouveau symbole, le cristal rouge.	25
10 460	4 avril 2007	Le gouverneur de l'État de Virginie, Tim Kaine a signé un projet de loi, vendredi 23 mars 2007, visant à interdire aux nouveaux conducteurs âgés de moins de 18 ans l'utilisation d'un téléphone mobile pendant la conduite d'un véhicule automobile.	24

TABLE 4.5 – Articles du corpus comptabilisant le plus de révisions.

```
<?xml version="1.0" encoding="UTF-8"?>
<document
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:noNamespaceSchemaLocation="http://www.uni-weimar.de/
      medien/webis/research/corpora/pan-pc-09/document.xsd"
  reference="suspicious-document00001.txt">
  <feature name="wikinews_metadata"
    page_id="4361" revision_id="55" source="Wikinews␣France"
    url="http://fr.wikinews.org/w/index.php?oldid=55"/>
  <feature name="language" value="fr"/>
  <!--Document level derivation information-->
  <feature name="annotated-doclevel-revision"
    revision_depth="1" source_reference="source-document00001
        .txt"/>
  <!--Local level derivation (whole document as it is a
      revision)-->
  <feature name="annotated-locallevel-revision"
    source_length="337" source_offset="0"
    source_reference="source-document00001.txt"
    this_length="2894" this_offset="0"/>
</document>
```

Listing 4.2 – Exemple d'annotation d'un document du corpus Wikinews.

une source donnée. Le tableau 4.6 présente la composition du corpus de manière synthétique.

4.3.3.4 Corpus réduit PAN (PANini)

La taille du corpus PAN et des documents qui s'y trouvent nous posent quelques soucis. Il a en effet été construit pour une évaluation globale de tous les systèmes de détection de dérivation :

- il est extrêmement volumineux : plus de 10 Go pour la version PAN-PC-10 complète, ce qui ne correspond pas à nos besoins expérimentaux ;
- il regroupe des documents de tailles très diverses : d'une cinquantaine de mots à plusieurs centaines de milliers, ce qui est assez éloigné des tailles des articles de presse des précédents corpus ;
- les dérivations mélangent des passages de plusieurs documents alors que nous travaillons plutôt avec des dérivés mono-source ;
- les obfuscations sur les passages ne correspondent pas aux intégrations manuelles, nous aurions préféré travailler avec les dérivations simulées mais les auteurs n'ont pas encore distribué les annotations sur cette partie du corpus ;
- les textes sont en anglais et de genre littéraire.

Malgré ces différences par rapport aux choix que nous avons fait pour les corpus Piithie et Wikinews, il nous semble primordial d'expérimenter nos méthodes sur ce corpus de référence.

Nous avons compilé un sous-ensemble du corpus PAN qui répond le mieux à nos besoins tout en restant globalement comparable au corpus original. Ainsi, nous l'avons restreint aux documents suspects de moins de 2 500 mots desquels nous avons filtré les traductions. Finalement, nous avons sélectionné les documents sources avec lesquels les documents restants avaient des liens de dérivation. Les annotations de ces

Composition Wikinews		
Genre des textes	articles de presse en ligne	
Langue	français	
Documents sources	221	7 %
Documents suspects	2 670	93 %
... dérivés	2 670	93 %
... non-dérivés	0	0 %
Caractéristiques des documents		
Taux de dérivation (sur 2 670 documents dérivés)		
5 %–20 %	0	0 %
20 %–50 %	0	0 %
50 %–80 %	0	0 %
>80 %	2 670	100 %
Taille des documents (sur 2 891 documents)		
0–2 500 mots	2 875	99 %
2 500–25 000 mots	16	1 %
25 000–250 000 mots	0	0 %
Caractéristiques des dérivations		
Obfuscation (sur 2 891 cas de dérivation)		
révision	2 891	100 %
Taille des passages dérivés		
moins de 50 mots	32	1 %
50 à 150 mots	485	18 %
150 à 300 mots	987	37 %
300 à 500 mots	712	27 %
500 à 3 000 mots	441	17 %
3 000 à 5 000 mots	13	1 %

TABLE 4.6 – Composition du corpus de révisions Wikinews

Composition de PANini		
Genre des textes Langue	Textes du projet Gutenberg : ouvrages littéraires principalement anglais	
Documents sources	940	41 %
Documents suspects	1 362	59 %
... dérivés	681	30 %
... non-dérivés	681	30 %
Caractéristiques des documents		
Taux de dérivation (sur 681 documents dérivés)		
5 %–20 %	181	27 %
20 %–50 %	136	20 %
50 %–80 %	179	26 %
>80 %	136	20 %
Taille des documents (sur 2 302 documents)		
0–2 500 mots	1 604	70 %
2 500–25 000 mots	377	16 %
25 000–250 000 mots	321	14 %
Caractéristiques des dérivations		
Obfuscation (sur 2 349 cas de dérivation estimés)		
aucune	1 145	49 %
artificielle légère	576	25 %
artificielle importante	628	26 %
Taille des passages dérivés		
moins de 50 mots	792	34 %
50 à 150 mots	854	36 %
150 à 300 mots	185	8 %
300 à 500 mots	289	12 %
500 à 3 000 mots	229	10 %

TABLE 4.7 – Composition du corpus réduit PAN (PANini)

documents n'ont pas été altérées, leurs contenus textuels non plus.

Au final, le corpus se compose de 940 textes sources qui sont impliqués dans des relations de dérivation avec 681 document dérivés. À ces dérivés s'ajoutent 681 textes tiers sélectionnés aléatoirement afin de revenir à la proportion 50/50 caractérisant le corpus PAN. Le tableau 4.7 présente la composition du corpus de manière synthétique.

4.3.3.5 Discussion

Le tableau 4.8 met en perspective la composition des corpus de référence PAN et METER par rapport aux corpus utilisés dans le cadre de cette thèse (Piithie, Wikinews et PANini).

Nos corpus sont une réelle contribution pour l'étude des détections de dérivation :

- Piithie et Wikinews sont à l'heure actuelle et à notre connaissance les seuls corpus pour l'étude de la dérivation en langue française ;
- Piithie et Wikinews ciblent des formes de dérivation précises. Piithie se veut, à l'image de METER, descriptif des dérivations mises en œuvre dans la presse écrite entre les dépêches d'agence et les articles de journaux. Wikinews décrit,

	PAN	METER	PIITHIE	WikiNews	PANini
Genre des textes	littéraire	presse	presse	presse	littéraire
Langue	90 % anglais, 10 % allemand ou espagnol	anglais	français	français	anglais
Taille	10 Go	31 Mo	23 Mo	54 Mo	189 Mo
Documents sources	50 %	14 %	8 %	7 %	41 %
Documents suspects	50 %	86 %	92 %	93 %	59 %
... dérivés	25 %	59 %	65 %	93 %	30 %
... non-dérivés	25 %	27 %	27 %	0 %	30 %
Nb. sources par dérivé	plusieurs	une	une	une	plusieurs
Caractéristiques des documents					
Taux de dérivation					
5 %–20 %	45 %	60 %	<1 %	0 %	27 %
20 %–50 %	15 %	32 %	<1 %	0 %	20 %
50 %–80 %	25 %	8 %	<1 %	0 %	26 %
>80 %	15 %	0 %	98 %	100 %	20 %
Taille des documents					
0–2 500 mots	50 %	98 %	99 %	99 %	70 %
2 500–25 000 mots	35 %	2 %	1 %	1 %	16 %
25 000–250 000 mots	15 %	0 %	0 %	0 %	14 %
Caractéristiques des dérivations					
Obfuscation					
aucune	40 %	0 %	55 %	0 %	49 %
artificielle légère	20 %	0 %	0 %	0 %	25 %
artificielle importante	20 %	0 %	0 %	0 %	26 %
réécriture journalistique	0 %	100 %	19 %	0 %	0 %
simulée	6 %	0 %	26 %	0 %	0 %
traduction	14 %	0 %	0 %	0 %	0 %
révision	0 %	0 %	0 %	100 %	0 %
Taille des passages dérivés					
moins de 50 mots	0 %	?	12 %	1 %	34 %
50 à 150 mots	34 %	?	24 %	18 %	36 %
150 à 300 mots	0 %	?	31 %	37 %	8 %
300 à 500 mots	33 %	?	23 %	27 %	12 %
500 à 3 000 mots	0 %	?	10 %	17 %	10 %
3 000 à 5 000 mots	33 %	?	0 %	1 %	0 %

TABLE 4.8 – Comparaison de la composition des corpus de référence et de nos corpus

toujours pour le genre de l'article de presse, les dérivations de type révision.
- Wikinews est librement redistribuable sous licence CC-by-sa et pourra donc aisément profiter aux autres chercheurs du domaine ;
- Chaque corpus utilisé a été organisé selon le format du corpus de référence PAN afin de simplifier la réutilisation des chaînes expérimentales et participer à la diffusion de ce format comme standard du domaine de la détection de dérivation de texte.

4.4 Recherche de résultats de référence

Nos travaux sont à notre connaissance les premiers à s'intéresser à la langue française. Par conséquent nous n'avons pas de résultats de référence auxquels comparer ceux de nos propositions. La littérature cite régulièrement la signature complète à base de n-grammes mots comme approche de référence pour l'anglais (Lyon et collab., 2001; Clough, 2003a; Yang, 2006a). Ainsi, Yang (2006a) regroupe des textes dérivés à l'aide de cette méthode avec une précision et un rappel oscillant entre 90 % et 100 %. Nous proposons d'utiliser une telle approche comme point de comparaison.

Nous explorons par la suite l'impact de la variation des différents paramètres de cette approche sur les résultats : tailles des n-grammes, mesures de similarité, structures de données et normalisation des éléments. Le détail des expérimentations des combinaisons extensives des différents paramètres est présenté dans l'annexe F. Nous retenons pour chaque corpus la configuration des paramètres qui offre les meilleurs résultats. Ceux-ci nous serviront de résultats de référence.

4.4.1 Paramètrage de la signature complète

4.4.1.1 Taille des n-grammes

La taille des n-grammes est le paramètre de la signature complète le plus évident. Une taille trop petite capture des combinaisons de mots que l'on retrouve communément dans les textes, tandis qu'une taille trop importante capture des combinaisons trop spécifiques qui passent sous silence certaines dérivations. Ainsi, Broder (1997) choisit des séquences de 10 mots, Lyon et collab. (2004) des trigrammes et Yang (2006a) des séquences dont le nombre de mots varie en fonction de la taille des textes. Nous explorons l'utilisation de n-grammes de 1 à 10 mots.

Nous considérons comme mots les suites de caractères alphanumériques séparés par des caractères blancs ou de contrôle. Pour l'anglais, nous découpons les mots aux endroits des caractères blancs. Pour le français, nous découpons les mots en tenant compte des articles et pronoms contractés (*l'*, *d'*, *j'*, *m'*...), des composés lexicaux à apostrophe (*aujourd'hui*...) et à traits d'union (*arc-en-ciel*, *peut-être*, *sauve-qui-peut*...) ainsi que des valeurs numériques (*14,18*, *30%*...).

Les n-grammes sont construits par composition séquentielle des mots découpés tout en filtrant les éléments de ponctuation (., « , ?...). Tout mot est représenté par la forme textuelle qu'il prend dans le texte, à la casse des caractères près.

Nos expérimentations laissent penser que les n-grammes de petite taille (unigrammes, bigrammes et trigrammes) sont les plus adaptés à la détection de dérivation sur les corpus Piithie et Wikinews. Par contre, les n-grammes de plus grandes tailles (6-grammes, 7-grammes) donnent de meilleurs résultats pour PANini.

4.4.1.2 Mesures de similarité

Les deux mesures de similarité proposées dans (Broder, 1997) et reprises par la suite dans toutes les approches par signature sont la *resemblance* (*cf. Équation 2.1*) et le *containment* (*cf. Équation 2.2*). Si la mesure de *resemblance* est symétrique ($r(a,b) = r(b,a)$), il n'en est pas de même pour le *containment* pour lequel l'ordre des paramètres influe sur le résultat final.

Nous énumérons quatre façons de calculer le containment dans les conditions de notre approche impliquant des paires composées d'un document cible (ci) — par la suite, nous l'appelerons document source par abus de langage — et d'un document candidat (ca).

- calculer le nombre d'éléments communs aux deux signatures par rapport à la signature du texte cible ($c(ci,ca)$), ou bien par rapport à celle du texte candidat ($c(ca,ci)$) ;
- calculer les deux combinaisons précédentes et retenir la valeur de containment la plus élevée ($c_{max} = \max(c(ci,ca), c(ca,ci))$), ou bien la plus basse ($c_{min} = \min(c(ci,ca), c(ca,ci))$).

Les deux premières façons de calculer consistent à définir une signature de référence. Cette approche peut être problématique lorsque nous ne connaissons pas le sens du lien de dérivation à identifier. Si l'on pose l'hypothèse que le texte cible est une source, il semble alors préférable d'utiliser la signature de ce dernier comme référence ($c(ci,ca)$). Les deux dernières approches sont préférables lorsque nous n'avons pas d'hypothèse sur le rôle du document cible et donc sur le sens du lien de dérivation. Le tout est alors de savoir s'il est préférable de conserver la plus haute valeur ou la plus basse afin d'obtenir le meilleur classement au final.

Nos expérimentations montrent que la mesure c_{max} donne les meilleurs résultats sur chacun de nos corpus. La mesure $c(ci,ca)$ donne également de bons résultats, notamment pour le corpus Wikinews.

4.4.1.3 Modèles de données

Dans son cadre théorique de la détection de dérivation par signature, (Broder, 1997) propose deux modèles : l'ensemble et le multiensemble. Ces deux modèles sont dépourvus d'ordre et s'apparentent à des collections. Il se différencient en ce que l'ensemble ne contient qu'une seule occurrence de chaque élément contrairement au multiensemble. La plupart des travaux font le choix d'une seule instance des éléments quel que soit leur nombre d'occurrences dans le document.

Nous avons comparé l'utilisation d'un multiensemble et d'un ensemble pour représenter les signatures. Le multiensemble permet de conserver toutes les occurrences des n-grammes et non une seule. Ce changement de structure nécessite de redéfinir la façon dont sont calculées les intersections et les unions des signatures pour le besoin de nos mesures de similarité :

- le multiensemble est défini par le couple (A,m) avec A l'ensemble des éléments (ensemble support) et m la fonction de multiplicité défini de A dans $\mathcal{N}$ et associant à chaque élément de la signature son nombre d'occurrences ;
- nous définissons l'intersection entre ces deux signatures $(A_1,m_1) \cap (A_2,m_2)$ comme le multiensemble $(A_1 \cap A_2, m_{\min})$ où la multiplicité $m_{\min}$ associe à chaque élément e son image minimum : $\forall e \in A_1 \cap A_2, m_{\min}(e) = \min(m_1(e), m_2(e)$;
- nous définissons l'union entre ces deux signatures $(A_1,m_1) \cup (A_2,m_2)$ comme le multiensemble $(A_1 \cup A_2, m_{\max})$ où la multiplicité $m_{\max}$ associe à chaque élément e son image maximum : $\forall e \in A_1 \cup A_2, m_{\max}(e) = \max(m_1(e), m_2(e)$;

Nos expérimentations montrent que les deux modèles donnent globalement des résultats comparables. Toutefois, les résultats obtenus pour les n-grammes de petite

taille avec le modèle ensembliste sont meilleurs que ceux obtenus avec le modèle multiensembliste.

4.4.1.4 Normalisation des éléments de la signature

Normalisation lexicale Les mots fonctionnels (ou mots outils) sont les mots les plus fréquemment utilisés dans une collection. Leur distribution dans les textes est uniforme. Ces mots ont en général un rôle de structuration de la phrase (articles, déterminants, prépositions...) ou de cohésion du texte (pronoms, adjectifs possessifs...), mais ne sont pas sémantiquement significatifs. En RI, les mots fonctionnels sont filtrés à l'aide de listes figées afin de contrôler le bruit généré par leur forte présence tout en privilégiant certaines classes grammaticales (noms, adjectifs...). Comme toutes les approches à base de dictionnaire hors contexte, la reconnaissance des mots fonctionnels à l'aide de liste se heurte aux problèmes d'homographie. Ainsi, le mot « son » en français correspond à la fois à l'adjectif possessif et le nom commun désignant un bruit harmonieux.

Les mots fonctionnels étant distribués uniformément entre les textes, leur retrait de la signature ne doit pas avoir d'impact sur les résultats. En outre, le fait de ne pas en tenir compte lors de la construction des n-grammes permet de les remplacer par des mots plus significatifs. Ainsi, si l'on considère la phrase « nous assistons au spectacle de la démocratie », nous n'extrayons plus qu'un seul trigramme *{assistons, spectacle, démocratie}* plutôt que les quatre trigrammes *{assistons, au, spectacle}*, *{au, spectacle, de}*, *{spectacle, de, la}*, *{de, la, démocratie}*. Le gain paraît important au regard du nombre d'éléments dans la signature et en représentativité de la signature.

Nous avons utilisé les listes de mots fonctionnels créés par Savoy (1999) et librement disponibles[13]. Les recherches en attribution d'auteur montrent que les mots fonctionnels sont très caractéristiques du style (Stamatatos, 2009b), un filtrage trop appuyé pourrait faire disparaître des indices pertinents. Par conséquent, nous avons regroupé les mots fonctionnels des listes en quatre catégories : (i) les articles et prépositions, (ii) les pronoms et adjectifs, (iii) les adverbes et auxiliaires et (iv) les autres mots fonctionnels non grammaticaux.

Nos expérimentations montrent que le fait d'ignorer les mots fonctionnels lors de la construction des n-grammes n'a aucun impact sur les résultats pour le corpus PANini. Le filtrage permet une légère amélioration des résultats pour les n-grammes de petites tailles. D'une manière générale, ils permettent de réduire légèrement le nombre de n-grammes dans les signatures ce qui a un impact positif sur le coût de la méthode.

Normalisation morphologique Nous nous intéressons à une autre forme de normalisation, morphologique celle-ci, la racinisation.

La racinisation (*stemming* en anglais) est un processus de déconstruction des mots par suppression des désinences afin d'obtenir le morphème radical. Cette méthode classique permet de regrouper les mots graphiquement proches afin de permettre des rapprochements sémantiques approximatifs à faible coût.

Nous souhaitons explorer la racinisation afin de normaliser les mots et par extension les n-grammes. Les n-grammes en l'état ne permettent de rapprocher que les séquences textuelles en correspondance exacte. La racinisation permet d'introduire une forme d'abstraction sémantique même si elle reste grossière. Ceci devrait notamment permettre de supprimer une grande partie des variations dans les flexions qui

13. `http://members.unine.ch/jacques.savoy/clef/`

proviennent d'erreurs orthographiques ou grammaticales. C'est également une tentative de capturer des concepts plutôt que des séquences de caractères.

Quelques travaux ont expérimenté la racinisation pour la détection de dérivation (Si et collab., 1997; Hoad et Zobel, 2002). En règle générale, celle-ci dégrade les résultats sur les textes en anglais. Il faut toutefois noter que l'anglais est une langue fusionnelle très faiblement fléchie par rapport au français.

Nous utilisons le racinisateur *Snowball* qui a l'avantage de supporter plusieurs langues dont le français et l'anglais. Il offre également la possibilité de gérer l'agressivité de la racinisation, c-à-d le nombre maximum de transformations morphologiques opérées. Nous la faisons varier de 1 à 5 opérations morphologiques maximum.

Nos expérimentations montrent que la racinisation dégrade légèrement les résultats pour les n-grammes de petites tailles sur le corpus PANini. Elle n'a aucun impact sur les n-grammes de grande taille qui donnent les meilleurs résultats sur ce corpus. La racinisation est profitable pour les corpus Piithie et Wikinews pour lesquels elle permet une amélioration de la capacité de discrimination pour les n-grammes de petites tailles.

4.4.2 Résultats de l'approche de référence

Nos expérimentations ont montré que les différentes formes de dérivation représentées par nos corpus obtiennent des résultats différents et réagissent différemment à la variation des paramètres. Cette observation concorde avec ce que nous avions anticipé, nous devons donc continuer à distinguer les expérimentations concernant ces différentes formes de dérivation.

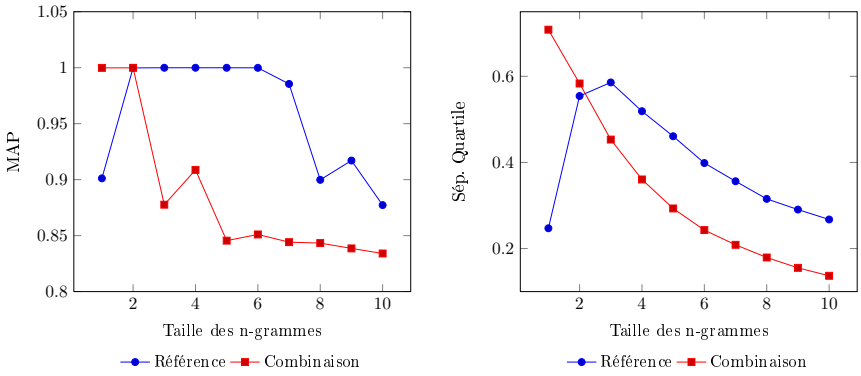
Nous notons tout de même deux paramètres stables. Le premier est le modèle utilisé pour les signatures. Nos expérimentations montrent que l'utilisation d'un ensemble donne généralement de meilleurs résultats qu'un multiensemble. Nous retenons donc cet unique modèle pour nos trois corpus. Le second est la mesure de similarité utilisée pour comparer les signatures. Nos expérimentations montrent que le *containment* c_{max} offre les meilleurs résultats sur les trois corpus. Les autres paramètrages (taille des n-grammes, filtrage des mots fonctionnels et niveau de racinisation) sont propres à chacune des formes de dérivation représentées par nos corpus.

4.4.2.1 Corpus Piithie

Le corpus Piithie se veut représentatif des dérivations opérées par les journalistes de la presse française écrite pour Internet.

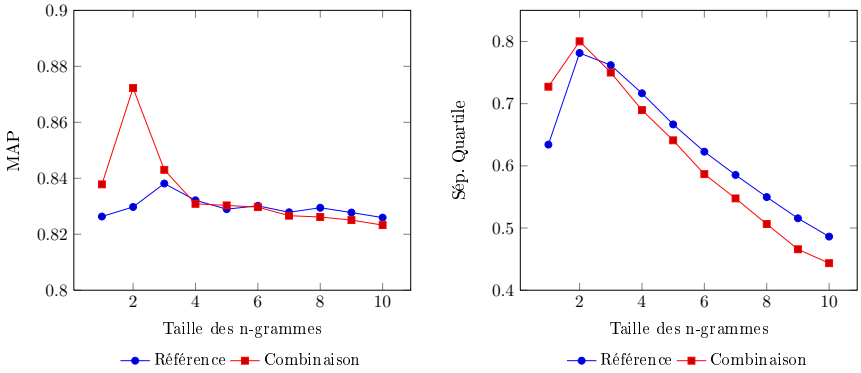
Nos expérimentations montrent que le filtrage des mots fonctionels ainsi que la racinisation améliorent les résultats. La figure 4.9 rapporte les résultats de la combinaison de ces deux opérations en termes de qualité de classification 4.9(a) et de capacité de discrimination 4.9(b). Notre approche de référence sera la signature par unigrammes qui se détache clairement en termes de discrimination et est parmi les meilleures en termes de qualité de classification avec un score proche de un.

La signature par unigrammes de mots racinisés et construits sans tenir compte des mots fonctionnels revient à représenter les textes par la collection de ce qui correspond grossièrement aux concepts (racines) qu'ils contiennent. L'articulation entre ces concepts ainsi que les constructions syntaxiques, voire stylistiques (exception faite des choix lexicaux) sont passées sous silence par le modèle. Les moins bons scores obtenus avec des n-grammes de moyennes et grandes tailles montrent que les structures syntaxiques sont soit peu spécifiques aux textes, soit modifiées lors du processus de dérivation. Nous privilégions la seconde explication. Nous pensons que les contraintes éditoriales propres à chaque journal impliquent des réécritures qui entraînent des



(a) MAP - Piithie (b) Sép. Q - Piithie

FIGURE 4.9 – Combinaison des normalisations sur le corpus Piithie.



(a) MAP - Wikinews (b) Sép. Q - Wikinews

FIGURE 4.10 – Combinaison des normalisations sur le corpus Wikinews.

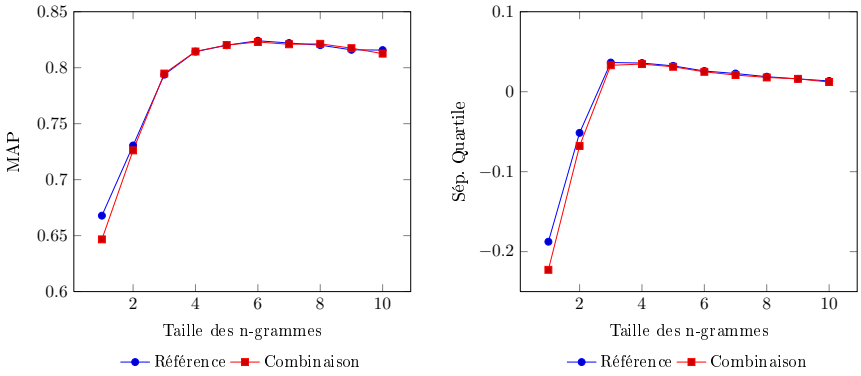
modifications lexicales et syntaxiques de sorte que les articles aient peu de longues sous-chaînes communes.

4.4.2.2 Corpus Wikinews

Le corpus Wikinews se veut représentatif des dérivations de type révisions, pour le genre article de presse en ligne et en français.

La normalisation a un impact hétérogène selon la taille des n-grammes. Le filtrage des seuls articles, prépositions et pronoms provoque un pic de la MAP pour les trigrammes tandis qu'une racinisation appuyée permet d'améliorer légèrement la discrimination et la qualité de classification. Nous avons donc expérimenté la combinaison d'un filtrage de ces mots fonctionnels (i+ii) avec une racinisation appuyée (5 itérations). La figure 4.10 rapporte les résultats de cette expérimentation en termes de qualité de classification 4.10(a) et de capacité de discrimination 4.10(b). Ceux-ci pointent clairement l'approche par bigrammes combinant lesdites normalisations comme la meilleure approche, tant en termes de qualité de classification que de capacité de discrimination. Nous retenons cette combinaison comme approche de référence pour Wikinews.

Nous pensons que l'impact des normalisations, plus important sur ce corpus que



(a) MAP - PANini (b) Sép. Q - PANini

FIGURE 4.11 – Combinaison des normalisations sur le corpus PANini.

sur les autres, est directement liée aux types de réécritures que l'on trouve dans les révisions. En effet, les révisions correspondent le plus souvent à des corrections d'idiosyncrasies, syntaxiques et orthographiques notamment, qui résident à notre avis principalement au sein des désinences. Celles-ci sont supprimées par le processus de racinisation. Nous pensons que les bigrammes capturent des constructions syntaxiques particulières conservées lors des dérivations.

4.4.2.3 Corpus PANini

Le corpus PANini est issu du corpus de référence PAN. Les dérivations qui y sont présentes ne sont pas naturelles mais simulées par des algorithmes ce qui le rend beaucoup plus difficile que les autres. Les résultats sont tout de même intéressants pour comparer les performances des méthodes sur un corpus d'un genre et d'une langue différente.

Les approches avec et sans normalisation obtiennent les mêmes résultats sur ce corpus. Toutefois ces normalisations permettent de réduire le coût de traitement induit ce qui justifie qu'à résultats équivalents on préfère les normalisations à leur absence. Les résultats de la figure 4.11 correspondent à la combinaison des normalisations. Les normalisations n'ont virtuellement aucun impact — positif ou négatif — sur les résultats. Nous retenons les 6-grammes avec normalisation comme approche de référence pour ce corpus.

La bonne performance des n-grammes de grandes tailles s'explique selon nous par la méthode de construction artificielle du corpus à base de fragments de texte complètement différents. Ces fragments, repris directement de leurs textes originaux respectifs, sont spécifiques au texte de par leur grande taille et sont conservés lors de la dérivation excepté lorsqu'ils sont altérés par les mécanismes d'obfuscation. La nette infériorité des résultats par rapport aux deux autres corpus s'explique à la fois par des niveaux de granularité partielle et par les obfuscations très appuyées qui mélangent complètement les mot de sorte que les n-grammes ne permettent pas d'identifier des éléments communs.

4.4.2.4 Synthèse des résultats

Le tableau F.1 fait la synthèse du paramétrage des approches de référence retenues et qui nous serviront de point de comparaison pour nos propositions.

Corpus	Éléments	Filtrage	Racinisation	MAP	Sép. Quartile
Piithie	unigrammes	i+ii+iii+iv	1	0,999	0,708
Wikinews	bigrammes	i+ii	5	0,872	0,800
PANini	6-grammes	i+ii+iii+iv	1	0,823	0,024

TABLE 4.9 – Approches de références pour nos différents corpus.

Clough (2003a) s'est intéressé au paramétrage des signatures complètes sur le corpus METER, un corpus comparable à Piithie mais en anglais, notamment le filtrage des mots fonctionnels et la racinisation. Ses travaux concluaient que le *containment* entre les unigrammes des textes était l'approche la plus performante et que les différentes normalisations n'amélioraient pas les résultats. Nos conclusions sont similaires pour le français (Piithie uniquement). Toutefois, si la normalisation n'a pas d'impact sur les résultats en termes de classification, elle en a un sur le coût de stockage et d'exécution.

La principale conclusion que nous pouvons tirer de ces expérimentations est que, comme nous le pensions, la dérivation est un phénomène ectoplasmique et par conséquent chaque forme de dérivation réagit différemment à différentes approches de détection.

4.5 Conclusion

Dans ce chapitre, nous avons proposé une méthode d'évaluation inspirée des protocoles de la RI. Les paires de documents comparés sont classés par ordre décroissant du score de similarité obtenu et ce classement est utilisé pour le calcul des mesures d'évaluation. Nous avons défini trois axes d'évaluation :

- la qualité de classification, mesurée par la MAP, qui reflète la bonne distribution des liens de dérivation en haut du classement ;
- la capacité de discrimination, mesurée par la séparation des quartiles, qui reflète l'existence d'un espace important entre les paires représentant les liens de dérivation et les autres ;
- le coût de la méthode, mesurée par la taille des signatures.

Nous avons également constitué trois corpus pour mener l'évaluation de nos méthodes :

- le corpus Piithie représente les dérivations opérées entre les dépêches et les articles de journaux en français et sur le Web ;
- le corpus Wikinews représente les dérivations de type révisions entre des versions d'articles de presse publiées en ligne ;
- le corpus PANini représente des dérivations de type plagiat générées artificiellement.

Finalement, nous avons recherché, pour chacune des formes de dérivation représentées par nos corpus, l'approche de référence la plus performante. Les résultats de celles-ci serviront de point de comparaison aux résultats obtenus par nos méthodes. Nous présentons nos propositions et leur évaluation dans le chapitre suivant.

Chapitre 5

Détection extrinsèque de dérivation : exploitation de descripteurs singuliers et invariants et leur combinaison

C'est avec la logique que nous prouvons et avec l'intuition que nous trouvons.

— Henri Poincaré

Sommaire

Nous avons présenté dans le chapitre 2 les différentes approches de la littérature pour la détection de dérivation. Si sur le plan théorique nous avons cherché à embrasser la notion de dérivation dans son intégralité et sa complexité, nous explorons expérimentalement la détection des seules formes de dérivation représentées par nos corpus [1] : dérivations entre dépêches et articles de presse (corpus Piithie), révisions d'articles de presse (corpus Wikinews) et plagiat artificiel de textes littéraires (corpus PANini). De plus, nous nous restreignons au seul paradigme de détection des dérivés d'un texte source parmi une collection fermée de candidats. Nous parlons de collection fermée par opposition aux collections, dites ouvertes, où la liste des textes évolue au fil du temps et où il est difficile d'avoir connaissance de tous les documents constituant le corpus à un moment donné (p. ex. le Web). En d'autres termes, nous cherchons à répondre à la question : « étant donné ce premier texte, quels sont ceux qui en dérivent parmi la liste arrêtée de ces autres textes ? ».

Le paradigme choisi correspond clairement à des besoins applicatifs. Il répond aux scénarios classiques de détection de plagiat (veille sur une œuvre particulière, vérification de l'originalité d'un rapport d'étudiant...) ou de suivi de l'impact d'un document (articles écrits suite à une conférence de presse...). Bien que nous ne la traitions pas, la généralisation à la détection de dérivation sans source identifiée est également possible [2]. Le choix d'une collection fermée est une simplification classique du problème. Elle revient à considérer qu'une liste finie de candidats a été sélectionnée en amont depuis une collection ouverte, ou bien que la tâche applicative prend en entrée une collection fermée (ensemble des rapports d'étudiants pour la détection de plagiat académique par exemple). À notre connaissance, seuls Bendersky et Croft (2009) ont travaillé sur une collection ouverte — le Web — dans le cadre de la détection de dérivation.

La méthode majoritairement employée pour répondre au paradigme choisi (la signature complète, *cf. Section 2.2.2.4*) consiste à comparer les textes sur la base de tous les éléments textuels qui y apparaissent (n-grammes mots). Cette approche reste très coûteuse. De plus, elle n'est pas réellement motivée sur le plan théorique ce qui ne permet pas d'avancer dans la compréhension du problème. Nous proposons dans un premier temps de cadrer la sélection des éléments de la signature en motivant le choix des descripteurs. Nous introduisons ainsi les propriétés de singularité et d'invariance qui correspondent respectivement à la spécificité des éléments capturés par le descripteur au texte modélisé et à leur conservation lors du processus de dérivation (*cf. Section 5.1*). Dans un second temps nous expérimentons plusieurs descripteurs correspondant à différents niveaux de singularité et d'invariance : des n-grammes mots sélectionnés selon des critères statistiques (*cf. Section 5.2*) et des unités linguistiques particulières (*cf. Section 5.3*). Nous détaillons pour chacun, et pour chaque corpus, les résultats obtenus en termes de qualité de classification (MAP) et de capacité de discrimination (Sép. Q) [3] tout en les comparant aux résultats de référence obtenus dans le chapitre précédent (*cf. Section 4.4*). Dans un troisième temps, nous expérimentons la combinaison de ces précédents descripteurs (*cf. Section 5.4*) afin de vérifier l'intérêt d'une approche multidimensionnelle pour la détection de dérivation. Enfin, nous concluons sur les résultats de ces expérimentations (*cf. Section 5.5*).

1. Ces corpus sont décrits dans la section 4.3.3.
2. Les similarités doivent être calculées non plus sur les paires de la source et des suspects mais sur le produit cartésien des suspects de manière à obtenir une matrice des distances sur laquelle sera appliquée un algorithme de clustering par exemple (Yang, 2006a).
3. Ces mesures sont introduites dans la section 4.3.2.

5.1 Tirer parti de la singularité et de l'invariance des descripteurs

Nous exploitons les propriétés de singularité et d'invariance des descripteurs pour détecter les dérivations à l'aide de signatures de taille réduite (*cf. Section 5.1.1*). Pour rappel (*cf. Section 2.3*), un descripteur est un ensemble de critères de sélection des éléments qui constituent la signature d'un texte et en garantit l'homogénéité. Nous introduisons quatre descripteurs qui présentent ces propriétés à différents degrés et que nous évaluons par la suite (*cf. Sections 5.2 et 5.3*). Nous les divisons en deux catégories : ceux dont la singularité est définie par un critère statistique (*cf. Section 5.1.2*) et ceux dont la singularité découle de leur rôle linguistique (*cf. Section 5.1.3*).

5.1.1 Améliorer les signatures par la singularité et l'invariance des descripteurs

Nous rappelons tout d'abord le fonctionnement de l'approche par signature complète sur laquelle nous nous reposons et qui est majoritairement employée pour la détection de dérivés d'un texte source. Nous repositionnons ensuite les spécificités de notre approche par rapport à ce cadre général. Enfin, nous introduisons les propriétés de singularité et d'invariance qui fondent notre démarche.

5.1.1.1 Principe et limite de l'approche par signature

Pour le paradigme de la détection des dérivés d'un texte source, l'approche dominante repose sur la comparaison de signatures (*cf. Section 2.2.2*). En règle générale, une telle approche se déroule en trois étapes (*cf. Algorithme 5.1*) : la sélection des éléments selon le descripteur, leur regroupement sous forme d'une modélisation et leur comparaison à l'aide d'une mesure de similarité.

Entrée: $texte1, texte2$
Entrée: $descripteur$
Sortie: $sim \in [0..1]$
{Étape 1 : sélectionner les éléments qui respectent le descripteur}
1: $elts1 \Leftarrow selection\_elements(texte1, descripteur)$
2: $elts2 \Leftarrow selection\_elements(texte2, descripteur)$
{Étape 2 : modéliser la signature à partir des éléments retenus}
3: $s1 \Leftarrow modelisation\_signature(elts1)$
4: $s2 \Leftarrow modelisation\_signature(elts2)$
{Étape 3 : mesurer la similarité entre ces signatures}
5: $sim \Leftarrow mesurer\_similarite(s1, s2)$

Algorithme 5.1: Déroulement de l'approche par signature

La signature d'un texte est la collection des éléments capturés par le descripteur dans le texte. Par exemple, la figure 5.1 montre les signatures obtenues pour deux textes étant donné le descripteur bigramme mots avec normalisation de la casse et filtrage de la ponctuation. Les signatures ainsi constituées sont comparées deux à deux à l'aide d'une mesure de similarité. L'hypothèse sous-jacente est que plus le score de similarité est élevé, plus il est probable que les deux textes dérivent l'un de l'autre.

Par exemple, l'emploi de la mesure c_{max} sur les signatures de la figure 5.1 donne une similarité de $\max(\frac{44}{55}, \frac{44}{75})$ (env. 0,586).

Dans le cas particulier de la signature complète, l'union de tous les éléments (n-grammes mots) d'une signature couvre complètement le texte. Cette représentation exhaustive a un coût en terme de stockage, mais pas seulement. La comparaison des signatures est réalisée par des opérations ensemblistes [4] dont la complexité est linéaire avec la taille des signatures ($O(|s_1| + |s_2|)$). En d'autres termes, la taille élevée des signatures impacte le coût de calcul de la méthode en plus du coût de stockage.

5.1.1.2 Approche par des signatures de taille réduite

Il est possible d'agir sur au moins trois leviers pour améliorer le coût des méthodes à base de signature : le nombre de candidats comparés, la taille des éléments composant la signature (p. ex. le nombre de caractères les constituant) et le nombre de ces éléments dans la signature.

- L'utilisation de méthodes de filtrage en amont du système de détection permet de réduire le nombre de candidats (Hose, 2003; Clough, 2003a). Cependant, ces méthodes, qui reposent sur des comparaisons de distribution des mots, augmentent le risque de négliger des dérivés.
- L'application de méthodes classiques de compression sans perte (codage de Huffman par exemple) aux éléments des signatures permet de réduire leur coût de stockage (Heintze, 1996; Schleimer, 2003; Stein, 2005). Cependant, le coût de calcul de leur comparaison, qui dépend du nombre d'éléments dans la signature, reste inchangé.
- Nous pensons que la réduction du nombre d'éléments est le levier le plus efficace pour réduire à la fois les coûts de stockage des signatures et de calcul des comparaisons des approches par signature. C'est notamment une piste qui a été peu étudiée au regard des deux autres solutions (filtrage en amont et compression des éléments).

Nous nous concentrons sur le choix du descripteur, soit la première étape du déroulement d'une approche par signature ce qui devrait permettre de répercuter les gains opérés sur les étapes suivantes et *in fine* sur la totalité de la méthode.

Nous devons figer les étapes restantes de l'approche par signature afin de pouvoir mesurer au mieux l'impact des choix opérés concernant les descripteurs. Ainsi, nous conservons par la suite les étapes sélectionnées lors du paramétrage des approches de référence détaillé dans l'annexe F :

- l'utilisation d'une modélisation ensembliste pour la signature dont la construction est décrite par l'algorithme 5.2 ;
- l'utilisation de la mesure de similarité c_{max} décrite par l'algorithme 5.3.

5.1.1.3 Singularité et invariance

Le choix de la réduction du nombre d'éléments dans les signatures semble contre-intuitif. L'augmentation de la taille de la signature pour un document augmente les chances d'obtenir des correspondances lorsque le vocabulaire — ensemble des mots ou combinaisons de mots disponibles — est figé (Jones, 2004). À l'inverse, la réduction que nous proposons risque de se faire au détriment de la qualité de la détection. Nous proposons de considérer deux propriétés des éléments afin de limiter cette détérioration : la singularité et l'invariance.

4. Il s'agit des mesures de *resemblance* (*cf. Équation 2.1*), de *containment* (*cf. Équation 2.2*) et leurs dérivées

Texte source

C'est une mauvaise relance de Fabien Barthez qui avait permis à Materazzi d'égaliser en finale du Mondial 2006. Le portier chauve, célèbre pour la bise de Laurent Blanc sur le crâne, aura vainement espéré la revanche en septembre au Stade de France mais Grégory Coupet était devenu entretemps numéro un des gardiens du football français.

Corpus Wikinews (source 163)

Signature du texte source

{(sur, le) ; (une, mauvaise) ; (devenu, entretemps) ; (espéré, la) ; (crâne, aura) ; (en, septembre) ; (gardiens, du) ; (chauve, célèbre) ; (finale, du) ; (le, crâne) ; (2006, le) ; (en, finale) ; (barthez, qui) ; (des, gardiens) ; (à, materazzi) ; (relance, de) ; (était, devenu) ; (de, fabien) ; (france, mais) ; (de, Laurent) ; (de, france) ; (coupet, était) ; (revanche, en) ; (au, stade) ; (pour, bise) ; (portier, chauve) ; (d', égaliser) ; (laurent, blanc) ; (un, des) ; (aura, vainement) ; (grégory, coupet) ; (fabien, barthez) ; (qui, avait) ; (mondial, 2006) ; (mauvaise, relance) ; (football, français) ; (blanc, sur) ; (numéro, un) ; (célèbre, pour) ; (égaliser, en) ; (materazzi, d') ; (septembre, au) ; (permis, à) ; (entretemps, numéro) ; (c', est) ; (bise, de) ; (le, portier) ; (la, revanche) ; (avait, permis) ; (est, une) ; (du, mondial) ; (vainement, espéré) ; (mais, grégory) ; (stade, de) ; (du, football)}

Texte suspect

C'est une mauvaise relance de Fabien Barthez qui avait permis à Materazzi d'égaliser en finale du Mondial 2006. Fabien Bartez, "... le plus grand gardien de but de toute l'histoire du football français et de loin...", selon Grégory Schneider de Libération, célèbre pour la bise de Laurent Blanc sur le crâne, n'aura pas eu sa revanche en septembre au Stade de France mais Grégory Coupet était devenu entretemps numéro un des gardiens du football français.

Corpus Wikinews (suspect 1990)

Signature du texte suspect

{(2006, fabien) ; (sur, le) ; (pas, eu) ; (une, mauvaise) ; (barthez, le) ; (histoire, du) ; (coupet, était) ; (libération, célèbre) ; (français, et) ; (de, loin) ; (grand, gardien) ; (fabien, bartez) ; (le, crâne) ; (stade, de) ; (en, finale) ; (des, gardiens) ; (gardien, de) ; (loin, selon) ; (barthez, qui) ; (gardiens, du) ; (devenu, entretemps) ; (à, materazzi) ; (relance, de) ; (schneider, de) ; (de, fabien) ; (crâne, n') ; (but, de) ; (france, mais) ; (aura, pas) ; (toute, l') ; (mais, grégory) ; (plus, grand) ; (grégory, schneider) ; (sa, revanche) ; (de, france) ; (c', est) ; (revanche, en) ; (au, stade) ; (en, septembre) ; (et, de) ; (du, football) ; (finale, du) ; (d', égaliser) ; (laurent, blanc) ; (grégory, coupet) ; (selon, grégory) ; (un, des) ; (célèbre, pour) ; (de, laurent) ; (football, français) ; (le, plus) ; (fabien, barthez) ; (qui, avait) ; (mondial, 2006) ; (mauvaise, relance) ; (était, devenu) ; (eu, sa) ; (blanc, sur) ; (la, bise) ; (de, toute) ; (égaliser, en) ; (materazzi, d') ; (septembre, au) ; (permis, à) ; (n', aura) ; (pour, la) ; (numéro, un) ; (bise, de) ; (de, libération) ; (avait, permis) ; (est, une) ; (du, mondial) ; (entretemps, numéro) ; (de, but) ; (l', histoire)}

Intersection des signatures :
{(c', est) ; (une, mauvaise) ; (des, gardiens) ; (numéro, un) ; (d', égaliser) ; (septembre, au) ; (égaliser, en) ; (gardiens, du) ; (revanche, en) ; (du, football) ; (barthez, qui) ; (de, fabien) ; (materazzi, d') ; (sur, le) ; (devenu, entretemps) ; (france, mais) ; (qui, avait) ; (grégory, coupet) ; (football, français) ; (stade, de) ; (célèbre, pour) ; (mondial, 2006) ; (était, devenu) ; (au, stade) ; (mauvaise, relance) ; (finale, du) ; (est, une) ; (en, finale) ; (avait, permis) ; (blanc, sur) ; (entretemps, numéro) ; (laurent, blanc) ; (bise, de) ; (un, des) ; (le, crâne) ; (coupet, était) ; (du, mondial) ; (en, septembre) ; (fabien, barthez) ; (de, france) ; (relance, de) ; (permis, à) ; (à, materazzi) ; (mais, grégory)}

Figure 5.1 – Descripteur : bigramme mot avec recouvrement, normalisation de la casse, suppression des ponctuations

---

Entrée: *elements*
Sortie: s une signature ensembliste
{Modélisation de la signature par un ensemble}
1: $s \Leftarrow \emptyset$
2: **for all** $elt \in elements$ **do**
3: $s \Leftarrow s \cup \{elt\}$
4: **end for**
5: **return** s

Algorithme 5.2: Modélisation comme un ensemble des éléments du texte correspondant au descripteur.

Entrée: $signature1$ une signature ensembliste
Entrée: $signature2$ une signature ensembliste
Sortie: score de similarité entre les deux signatures
{Utilisation de la mesure c_{max}}
1: $intersection \Leftarrow signature1 \cap signature2$
2: $c1 \Leftarrow \frac{|intersection|}{|signature1|}$
3: $c2 \Leftarrow \frac{|intersection|}{|signature2|}$
4: **return** $\max(c1, c2)$

Algorithme 5.3: Algorithme de calcul du score de similarité entre deux signatures ensemblistes tels qu'obtenus par l'algorithme 5.2.

Les propriétés de singularité et d'invariance sont plus faciles à appréhender au niveau des éléments. Pour des raisons de généralisation et afin de nous détacher de ces éléments individuels pour raisonner à l'échelle des signatures, nous considérons toutefois ces propriétés au niveau des descripteurs. Il suffit alors de les appréhender comme le degré moyen de ces propriétés parmi les éléments capturés par le descripteur considéré.

SINGULARITÉ

Singularité La propriété de singularité d'un descripteur caractérise la probabilité de trouver des éléments d'un texte correspondants au descripteur dans d'autres textes. On parlera de descripteur singulier lorsque la probabilité de trouver des éléments du texte caractéristiques du descripteur dans d'autres textes est faible, exception faite des dérivés dudit texte. Nous pensons que des textes qui partagent des éléments singuliers, décrits par un descripteur singulier, ont de fortes chances d'être impliqués dans une relation de dérivation.

Le principe de l'approche par signature classique est de repérer les combinaisons peu probables impliquant de nombreux éléments. Deux textes qui partageraient une telle combinaison seraient alors probablement dérivés. Nous proposons d'exploiter des éléments dont la présence individuelle est elle-même peu probable. En effet, nous pensons que les liens de dérivation sont plus vraisemblables lorsque les éléments partagés par la source et le candidat sont des éléments dont l'usage est spécifique au texte source.

Par exemple, dans notre corpus anglais, des trigrammes tels que *lot of the*, *order to prevent* ou encore *being allowed to* se retrouvent communément indépendamment de la présence d'un lien de dérivation. Par conséquent, le partage de ces éléments entre deux textes n'informe en rien d'un éventuel lien de dérivation. Ils ne participent pas à discriminer les dérivés des non-dérivés et peuvent donc être ignorés au profit d'autres qui sont plus singuliers. À l'opposé, certains éléments sont beaucoup plus discriminants du texte dans lequel ils apparaissent. Ainsi, les séquences de mots propres à un texte (n-grammes hapax) ou de fort poids informatif sont spécifiques à un texte puisque, statistiquement, elles apparaissent principalement dans ledit texte. De même, les entités nommées ou encore les composés nominaux sont à notre sens singuliers car ils font référence à des entités ou des concepts uniques et précis. Les idiosyncrasies (fautes d'orthographes, maladresses stylistiques ou syntaxiques, particularités graphiques...) sont également spécifiques aux textes de par leur caractère atypique. Elles sont toutefois plus complexes à exploiter dans le cadre de la détection de dérivation car difficiles à identifier.

INVARIANCE

Invariance La propriété d'invariance d'un descripteur caractérise la probabilité que les éléments d'un texte correspondants au descripteur soient conservés lors d'un processus de dérivation. L'invariance complète ainsi la singularité car un élément ne participe pas à la détection de dérivation s'il n'est pas présent dans le texte dérivé, indépendamment de sa singularité.

Nous parlerons de descripteur invariant lorsque la probabilité de conservation des éléments correspondants dans les textes dérivés est forte, mais nous considérons différents degrés d'invariance. L'invariance peut se percevoir le long d'un continuum borné d'un côté par l'invariance textuelle et de l'autre par l'invariance de la référence. La propriété d'invariance textuelle signifie que l'élément est repris dans le texte dérivé avec la même forme textuelle qu'il prend dans le texte source. À l'opposé l'invariance de la référence signifie que si l'on retrouve mention de l'entité dans le texte source, la forme n'est plus la même dans le texte dérivé.

L'invariance ne semble pas être une propriété intrinsèque des descripteurs. En effet, la conservation d'un élément dans le processus de dérivation relève majoritairement

du choix de l'auteur. Toutefois, certains éléments sont nécessaires à l'expression du contenu et pourraient être invariants (concepts spécifiques, entités...). Pour autant l'auteur peut, dans les cas de granularité partielle par exemple, ne pas reprendre l'intégralité du contenu et donc délaisser de tels éléments. Prenons l'exemple d'un texte traitant de la situation au Proche-Orient en début d'année 2011 et détaillant les émeutes en Tunisie ainsi que la démission des ministres du Hezbolla au Liban. Une dérivation de ce texte focalisée sur la situation en Tunisie délaisserait les éléments concernant le Liban quand bien même ceux-ci semblent nécessaires du point de vue du texte source.

Dans la suite nous proposons plusieurs descripteurs qui respectent ces propriétés de singularité et d'invariance à différents degrés.

5.1.2 Singularité des n-grammes mots selon un critère statistique

Nous proposons de sélectionner, sur la base de propriétés statistiques, les plus singuliers des n-grammes mots de la signature complète[5] et de supprimer les autres. La sélection selon des propriétés statistiques plutôt que sémantique est commune en RI depuis les travaux de Jones (2004). La spécificité d'un objet est fonction de l'utilisation : les expressions les plus générales sont les plus représentées, à l'inverse des expressions les plus spécifiques.

Nous formulons deux propositions fondées sur un critère statistique pour sélectionner les n-grammes singuliers. La première consiste à exploiter les n-grammes rares et la seconde les n-grammes de fort poids informatif.

5.1.2.1 Les n-grammes rares

D'après Church et Gale (1995), plus un élément dérive de la loi de Poisson, plus cet élément est utile pour discriminer les documents sur la base des dépendances cachées à faire émerger. La loi de Poisson décrit le comportement du nombre d'évènements se produisant dans un intervalle fixé. Dans notre cas cela revient à étudier le nombre d'occurrence d'un n-gramme mot dans les différents documents de la collection, c-à-d son df (*document frequency*).

Nous proposons de tirer parti des n-grammes qui n'apparaissent que dans un seul document. Ils sont par définition extrêmement singuliers étant donné qu'ils sont à l'extrémité droite[6] de la courbe de la loi de Zipf, qui peut s'apparenter à une distribution de Poisson. Un *hapax* désigne généralement un mot qui n'a qu'une seule occurrence. Dans le cadre de nos travaux, un *hapax* désigne un élément d'une signature qui n'apparaît que dans un seul document d'une collection de textes de référence. Le nombre d'occurrences de cet *hapax* dans ce document n'a pas d'incidence. En d'autres termes, il s'agit des éléments qui ont un df de 1 dans cette collection de référence.

Les expérimentations autour de cette approche sont détaillées dans la section 5.2.2.

5.1.2.2 Les n-grammes de fort poids informatif

Nous proposons de tirer parti des n-grammes qui ont un fort poids informatif. Il s'agit des éléments qui jouent un rôle important dans l'expression du contenu transversal au texte. Plusieurs mesures de pondération ont été discutées dans la littérature

5. Pour rappel, la signature complète, qui nous sert d'approche de référence (*cf. Section 5.1.1*), contient tous les n-grammes présents dans le texte modélisé.

6. L'extrémité droite correspond aux rangs les plus élevés, c-à-d les éléments qui apparaissent le moins fréquemment.

afin d'estimer ce poids informatif (Ibekwe-SanJuan, 2007). Nous retenons la plus classique d'entres elles, le tf · idf (Salton et Yang, 1973), car elle donne de très bons résultats et sa mise en œuvre est des plus simples. Nous avons notamment délaissé les mesures telle qu'OKAPI car nous n'avons pas de données concernant la pertinence des documents.

La pondération d'un élément par tf · idf (*cf. Équation 5.4*) permet de valoriser à la fois son exhaustivité au sein d'un document d mesurée par le tf (*cf. Équation 5.1*) et sa spécificité par rapport à une collection D mesurée par l'idf (*cf. Équation 5.2*). Nous lissons à l'aide d'un logarithme les valeurs de l'idf qui pénaliseraient les mots très répandus dans la collection. De plus, nous valorisons les hapax présents dans les textes en ajoutant 1 à l'idf. En effet, par définition ces derniers n'apparaissent que dans un seul document (df = 1), la valeur de leur tf · idf est donc figée à 0 indépendamment de leur nombre d'occurrences dans le texte.

$$\mathrm{tf}(m,d) = \frac{n_{m,d}}{\sum_{i \in d} n_{i,d}} \tag{5.1}$$

m le terme considéré
d le document considéré
$n_{i,d}$ nombre d'occurrences de i dans d

$$\mathrm{idf}(m,D) = \log \frac{|D|}{|\{d : \forall d \in D : m \in d\}|} \tag{5.2}$$

m le terme considéré
D la collection considérée

(5.3)

$$\mathrm{tf} \cdot \mathrm{idf}(m,d,D) = \mathrm{tf}(m,d) \cdot (1 + \mathrm{idf}(m,D)) \tag{5.4}$$

Les expérimentations autour de cette approche sont détaillées dans la section 5.2.3.

5.1.3 Singularité inhérente aux unités linguistiques

Nos propositions précédentes se fondent, comme la signature complète, sur les n-grammes mots. Ces n-grammes sont définis par une taille figée qui ne correspond à aucune réalité linguistique et constituent un choix discutable pour la détection de dérivation. Premièrement, d'un point de vue purement opérationnel, il est nécessaire que les n-grammes se recouvrent afin de pallier un éventuel décalage provoqué par l'ajout d'un mot. La redondance qui en résulte impacte négativement les performances. Deuxièmement, nous pensons que les éléments porteurs de contenu ou de la structure du texte source sont les mieux à même d'être conservés lors d'une dérivation. Cependant, ces éléments sont de taille variable et ne sont par conséquent pas capturés efficacement par des n-grammes de taille fixe.

Nous proposons de considérer deux descripteurs ayant une réalité linguistique et qui permettent de capturer des éléments de contenu du texte : les entités nommées et les composés nominaux.

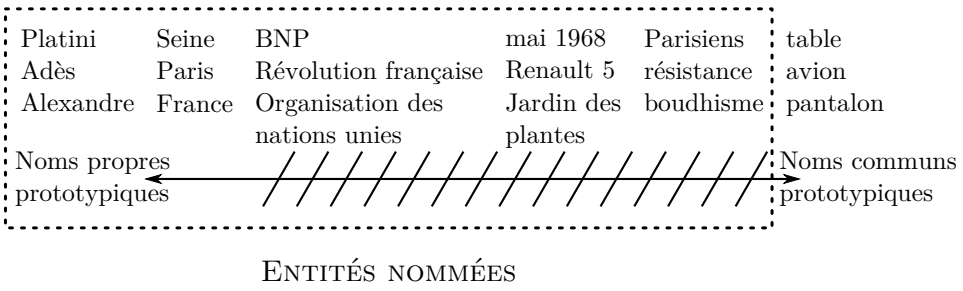


FIGURE 5.2 – Une entitée nommée est un nom propre au sens le plus large. Schéma tiré de Fourour (2004)

5.1.3.1 Les entités nommées

Les entités nommées, telles qu'introduites dans MUC-6 (muc, 1995), regroupent les noms propres[7], divers identificateurs uniques tels que les acronymes, les noms d'organisation et de personnes et les lieux politiquement ou géographiquement définis (toponymes). Les entités nommées couvrent donc une grande partie du continuum qui sépare les prototypes des noms propres des prototypes des noms communs comme l'illustre la figure 5.2.

Fourour (2004) recense des critères graphiques (majuscule des initiales, orthographe variable), morphologiques (absence de dérivations en genre et en nombre, pas de déterminant) et syntaxiques (noms propres modifiés et non modifiés) qui se retrouvent dans plusieurs entités nommées. Par exemple, et pour illustrer la notion, on trouve dans le premier texte de la figure 5.1 les entité nommées : *Fabien Barthez, Materazzi, Mondial 2006, Laurent Blanc, Stade de France, Grégory Coupet* mais également *le portier chauve* qui désigne spécifiquement Fabien Barthez dans le contexte de l'énonciation. Ainsi, le critère le plus stable et qui justifie selon nous l'utilisation des entités nommées pour la détection de dérivation est celui du référent unique.

Les entités nommées sont parmi les entités les plus signifiantes car un lien conventionnel propre à la situation d'énonciation les relie à un référent unique. Nous pensons que l'unicité de ce référent fait des entités nommées des éléments singuliers. De plus, il semble difficile de rapporter un propos qui met en scène de telles entités uniques sans les mentionner, signe de l'invariance de ces référents uniques. Les entités nommées constituent un bon descripteur car elles sont à la fois spécifiques (propriété de singularité) et nécessaires (propriété d'invariance) à l'expression du contenu.

Pour autant, si le référent unique derrière une entité nommée est conservé celle-ci accepte de nombreuses variations. Plusieurs formes textuelles peuvent ainsi être utilisées pour faire référence à la même entité. Par exemple, il est possible de faire référence à *Fabien Barthez* par ledit patronyme, mais également au moyen du titre *le portier chauve* ou encore d'une forme contractée du patronyme *Barthez*. Le choix de la forme textuelle est du ressort de l'auteur. Cette propriété est à double tranchant. L'auteur peut choisir une forme très précise pour une référence commune qui devient singulière du texte et par conséquent est discriminante pour la détection de dérivation. Mais si l'auteur de la source a le choix de la forme textuelle, il en est de même pour l'auteur du dérivé. L'invariance ne porte plus sur la forme mais sur la référence. L'exploration des variations textuelles des entités nommées serait trop coûteuse. Nous

7. Molino (1982) propose une typologie exhaustive des noms propres : les noms de personnes ou anthroponymes, les noms d'animaux, les appellatifs et titres, les noms de lieux, les noms de temps, les noms d'institutions, les noms de produit de l'activité humaine, les noms de symboles mathématiques et scientifiques et les autres noms propres.

nous limitons à considérer les formes textuelles des entités nommées choisies par l'auteur du texte source.

Les expérimentations autour de cette approche sont détaillées dans la section 5.3.1.

5.1.3.2 Les composés nominaux

Les composés nominaux sont des constructions syntaxiques de plusieurs mots incluant au moins un nom. Ils désignent une chose ou une notion spécifique et sont privilégiés pour exprimer des idées et des concepts précis. Ainsi, d'après Cerbah (2000) ils représentent plus de 80 % des termes spécifiques à un domaine pour les langages de spécialité. Ils sont également communément employés dans le langage commun. Par exemple, on trouve dans le premier texte de la figure 5.1 les composés nominaux : *mauvaise relance, finale du Mondial, portier chauve, numéro un, gardiens du footbal français.*

Nous considérons avec intérêt les composés nominaux. Leur construction comme une composition de plusieurs mots rappelle le principe des n-grammes mots et leur confère une certaine singularité. La précision des concepts qu'ils désignent nous laisse penser qu'ils ont un rôle central dans l'expression du contenu et que ces concepts seront probablement repris lors d'une dérivation (propriété d'invariance). Les propriétés de singularité et d'invariance qui semblent associées aux composés nominaux en font un bon descripteur.

Tout comme les entités nommées, les composés nominaux acceptent un certain nombre de variations qui se limitent cependant à des précisions ou des réductions. Par exemple, le composé nominal *aménagement de la forêt* peut apparaître sous la forme *aménagement de l'agriculture et de la forêt*. Il serait envisageable de tenter une normalisation en ne retenant que les éléments de tête ou de queue par exemple. Nous ne prenons pas en compte ces variations dans le cadre de cette thèse et nous nous limitons à considérer les formes textuelles choisies par l'auteur du texte source.

Les expérimentations autour de cette approche sont détaillées dans la section 5.3.2.

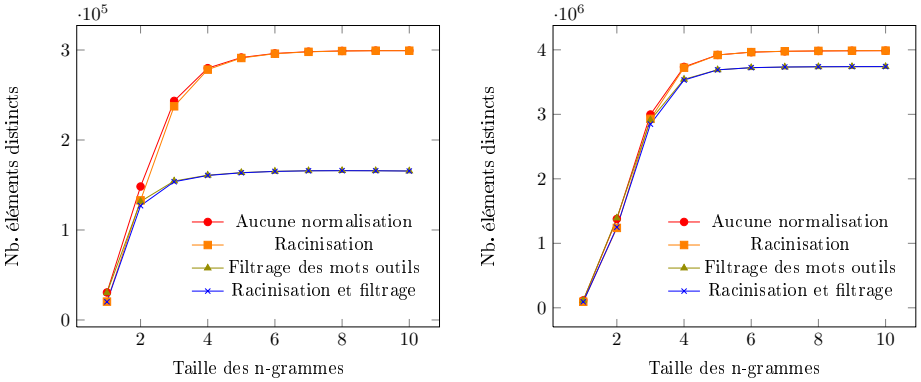
5.2 Exploitation des critères statistiques

À la section 5.1.2, nous avons proposé deux méthodes fondées sur la sélection statistique des n-grammes. Nous décrivons dans cette section leur mise en œuvre sur nos corpus. Dans une première section nous présentons le calcul des distributions statistiques de référence (*cf. Section 5.2.1*) que nous utilisons pour la sélection des n-grammes rares (*cf. Section 5.2.2*) et des n-grammes au fort poids informatif (*cf. Section 5.2.3*).

5.2.1 Calcul des distributions de référence des n-grammes

La sélection statistique nécessite des données de référence concernant la distribution des n-grammes. Ces données doivent être extraites de corpus de même genre et dans la même langue que ceux que nous utilisons pour nos expérimentations : des articles de presse en français pour les corpus Piithie et Wikinews et des œuvres littéraires en anglais pour le corpus PANini.

L'utilisation des textes de nos corpus introduirait un biais dans notre méthode. Tout d'abord, la distribution des éléments de nos corpus et des corpus de référence auraient été strictement identique ce qui réduit l'intérêt de nos méthodes de filtrage puisqu'elles tirent justement parti des variations entre les distributions. Ensuite, nos corpus ont la particularité d'être composés en majorité de dérivations. Cela a pour effet d'augmenter artificiellement la présence des éléments des textes dérivés, ce qui



(a) Français (b) Anglais

FIGURE 5.3 – Comparaison du nombre d'éléments distincts par taille des n-grammes et par normalisation des mots les constituants pour les corpus de référence (a) français et (b) anglais

biaise nos données et va à l'encontre de notre objectif. Pour ces raisons, nous avons constitué deux nouveaux corpus sur lesquels nous avons calculé les distributions de référence des descripteurs étudiés.

Les corpus Piithie et Wikinews sont constitués de courts articles de presse en français (386 mots/article en moyenne). Nous n'avons exploité que les articles de Wikinews de 10 révisions ou plus (*cf. Section 4.3.3*). Nous choisissons d'exploiter le reliquat pour constituer le corpus de référence des corpus Piithie et Wikinews. Nous n'avons retenu que la dernière révision des articles de ce reliquat pour deux raisons. D'une part, il s'agit généralement de la version la plus longue et la plus correcte de l'article. D'autre part, nous n'avons conservé qu'une révision par article afin d'éviter de biaiser les distributions pour les raisons que nous avons discutées au paragraphe précédent. Au final, le corpus ainsi constitué compte 1 027 articles de presse en français pour un total de 289 288 mots.

Le corpus PANini est constitué d'œuvres littéraires en anglais. Nous l'avons construit en prélevant tous les textes candidats du corpus PAN d'au plus 2 500 mots et en le complétant par les documents sources liés. Comme précédemment, nous pouvons tirer du reliquat du corpus PAN de quoi constituer un corpus de référence. Nous choisissons de piocher dans les sources du corpus PAN, à l'exception de celles présentes dans PANini. En effet, de par la méthode de construction employée, les textes candidats peuvent contenir des passages qui ne sont pas syntaxiquement correct, et plus ennuyeux on peut y trouver des dérivés des textes composant notre corpus. Nous ne retenons que les sources d'au plus 2 500 mots afin d'obtenir une collection comparable à celle de PANini et filtrons les textes qui ne sont pas de l'anglais. Au final, le corpus ainsi constitué compte 3 508 textes pour un total de 3 987 663 mots.

Nous avons utilisé ces corpus pour calculer le df (*document frequency*) de chaque forme de n-gramme que nous utilisons, comme pour l'approche de référence (*cf. Section 4.4*) : pour n variant de 1 à 10 et après normalisation des corpus par racinisation des mots et filtrage des mots outils.

La figure 5.3 décrit l'évolution du nombre d'éléments distincts extraits des deux corpus de référence, selon la taille des n-grammes et le type de normalisation du corpus[8]. Nous pouvons observer que le nombre d'éléments distincts augmente rapi-

8. Pour des raisons de clarté du graphique, et étant donné que les valeurs variaient peu, nous ne montrons qu'un niveau de racinisation (le plus élevé) et qu'un niveau de filtrage des mots outils (la

dement avec la taille des n-grammes, indépendamment de la normalisation, pour se stabiliser à partir des 5-grammes. C'est une observation attendue puisque la racinisation permet principalement de supprimer les variations flexionnelles, or l'ordre et la cooccurrence des mots étant figée par la structure en n-grammes, ces variations sont peu nombreuses. Le filtrage des mots outils a un impact beaucoup plus important car il retire complètement ces mots qui sont très présents dans les textes. Cette diminution du nombre de mots entraîne mécaniquement une diminution des n-grammes distincts. Il est important de noter que l'échelle des ordonnées du graphique 5.3(b) est dix fois supérieure à celle du graphique 5.3(a), ce qui peut expliquer la différence d'écart observée entre les courbes avec et sans filtrage des mots outils.

5.2.2 Exploitation des n-grammes rares

Dans cette section nous évaluons l'utilisation d'une signature à base d'éléments rares (les n-grammes hapax) telle que décrite en section 5.1.2.1. Nous détaillons tout d'abord nos choix expérimentaux pour la mise en œuvre de cette approche. Nous présentons puis discutons ensuite les résultats obtenus sur les corpus Piithie, Wikinews et PANini. Le détail de l'analyse des résultats est présenté en annexe (*cf. Annexe G.1*).

5.2.2.1 Mise en œuvre expérimentale

L'approche par exploitation des n-grammes hapax consiste à ne retenir que les éléments qui apparaissent zéro ou une seule fois dans un corpus de référence. L'objectif est d'éliminer les éléments communs pour ne retenir que ceux qui sont spécifiques au texte. La figure 5.4 illustre la différence entre une signature complète et une signature ne contenant que les n-grammes hapax.

Nous sélectionnons les n-grammes hapax (*cf. Section 5.1.2.1*) en prenant pour référence les distributions du corpus de référence français pour les corpus Piithie et Wikinews, et du corpus de référence anglais pour PANini. Pour chaque type de n-gramme considéré nous obtenons un index associant à chaque n-gramme son df qui constitue la distribution de référence du descripteur correspondant. Une fois les distributions de référence calculées, nous considérons pour chaque texte de nos corpus la signature complète correspondante (p. ex. unigrammes avec filtrage des mots outils et racinisation). Les éléments de cette signature qui ont un $df = 1$ dans la distribution de référence ou bien qui n'y apparaissent pas sont retenus dans la signature hapax. Les autres éléments sont supprimés.

5.2.2.2 Résultats

Le tableau 5.1 présente, pour chacun de nos corpus, les résultats de l'exploitation des n-grammes rares en termes de MAP et de Sép. Q (ces mesures ont été introduites à la section 4.3.2). Nous avons exploré différentes tailles de n-grammes mots (des unigrammes aux 10-grammes) ainsi que la mise en œuvre ou non de normalisation sur les textes (racinisation et filtrage des mots outils). Nous retenons dans le tableau, pour chaque corpus, uniquement la meilleure configuration en termes de MAP (max(MAP)) et la meilleure configuration en termes de Sép. Q (max(Sép. Q)).

Corpus Piithie Les meilleurs scores de la MAP et de la Sép. Q sont obtenus avec les bigrammes mots sans normalisation sur les textes dans le premier cas, les unigrammes mots avec normalisation des textes dans le second cas. Les performances obtenues avec ces descripteurs sont similaires à celles de l'approche de référence puisque qu'ils

liste la plus complète)

C'est une mauvaise relance de Fabien Barthez qui avait permis à Materazzi d'égaliser en finale du Mondial 2006. Fabien Bartez, "... le plus grand gardien de but de toute l'histoire du football français et de loin...", selon Grégory Schneider de Libération, célèbre pour la bise de Laurent Blanc sur le crâne, n'aura pas eu sa revanche en septembre au Stade de France mais Grégory Coupet était devenu entretemps numéro un des gardiens du football français.
Au lendemain de sa décision de retraite annoncée en direct sur TF1, toute la presse fait revivre sa carrière marquée par deux points d'orgue : * la finale européenne offerte à l'Olympique de Marseille à l'époque de Bernard Tapie, * celle de la victoire du mondiale en 1998.

Corpus Wikinews (source 163)

Signature complète du texte source (126 éléments)

{ (pas, eu) ; (de, sa) ; (points, d') ; (le, crâne) ; (stade, de) ; (du, football) ; (la, victoire) ; (gardien, de) ; (relance, de) ; (crâne, n') ; (but, de) ; (presse, fait) ; (tf, 1) ; (français, au) ; (européenne, offerte) ; (c', est) ; (de, france) ; (et, de) ; (de, libération) ; (1, toute) ; (eu, sa) ; (de, toute) ; (qui, avait) ; (mondial, 2006) ; (blanc, sur) ; (tapie, celle) ; (entretemps, numéro) ; (en, direct) ; (toute, la) ; (finale, européenne) ; (direct, sur) ; (de, but) ; (de, retraite) ; (annoncée, en) ; (libération, célèbre) ; (septembre, au) ; (offerte, à) ; (en, finale) ; (marseille, à) ; (barthez, qui) ; (sur, tf) ; (toute, l') ; (grégory, schneider) ; (finale, du) ; (laurent, blanc) ; (des, gardiens) ; (aura, pas) ; (de, laurent) ; (de, la) ; (victoire, du) ; (mauvaise, relance) ; (de, bernard) ; (carrière, marquée) ; (par, deux) ; (numéro, un) ; (au, lendemain) ; (l', histoire) ; (2006, fabien) ; (de, marseille) ; (coupet, était) ; (bartez, le) ; (fait, revivre) ; (gardiens, du) ; (décision, de) ; (à, l') ; (mondiale, en) ; (loin, selon) ; (à, materazzi) ; (la, finale) ; (schneider, de) ; (olympique, de) ; (france, mais) ; (en, septembre) ; (mais, grégory) ; (selon, grégory) ; (celle, de) ; (marquée, par) ; (bernard, tapie) ; (grégory, coupet) ; (fabien, barthez) ; (histoire, du) ; (l', olympique) ; (époque, de) ; (célère, pour) ; (égaliser, en) ; (pour, la) ; (en, 1998) ; (bise, de) ; (un, des) ; (est, une) ; (du, mondial) ; (la, bise) ; (sur, le) ; (deux, points) ; (une, mauvaise) ; (devenu, entretemps) ; (de, loin) ; (du, mondiale) ; (grand, gardien) ; (fabien, bartez) ; (retraite, annoncée) ; (n', aura) ; (était, devenu) ; (de, fabien) ; (plus, grand) ; (sa, revanche) ; (orgue, la) ; (d', orgue) ; (revanche, en) ; (au, stade) ; (d', égaliser) ; (l', époque) ; (revivre, sa) ; (football, français) ; (le, plus) ; (avait, permis) ; (sa, carrière) ; (1998, sources) ; (la, presse) ; (materazzi, d') ; (sa, décision) ; (permis, à) ; (lendemain, de) }

Signature hapax du texte source (63 éléments)

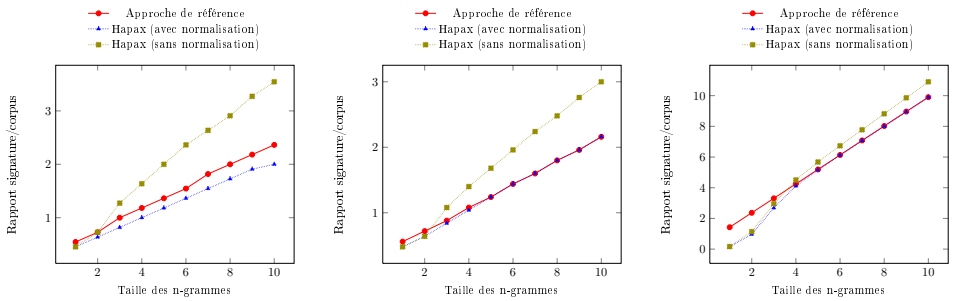
{ (2006, fabien) ; ; (de, retraite) ; (une, mauvaise) ; (devenu, entretemps) ; (annoncée, en) ; (coupet, était) ; (bartez, le) ; (fait, revivre) ; (gardiens, du) ; (grand, gardien) ; (fabien, bartez) ; (le, crâne) ; (la, bise) ; (retraite, annoncée) ; (loin, selon) ; (barthez, qui) ; (à, materazzi) ; (schneider, de) ; (de, fabien) ; (crâne, n') ; (presse, fait) ; (grégory, schneider) ; (français, au) ; (sa, revanche) ; (orgue, la) ; (offerte, à) ; (d', orgue) ; (revanche, en) ; (libération, célère) ; (d', égaliser) ; (européenne, offerte) ; (laurent, blanc) ; (sur, tf) ; (grégory, coupet) ; (selon, grégory) ; (du, mondiale) ; (époque, de) ; (1, toute) ; (revivre, sa) ; (bernard, tapie) ; (football, français) ; (de, laurent) ; (avait, permis) ; (mondial, 2006) ; (mauvaise, relance) ; (eu, sa) ; (blanc, sur) ; (carrière, marquée) ; (célèbre, pour) ; (égaliser, en) ; (materazzi, d') ; (tapie, celle) ; (septembre, au) ; (entretemps, numéro) ; (bise, de) ; (mais, grégory) ; (marseille, à) ; (finale, européenne) ; (du, mondial) ; (deux, points) }

N-grammes retirés de la signature complète (non hapax) : { (marquée, par) ; (par, deux) ; (au, lendemain) ; (l', histoire) ; (du, football) ; (et, de) ; (était, devenu) ; (de, marseille) ; (direct, sur) ; (en, septembre) ; (relance, de) ; (points, d') ; (qui, avait) ; (l', olympique) ; (de, but) ; (est, une) ; (c', est) ; (de, bernard) ; (1998, sources) ; (toute, la) ; (aura, pas) ; (à, l') ; (un, des) ; (sa, décision) ; (fabien, barthez) ; (histoire, du) ; (l', époque) ; (pas, eu) ; (de, libération) ; (décision, de) ; (but, de) ; (plus, grand) ; (n', aura) ; (en, direct) ; (numéro, un) ; (toute, l') ; (olympique, de) ; (des, gardiens) ; (de, sa) ; (mondiale, en) ; (le, plus) ; (de, la) ; (de, france) ; (pour, la) ; (tf, 1) ; (en, 1998) ; (en, finale) ; (stade, de) ; (permis, à) ; (france, mais) ; (la, victoire) ; (celle, de) ; (finale, du) ; (au, stade) ; (de, loin) ; (gardien, de) ; (la, presse) ; (lendemain, de) ; (victoire, du) ; (la, finale) ; (sa, carrière) ; (sur, le) ; (de, toute) }

Figure 5.4 – Descripteur : bigramme mot avec recouvrement, normalisation de la casse, suppression des ponctuations

Corpus	Critère	Descripteur	MAP	Sép.Q.
Piithie	Approche de référence (unigrammes)		0,999	0,708
	max(MAP)	hapax bigrammes sans normalisation	0,999	0,590
	max(Sép. Q)	hapax unigrammes avec normalisation	0,983	0,8
Wikinews	Approche de référence (bigrammes)		0,872	0,800
	max(MAP)	hapax bigrammes avec normalisation	0,856	0,807
	max(Sép. Q)	hapax unigrammes avec normalisation	0,849	0,866
PANini	Approche de référence (6-grammes)		0,823	0,024
	max(MAP)	hapax bigrammes sans normalisation	0,834	0,048
	max(Sép. Q)	hapax unigrammes sans normalisation	0,790	0,090

TABLE 5.1 – Comparaison des meilleurs résultats pour l'approche par hapax par rapport aux approches de référence pour nos trois corpus décrites en section 4.4.



(a) Piithie (b) Wikinews (c) PANini

FIGURE 5.5 – Évolution du coût de stockage en fonction de la taille des n-grammes.

permettent d'égaler ou de dépasser chacun de ces scores individuellement. Ils ne permettent toutefois pas de les dépasser conjointement.

Corpus Wikinews Le meilleur score de la MAP, obtenu avec les bigrammes mots sans normalisation sur les textes, reste inférieur à celui de l'approche de référence (−0,02 points). Cependant, les scores de la Sép. Q pour les unigrammes et les bigrammes avec normalisation sont tous les deux supérieurs à celui de l'approche de référence.

Corpus PANini L'utilisation des bigrammes mots sans normalisation sur les textes offre de meilleurs résultats que l'approche de référence : le score de la MAP y est supérieur de 0,01 points et le score de la Sép. Q est le double de celui de l'approche de référence. L'utilisation d'unigrammes mots sans normalisation sur les textes permet même de plus que tripler la Sép. Q de l'approche de référence mais entraîne une dégradation de 0,03 points de la MAP.

Synthèse Dans le cas des corpus Piithie et Wikinews, la mise en œuvre de l'approche par exploitation des n-grammes mots rares permet de conserver globalement le niveau de performance de l'approche de référence sans réellement pouvoir faire mieux en termes de qualité de classification (MAP) et de capacité de discrimination (Sép. Q). Pour le corpus PANini, les résultats sont meilleurs que ceux de l'approche de référence.

Pour tous les corpus, cette approche entraîne automatiquement une réduction de la taille des signatures pour les n-grammes de petite taille qui obtiennent les meilleurs résultats, comme le montre la figure 5.5, et par conséquent une réduction du coût de traitement. Le gain en coût est d'autant plus important pour le corpus PANini que les meilleurs résultats sont obtenus avec des unigrammes ou des bigrammes contre des 6-grammes pour l'approche de référence.

5.2.2.3 Analyse et discussion

L'exploitation des n-grammes mots hapax permet de maintenir la qualité des résultats par rapport aux approches de références respectives des corpus Piithie et Wikinews, voire d'améliorer ces résultats pour le corpus PANini. Les signatures issues de cette approche sont composées d'un nombre réduit d'éléments par rapport aux signatures complètes ce qui se traduit par une amélioration du coût de stockage et par conséquent d'une amélioration du coût de comparaison des signatures. À résultats comparables, cette approche est donc préférable.

Nos expérimentations ont révélé principalement deux phénomènes : l'approche est particulièrement meilleure sur le corpus PANini et les meilleurs scores sont obtenus en général avec des unigrammes ou des bigrammes.

En premier lieu, l'approche par exploitation des n-grammes mots rares améliore plus nettement les résultats sur le corpus PANini que sur les autres. Les textes de PANini sont les plus volumineux et les dérivations qu'ils contiennent sont de granularité partielle contrairement aux deux autres corpus (*cf. Tableau 4.8, page 120*). La combinaison de ces deux paramètres engendre des signatures complètes très bruitées qui dégradent le score de similarité et donc mécaniquement les résultats. La restriction aux seuls hapax réduit le bruit occasionné en supprimant les n-grammes communs des signatures qui sont les premières causes de ces mauvais scores de par leur présence dans des textes non-dérivés.

Les n-grammes de petite taille (unigrammes à trigrammes) donnent de médiocres résultats sur le corpus PANini pour l'approche par signature complète. Ainsi, l'approche de référence exploite des 6-grammes mots alors que pour les corpus Piithie et Wikinews il s'agit respectivement d'unigrammes et de bigrammes. L'exploitation des n-grammes hapax n'améliore pas les résultats obtenus avec des n-grammes de grande taille (à partir des 5-grammes) puisque ceux-ci sont majoritairement des hapax mais les résultats obtenus avec les n-grammes de petites tailles (unigrammes à trigrammes) progressent fortement jusqu'à dépasser ceux de l'approche de référence. Ces résultats semblent confirmer que la suppression des éléments communs, qui provoquent des correspondances entre des textes non dérivés, améliore les performances.

En second lieu, les n-grammes de petite taille (unigrammes et bigrammes) semblent les plus appropriés à la détection des dérivations. L'approche de référence pour le corpus Piithie repose sur des unigrammes, celle pour le corpus Wikinews sur des bigrammes. Pour l'approche par exploitation des n-grammes rares les meilleurs scores de la MAP sont obtenus à l'aide de bigrammes tandis que les meilleurs scores de la Sép. Q sont obtenus à l'aide d'unigrammes.

Les bigrammes semblent le meilleur compromis entre la spécificité par rapport au texte modélisé et l'invariance par rapport au processus de dérivation. Les unigrammes ne permettent pas de capturer des éléments spécifiques et les n-grammes de plus grande taille ne sont pas conservés par les processus de dérivation impliquant beaucoup de réécritures.

La raison pour laquelle les unigrammes offrent la meilleure capacité de discrimination n'est pas très claire. Les expérimentations montrent qu'en général la capacité de discrimination baisse avec l'augmentation de la taille des n-grammes et que son maximum est en général atteint pour les unigrammes. Dans le cas particulier de notre approche, le nombre de mots hapax, et par conséquent d'éléments dans les signatures, est assez réduit ce qui limite les valeurs accessibles par la mesure de similarité. Un nombre réduit de valeur sur une même amplitude (de 0 à 1) creuse mécaniquement les écarts ce que reflète la Sép. Q.

En conclusion, les expérimentations sur l'exploitation des n-grammes rares nous ont permis de tirer deux nouvelles conclusions quant à la détection des dérivations. Premièrement, la présence d'éléments communs dans les signatures provoque du bruit lors des comparaisons ce qui réduit les performances. Le choix d'éléments spécifiques est préférable et l'utilisation des hapax y répond en partie. Cet argument joue en faveur du choix de la propriété de singularité. Deuxièmement, les bigrammes semblent la taille la mieux adaptée à la détection de dérivation : les unigrammes sont trop peu spécifiques et les n-grammes de plus grande taille sont plus à même d'être modifiés par les réécritures du processus de dérivation.

5.2.3 Exploitation des n-grammes de fort poids informatif

L'approche par sélection des hapax est une mise en œuvre totale de la propriété de singularité. Nous explorons dans cette section une vision plus modérée de la singularité en exploitant les n-grammes de fort poids informatif à l'aide du $\mathrm{tf} \cdot \mathrm{idf}$ (*cf. Section 5.1.2.2*). Nous détaillons tout d'abord nos choix expérimentaux pour la mise en œuvre de cette approche. Nous présentons puis discutons ensuite les résultats obtenus sur les corpus Piithie, Wikinews et PANini. Le détail de l'analyse des résultats est présenté en annexe (*cf. Annexe G.2*).

Classe	tf · idf	Élément	Classe	tf · idf	Élément
0	0,004 514	planeta	2	0,002 078	ogle-2005
1	0,002 257	austral	3	0,001 385	2,6
1	0,002 257	gazeuse	3	0,001 385	gravitationnelle
1	0,002 257	jean-philippe	3	0,001 385	blg-390lb
1	0,002 257	hémisphère	3	0,001 385	exoplanète
1	0,002 257	223	4	$6{,}926 \cdot 10^{-4}$	radioactivité
1	0,002 257	scoperto	4	$6{,}926 \cdot 10^{-4}$	lentille
1	0,002 257	pianeta	4	$6{,}926 \cdot 10^{-4}$	tellurique
1	0,002 257	lació	4	$6{,}926 \cdot 10^{-4}$	descubren
1	0,002 257	anomalies	4	$6{,}926 \cdot 10^{-4}$	rassemblées
1	0,002 257	constel	4	$6{,}926 \cdot 10^{-4}$	nuevo
1	0,002 257	micro-lentille	4	$6{,}926 \cdot 10^{-4}$	blg-390l
1	0,002 257	détection	4	$6{,}926 \cdot 10^{-4}$	rapproche

TABLE 5.2 – Catégorisation des éléments en classes de tf · idf.

5.2.3.1 Mise en œuvre expérimentale

L'approche par sélection des n-grammes de plus fort poids informatif consiste à construire une signature composée des n-grammes qui obtiennent les meilleurs scores de tf · idf.

Tout comme pour la rareté où nous avions fixé un seuil (df = 1), nous devons fixer un seuil de tf · idf pour sélectionner les n-grammes. Ni une sélection empirique des n plus hauts scores, ni l'utilisation d'un score absolu comme seuil empirique ne seraient satisfaisants. En effet, nos expérimentations préliminaires nous montrent que les éléments se concentrent autour de scores similaires aux valeurs variables selon les textes et les distributions de référence. Pour palier cela, nous proposons de regrouper les éléments de même score afin de former des classes et d'opérer la sélection sur celles-ci. Par abus de langage, lorsque nous parlerons des éléments d'une classe nous ferons référence aux éléments contenus dans ladite classe et dans les classes de scores supérieurs. Le tableau 5.2 illustre quelques unigrammes et leurs classes respectives ordonnées sur deux colonnes. La classe 0 contient uniquement « planeta » tandis que la classe 3 regroupe les unigrammes « planeta » à « ogle-2005 ». Les éléments de la classe 0 ont un tf · idf plus fort que la classe 1. Lorsque nous parlons de la classe ≤ 1 nous faisons référence aux éléments des classes 0 et 1.

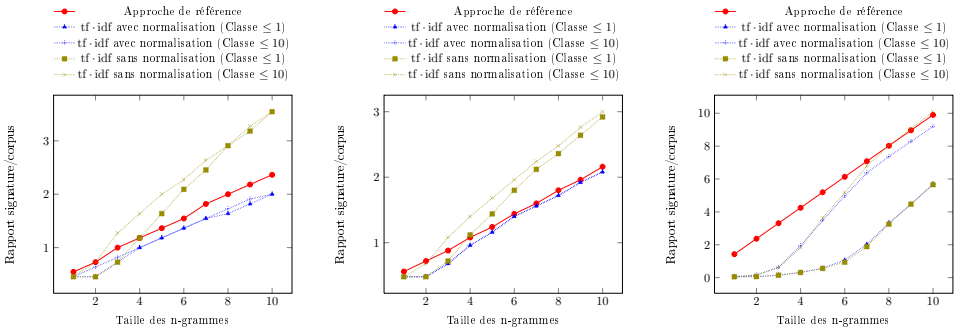
5.2.3.2 Résultats

Le tableau 5.3 présente les résultats, pour chacun de nos corpus, de l'exploitation des n-grammes de plus fort poids informatif en termes de MAP et de Sép. Q. Nous avons exploré la sélection de la première à la dixième classe de tf · idf pour les différentes tailles de n-grammes avec la mise en œuvre ou non de normalisation sur les textes (racinisation et filtrage des mots outils). Nous retenons dans le tableau, pour chaque corpus, uniquement la meilleure configuration en termes de MAP et la meilleure configuration en termes de Sép. Q.

Corpus Piithie Le meilleur score de la MAP est obtenu avec les bigrammes mots des classes ≤ 7 avec normalisation sur les textes. Le meilleur score de la Sép. Q est obtenu avec les bigrammes mots des classes ≤ 1 sans normalisation. Les performances obtenues avec ces descripteurs sont inférieures à celles de l'approche de référence, en

Corpus	Critère	Descripteur	MAP	Sép. Q.
Piithie	Approche de référence (unigrammes)		0,999	0,708
	max(MAP)	tf · idf bigrammes avec normalisation (classes ≤ 7)	0,999	0,572
	max(Sép. Q)	tf · idf bigrammes sans normalisation (classes ≤ 1)	0,876	0,666
Wikinews	Approche de référence (bigrammes)		0,872	0,800
	max(MAP)	tf · idf bigrammes avec normalisation (classes ≤ 10)	0,880	0,794
	max(Sép. Q)	tf · idf unigrammes avec normalisation (classes ≤ 0)	0,813	1,0
PANini	Approche de référence (6-grammes)		0,823	0,024
	max(MAP)	tf · idf 7-grammes avec normalisation (classes ≤ 10)	0,819	0,018
	max(Sép. Q)	tf · idf unigrammes sans normalisation (classes ≤ 10)	0,796	0,02

TABLE 5.3 – Comparaison des meilleurs résultats pour l'approche par sélection des n-grammes de plus fort poids informatif par rapport à l'approche de référence pour nos trois corpus.



(a) Piithie (b) Wikinews (c) PANini

FIGURE 5.6 – Évolution du coût de stockage en fonction de la taille des n-grammes.

particulier pour la Sép. Q. Le meilleur résultat en termes de MAP est égal à celui de l'approche de référence mais la Sép. Q est alors inférieure de 0,13 points.

Corpus Wikinews Contrairement au corpus Piithie, le meilleur score de la MAP, obtenu avec les bigrammes mots des classes ≤ 10 avec normalisation sur les textes, est supérieur de 0,008 points à celui de l'approche de référence pour une dégradation du score de la Sép. Q de seulement 0,006 points. Les unigrammes de la classe ≤ 0 avec normalisation sur les textes permettent d'atteindre le score de Sép. Q maximum. Dans cette configuration les scores de similarité entre les textes dérivés sont de 1 contre 0 pour les non dérivés.

Corpus PANini Les meilleurs scores de la MAP et de la Sép. Q sont obtenus avec les 7-grammes mots des classes ≤ 10 avec normalisation sur les textes dans le premier cas et avec les unigrammes mots des classes ≤ 10 sans normalisation dans le second cas. Les résultats sont inférieurs à ceux de l'approche de référence tout en restant comparables : de $-0{,}005$ points à $-0{,}027$ points pour la MAP et de $-0{,}004$ à $-0{,}006$ points pour la Sép. Q.

Synthèse Dans le cas des corpus Piithie et PANini, la mise en œuvre de l'approche par exploitation des n-grammes mots de fort poids informatif dégrade légèrement les performances par rapport à l'approche de référence. Pour le corpus Wikinews, cette approche permet de faire légèrement mieux en termes de MAP sans impact majeur sur la Sép. Q.

Pour tous les corpus, l'approche entraîne automatiquement une réduction de la taille des signatures pour les n-grammes de petites tailles, comme le montre la figure 5.6, et par conséquent une réduction du coût de traitement.

Les résultats semblent indiquer, plus particulièrement pour Piithie et Wikinews, que la meilleure Sép. Q est obtenue pour un nombre de classe de score réduit (≤ 1 ou ≤ 0) contrairement aux configuration des meilleurs MAP qui sont toutes élevées (≤ 7 ou ≤ 10).

5.2.3.3 Analyse et discussion

L'exploitation des n-grammes mots de plus fort poids informatif permet, à l'instar de l'approche par exploitation des n-grammes rares, de maintenir la qualité des résultats en termes de qualité de la classification par rapport aux approches de références respectives pour les corpus Piithie et PANini, voire d'améliorer ces résultats pour le corpus Wikinews. Les résultats en termes de capacité de discrimination sont légèrement dégradés pour les corpus Piithie et PANini. Pour autant, les signatures issues de cette approche sont composées d'un nombre réduit d'éléments par rapport aux signatures complètes ce qui se traduit par une amélioration du coût de stockage et ainsi d'une amélioration du coût de comparaison des signatures. Cette approche peut être préférable selon l'intérêt porté à chacune des dimensions de l'évaluation.

Tout comme pour l'approche par exploitation des n-grammes mots rares, nous avons observé que les unigrammes et les bigrammes étaient particulièrement adaptés à la détection des liens de dérivation. Nous ne le détaillons pas de nouveau. Nous observons par contre que le corpus Wikinews réagit mieux à cette approche que les deux autres, alors que c'était l'inverse précédemment et que l'augmentation du nombre de classes de scores considérées se traduit en général par une amélioration des résultats.

En premier lieu, l'approche est plus efficace sur le corpus Wikinews que sur les deux autres corpus. En effet, elle permet de dépasser les scores de la MAP ou de la Sép. Q de l'approche de référence, sans pour autant faire mieux que les deux conjointement.

	Approche	Éléments	Normalisation	Rang	MAP	Sép.Q.	Stockage
Piithie	Référence	unigrammes	oui		0,999	0,708	100 %
	Hapax	unigrammes	non		0,989	0,769	90 %
	tf · idf	bigrammes	oui	≤ 7	0,999	0,572	87 %
Wikinews	Référence	bigrammes	oui		0,872	0,800	100 %
	Hapax	bigrammes	oui		0,856	0,807	87 %
	tf · idf	bigrammes	oui	≤ 10	0,880	0,794	89 %
PANini	Référence	6-grammes	oui		0,823	0,024	100 %
	Hapax	bigrammes	non		0,834	0,048	18 %
	tf · idf	7-grammes	oui	≤ 10	0,819	0,018	103 %

TABLE 5.4 – Comparaison des maximums sélectionnés pour les différentes approches et les différents corpus. La MAP a été privilégiée par rapport à la Sép.Q.

Ce résultat nous laisse penser que les dérivations de type révision réagissent mieux à l'exploitation des n-grammes mots de poids forts qu'aux n-grammes rares.

En second lieu, les expérimentations montrent qu'en général plus l'on considère un grand nombre de classe de scores de tf · idf, plus les scores de la MAP sont élevés tandis qu'ils sont plus mauvais lorsque ce nombre est réduit. Un nombre réduit de classes se traduit par une signature contenant trop peu d'éléments pour être discriminante : elle n'offre pas suffisamment de matériau permettant de distinguer les textes. Ainsi, la seule considération des n-grammes mots de la classe ≤ 0 donne les plus mauvais résultats excepté pour le corpus Wikinews où cette configuration offre la meilleure Sép. Q. À l'inverse, la prise en compte des n-grammes mots des classes ≤ 10, soit la plus haute classe expérimentée, donne les meilleurs résultats. Cette observation est vraie également pour le corpus Piithie, même si la MAP atteint sa valeur maximale dès la classe ≤ 7. Des expérimentations complémentaires non rapportées montrent que lorsque l'on considère des classes de scores au-delà de 10 pour les corpus Piithie et PANini, les résultats en termes de MAP rejoignent asymptotiquement ceux de l'approche de référence.

En conclusion, les dérivations de type révision réagissent mieux à l'exploitation des n-grammes mots de poids forts qu'aux n-grammes rares. De plus, les signatures composées d'un nombre trop réduit d'éléments donnent de mauvais résultats, et ce indépendamment de la taille des n-grammes. Par extension, nous pensons que cette observation est vraie quels que soient les descripteurs et découle directement de la théorie de l'information (Shannon, 1948). À l'inverse, plus l'on augmente le nombre de n-grammes retenus dans la signature et meilleurs sont les résultats en terme de qualité de classification mais plus l'on se rapproche également de la configuration de la signature complète. Le choix des classes de n-grammes retenus est donc un paramètre de compromis intéressant entre résultats et coût de stockage.

5.2.4 Conclusion

Le tableau 5.4 fait la synthèse des résultats, en termes de MAP puis de Sép. Q, obtenus en exploitant les n-grammes rares et ceux de fort poids informatif. Il est difficile de parler de meilleurs résultats étant donné les différents objectifs à couvrir. Nous notons notamment que les économies en coûts de stockage et de calcul ne sont pas aussi prononcés que pour d'autres configurations non rapportées [9].

9. Nous avons en effet privilégiés les résultats en termes de MAP et de Sép. Q, mais des configurations qui donnent des résultats très proches, quoi qu'inférieurs, permettent une meilleure contraction

La première conclusion que nous tirons de ces expérimentations concerne la taille des n-grammes. Au regard des résultats, les n-grammes de petites tailles (notamment les bigrammes) semblent les mieux adaptés à la tâche de détection de dérivation. Ils donnent en effet de bons résultats en termes de qualité de classification ainsi que de capacité de discrimination.

La seconde conclusion que nous tirons de ces expérimentations est l'inaptitude des approches dérivées de la signature complète à appréhender les dérivations à granularité partielle du corpus PANini. Contrairement aux deux autres corpus, les meilleurs résultats sont obtenus avec des n-grammes de grande taille. Nous avions supposé que cela était dû à l'absence de modifications importantes lors du processus de dérivation mais les résultats obtenus avec les bigrammes hapax semblent contredire cette hypothèse.

5.3 Exploitation des éléments linguistiques

Nous explorons l'utilisation de descripteurs linguistiques. Nous présentons tout d'abord les expérimentations sur les entités nommées, puis sur les composés nominaux. En plus des résultats expérimentaux, nous discutons pour chacun des descripteurs les erreurs rencontrées en réalisant un retour aux textes. Finalement nous concluons sur les résultats de ces deux approches par rapport aux résultats précédents.

5.3.1 Exploitation des entités nommées

Dans cette section nous évaluons l'utilisation d'une signature à base d'entités nommées telle que décrite en section 5.1.3.1. Nous détaillons tout d'abord nos choix expérimentaux pour la mise en œuvre de cette approche. Nous présentons puis discutons ensuite les résultats obtenus sur les corpus Piithie, Wikinews et PANini.

5.3.1.1 Mise en œuvre expérimentale

L'extraction d'entités nommées est une technologie mature qui obtient des niveaux de performance de l'ordre de 90 % en précision et en rappel sur les articles de presse en anglais (MUC, 1993). Nous utilisons deux systèmes d'extraction d'entités nommées pour ces expérimentations : *Némesis* pour le français et l'*Illinois Named Entity Tagger* pour l'anglais. Chaque texte est représenté par l'ensemble des formes textuelles détectées comme entités nommées par ces outils, indépendamment de leur type (anthroponyme, toponyme...).

Némesis (Fourour, 2004) a été développé exclusivement pour le français. Il effectue une reconnaissance des entités nommées à l'aide de règles lexicales et syntaxiques complétées par un processus d'apprentissage permettant d'enrichir automatiquement le lexique. Il a été évalué sur des articles de presse (Fourour et collab., 2002, p. 1073) où il extrait les anthroponymes et les toponymes avec une précision de 95 % et un rappel de 90 %.

L'*Illinois Named Entity Tagger*[10] (Ratinov et Roth, 2009) est un extracteur d'entités nommées multilingue. Il utilise des nomenclatures géographiques (*gazetteers*) tirées de Wikipédia, un modèle de classes de mots ainsi que des traits d'expression non liés à la localité. C'est un extracteur robuste qui a été évalué sur plusieurs types

des signatures.

10. Le logiciel peut-être téléchargé librement à l'adresse : `http://cogcomp.cs.illinois.edu/page/software_view/4`

Corpus	Approche	MAP	Sép. Q	Stockage
Piithie	Référence	0,999	0,708	100 %
	Entités nommées	0,839	0,628	87 %
Wikinews	Référence	0,872	0,800	100 %
	Entités nommées	0,646	0,833	61 %
PANini	Référence	0,823	0,024	100 %
	Entités nommées	0,774	0,036	2 %

TABLE 5.5 – Comparaison des résultats de l'approche par entités nommées par rapport aux approches de référence respectives des différents corpus.

de corpus et notamment sur celui de la conférence CoNLL03 [11] pour lequel il a obtenu un F-score de 90,8 %. Il est livré avec plusieurs modèles d'apprentissage. Nous avons choisi d'utiliser le modèle à une couche *allLayer1* pour l'anglais qui offre de très bons résultats pour un temps de traitement raisonnable. Malheureusement, l'outil n'a pas été en mesure de produire de résultats pour une trentaine de sources (sur presque mille) du corpus PANini. Les résultats ont été calculés sans tenir compte de ces sources.

5.3.1.2 Résultats

Le tableau 5.5 présente, pour chacun de nos corpus, les résultats de l'exploitation des entités nommées en termes de MAP et de Sép. Q. Nous pouvons observer que la qualité de la classification est nettement inférieure à celle de l'approche de référence quel que soit le corpus. Ainsi, la MAP se dégrade de 0,16 points pour Piithie, de 0,22 points pour Wikinews et de 0,04 points pour PANini. À l'inverse, la capacité de discrimination est meilleure de 0,03 points pour Wikinews et 0,01 pour PANini mais est 0,08 points inférieure pour Piithie. En réalité, le seul véritable gain de cette approche concerne le coût de stockage de la signature [12] qui est réduit de 20 % pour Piithie, 30 % pour Wikinews et de 98 % [13] pour PANini par rapport à l'approche de référence.

Il est intéressant de noter que la compression la plus spectaculaire obtenue pour PANini correspond également à la plus faible dégradation des résultats.

Pour l'extraction des signatures précédentes, nous avions considéré que le coût de construction de la signature était négligeable au regard du nombre d'opérations de comparaisons potentielles (*cf. Section 5.1.1*). Si cette remarque est toujours fondée, force est de constater que le coût de l'extraction des entités nommées est bien supérieur à celui des signatures précédentes — nous estimons l'écart à un facteur 10 — et que par conséquent le nombre de comparaisons nécessaires à équilibrer les coûts d'extraction et de comparaison est d'autant plus élevé.

11. `http://www.cnts.ua.ac.be/conll2003/ner/`

12. Pour le corpus PANini, le rapport entre taille du corpus et de la signature a été mesuré en retirant le poids des vingt-huit textes pour lesquels nous n'avons pas été en mesure d'extraire les entités nommées.

13. La différence importante entre les gains pour PANini et les deux autres corpus s'explique par la nature du premier corpus. PANini est composé de beaucoup de fichiers beaucoup plus volumineux que les autres corpus mais d'un genre différent qui est beaucoup moins riche en entités nommées. La compression est d'autant plus importante.

5.3.1.3 Analyse et discussion

Nous avons opéré un retour au texte pour analyser de manière qualitative les erreurs de l'approche basée sur les entités nommées. Nous nous sommes seulement intéressé aux erreurs correspondant aux négatifs (absence de lien de dérivation avéré) obtenant les scores de similarité les plus élevés ainsi qu'aux positifs (présence d'un lien de dérivation avéré) obtenant les scores de similarité les plus bas. Nous excluons de cette sélection les textes obtenant des scores de similarité de 0 ou NaN qui correspondent à l'absence d'entités nommées en commun ou à l'absence d'entités nommées du texte.

Négatifs avec un score de similarité élevé Les paires de textes qui ne correspondent pas à des liens de dérivation mais qui sont toutefois hautes placées dans les classements correspondent à quatre cas : (i) les signatures sont de tailles très différentes ce qui déséquilibre le calcul de la similarité, (ii) les textes exposent peu d'entités nommées, (iii) les textes traitent de sujets distincts impliquant les mêmes entités et (iv) l'annotation du corpus est discutable.

Le premier cas est le plus répandu et couvre la majorité des erreurs. Le mauvais classement est dû à un déséquilibre du nombre d'éléments des signatures : un des textes comporte un très grand nombre d'entités nommées tandis que le second n'en a qu'un nombre réduit. Ce scénario se produit lorsque les deux textes sont de tailles très différentes et l'un utilise donc mécaniquement plus d'entités nommées que l'autre. Le score de similarité élevé entre les deux signatures est un effet de bord de l'utilisation de la mesure de similarité c_{max} (*cf. Section 4.4.2*). Plusieurs des entités nommées de la plus grande signature sont communes à la plus petite, ce qui rapporté à la taille de cette dernière donne un score élevé. Les toponymes de lieux communs sont en parti fautifs puisque « France », « Londres », « New-York », « Suisse », « Washington » et « United States » apparaissent plusieurs fois dans les intersections des signatures de textes non dérivés.

Le deuxième cas, qui a des conséquences similaires au premier mais pour des causes différentes, a été observé dans les corpus Wikinews et PANini. Les textes incriminés utilisent très peu d'entités nommées (de l'ordre de la demi-douzaine). Il suffit alors de quelques éléments en commun pour faire bondir le score de similarité. Encore une fois, les toponymes, tels que « France » ou « Nord », se retrouvent dans ces intersections. On y trouve également le nom « God » pour PANini, duquel on pourrait questionner le statut d'entité nommée.

Le troisième cas n'a été observé que pour le corpus Piithie. Les deux textes traitent de deux sujets distincts mais qui impliquent les mêmes entités nommées. Ainsi, un premier texte traite du détournement d'un fleuve vers la ville de Pékin afin d'alimenter cette dernière en eau en prévision des Jeux Olympiques. Le second relate la colère des journalistes dépêchés à Pékin pendant les Jeux Olympiques concernant la censure d'Internet. Chacun de ces textes fait largement appel aux toponymes de la région de Pékin ainsi qu'aux entités liées aux Jeux Olympiques, de sorte que les deux signatures possèdent un grand nombre d'éléments en commun et donc un score de similarité élevé.

Finalement, le dernier cas n'a également été observé que dans le corpus Piithie. La paire de textes src00049–cand00582 a été annotée comme non-dérivée, mais nous questionnons ce choix étant donné que les deux textes traitent du même sujet, relatent les mêmes citations et sont structurés de la même façon.

Positifs avec un score de similarité bas Nous avons identifié trois cas pour les paires correspondant à de réels liens de dérivation mais obtenant de mauvais scores de similarité : (i) peu d'entités communes pour des textes de taille similaires, (ii) entités

peu visibles ou modifiées par la différence de taille des textes et (iii) bruit introduit par des entités puisées hors du contenu.

Le premier cas n'est observé que pour le corpus Piithie. Les liens de dérivation ne sont pas identifiés car les textes dérivés ont peu d'entités nommées en commun avec leur texte source. Les causes semblent être doubles. La première est purement technique : les extracteurs ne sont pas en mesure d'identifier toutes les entités nommées. La seconde remet partiellement en cause l'invariance des entités nommées. Nous avons pu observé que si la référence de l'entité nommée est présente, la forme textuelle du document source n'est pas forcément préservée lors du processus de dérivation. Ainsi, « Directeur Général du groupe », « Groupe CMA CGM » ou encore « Advens » mutaient en « PDG du groupe », « Compagnie malienne pour le développement des textiles » et « Groupe Advens ». Dans d'autres cas, l'auteur glisse d'un concept à un autre en détaillant ou en généralisant. Ainsi, plutôt que d'indiquer qu'un évènement « prend place au Mexique », il écrira qu'il « prend place à Mexico ».

Le deuxième cas est commun à tous nos corpus mais plus particulièrement dans PANini. Les textes de tailles très différentes ont un volume aussi différent d'entités nommées ce qui trompe notre approche. Le phénomène se manifeste d'une part par la dilution des entités nommées communes à la source et son dérivé dans la masse des entités distinctes. Par exemple un extrait de 240 caractères de PANini est dérivé d'un texte originale qui en fait 38 000. La seule entité nommée, « Desmonds », dérivé du texte source est noyée dans la centaine présentes dans le texte original. Il se manifeste d'autre part par l'utilisation de formes textuelles distinctes. Par exemple, lors de la synthèse d'un article de Piithie, l'auteur choisit alors de remplacer les formes textuelles employées par d'autres, sans doute sous des contraintes d'écriture. Ainsi, « Arnaud Apoteker de Greenpeace France » se contracte en « Greenpeace » tandis que « UMP » est nominalisé en « parti majoritaire ».

Finalement, le troisième cas est un cas particulier du premier où au nombre réduit d'entités nommées s'ajoutent des entités nommées qui ne relèvent pas du contenu. Ces dernières gonflent artificiellement la signature, entraînant la baisse du score de similarité. Le cas le plus éloquent est celui de l'entité « AFP » qui apparaît dans tous les articles de l'agence et qui se retrouve dans les signatures des textes sources mais pas dans les textes dérivés. La modélisation de l'organisation du document permettrait de prendre en compte ces particularités.

Synthèse L'analyse des résultats montre que cette approche n'est pas adaptée, en l'état, aux cas où les textes sont de tailles très différentes ou lorsque la granularité de la dérivation est partielle. Nous avons également noté que certaines entités nommées, les toponymes de lieux communs notamment, ne sont pas assez singulières et créent du bruit plus qu'elles ne participent à la discrimination des liens de dérivation. Il serait envisageable de les filtrer, mais les entités sont assez rares (d'où la faible taille des signatures) et nous risquerions d'appauvrir fortement les signatures. La solution d'une pondération des entités selon leur représentation de la texte reste à explorer.

5.3.2 Exploitation des composés nominaux

Dans cette section nous évaluons l'utilisation d'une signature à base de composés nominaux telle que décrite en section 5.1.3.2. Nous détaillons tout d'abord nos choix expérimentaux pour la mise en œuvre de cette approche. Nous présentons puis discutons ensuite les résultats obtenus sur les corpus Piithie, Wikinews et PANini.

Motif syntaxique	Exemples	Nb. instances Piithie	Wikinews
N A	*mission première, Assemblée Parlementaire, aviation civile*	17 439	31 516
N N	*ministre francais, lobbyistes pro-brevets, équipement anti-émeute*	4 488	19 324
N à Vinf	*impôt à acquitter, fonds à venir*	452	515
N P D N	*annulation d'une TVA, appartements de certains immeubles, candidats à l'investiture*	10 555	19 595

TABLE 5.6 – Motifs syntaxiques utilisés pour le français et composés correspondant extraits des corpus Piithie et Wikinews

5.3.2.1 Mise en œuvre expérimentale

L'identification des composés nominaux a fait l'objet de nombreux travaux. Elle a reçu une attention proportionnelle à celle dédiée à l'extraction terminologique qui reste une problématique majeure en Traitement du Langage Naturel (Sag et collab., 2002). Les premiers systèmes d'extraction des syntagmes nominaux étaient stochastiques et reposaient sur les statistiques de colocations des rôles grammaticaux (Church, 1988). Ils ont évolué par la suite vers une analyse morphologique et syntaxique (Seretan et Wehrli, 2008). Ainsi, l'outil LEXTER (Bourigault, 1992) extrait les termes candidats à partir de motifs syntaxiques, spécifiques à chaque langue, exprimés à l'aide d'étiquettes de rôles grammaticaux.

Pour le français Daille (2007) propose trois motifs syntaxiques[14] de base :

N A *emballage biodégradable*, *protéine végétale* ;

N (Prep (D)) N *ions calcium*, *protéine de poissons*, *chimioprophylaxie au rifampine*, *lait de brebis* ;

N à Vinf *viande à griller*.

À ces motifs de base s'ajoutent certaines variations syntaxiques telles que l'ajout d'un adjectif qualificatif au sein du motif N (Prep (D)) N (*lait cru de brebis*) ou à l'inverse d'une précision nominale dans N A (*protéine d'origine végétale*). Il est également possible de combiner récursivement ces motifs de base.

Nous considérons les formes de base en acceptant des possibles recouvrement lors des variations. Les expérimentations précédentes montrent que les grandes tailles sont peu conservées alors que les bigrammes obtiennent de bons résultats. Pour les mêmes raisons, nous autorisons le recouvrement des instances, p. ex. la reconnaissance par N N A de « forces armées britanniques » n'empêche pas N N d'extraire « forces armées ». L'augmentation artificielle du nombre d'éléments dans la signature permet de capter les cas de réécritures partielles.

En pratique, nous utilisons le composant *HMM Tagger* d'Apache UIMA[15] que nous avons entraîné[16] sur le corpus French Treebank (Abeillé et collab., 2003). Par

14. Les étiquettes de rôles grammaticaux utilisés sont A (Adjectif), N (Nom), D (Déterminant), Prep (Préposition) et Vinf (Verbe à l'infinitif).

15. `http://uima.apache.org/downloads/sandbox/hmmTaggerUsersGuide/hmmTaggerUsersGuide.html`

16. Notre méthode d'entraînement sur le French Treebank est similaire à celle présentée dans notre article Charles Dejean et collab. (2010).

conséquent, les étiquettes utilisées sont les mêmes que celles du French Treebank [17] et offrent une analyse syntaxique plus fine que nécessaire. Elles distinguent ainsi différents types d'adjectifs (qualificatifs `A:qual`, possessifs `A:poss`, cardinaux `A:card`...), de déterminants (démonstratifs `D:dem`, définis `D:def`, indéfinis `D:ind`...) ou encore de noms (communs `N:C`, cardinaux `N:card`, propres `N:P`). Nous nous limitons aux seuls noms communs (`N:C`) et délaissons les noms propres qui sont couverts par les entités nommées. Nous conservons les différentes variations des adjectifs et des déterminants. Nous avons donc écrit autant de règles syntaxiques que de combinaisons possibles de ces différentes sous-catégories. Par exemple pour le motif **A N** nous avons écrit des règles du type `N:C A:qual`, `N:C A:poss` et `N:C A:card`. Le tableau 5.6 illustre quelques instances des motifs de base capturés dans nos corpus.

Pour l'anglais Nous reprenons largement l'approche édifiée pour le français. Nous repartons des motifs syntaxiques de base pour l'anglais proposés par Bowker (2002) qui n'acceptent pas de variation syntaxique contrairement au français :

N N *antivirus software*, *virus signature* ;

A N *firm resolution*, *full range*, *powerful force* ;

Nous reprenons le même système d'extraction que pour le français, au choix du modèle et des règles syntaxiques près. Nous utilisons le modèle du composant *HMM Tagger* entraîné sur le Brown Corpus (Francis et Kucera, 1979) et par conséquent nous utilisons les étiquettes du Brown Corpus [18].

5.3.2.2 Résultats

Le tableau 5.7 présente, pour chacun de nos corpus, les résultats de l'exploitation des composés nominaux en termes de MAP et de Sép. Q. À l'instar de ce que nous avions pu observer pour les entités nommées, la qualité de la classification obtenue avec les composés nominaux est nettement moins bonne que pour l'approche de référence. Ainsi, la MAP se dégrade de 0,11 points pour Piithie, de 0,04 points pour Wikinews et de 0,03 points pour PANini. Les écarts sont toutefois moins importants que ce que nous avions observé pour les entités nommées. La capacité de discrimination varie assez peu par rapport à l'approche de référence. Elle est équivalente pour le corpus PANini et dégradée de 0,06 point environ pour Piithie et Wikinews. Encore une fois, le véritable gain de cette approche concerne le coût de stockage de la signature qui est réduit de 17 % pour Piithie, 25 % pour Wikinews et de 97 % pour PANini par rapport à l'approche de référence.

Il est intéressant de noter que, comme nous l'avions observé pour les entités nommées, la compression la plus spectaculaire obtenue pour PANini correspond également à la plus faible dégradation des résultats.

5.3.2.3 Analyse et discussion

Tout comme nous l'avons fait pour les entités nommées, nous opérons une retour au texte pour les erreurs correspondant aux négatifs (absence de lien de dérivation avéré) obtenant les scores de similarité les plus élevés ainsi que les positifs (présence d'un lien de dérivation avéré) obtenant les scores de similarité les plus bas.

17. `http://www.llf.cnrs.fr/Gens/Abeille/French-Treebank-fr.php`
18. La liste des étiquettes est disponible à l'adresse : `http://khnt.aksis.uib.no/icame/manuals/brown/INDEX.HTM#bc6`

Corpus	Approche	MAP	Sép. Q	Stockage
Piithie	Référence	0,999	0,708	100 %
	Comp. Nom.	0,889	0,640	90 %
Wikinews	Référence	0,872	0,800	100 %
	Comp. Nom.	0,831	0,746	67 %
PANini	Référence	0,823	0,024	100 %
	Comp. Nom.	0,793	0,027	3 %

TABLE 5.7 – Comparaison des résultats de l'approche par composés nominaux par rapport aux approches de référence respectives des différents corpus.

Négatifs avec un score de similarité élevé On retrouve dans l'analyse des paires de textes qui ne correspondent pas à des liens de dérivation mais qui sont toutefois hautes placées dans les classement le scénario majoritaire que nous avions discuté pour les entités nommées, à savoir des signatures de tailles très différentes. Nous distinguons toutefois deux cas : (i) les textes traitent de sujets différents et (ii) les textes traitent de la même thématique sans être dérivés.

Le premier cas constitue la totalité des erreurs analysées pour les corpus Wikinews et PANini. Les contextes sont différents puisque pour les corpus Piithie et Wikinews les textes sont de tailles comparables mais pas les signatures tandis que pour PANini les textes sont de tailles très différentes et il en est donc mécaniquement de même pour les signatures. Les éléments en commun des signatures qui provoquent l'augmentation du score de similarité sont de deux types. Il y a d'une part les composés nominaux très peu spécifiques, tel que « couleur rouge » que l'on retrouve dans une paire du corpus Piithie, et d'autre part des éléments qui ne relèvent pas directement du contenu mais du modèle de document, tels que « droit réservé » ou « sources * » [19]. Dans le premier cas, nous pourrions envisager une sélection statistique mais elle risquerait d'entraîner une diminution drastique de la taille des signatures et probablement l'augmentation probable de signatures vides. Dans le second cas, comme nous l'avions également évoqué pour les entités nommées, l'appel à un modèle de document nous permettrait de supprimer le texte qui ne relève pas directement du contenu.

Le second cas consiste en des textes traitant de la même thématique sans être des dérivés. Pour deux des erreurs observés, les textes traitent même du même sujet : l'implication des États-Unis et de l'Europe dans la gestion du conflit russo-géorgien et la censure du réseau Internet par les autorités chinoises pendant les jeux olympiques. Les éléments en commun des signatures sont caractéristiques de ces sujet, soit respectivement {*troupes russes*, *état américain*, *intégrité territoriale*, *union européenne*, *président russe*} et {*liberté totale*, *mouvement spirituel*, *comité international*, *comité olympique*, *autorités chinoises*, *radio allemande*}. Ces éléments en commun, au regard des tailles des signatures, entraînent un score de similarité élevé. Une autre paire de texte analysée traite du lancement d'une offre de vidéo à la demande, mais chaque texte relate un sujet différent. D'un côté la négociation d'un accord entre SFR et TF1, et de l'autre le lancement d'une offre par Carrefour. La combinaison de ces composés nominaux avec les entités nommées permettrait certainement de résoudre cette erreur.

Positifs avec un score de similarité bas On retrouve dans l'analyse des paires de textes qui correspondent à des liens de dérivation mais qui sont dans le bas du

19. Le caractère « * »correspond à une puce de liste.

classement globalement les mêmes cas que pour les entités nommées : (i) peu de composés communs pour des textes de taille similaire, (ii) composés peu présents ou modifiés par la différence de taille des textes et (iii) composés nominaux non conservés à cause d'un changement de langue. Contrairement à ce que nous avions observé précédemment, chacun de ces cas est spécifique à un corpus.

Le premier cas caractérise les erreurs du corpus Piithie. Nous avons observé deux causes du nombre réduit de composés nominaux en commun. En premier lieu, les composés nominaux ne sont pas forcément les mieux à même de décrire les textes dont le sujet s'articule autour d'entités plutôt que de concepts. Par exemple, lorsque les articles traitent d'une enquête judiciaire impliquant une personnalité politique et une entreprise. Les composés nominaux caractérisent uniquement le fait qu'il s'agisse d'une enquête alors que la spécificité réside dans les parties impliquées. En second lieu, les composés nominaux ne se recoupent pas à cause de variations diverses : idiosyncrasies (*ministre de l'economie* vs. *ministre de l'économie*), variations lexicales (*fonds propres* vs. *capitaux propres*), glissement sémantique (*secondes dans un sauna* vs. *minutes dans un sauna*) ou modification de l'angle de présentation (*fournaise pendant 5 minutes* vs. *résistant pendant 5 minutes*).

Le deuxième cas est spécifique des erreurs rencontrées pour le corpus PANini. Encore une fois, de par sa construction, PANini expose plus particulièrement les problèmes liés à la granularité de la dérivation. La différence de taille des passages dérivés par rapport à la taille des textes noie le nombre de composés nominaux en commun dans la masse des composés distincts. Nous avons toutefois observé ce même scénario pour un article du corpus Piithie qui résumait plusieurs autres articles.

Enfin, le dernier cas a été observé sur le corpus Wikinews. Des révisions dans d'autres langues que le français ont été conservées dans le corpus. Elles proviennent probablement de mises-à-jour depuis la version anglaise. Les composés nominaux présents en français et en anglais ne sont bien sûr pas alignables dans leurs graphies respectives.

Synthèse L'analyse des erreurs montre que cette méthode n'est pas adaptée aux cas où les textes sont de tailles très différentes ou lorsque la granularité de la dérivation est partielle. Nous nous sommes interrogés sur la part de responsabilité du score de similarité étant donnés les effets de bords constatés de c_{max} (*cf. Section 5.1.1*). Comme le montre le tableau 5.8, le choix d'autres mesures de similarité n'a que très peu d'impact sur les résultats finaux. Il nous semble que le seul moyen de répondre correctement aux cas problématiques des granularités partielles est de considérer des niveaux de signatures inférieurs à l'échelle du texte, ce que nous n'explorons pas dans le cadre de cette thèse. Nous faisons toutefois plusieurs propositions dans ce sens dans le chapitre 5.5.

Notre analyse a également montré que dans plusieurs cas le nombre de composés nominaux communs aurait pu être augmenté en tenant compte de leurs variations. La détection des variations des composés nominaux fait l'objet de nombreuses recherches étant donné leur utilité pour l'extraction terminologique. Une approche classique consiste à générer les différentes variations d'un composé afin d'en rechercher les correspondances (Daille, 2007). Sa mise en œuvre dans le cadre de nos travaux nous pose problème puisqu'une telle normalisation retirerait en partie la spécificité des choix de l'auteur ce qui pourrait augmenter le nombre de faux positifs. L'utilisation d'une mesure de similarité entre formes textuelles pourrait nous permettre de rapprocher certaines variations sans modifier lesdites formes. Un cas particulier de variation nécessiterait d'être considéré différemment. Il s'agit des variations liées aux modifications des référentiels spatio-temporels. Dans le cas des articles de presse, le référentiel spatio-temporel de l'article peut être repositionné lors du processus de dérivation et

Corpus	Mesure	MAP	Sép. Q	Corpus	Mesure	MAP	Sép. Q
Piithie	$c_{ci,ca}$	0,866	0,384	PANini	$c_{ci,ca}$	0,797	0,002
	$c_{ca,ci}$	0,879	0,6		$c_{ca,ci}$	0,789	0,023
	c_{max}	**0,889**	**0,64**		c_{max}	**0,793**	**0,027**
	c_{min}	0,852	0,342		c_{min}	0,798	0,002
	r	0,869	0,279		r	0,799	0,002
Wikinews	$c_{ci,ca}$	0,838	0,672				
	$c_{ca,ci}$	0,833	0,616				
	c_{max}	**0,831**	**0,746**				
	c_{min}	0,836	0,534				
	r	0,838	0,476				

TABLE 5.8 – Variation des résultats selon la mesure de similarité utilisée pour l'approche par composés nominaux.

Corpus	Approche	MAP	Sép. Q	Stockage
Piithie	Référence	0,999	0,708	100 %
	Entités nommées	0,839	0,628	87 %
	Comp. nominaux	0,889	0,640	90 %
Wikinews	Référence	0,872	0,800	100 %
	Entités nommées	0,646	0,833	61 %
	Comp. nominaux	0,831	0,746	67 %
PANini	Référence	0,823	0,024	100 %
	Entités nommées	0,774	0,036	2 %
	Comp. nominaux	0,793	0,027	3 %

TABLE 5.9 – Comparaison des maximums sélectionnés pour les différentes approches et les différents corpus.

entraîner par conséquent des variations dans l'énonciation des dates (*aujourd'hui* vs. *hier* vs. *mardi*) ou des lieux (*place de la Bastille* vs. *la capitale* vs. *Paris*).

5.3.3 Conclusion

Le tableau 5.9 fait la synthèse des résultats, en termes de MAP puis de Sép. Q, obtenus en exploitant les entités nommées et les composés nominaux. Les résultats sont inférieurs à ceux des approches de référence respectives de chaque corpus. Pour autant, elles peuvent être encore largement améliorées, notamment en tenant compte des différentes variations des objets étudiés.

Nous pouvons observer que l'approche par entités nommées n'est pas adaptée aux dérivations de type révisions pour lesquelles les composés nominaux donnent de meilleurs résultats. En ce qui concerne les dérivations représentées par Piithie et PANini, les deux approches sont comparables. Les résultats étant tout de même plus proches de ceux de l'approche de référence pour le corpus PANini alors qu'ils offrent le plus fort taux de compression des signatures.

5.4 Combinaison des approches

Nous avons observé lors de l'analyse des erreurs de l'approche par composés nominaux que ceux-ci ne capturaient pas toujours le sujet des textes notamment lorsque la thématique du texte repose principalement sur des entités nommées. Nous pensons que la combinaison de ces descripteurs permettrait de pallier leurs faiblesses respectives. D'une manière générale, nous souhaitons explorer les combinaisons des différents descripteurs comme piste d'amélioration des performances.

Une telle combinaison peut s'envisager de deux façons :
- la génération d'une seule signature contenant les différents descripteurs ;
- la combinaison arithmétique des scores de similarité obtenus par les approches individuelles.

La première solution qui consiste à opérer la combinaison lors de l'étape de modélisation peut être problématique lorsque le nombre d'éléments correspondant à chaque descripteur n'est pas comparable. La mesure de similarité donnera plus de poids aux descripteurs ramenant le plus d'éléments. La deuxième solution qui consiste à opérer la combinaison lors du calcul de similarité supprime ce déséquilibre potentiel. Dans ce cas, la similarité est calculée comme une fonction linéaire des scores de similarité des approches. Non seulement les différents descripteurs sont par défaut sur un pied d'égalité mais il nous est également possible de contrôler le poids de chacun.

Nous expérimentons ces deux formes de combinaison pour chacune de nos propositions : n-grammes rares, n-grammes de fort poids informatif, entités nommées et composés nominaux.

5.4.1 Combinaison des signatures

Soient les fonctions de projection d'un texte en sa signature :
- Π_H^n pour les n-grammes hapax, avec n la taille des n-grammes (*cf. Section 5.2.2*) ;
- $\Pi_R^{n,r}$ pour les n-grammes de fort poids informatif, avec n la taille des n-grammes et r la classe seuil (*cf. Section 5.2.3*) ;
- Π_E pour les entités nommées (*cf. Section 5.3.1*) ;
- Π_C pour les composés nominaux (*cf. Section 5.3.2*).

Nous explorons dans cette section la combinaison de ces approches lors de l'étape de modélisation, c-à-d l'union dans une même signature des éléments composant les signatures des approches combinées. Nous notons la combinaison comme une fonction de projection du produit des indices. Par exemple, $\Pi_{H \cdot R}$ correspond à la combinaison de Π_H et de Π_R, soit la signature $\Pi_H \cup \Pi_R$.

Nous expérimentons plus particulièrement la combinaison des approches linguistiques ($\Pi_{E \cdot C}$) puisque nous avons observé dans l'analyse de leurs erreurs que ces deux approches pouvaient être complémentaires. Nous nous intéressons également à la combinaison de l'approche par sélection des hapax avec chacune des approches linguistiques ($\Pi_{H \cdot E}^n$, $\Pi_{H \cdot C}^n$) ainsi qu'avec les trois réunies ($\Pi_{H \cdot E \cdot C}^n$) car nous pensons que ces descripteurs respectent le mieux la propriété de singularité. Enfin, la combinaison de l'approche par sélection des éléments de fort poids informatif nous semble complémentaire des approches linguistiques en ce que les entités nommées ($\Pi_{R \cdot E}^{n,r}$) particulièrement, mais également les composés nominaux ($\Pi_{R \cdot C}^{n,r}$) sont par nature représentatifs des thématiques des textes alors qu'ils ne sont pas forcément correctement capturés par les n-grammes. La combinaison des deux approches par sélection statistique n'est pas pertinente car les signatures résultantes seraient plus volumineuses que la signature complète.

Le tableau 5.10 fait la synthèse des résultats de ces différentes combinaisons. Bien que les expérimentations aient été menées en faisant varier la majorité des paramètres

	Approche	Norm.	MAP		Sép.Q		Stockage	
Piithie	Référence	oui	0,999		0,708		100 %	
	Π_{H}^{1}	non	0,989	↘	0,769	↗	90 %	↗
	$\Pi_{R}^{2,7}$	oui	0,999	=	0,572	↘	87 %	↗
	Π_{E}		0,839	↘	0,628	↘	87 %	↗
	Π_{C}		0,889	↘	0,640	↘	90 %	↗
	$\Pi_{E \cdot C}$		0,874	↘	0,663	↘	89 %	↗
	$\Pi_{H \cdot E}^{1}$	oui	0,999	=	0,778	↗	88 %	↗
	$\Pi_{H \cdot C}^{1}$	non	0,999	=	0,72	↗	91 %	↗
	$\Pi_{H \cdot E \cdot C}^{1}$	non	0,999	=	0,716	↗	92 %	↗
	$\Pi_{R \cdot E}^{1,10}$	non	0,999	=	0,659	↘	87 %	↗
	$\Pi_{R \cdot C}^{1,10}$	non	0,999	=	0,650	↘	90 %	↗
	$\Pi_{R \cdot E \cdot C}^{1,5}$	oui	0,999	=	0,664	↘	90 %	↗
Wikinews	Référence	oui	0,872	0,800	100 %			
	Π_{H}^{2}	non	0,856	↘	0,807	↗	87 %	↗
	$\Pi_{R}^{2,10}$	oui	0,880	↗	0,794	↘	89 %	↗
	Π_{E}		0,646	↘	0,833	↗	61 %	↗
	Π_{C}		0,831	↘	0,746	↘	67 %	↗
	$\Pi_{E \cdot C}$		0,836	↘	0,752	↘	71 %	↗
	$\Pi_{H \cdot E}^{2}$	oui	0,874	↗	0,803	↗	90 %	↗
	$\Pi_{H \cdot C}^{1}$	oui	0,850	↘	0,793	↘	67 %	↗
	$\Pi_{H \cdot E \cdot C}^{1}$	oui	0,855	↘	0,792	↘	68 %	↗
	$\Pi_{R \cdot E}^{1,5}$	non	0,858	↘	0,774	↘	67 %	↗
	$\Pi_{R \cdot C}^{2,5}$	oui	0,866	↘	0,769	↘	93 %	↗
	$\Pi_{R \cdot E \cdot C}^{1,5}$	oui	0,862	↘	0,739	↘	67 %	↗
PANini	Référence	oui	0,823	0,024	100 %			
	Π_{H}^{2}	non	0,834	↗	0,048	↗	18 %	↗
	$\Pi_{R}^{7,10}$	oui	0,819	↘	0,018	↘	103 %	↘
	Π_{E}		0,774	↘	0,036	↗	2 %	↗
	Π_{N}		0,793	↘	0,027	↗	3 %	↗
	$\Pi_{E \cdot C}$		0,803	↘	0,039	↗	4 %	↗
	$\Pi_{H \cdot E}^{2}$	non	0,828	↗	0,048	↗	19 %	↗
	$\Pi_{H \cdot C}^{1}$	oui	0,837	↗	0,054	↗	4 %	↗
	$\Pi_{H \cdot E \cdot C}^{1}$	non	0,829	↗	0,056	↗	6 %	↗
	$\Pi_{R \cdot E}^{2,10}$	oui	0,824	↗	0,012	↘	3 %	↗
	$\Pi_{R \cdot C}^{3,10}$	oui	0,818	↘	0,009	↘	12 %	↗
	$\Pi_{R \cdot E \cdot C}^{3,10}$	non	0,824	↗	0,013	↘	13 %	↗

TABLE 5.10 – Résultats des combinaisons des signatures : ↗ indique une amélioration des résultats par rapport à l'approche de référence, ↘ une dégradation et = des résultats équivalents.

pour chaque combinaison, nous n'y que les meilleurs résultats obtenus.

5.4.1.1 Corpus Piithie

Tout d'abord, nous pouvons observer que quelque soit la combinaison, elle ne se concrétise que par une légère augmentation du coût de stockage par rapport aux descripteurs considérés isolément. Ensuite, toutes les combinaisons expérimentées, à l'exception de $\Pi_{E \cdot C}$, permettent d'atteindre le niveau de qualité de la classification de l'approche de référence. Enfin, les combinaisons impliquant l'approche par sélection des hapax permettent également d'obtenir une meilleure capacité de discrimination que l'approche de référence. Ces combinaisons permettent donc de faire mieux que l'approche de référence pour un coût moins élevé.

5.4.1.2 Corpus Wikinews

La combinaison des méthodes n'entraîne qu'une légère augmentation du coût de stockage de la signature. Dans certains cas ($\Pi_{H \cdot C}$, $\Pi_{H \cdot E \cdot C}$, $\Pi_{R \cdot E}$, $\Pi_{R \cdot E \cdot C}$), le décalage du maximum vers une taille de n-grammes inférieure permet même de réduire le coût de la signature. Toutefois, contrairement à Piithie, une seule combinaison ($\Pi^2_{H \cdot E}$) permet de faire mieux que l'approche de référence en termes de qualité de classification et de capacité de discrimination. Cette meilleure combinaison implique les hapax, comme pour la meilleure combinaison pour Piithie. Alors que l'approche individuelle par représentativité permettait d'obtenir une MAP supérieure à l'approche de référence, aucune combinaison impliquant cette approche ne permet de maintenir ce résultat.

5.4.1.3 Corpus PANini

Tout comme pour les corpus précédents, la combinaison des descripteurs n'entraîne pas de forte augmentation du coût de stockage de la signature. De plus, comme nous l'avons observé sur Wikinews, le glissement vers des n-grammes de petites tailles permet même de diminuer les coûts par rapport aux approches individuelles. La combinaison avec l'approche hapax donne de meilleurs résultats que l'approche de référence mais ne permet pas forcément de faire mieux que les hapax seuls. Seule la combinaison avec les composés nominaux permet de dépasser le score individuel obtenu avec les hapax. Enfin, la combinaison des descripteurs linguistiques permet d'augmenter à la fois la qualité de la classification et la capacité de discrimination, ce que nous n'avons pas observé sur les autres corpus.

5.4.1.4 Synthèse

En résumé, la combinaison des descripteurs lors de l'étape de modélisation permet dans certaines configurations de faire mieux qu'en exploitant les descripteurs isolément et parfois de dépasser les résultats obtenus avec l'approche de référence. Ainsi, les configurations qui permettent d'obtenir les plus hauts scores en termes de qualité de classification et de capacité de discrimination sont :

- $\Pi^1_{H \cdot E}$ pour Piithie ;
- $\Pi^2_{H \cdot E}$ pour Wikinews, même si $\Pi^{2,10}_{R}$ permet d'obtenir une MAP plus élevée ;
- $\Pi^1_{H \cdot C}$ pour PANini.

Les approches Π^2_H et Π_E sont remarquablement complémentaires pour Wikinews alors que Π_E donnait de loin les plus mauvaises performances.

Nous retiendrons également que l'ajout de descripteurs non complémentaires entraîne une dégradation des résultats. Ainsi, l'ajout des deux types de descripteurs

linguistiques à l'approche par sélection des n-grammes hapax donne de moins bon résultats que l'ajout d'un seul. Il est donc préférable de sélectionner les descripteurs les plus appropriés selon le type de dérivation plutôt que tous les combiner.

5.4.2 Combinaison des scores de similarité

L'opération de combinaison lors de l'étape du calcul de similarité plutôt que de la modélisation permet de contrôler le poids de chaque descripteur. Nous définissons une fonction linéaire permettant de combiner les scores de similarité obtenus avec nos différentes descripteurs :

Soient t_1 et t_2 les textes considérés,

Soient sim_H, sim_R, sim_E et sim_C les mesures de similarité associées à chaque descripteur,

$\text{sim}_{a,b,c,d}$ est la fonction de combinaison de ces mesures de similarité telle que :

$$\text{sim}_{a,b,c,d}(t_1,t_2) = a.\,\text{sim}_H(t_1,t_2)+b.\,\text{sim}_R(t_1,t_2)+c.\,\text{sim}_E(t_1,t_2)+d.\,\text{sim}_N(t_1,t_2)$$

Nous nous intéressons aux mêmes combinaisons que précédemment : les deux approches linguistiques, l'approche par sélection des n-grammes rares avec les deux approches linguistiques et l'approche par sélection des n-grammes de fort poids informatif avec les deux approches linguistiques. L'exploration de la totalité des combinaisons n'est pas envisageable dans le cas présent car en plus de la totalité des paramètres habituels (taille des n-grammes, normalisation, classe seuil) s'ajoutent les coefficients de la fonction linéaire (a, b, c et d). Nous faisons donc le choix de nous limiter aux meilleures configurations de chaque approche. Soit :

- pour Piithie : $\text{sim}_H \equiv \Pi_H^1$ sans normalisation et $\text{sim}_R \equiv \Pi_R^{2,7}$ avec normalisation ;
- pour Wikinews : $\text{sim}_H \equiv \Pi_H^2$ avec normalisation et $\text{sim}_R \equiv \Pi_R^{2,10}$ avec normalisation ;
- pour PANini : $\text{sim}_H \equiv \Pi_H^2$ sans normalisation et $\text{sim}_R \equiv \Pi_R^{7,10}$ avec normalisation ;

Le choix des scores de similarité maximum est une hypothèse raisonnable étant donné que la combinaison s'opère après le calcul de similarité et qu'elle est linéaire.

La combinaison algébrique des scores de similarité nécessite de porter une attention particulière sur les signatures vides. Un certain nombre de textes de Piithie et Wikinews sont dépourvus d'entités nommées ou de composés nominaux ce qui résulte en une division par zéro et un score *NaN*. Nous remplaçons ces scores par 0 afin de les combiner algébriquement.

Le tableau 5.11 présente les meilleurs résultats [20] obtenus en termes de MAP et de Sép. Q pour chaque combinaison expérimentée en regard des résultats de l'approche de référence et des approches individuelles utilisées. Nous y rapportons les combinaisons des deux approches linguistiques ainsi que les combinaisons de chaque approche par sélection statistique avec chaque approche linguistique puis les deux ensembles.

5.4.2.1 Corpus Piithie

Toutes les combinaisons qui impliquent une approche par sélection statistique et l'approche par composés nominaux permettent d'atteindre les performances de

20. La combinaison algébrique entraîne mécaniquement une amélioration de la capacité de discrimination. La séparation des quartiles est en théorie égale à la somme pondérée des séparations des quartiles.

		sim_H	sim_R	sim_E	sim_C	MAP		Sép.Q	
Piithie	Référence					0,999		0,708	
	sim_H					0,989	↘	0,769	↗
	sim_R					0,999	=	0,572	↘
	sim_E					0,839	↘	0,628	↘
	sim_C					0,889	↘	0,640	↘
	$\text{sim}_{0,0,1,4}$	0	0	1	4	0,956	↘	3,16	↗
	$\text{sim}_{1,0,0,1}$	1	0	0	1	0,999	=	1,40	↗
	$\text{sim}_{1,0,1,0}$	1	0	1	0	0,998	↘	1,38	↗
	$\text{sim}_{1,0,1,1}$	1	0	1	1	0,999	=	1,98	↗
	$\text{sim}_{0,1,0,1}$	0	1	0	1	0,999	=	1,21	↗
	$\text{sim}_{0,3,1,0}$	0	3	1	0	0,997	↘	2,31	↗
	$\text{sim}_{0,2,1,1}$	0	2	1	1	0,999	=	2,40	↗
Wikinews	Référence					0,872		0,800	
	sim_H					0,856	↘	0,807	↗
	sim_R					0,880	↗	0,794	↘
	sim_E					0,646	↘	0,833	↗
	sim_C					0,831	↘	0,746	↘
	$\text{sim}_{0,0,1,3}$	0	0	1	3	0,848	↘	2,75	↗
	$\text{sim}_{1,0,0,2}$	1	0	0	2	0,862	↘	2,31	↗
	$\text{sim}_{1,0,3,0}$	1	0	3	0	0,864	↘	2,57	↗
	$\text{sim}_{2,0,2,2}$	2	0	2	2	0,866	↘	3,83	↗
	$\text{sim}_{0,3,0,1}$	0	3	0	1	0,860	↘	3,10	↗
	$\text{sim}_{0,3,3,0}$	0	3	3	0	0,878	↗	4,08	↗
	$\text{sim}_{0,1,1,2}$	0	1	1	2	0,864	↘	2,75	↗
PANini	Référence					0,823		0,024	
	sim_H					0,834	↗	0,048	↗
	sim_R					0,819	↘	0,018	↘
	sim_E					0,774	↘	0,036	↗
	sim_C					0,793	↘	0,027	↗
	$\text{sim}_{0,0,2,4}$	0	0	2	4	0,811	↘	0,287	↗
	$\text{sim}_{1,0,1,0}$	1	0	1	0	0,839	↗	0,115	↗
	$\text{sim}_{2,0,0,1}$	2	0	0	1	0,843	↗	0,137	↗
	$\text{sim}_{3,0,2,1}$	3	0	2	1	0,833	↗	0,338	↗
	$\text{sim}_{0,3,0,2}$	0	3	0	2	0,817	↘	0,154	↗
	$\text{sim}_{0,2,1,0}$	0	2	1	0	0,820	↘	0,115	↗
	$\text{sim}_{0,3,1,3}$	0	3	1	3	0,823	=	0,264	↗

TABLE 5.11 – Résultats des combinaisons des scores de similarité : ↗ indique une amélioration des résultats par rapport à l'approche de référence, ↘ une dégradation et = des résultats équivalents.

l'approche de référence. La combinaison algébrique permet donc effectivement d'améliorer les résultats puisque seuls les n-grammes de fort poids informatif permettent d'atteindre ces résultats individuellement.

La combinaison des deux approches linguistiques résulte en une amélioration de pratiquement 0,1 point par rapport aux résultats individuels. Cette performance nous conforte dans l'idée de la complémentarité des composés nominaux et des entités nommées d'une part, et dans la nécessité d'une approche multidimensionnelle pour la détection de dérivation d'autre part.

La combinaison de l'approche par représentativité et de l'approche par entités nommées donne de moins bons résultats que l'approche par représentativité seule. Une combinaison infructueuse peut donc entraîner une dégradation des résultats. La sélection parcimonieuse des approches combinées est préférable à la combinaison de toutes les approches.

5.4.2.2 Corpus Wikinews

Les performances des approches individuelles sur le corpus Wikinews sont comparables à celles obtenues sur le corpus Piithie : seule l'approche par représentativité permet d'obtenir une meilleure MAP que l'approche de référence. En comparaison, les combinaisons donnent de moins bonnes performances puisqu'aucune, à l'exception de celle impliquant les n-grammes de fort poids informatif et les entités nommées, ne permet d'égaler les résultats de l'approche de référence.

Les combinaisons des approches linguistiques seules ou avec les hapax permettent de faire mieux que les approches individuelles correspondantes. À l'opposé, toutes les combinaisons impliquant les n-grammes de fort poids informatif donnent de moins bons résultats que l'approche individuelle.

5.4.2.3 Corpus PANini

Les combinaisons impliquant les hapax sont les seules à dépasser les résultats de l'approche de référence. La combinaison impliquant les n-grammes de fort poids informatif et les deux approches linguistiques permettent toutefois d'égaler ces résultats.

Si les trois combinaisons impliquant les hapax dépassent l'approche de référence, elles n'entraînent pas automatiquement une augmentation de la MAP par rapport à l'approche hapax seule. La combinaison des hapax et des deux approches linguistiques résulte même en une dégradation. À l'opposé, la combinaison des hapax et des composés nominaux seuls est profitable avec une amélioration de presque 0,01 points.

Les combinaisons restantes permettent généralement d'améliorer les performances par rapport aux approches individuelles. Nous notons encore une fois la bonne complémentarité des approches linguistiques puisque leur combinaison permet d'augmenter la MAP de 0,02 à 0,04 points par rapport à leurs scores individuels.

5.4.3 Conclusion

Nous avons souhaité combiner les différents types de descripteurs qui ont été expérimentés individuellement dans l'hypothèse que les qualités des uns pourraient pallier les défauts des autres. L'objectif étant bien entendu de faire émerger des combinaisons permettant d'obtenir des résultats meilleurs que ceux des approches individuelles.

Nous avons proposé deux méthodes de combinaison. La première prend place lors du processus de modélisation des textes. Elle consiste à construire une signature correspondant à l'union des signatures obtenues pour chacun des descripteurs. La seconde prend place lors du calcul du score de similarité. Elle consiste à combiner les

scores obtenus individuellement par chaque descripteur à l'aide d'une fonction linéaire simple.

Les deux méthodes offrent des combinaisons permettant de faire mieux que les approches individuelles, et dans de plus rares cas que l'approche de référence. La première méthode a permis les plus fortes progressions. Nous pensions que l'inverse se produirait puisque la combinaison linéaire offre un meilleur contrôle du poids de chaque approche. Nous relativisons toutefois du fait que pour la première méthode les paramètres de taille et de normalisation ont été explorés contrairement à la seconde méthode pour laquelle nous avions fixé les approches individuelles à combiner. Ainsi, l'espace de recherche des combinaisons étant plus restreint pour la combinaison algébrique, nous pensons qu'elle offre une marge de progression plus importante.

L'expérience nous confirme que nos deux approches linguistiques sont complémentaires. Leur combinaison, par l'une ou l'autre méthode, résulte en une MAP supérieure aux résultats obtenus individuellement. Seule la combinaison selon la première méthode pour Piithie déroge à cette règle. Nous constatons également que ces approches linguistiques peuvent complèter positivement les approches par sélection statistique sous réserve de les accorder correctement. En effet, les combinaisons inadéquates dégradent les résultats. Le choix parcimonieux des approches à combiner est donc préférable à une combinaison aveugle de toutes les approches.

En conclusion, les bons résultats obtenus pour plusieurs combinaison soutiennent notre idée selon laquelle la dérivation ne peut être capturée correctement qu'en l'abordant par plusieurs dimensions. Nous notons toutefois qu'il faut être attentif à la complémentarité des dimensions utilisées et éviter leur mise en concurrence sous peine d'une dégradation des résultats.

5.5 Conclusion

Nous avons exploré la détection extrinsèque de dérivation en exploitant des descripteurs aux différents degrés de singularité et d'invariance. Nous avons fait l'hypothèse que la réduction du nombre d'éléments des signatures permettraient de maintenir les résultats au niveau de ceux de la signature complète à l'aide des propriétés de singularité et d'invariance. Notre intuition était double. Premièrement, les éléments qui sont très spécifiques à un texte (propriété de singularité) ne sont éventuellement présents que dans les reprises de ce texte alors qu'ils sont virtuellement absents des autres textes. Deuxièmement, le processus de dérivation conserve forcément des éléments du texte source (propriété d'invariance) qui constituent les traces de la dérivation et sur lesquels il faut se focaliser.

Nous avons expérimenté quatre descripteurs sur trois corpus (Piithie, Wikinews et PANini) : n-grammes rares, n-grammes de plus fort poids informatif, entités nommées et composés nominaux ainsi que leurs combinaisons. Nous avons globalement validé notre hypothèse puisque, comme l'illustrent les tableaux 5.4 et 5.11, les signatures de taille réduite expérimentées permettent de maintenir la qualité des résultats au niveau des approches de référence et éventuellement de faire mieux. Nous avons notamment observé que les n-grammes de petites tailles (notamment les bigrammes) semblent les mieux appropriés à la tâche de détection de dérivation. Ils donnent en effet de bons résultats en termes de qualité de classification ainsi que de capacité de discrimination. Pour autant nous avons identifié plusieurs pistes d'amélioration.

Premièrement, nos approches ne semblent pas aptes à appréhender efficacement les dérivations à granularité partielle représentées par le corpus PANini. Si les résultats en terme de qualité de classification (MAP) sont comparables à ceux obtenus pour Piithie et Wikinews, les résultats en termes de capacité de discrimination (Sép. Q)

sont bien en deçà. De plus, les meilleurs résultats sont majoritairement obtenus avec des n-grammes de grande taille (6-grammes ou 7-grammes). Nous avions supposé que cela était dû à l'absence de modifications importantes lors du processus de dérivation mais les résultats obtenus avec les bigrammes hapax (*cf. Section 5.2.2*) semblent aller à l'encontre de cette hypothèse. La détection des dérivation de granularité partielle nécessite très certainement de travailler à une échelle de modélisation inférieure au texte (passages thématiques, paragraphes...).

Deuxièmement, les résultats rapportées dans le tableau 5.9 (p. 159) montrent que l'exploitation des descripteurs linguistiques offre en général des résultats inférieurs aux autres méthodes (approche de référence ou descripteurs statistiques). Les composés nominaux permettent d'approcher les résultats des autres approches mais pas les entités nommées. Nous retenons notamment que l'exploitation des seules entités nommées n'est pas pertinente pour l'identification des révisions puisqu'elles offrent de loin la plus mauvaise des performances pour le corpus Wikinews. Malgré cela, nous maintenons que les entités nommées et les composés nominaux constituent des descripteurs singuliers. Toutefois, si nous pensons que les concepts et les entités sont invariants, leur forme textuelle ne l'est pas forcément. La prise en compte de leurs variations pourrait permettre d'améliorer les résultats obtenus avec ces descripteurs.

Troisièmement, la détection des liens de dérivation passe par la prise en compte du caractère multidimensionnel de la dérivation. Les bons résultats obtenus par combinaison des descripteurs soutiennent cette proposition. Elle s'illustre particulièrement pour la combinaison des approches linguistiques qui offre des résultats supérieurs aux résultats obtenus individuellement par chacune. Nous constatons également que ces approches linguistiques peuvent complèter positivement les approches par sélection statistique sous réserve de les accorder correctement. L'évaluation de la complémentarité des dimensions combinées permettra d'éviter leur mise en concurrence et par conséquent les combinaisons inadéquates qui dégradent les résultats.

Finalement, nous nous interrogeons sur la propriété d'invariance des descripteurs. Si nous avons été en mesure de caractériser statistiquement la singularité, la propriété d'invariance nous échappe. Nous avons l'intuition que les éléments qui fondent le texte seront conservés par les processus de dérivation. Nous pourrions profiter des travaux en résumé automatique qui cherchent à pondérer les éléments selon leur importance dans le texte. Il nous semble toutefois nécessaire de ne pas faire reposer les comparaisons sur la forme textuelle de ces éléments mais plus vraisemblablement à un niveau sémantique. Une telle caractérisation ne serait de toute façon que partielle puisque le choix final de l'auteur de la dérivation prévaut et il ne nous est pas accessible.

Conclusion générale

Nearly every man who develops an idea works at it up to the point where it looks impossible, and then gets discouraged. That's not the place to become discouraged.

— Thomas Edison

L'étude qui vient d'être présentée porte sur la question de la détection des dérivations de texte. L'essor de la diffusion numérique en masse des documents (livres, articles de presse, rapports scientifiques, billets d'humeurs, interactions sociales...) offre la possibilité de facilement recopier et remodeler à l'envie des textes préexistant afin d'en produire de nouveaux. Si la démarche n'est pas forcément condamnable et stimule même la créativité, il est nécessaire de mettre en place des outils permettant de suivre ces dérivations afin de protéger les intérêts des auteurs ou de filtrer les informations redondantes. L'objectif principal de cette thèse était d'explorer des méthodes de détection de ces dérivations qui aient un coût opératoire plus faible que l'approche par signature complète tout en produisant des résultats de qualité similaire.

Nous présentons dans la suite de ce chapitre un bilan de chacune des différentes parties abordées dans ce document puis les perspectives de ce travail.

1 Bilan

Notre travail de thèse s'est divisé en trois grandes parties dont nous faisons le bilan ci-après : (i) l'unification des problématiques de la littérature autour de la notion de dérivation de texte, (ii) la détection intrinsèque des citations dans la presse francophone et (iii) la détection extrinsèque des dérivations à l'aide de signatures de taille réduite.

1.1 La dérivation de texte

L'état de l'art des travaux autour de la détection de reprise (*cf. Section 1.1*) a montré que ces travaux s'articulaient autour d'un même socle commun. Duplications, versions, résumés, plagiats, citations, transpositions de genre et traductions sont des processus de production d'un texte, ou d'une partie d'un texte, à partir d'une œuvre préexistante. Nous avons nommé ce processus de production particulier la dérivation de texte. Le processus de dérivation permet de créer un texte dérivé à partir d'un texte source de sorte que l'obtention du premier n'est possible qu'à la condition de la préexistance du second (*cf. Définition 1*). Cette généralisation du problème permet de désenclaver les recherches de la détection de plagiat et d'envisager de nouvelles mises en œuvre applicatives (filtrage des informations redondantes, agglomération de révisions, suivi d'une information...).

Nous avons ensuite cherché à identifier ce qui différencie les différentes formes de dérivation qui ont été étudiées dans la littérature. En nous appuyant sur des classifications existantes, nous avons abouti à un modèle multidimensionnel permettant de caractériser ces formes (*cf. Section 1.4*). Ce modèle dévoile toute la complexité de la dérivation en faisant apparaître un nombre important de variations du processus [21].

Ce modèle est une première ébauche qui nécessitera certainement des adaptations. Certaines dimensions telles que l'intention ou la similarité méritent notamment d'être affinées. Toutefois, il a le mérite de ne pas uniquement reposer sur les objets observés mais d'être déduit du concept de dérivation. Il devrait permettre de compiler des ressources expérimentales plus nombreuses tout en étant mieux cadrées théoriquement, ou bien guider la synthèse automatique de dérivation telle que réalisée pour les compétitions PAN.

1.2 Détection intrinsèque des citations

Les citations sont une forme de dérivation particulière, notamment de par le nombre de marques laissées par l'auteur dans le texte dérivé pour les identifier. Nous avons développé une méthode de détection intrinsèque pour les détecter qui repose sur l'apprentissage artificiel. L'apprentissage offre une certaine portabilité de la méthode à d'autres langues ou d'autres genres de texte en déduisant automatiquement les règles de combinaison des marques à partir d'un corpus annoté. Nous nous sommes restreint dans le cadre de cette thèse au français et à l'anglais pour le genre des articles de presse en ligne.

Nous avons fait le choix de diviser le problème de catégorisation des citations en deux sous-problèmes : la catégorisation des composants source contenant une entité nommée et la catégorisation des composants discours rapporté contenant un passage entre guillemets (*cf. Section 3.2*). Les classifieurs obtenus donnent des résultats variables bien que meilleurs en précision que l'approche naïve consistant à considérer tous les candidats (tous les segments entre guillemets ou toutes les entités nommées) comme des composants. Ils permettent d'identifier le premier composant avec une précision de 60 % pour le français et 70 % pour l'anglais, et le second avec une précision de 90 % pour le français et de 85 % pour l'anglais. L'évaluation des traits les plus pertinents pour chaque catégorisation montre que les marques retenues ne sont pas assez discriminantes pour notre tâche.

1.3 Détection extrinsèque des dérivations

Nous nous sommes intéressés dans cette thèse à la recherche des dérivés d'un texte source identifié parmi une collection fermée de textes. Les méthodes proposées dans la littérature n'exploitent qu'une seule dimension du texte : les méthodes par couverture de texte et par alignement ne considèrent que la forme tandis que les méthodes par modélisation vectorielle ne considèrent que le contenu général. Notre modèle montre que la dérivation porte sur des éléments de diverses natures (forme, contenu, structure, style). Nous avons voulu explorer l'exploitation de ces différentes dimensions. De plus, plutôt que chercher à classer en dérivé/non-dérivé comme le font ces méthodes, nous avons privilégié un classement des paires de texte selon la probabilité de l'existence d'un lien de dérivation entre eux.

Nous avons proposé deux propriétés permettant de sélectionner les descripteurs les plus pertinents pour notre tâche (*cf. Section 5.1*). La singularité caractérise la probabilité de trouver des éléments correspondants au descripteur dans d'autres textes.

21. Le produit cartésien des valeurs associées à chaque dimension du modèle permet de se donner une idée de ce nombre de variations même s'il ne s'agit que d'une estimation grossière.

L'invariance caractérise la probabilité que les éléments correspondants au descripteur soient conservés lors d'un processus de dérivation. Nous avons fait l'hypothèse que ces propriétés de singularité et d'invariance devraient permettre de maintenir des résultats comparables à ceux obtenus avec la signature complète mais en étant composé d'un nombre inférieur d'éléments. Nous avons identifié quatre descripteurs présentant ces propriétés à différents degrés : les n-grammes rares, les n-grammes de fort poids informatif, les entités nommées et les composés nominaux. Les deux premiers caractérisent la dimension textuelle du texte et éventuellement la dimension thématique (*cf. Section 5.2*). Les deux autres caractérisent le contenu (*cf. Section 5.3*). Nous avons également cherché à combiner ces différentes dimensions (*cf. Section 5.4*).

Concernant l'évaluation des méthodes expérimentées, nous nous sommes tournés vers un classement des paires de texte selon leur score de similarité à partir duquel nous avons pu estimer la qualité de la classification et la capacité de discrimination des méthodes à l'aide respectivement des mesures de MAP et de Sép. Q. Nous avons complété cette évaluation par l'estimation du coût opératoire qui repose principalement sur la taille des signatures obtenus. Les expérimentations ont été menées sur trois corpus : (i) Piithie pour les dérivations opérées entre les dépêches et les articles de journaux en français et sur le Web, (ii) Wikinews pour les dérivations de type révisions entre des versions d'articles de presse publiées en ligne et PANini pour les plagiats générés artificiellement.

Le tableau ci-dessous fait la synthèse des différentes caractéristiques que nous avons explorées expérimentalement :

Caractéristique théorique explorée	Mise en œuvre expérimentale
poids informatif / thématique	$\mathrm{tf} \cdot \mathrm{idf}$
spécificité / rareté	df
entités	entités nommées
concepts	composés nominaux
qualité de la classification	MAP
capacité de discrimination	Sép. Q
fr articles de presse	Corpus Piithie
fr révisions	Corpus Wikinews
en plagiats littéraire	Corpus PANini

Nous avons globalement validé notre hypothèse de la capacité des descripteurs singuliers et invariants à identifier les liens de dérivation avec la même qualité que l'approche par signature complète. Dans certains cas ils ont même permis de faire mieux. Nous avons également montré que la combinaison de plusieurs dimensions du texte pouvait être profitable à la détection de dérivation lorsque les descripteurs incriminés se complétaient. Toutefois cette combinaison pouvait entraîner une dégradation des résultats si ces mêmes descripteurs entraient en concurrence.

2 Perspectives

À l'instar des problématiques populaires telles que la terminologie, la traduction, l'extraction de connaissance ou l'analyse syntaxique, la détection de dérivation de texte n'est qu'une composante du problème organique de compréhension du langage. Nous avons cherché à faire avancer les connaissances concernant cette composante comme l'illustre la section précédente. Ce travail de thèse s'ouvre sur un large éventail

de perspectives.

2.1 Détection intrinsèque des citations

Notre proposition n'a pas donné de résultats probants, néanmoins une méthode de détection des citations par apprentissage reste envisageable. Une telle méthode nécessiterait de remettre en cause notre formalisation du problème qui consiste en la catégorisation de certains objets linguistiques (entités nommées et passages entre guillemets) comme des composants constitutifs de la citation (source et discours rapporté respectivement).

Nous avons choisi d'identifier indépendamment les composants source et les composants discours rapporté. L'expérience nous montre que ces derniers sont assez aisément identifiables même en considérant simplement les passages entre guillemets à l'inverse des premiers. L'identification des composants source pourrait certainement profiter de l'identification *a priori* des composants discours rapporté qui auraient alors un rôle de marque.

Plus généralement, notre approche revient à identifier les éléments les plus spécifiques (les composants constitutifs) en premier lieu avant de remonter aux éléments les plus généraux (les citations). Nous envisageons d'explorer l'approche inverse : repérage des citations puis recherche en leur sein des composants constitutifs.

2.2 Détection extrinsèque des dérivations

Nous voyons de nombreuses perspectives à notre travail pour la détection extrinsèque des dérivations, notamment en ce qui concerne l'invariance des descripteurs, leur complémentarité, les problématiques particulières de dérivations à granularité partielle et de dérivation multilingue et enfin sur la méthodologie même de notre démarche.

2.2.1 Invariance des descripteurs

L'exploitation des descripteurs linguistiques donne de moins bons résultats en termes de qualité de classification par rapport aux autres méthodes (approche de référence ou descripteurs statistiques). La propriété d'invariance de ces descripteurs semble être la principale cause de cette contre-performance. Nous maintenons que les concepts et les entités sont potentiellement invariants notamment lorsqu'ils participent à l'expression de la thématique. Leur réalisation textuelle, composés nominaux et entités nommées dans notre cas, ne l'est pas forcément. Le problème est double. Peut-on caractériser la propriété d'invariance d'un descripteur comme nous l'avons fait pour la singularité ? Comment prendre en compte les variations des concepts et des entités ?

Nous nous interrogeons sur la propriété d'invariance des descripteurs. Nous avons émis l'hypothèse selon laquelle les éléments qui fondent le texte sont conservés. Il faudrait toutefois pouvoir identifier ces éléments et considérer la forme dans laquelle ils sont conservés. Nos expérimentations montrent qu'il est préférable de se détacher de la dimension textuelle qui a tendance à être altérée pour viser plutôt des éléments de contenu (les idées, les informations. . .) lorsque l'invariance des descripteurs repose sur ledit contenu. N'importe quelle caractérisation de l'invariance restera partielle puisque le choix final de l'auteur de la dérivation prévaut et il ne nous est pas accessible.

En parallèle de la caractérisation de la propriété d'invariance, il est primordial d'arriver à traiter dans nos signatures les variations des éléments linguistiques. Les travaux spécifiques à l'extraction des entités nommées et des composés nominaux

proposent des solutions à explorer. Nous pourrions également nous inspirer des travaux sur la recherche de formes canonique des phrases (Chandrasekar et collab., 1996).

2.2.2 Complémentarité des descripteurs

La prise en compte du caractère multidimensionel de la dérivation est profitable à la détection comme le montrent les résultats obtenus par combinaison des descripteurs (*cf. Section 5.4*). L'amélioration des résultats par combinaison est toutefois soumise à la condition de la complémentarité des descripteurs. Ainsi, les descripteurs statistiques s'accordent généralement assez bien avec les descripteurs linguistiques mais les combinaisons inadéquates dégradent les résultats.

Une analyse approfondie des résultats des combinaisons qui fonctionnent et de celles qui ne fonctionnent pas pourraient nous donner des pistes pour déterminer pourquoi certains descripteurs se combinent bien et d'autres non. La réponse est certainement à chercher dans un premier temps dans le recouvrement textuel ou sémantique des éléments capturés par ces différents descripteurs. Le rôle de leur coordination locale (au sein d'une même fenêtre de texte) est également à étudier.

2.2.3 Dérivations à granularité partielle

Nos expérimentations ont montré les limites de l'approche par signature pour détecter les dérivations à granularité partielle. Les résultats en termes de capacité de discrimination (Sép. Q) sont notamment très inférieurs à ceux obtenus sur les corpus Piithie et Wikinews qui n'exposent pas ce type de granularité. La détection des dérivations à granularité partielle nécessite de remettre potentiellement en cause notre cadre général de travail basé sur les approches par signature à l'échelle du document.

Une première piste de travail est de passer à une échelle de modélisation inférieure et de modéliser les différentes parties du texte. Nous envisageons plusieurs solutions pour ce changement d'échelle. La plus évidente est le *découpage visuel* en phrases ou paragraphes par exemple. Un tel découpage ne repose cependant sur aucun argument théorique. Une autre solution serait d'opérer un découpage hiérachique selon la structure discursive du texte. Chaque feuille de l'arbre, qui correspondrait à des sections, correspond à une signature. Chaque nœud non feuille a pour signature l'union des signatures des feuilles auxquelles il mène. La comparaison peut ainsi s'effectuer à différents niveaux de l'arbre selon les besoins. Nous pourrions pour ce découpage tirer parti des travaux d'extraction de patrons de texte (Filatova et collab., 2006). Une dernière solution serait d'opérer un découpage thématique du texte. Le découpage par cohésion lexicale est une piste à explorer (Hearst, 1997). Il permettrait de ne générer qu'une signature par thématique du document. L'hypothèse que la reprise partielle d'un texte se concentre sur une thématique particulière de ce texte est tout à fait probable. La faiblesse principale du passage à une échelle de modélisation inférieure au texte est le coût que représentent les comparaisons de chacune des signatures. Étant donné deux textes découpés en n segments, le test de dérivation nécessitera n^2 comparaisons de signatures.

Une deuxième piste de travail est de considérer des modélisations plus représentatives du texte qu'un simple ensemble. La prise en compte des cooccurrences des éléments manipulés ou des relations discursives sont à explorer. La prise en compte dans le modèle de nos signatures de ces données structurelles permettrait de mieux refléter la complexité de l'organisation du texte. Ainsi, nos expérimentations ont montré que les bigrammes mots sont particulièrement adaptés pour la détection de dérivation. Cependant nous ne sommes pas certains des caractéristiques des bigrammes responsables de ces bons résultats. Nous pouvons nous demander si la simple collocation ne

suffirait pas à expliquer cette bonne performance. Si telle était le cas nous pourrions envisager de relâcher la contrainte d'ordre des bigrammes et observer l'effet sur les résultats.

Une dernière piste qui nous semble la plus prometteuse est de procéder par projection d'éléments extrêmement singuliers et d'extraire les passages où ces éléments sont denses dans le texte. Certains éléments sont suffisamment singuliers pour discriminer par leur seul présence une dérivation. C'est le cas des n-grammes de grande taille, de n-grammes hapax ou plus intéressant de certaines entités numéraires ou temporelles. À l'aide d'une mesure de densité il serait envisageable d'isoler des passages d'un texte où ces éléments se concentrent. Notre hypothèse est que ces dits passages sont dérivés. Il serait alors envisageable d'identifier des dérivations partielles sans modéliser l'intégralité des textes, ce qui est inefficace, ni sans devoir calculer des signatures à différentes échelles du texte.

2.2.4 Dérivations multilingues

La dérivation multilingue, si elle est hors de propos dans le cadre de cette thèse, pourrait en être une extension.

Outre les très récents travaux de Potthast et collab. (2010a) et de Barrón-Cedeño et collab. (2010), aucune étude n'a porté sur la détection de dérivation multilingue alors que la tâche a été proposée lors de la campagne d'évaluation PAN'09 (Potthast et collab., 2009). Nous voyons deux grandes approches pour cette tâche : (i) la traduction automatique du texte dérivé dans la langue du texte source ou (ii) la traduction partielle des signatures dans une langue commune. La première solution serait certainement la plus efficace si nous étions capables de produire des traductions correctes et fidèles de manière automatique. Si les récentes percées en traduction automatique permettent d'envisager atteindre le premier objectif, le second semble encore hors de portée. En ce qui concerne la seconde approche, nous pourrions nous inspirer des méthodes utilisées en extraction de lexiques (Prochasson, 2009) qui reposent sur l'approche par traduction directe (Fung, 1998), ou encore les approches par alignement qui reposent sur des cognats (Covington, 1996).

2.2.5 Méthodologie

Dans ce travail de thèse, l'effort a été également réparti entre modélisation et expérimentation. Pour autant, certains points méthodologiques pourraient être précisés. Ainsi, nous prétendons faire de la détection de dérivation, soit identifier les textes écrits à partir d'autres, ou du moins estimer une probabilité concernant cette identification. Les résultats de nos expérimentations montrent que nous y arrivons assez bien en moyenne. Mais nos méthodes identifient-elles réellement un lien de dérivation ou bien ne capturraient-elles pas une autre forme de relation entre les textes, concordante avec un lien de dérivation mais sans l'être réellement. Par exemple, n'identifieraient-elles pas une forme de relation thématique ?

Notre évaluation revient en effet à distinguer des liens de dérivation de textes aléatoires, simplement homogènes en genre. Dans ce cas, l'identification de lien thématique n'est-elle pas une réduction suffisante du problème qui permet de mimer l'identification d'un lien de dérivation ? Nos approches pourraient capturer de tels liens puisque les entités nommées et les composés nominaux sont des porteurs de contenus et par conséquent caractérisent la thématique du texte. Certes la corrélation des résultats avec l'approche par recouvrement de n-grammes nous laisse penser que des liens de dérivations sont bel et bien identifiés. Pour autant il nous semble nécessaire d'étendre ces travaux en considérant la compilation de textes dérivés et non-dérivés mais thématiquement homogènes. Une telle compilation est complexe et coûteuse mais nécessaire

selon nous pour éliminer ce doute et s'assurer d'une capture correcte des phénomènes qui nous intéressent.

Il serait intéressant de constituer un corpus contenant des articles traitant d'une seule thématique et d'y expérimenter nos différentes méthodes. Étant donné que les différents articles seraient thématiquement homogènes, nous pourrions vérifier que nos méthodes détectent bien des liens de dérivation et non des liens thématiques. L'expérience de la construction du corpus Piithie nous montre toutefois que la constitution de ce genre de corpus est très difficile.

Annexe A

Annotation du corpus des citations

Une première étape d'annotation expliquée précédemment a consisté à structurer logiquement les articles constituant le corpus. Nous avons ensuite prélevé au sein de ces articles des phrases dont nous considérions qu'elles contenaient des citations. Afin de permettre l'utilisation et la diffusion du corpus, il a fallu intégrer ces informations au sein du corpus à la main, aucun outil n'étant encore assez fiable pour le faire. La première section expose notre première tentative et les conclusions que nous en avons tirées pour finalement aboutir au format de la seconde section.

A.1 Démarche de l'annotation

Découvrant la formalisation de Giguet et Lucas (2004), nous avons tenté de l'appliquer telle quelle au corpus afin d'annoter les citations.

A.1.1 Première tentative d'annotation

A.1.1.1 Définition d'un schéma d'annotation

Nous avons fait le choix depuis le début de l'utilisation du métalangage XML pour structurer le corpus, et nous conservons ce choix pour ce qui est de l'annotation. Le premier avantage d'XML est qu'il est lisible et compréhensible par l'humain si l'on définit avec précision les noms des balises et des attributs. Étant donné que le XML est basé sur des balises textes, et qu'il supporte par défaut l'encodage unicode, il semble prédestiné à la structuration des textes. Un autre avantage du XML, lié à sa popularité, est le nombre de bibliothèques performantes et libres qui sont disponibles pour la manipulation de ce type de fichiers. Cela nous permet d'économiser un temps précieux lors du traitement automatisé des fichiers. Finalement, le XML étant largement utilisé par les autres chercheurs en TALN, il nous permettra d'échanger le corpus avec d'autres équipes ou d'intégrer d'autres corpus au notre.

L'annotation reprend la séquence canonique de la citation. Elle se répartit donc en trois balises XML : un pour la source (*<source/>*), un pour le relateur (*<linker/>*) et un dernier pour le discours rapporté (*<speech/>*). La réunion en objet citationnel s'effectue grâce à l'attribut commun aux trois balises : *citation*. L'affectation d'un identifiant identique aux balises d'une même citation permet de regrouper les objets citationnels d'une même séquence canonique. Cela permet de reconstituer les citations une fois l'annotation terminée. De plus, afin de palier le problème d'éclatement

des objets citationnels, notamment des relateurs qui peuvent s'étaler autour d'une expression locuteur, chaque balise est accompagnée d'un attribut *id* permettant de relier sous un même objet citationnel des segments textuels balisés. Finalement, afin de cloisonner l'annotation des citations de la structuration logique, nous avons choisi de placer les balises dans un espace de nom différent.

```
<cite:source citation="1" degree="0" id="1">
  Gerard Kleisterlee
</cite:source>
<cite:linker citation="1" source="1" id="1">
  parle avec fierté de ce site
</cite:linker>
censé illustrer la résurrection du groupe qu'il dirige depuis
    2001 :
<cite:speech citation="1" id="1">
  "Il y a six ans, cet endroit était entouré de grillages,
     peu de gens y avaient accès... et peu de choses en
     sortaient."
</cite:speech>
```

Les balises *source* et *linker* possèdent des attributs supplémentaires. Ainsi la balise *source* s'accompagne de l'attribut *degree* particulièrement utile lorsque plusieurs locuteurs sont présents pour une même citation. Sa valeur correspond à la valeur supposée du degré de la source correspondante, à savoir 0 pour la source originale, 1 pour la source ayant rapporté la source originale...

L'attribut *source* de la balise *linker* permet de préciser à quelle balise *source* il est rattaché. Cela est notamment utilisé lorsque plusieurs sources sont citées pour un même texte englobé.

```
<cite:speech citation="2" id="2">
  "L'orientation privilégiée de l'enquête est d'ordre
     familial"
</cite:speech>
<cite:linker citation="2" source="2">
  , indique mardi
</cite:linker>
<cite:source citation="2" degree="0" id="2">
  une source proche de l'enquête
</cite:source>
<cite:linker citation="2" source="3">
  , citée par
</cite:linker>
<cite:source citation="2" degree="1" id="3">
  l'AFP
</cite:source>
```

A.1.1.2 Tentative d'annotation

Une fois le schéma d'annotation défini, nous l'avons essayé sur deux articles du corpus afin de tester son efficacité. Trois personnes se sont donc attelées à annoter les objets citationnels de ces articles. Aucune communication sur l'annotation n'a été effectuée entre les personnes jusqu'à ce que l'on compare les résultats. Le schéma d'annotation s'est révélé inadapté.

La séquence canonique *source+relateur+discours rapporté* fonctionne bien pour les citations dont le texte englobé est clairement délimité par des marques typographiques ou ponctuatives et tel qu'un locuteur introduise ou conclue la citation. En d'autres termes, l'annotation fonctionne correctement lorsque les séquence canoniques correspondent au modèle proposé par Giguet et Lucas (2004) comme dans l'extrait ci-dessous :

> “ « Philips était très compartimenté, avec des métiers très différents. Les gens se rentraient dedans par hasard », raconte-t-il.
>
> ©*Challenges - 15 Février 2007*

Les choses se compliquent lorsque la structure de la phrase ne nous fournit pas de locuteur. Ainsi, dans l'exemple ci-dessous, l'on peut supposer que la source est *le tribunal administratif de Marseille*, sous la forme très certainement d'une communication par le biais d'un représentant officiel, mais aucun indice ne nous l'indique, et notamment pas un relateur.

> “ le tribunal administratif (TA) de Marseille a « enjoint au préfet des Bouches-du-Rhône de délivrer à M. Aït Baloua un titre de séjour »de dix ans dans les deux mois.
>
> ©*Libération - 22 Février 2007*

La difficulté croît avec la diffusion du texte englobé dans le texte englobant. En effet, les limites du texte englobé étant plus difficiles à définir, les balises sont proportionnellement plus difficiles à positionner dans le texte. L'éclatement du relateur autour du locuteur, ou l'abscence pure et simple de ce dernier compliquent d'autant l'annotation.

En résumé, le premier choix d'annotation se voulait trop optimiste sur notre capacité à définir distinctement les bornes des objets citationnels au sein des articles. L'expérience de l'annotation par différentes personnes a mis en valeur une définition différente des bornes par les différentes personnes sur certaines citations. Nous avons donc décidé d'adapter cette annotation en réduisant sa précision.

A.1.2 Nouveau schéma d'annotation : le segment citationnel

Le premier schéma d'annotation se voulait trop complet pour être efficacement appliqué à notre corpus. Le relateur est par exemple un véritable problème par la grande variation de ses formes.

La première décision prise pour le nouveau schéma d'annotation du corpus fût de supprimer la balise *linker* destinée au repérage du relateur. Nous avons en effet décidé de laisser de côté cet élément afin de nous concentrer plus particulièrement sur le repérage du locuteur et du texte englobé. La collection de balises s'est donc réduite aux deux balises :

- source : destiné à marquer les expressions locuteurs. Avec le recul le nom de la balise paraît mal choisi ;
- discours : destiné à marquer les segments de textes qui contiennent globalement la totalité du texte englobé ;

Les attributs de la balise *source* ont tous été supprimés à part l'attribut *id* qui permet d'identifier l'expression locuteur et éventuellement de segmenter l'annotation d'une expression locuteur. Cette capacité n'a cependant pas été extrêmement utilisée au sein du corpus.

Les attributs de la balise *discours* ont également tous été supprimés à l'exception cette fois de *source* qui permet toujours de relier le texte englobé à son expression

locuteur associée. Il n'est plus question désormais d'annoter le texte englobé, mais plutôt de relier tous les segments de texte rapportés à l'expression locuteur à laquelle ils se réfèrent. Si le cas l'impose, il est possible de spécifier que le segment de texte n'est relié à aucune expression locuteur avec l'identifiant *?*.

La reconstitution de la citation selon que l'on considère comme telle la combinaison d'une expression locuteur et du texte qui y est rattaché s'effectue alors en trois temps :

1. réunification des balises sources de même identifiant ;
2. compilation des balises discours référant à la dite source ;
3. englobement des balises récoltés sous une forme approximative de la citation : le *segment citationnel*.

Ce schéma d'annotation, illustré par l'extrait ci-après, a parfaitement fonctionné pour l'annotation complète du corpus. Les prises de décision quant aux bornes du discours englobé sont toujours présentes, mais contournées par la possibilité de sélectionner au plus large lors de l'apposition des balises ainsi que la segmentation en fragments de texte englobé reliés à une même expression locuteur. L'approximation de la citation en *segment citationnel* nous a donc permis d'annoter la totalité des informations qui nous intéressaient au sein du corpus. Cette approximation ne semble pas avoir causé la dégradation de ces dites informations.

```
<cite:discours source="1">
  "La question est de savoir si l'économie a aujourd'hui un
     problème de demande ou des difficultés du côté de l'
     offre"
</cite:discours>
,  résume
<cite:source id="1">
  Lionel Fontagné, professeur à Paris-I et membre du Conseil
     d'analyse économique
</cite:source>
, pour qui
<cite:discours source="1">
  "la plupart des économistes estiment qu'il y a d'abord un
     problème d'offre"
</cite:discours>
.
```

Petit effet de bord ennuyeux, l'ajout des balises délimitant les locuteurs et le texte englobé a parfois entraîné un mauvais chevauchement avec les balises utilisées pour la typographie. Il nous a été nécessaire de repositionner quelques unes des balises de marquage typographique (italique, gras et emphase). Ces décalages se sont toutefois réduits à décaler la fermeture ou l'ouverture des balises de typographie avant ou bien après un signe de ponctuation afin de concorder avec l'ouverture ou la fermeture d'une des balises *discours*.

La simplification du format d'annotation des citations en s'appuyant uniquement sur les expressions textuelles concernant le texte englobé et les sources, nous a permis de compléter plus efficacement la phase d'annotation. De plus, en délaissant la reconstitution des citations à la charge de l'utilisateur du corpus, nous lui donnons un certain degré de liberté quant à ce qu'il veut considérer comme citation.

A.2 Guide d'annotation

A.2.1 Type du discours rapporté

Dans Mourad (2001), les auteurs proposent de considérer deux catégories de citations. La première, reconnaissable à ces marques typographiques, correspondant au discours direct. La deuxième pour la forme indirecte du discours.

D'autres auteurs ont proposé d'introduire la catégorie du discours indirect libre aux deux précédentes.

Enfin, après avoir parcouru quelques articles de presse, il m'a paru judicieux de considérer la possibilité de voir plusieurs styles mélangés pour une même citation. Une sorte de discours hybride à mi-chemin entre le discours direct et le discours indirect.

A.2.1.1 Discours direct

Dans les articles de la bibliographie, le discours direct n'est clairement défini que par Mourad et Minel (2000) où ils indiquent que le discours direct se différencie du discours indirect par la présence de marqueurs typographiques.

Cette définition est renforcée par la présentation du discours direct par des sites généralistes sur l'enseignement du français. En effet, ces derniers le définissent comme un type de discours rapporté dans lequel les paroles ou les pensées sont rapportées directement, entre guillemets.

J'ai donc considéré comme citations directes les segments citationnels où les propos rapportés étaient introduits par l'ouverture d'un guillemet et conclus par la fermeture d'un guillemet.

> “
> Mais Brice Hortefeux précise que « le seul objectif de cette fusion, c'est de parvenir à une plus grande synergie et à une plus grande coordination de ceux qui s'expriment au nom du candidat ».

Exemple de discours direct extrait du corpus (©Le Figaro)

A.2.1.2 Discours indirect (ou discours indirect lié)

Le discours indirect n'est pas clairement défini, en réalité, il est souvent assimilé comme la catégorie complémentaire au discours direct. Cependant, cette définition ne peut pas nous satisfaire pleinement étant donné que nous avons défini quatre catégories de discours.

Il s'agit d'un type de discours rapporté par lequel les paroles ou les pensées sont rapportées indirectement, à l'aide de subordonnées. Contrairement au dicours direct, le discours indirect n'est pas délimité par des éléments typographiques. Si toutefois des guillemets étaient utilisés en son sein, l'utilisation serait celle de mise en valeur de l'entité entre guillemets, et non la délimitation des propos rapportés.

> “
> il a déclaré qu'il pourrait nommer un premier ministre de gauche, s'il était élu président de la République.

Exemple de discours indirect extrait du corpus (©Le Monde)

A.2.1.3 Discours indirect libre

Le discours indirect libre est une catégorie particulière. Certains (Mourad et Minel (2000), Mourad (2001) ou Mourad et Desclés (2002)) la considèrent comme une sous-catégorie du discours indirect. D'un autre côté, le linguiste Roman Jakobson, considère cette catégorie comme une catégorie à part entière (vérifier sources).

Le discours indirect libre se détache du discours indirect par la suppression des verbes introducteurs et des subordonnées. De plus, les pronoms, adverbes et les temps grammaticaux étant ceux du récit, cela en fait un type de discours rapporté ardu à distinguer du récit.

Contrairement aux deux discours précédents qui sont – dans la plupart des cas – introduits dans un contexte phrastique par une subordonnée issue du récit, dans le discours indirect libre les propos rapportés sont les éléments principaux de la phrase (subordonnée principale).

“ Il met bas son fagot, il songe à son malheur. / Quel plaisir a-t-il eu depuis qu'il est au monde ?

Exemple tiré de la fable La Mort et le Bûcheron de Jean de La Fontaine

“ Selon un policier binchois, habitué de l'événement, 80 000 à 100 000 personnes auraient rejoint la Cité du Gille.

Exemple de discours indirect libre extrait du corpus (© Le Soir)

A.2.1.4 Styles hybrides

La forme des articles de presse et les règles éditoriales spécifiques aux différents journaux contraignent souvent les journalistes à adapter les citations pour pouvoir les réduire en taille, tout en conservant l'idée originale et les propos forts qui illustrent les idées qu'ils veulent mettre en avant.

Cette combinaison semble faire apparaître un style de discours particulier où l'idée générale du propos rapporté est reformulée de manière synthétique au discours indirect, mais accompagnée d'expressions exactes entre guillemets – discours direct – afin d'appuyer la véracité et l'authenticité des dits propos.

“ c'est depuis vingt-cinq ans " l'une des économies les plus dynamiques du monde ", note l'OCDE.

Exemple de discours rapporté mélangeant les styles extrait du corpus (©Challenges)

A.2.1.5 Cas litigieux

Dans certains cas, les extrémités du propos rapporté sont clairement identifiées par des guillemets comme par exemple :

“ un de ses proches se réjouit : « Il sait écouter et accorde le droit à l'erreur. »

Exemple extrait du corpus (©Challenges)

Cependant, dans certains segments citationnels, le début — ou la fin — est moins bien marqué. Dans l'exemple ci-dessous, nous pouvons nous demander si "avec fierté" devrait faire partie du discours rapporté et auquel cas il s'agirait d'un mélange de styles plutôt que d'un discours direct :

“ Agon vante avec fierté « les résultats spectaculaires ».

Exemple extrait du corpus (©Challenges)

Cependant, l'expression "avec fierté" décrit la manière dont les propos sont introduits par la source, mais elle n'appartient pas au discours. Il ne faut donc par l'inclure dans le discours rapporté et considérer la citation comme appartenant au discours direct.

Dans le cas ci-dessous, toutefois, l'expression "c'est depuis vingt-cinq ans" appartient potentiellement aux propos tenus par la source :

> “ c'est depuis vingt-cinq ans « l'une des économies les plus dynamiques du monde », note l'OCDE.
>
> *Exemple extrait du corpus (©Challenges)*

Il est donc nécessaire de classer le discours rapporté comme un mélange de style. En effet, la partie "c'est depuis vingt-cinq ans" est une reformulation des propos de la source, alors que "l'une des économies les plus dynamiques du monde" est une reprise verbatim. La concaténation des deux consiste donc à un mélange des styles du discours dans le même segment citationnel.

L'oubli de la fermeture d'un guillemet est une chose assez fréquente. Cependant, parfois il semble que le journaliste est oublié d'ouvrir et de fermer les guillemets pour rapporter des propos exacts :

> “ Pour faire simple, dans quatre ans, un microprocesseur contiendra 32 milliards de transistors (100 fois plus qu'aujourd'hui) et sera doté d'une puissance phénoménale, nous explique le directeur du management des technologies d'Intel.
>
> *Exemple extrait du corpus (©Challenges)*

Bien que cette citation ressemble fortement à une reprise verbatim des propos de la source, l'auteur ne l'a pas marquée typographiquement. Dans ce cas précis, d'après les règles présentées plus haut, le discours est considéré comme indirect, car rien ne nous indique qu'il s'agit effectivement des propos exacts, et non pas d'une reformulation s'approchant de ces propos exacts.

A.2.2 Source de la citation

La source est un élément clef de la citation. En effet, elle permet l'identification de l'origine des propos rapportés par l'auteur. Cependant, les sources sont polymorphiques, ce sont donc des entités dont il nécessaire de réaliser une description.

Un des objectifs de la caractérisation du corpus était de déterminer les proportions de cette polymorphie et donc connaître quelles étaient les formes les plus usitées.

Il est important de noter que la source comprend non seulement la référence au locuteur mais également tous les éléments apportant des informations sur ce dernier et étant situés en dehors des propos rapportés et du relateur. Les caractérisations ci-dessous sont non-exclusives, i.e. il est possible qu'une source soit constituée de plusieurs des objets énoncés.

> “ Pour Jacques Delpla, économiste à BNP Paribas, il ne s'agit que de « signaux politiques »,
>
> *Exemple extrait du corpus (©Challenges)*

Dans l'exemple ci-dessus, l'on dénomme comme source, l'extrait : *Jacques Delpla, économiste à BNP Paribas.*

A.2.2.1 Source nommée

On entend "source nommée" dans le sens où la source est constituée d'au moins un élément capitalisé considéré comme le nom d'une personne ou d'une entité.

“
Steve Jobs estimait...
...estime Marc Menesguen, directeur général de la division produits de luxe.
Un porte-parole de la Maison Blanche a indiqué...

Exemples extraits du corpus (©Challenges, ©Le Soir)

A.2.2.2 Source pronominale

Une source est considérée "source pronominale" lorsqu'elle contient au moins un pronom personnel – pas impersonnel – sujet, complément d'objet direct ou indirect. Les pronoms personnels réfléchis ne sont pas pris en considération.

“
Le prochain chapitre à Bassora sera écrit par les Irakiens eux-mêmes, a-t-il dit
Elle a en revanche confirmé qu'il ne [...]
Pour lui, « toute l'administration [...]

Exemples extraits du corpus (©Le Soir)

A.2.2.3 Source nominale

Une source est considérée "source nominale" lorsqu'elle contient au moins un groupe nominal composé d'un nom commun. Les noms communs peuvent être connectés entre eux à l'aide de conjonctions et/ou de déterminants.

“
[...] a admis hier, le procureur Brice Raymondeaud-Castanet.
le magistrat avait pointé sans relâche [...]
Comme le conclut Françoise, 61 ans et deux paquets par jour depuis quarante-quatre ans, « ça n'est pas drôle de fumer ».

Exemples extraits du corpus (©Libération)

A.2.2.4 Source inconnue

Dans le cadre de la caractérisation, la recherche de la source se limite au contexte phrastique. Une source est donc considérée inconnue lorsqu'il n'est pas fait mention de cette dernière de manière explicite dans la phrase introduisant le discours rapporté.

“
« Pour la première fois était posée la question des "saisonniers" sous cet angle-là : leur qualité de travailleur permanent. »
l'affaire dite des bagagistes de Roissy révélait qu'un groupe de salariés soupçonnés de liens avec l'islamisme radical avait été écarté parce que présentant « une vulnérabilité incompatible avec une habilitation d'accès en zone réservée ».

Exemples extraits du corpus (©Libération)

A.2.3 Motif de la citation

L'expérimenation du corpus fût également une opportunité pour tester les méthodes proposées dans des travaux antérieurs comme la méthode des motifs proposée par Giguet et Lucas (2004).

A.2.3.1 Schéma du motif

Contrairement à la proposition de Giguet et Lucas (2004) qui incluait la typographie au sein des entités *source*, *relateur* et *discours*. Dans le cadre de l'expérimentation, j'ai quelque peu modifié cet aspect en intégrant aux motifs les éléments typographiques qui faisaient la jointure entre deux objets.

Ainsi, dans l'exemple ci-dessous, le relateur est *Selon*, la source *lui* et le discours rapporté *les Français ont [...] des maires de France*. Les éléments typographiques "**,**" et "**«** " joignent la source et le discours alors que les éléments " **»**" et "**.**" clôturent le discours. Le motif extrait est donc : *<relateur><source>, « <discours> ».* :

> Selon lui, « les Français ont [...] des maires de France ».
>
> *Exemple extrait du corpus (©Le Figaro)*

A.2.3.2 Schéma du relateur

Le relateur est un objet introduit par Giguet et Lucas (2004) qui permet de relier la source et ses propos au sein du texte englobant. Les relateurs peuvent être de différentes tailles et de différentes formes. Le but de la caractérisation était de se faire une idée des formes les plus employées.

Les conventions utilisées pour le schéma du relateur sont de spécifier les catégories linguistiques des mots ou expressions lorsque son utilisation n'est pas spécifique, ou bien le mot lui même – dans sa version lemmatisée – autrement.

Ainsi, dans l'exemple ci-dessous, le relateur est le verbe *assure*, cependant d'autres verbes auraient pû être employés : dire, penser,...Nous choisissons donc de le représenter par "Verbe" :

> « Il ne s'agit pas d'une augmentation déguisée de la taxe automobile », assure le ministre des Transports, Wolfgang Tiefensee,
>
> *Exemple extrait du corpus (©Libération)*

Cependant, dans le cas ci-dessous, le relateur est *a déclaré qu'* correspondant au verbe *déclarer* suivi de *que*. Bien qu'on puisse remplacer le verbe *déclarer* par *dire* ou *penser*, aucun remplacement n'est évident pour *que* que l'on conserve donc tel quel dans le schéma : "Verbe+que".

> il a déclaré qu'il pourrait nommer un premier ministre de gauche, s'il était élu président de la République.
>
> *Exemple extrait du corpus (©Le Monde)*

A.2.4 Concordance des temps

La caractérisation des temps employés concerne tous les temps et modes qui peuvent être présents au sein du segment citationnel.

Cette caractérisation du corpus se limitant à l'étude du contexte phrastique, il est possible qu'aucun verbe du récit ne soit présent dans le segment citationnel considéré, ni qu'aucun verbe ne soit présent au sein des propos rapportés. Dans ces cas, il suffit d'appliquer la valeur *NA*.

A.2.4.1 Temps du récit

On considère comme verbes du récit, tous les verbes situés à l'extérieur des propos rapportés par la source, les éventuels verbes du relateur y compris.

A.2.4.2 Temps du discours rapporté

On considère comme verbe du discours rapporté tous les verbes présents au sein des propos rapportés par la source, y compris lorsqu'il s'agit du discours indirect.

Annexe B

Extraits des corpus

B.1 Extrait du corpus de citations

L'extrait ci-dessous est tiré du corpus que nous avons constitué pour la détection de citations. La première partie de l'arbre XML donne des informations sur la provenance du texte. La seconde partie consiste en son contenu agrémenté des balises d'annotation :

- des segments textuels dérivés (<citation:discoursrepris />)
- et des expressions locutrices (<citation:locuteur />).

Nous avons également conservé des balises informant sur la mise en forme du texte et l'organisation des paragraphes et des titres. Nous n'en avons toutefois pas tiré parti.

```
<?xml version="1.0" encoding="utf-8"?>
<?xml-stylesheet href="annotation.css" type="text/css" media=
    "screen"?>
<!DOCTYPE article SYSTEM "DTDCorpus.dtd">
<corpus:article
  xmlns:corpus="http://www.fabienpoulard.info/xmlns/corpus"
  xmlns:typo="http://www.fabienpoulard.info/xmlns/typo"
  xmlns:ht="http://www.fabienpoulard.info/xmlns/ht"
  xmlns:citation="http://www.fabienpoulard.info/xmlns/
      citation">
  <corpus:metadata>
    <corpus:journal>Challenges</corpus:journal>
    <corpus:url>http://www.challenges.fr/business/
        chall_1001822.html</corpus:url>
    <corpus:authoring>
      <corpus:author>Gilles Fontaines</corpus:author>
    </corpus:authoring>
    <corpus:edition>Challenges.fr</corpus:edition>
    <corpus:publicationdate>15.02.2007</
        corpus:publicationdate>
  </corpus:metadata>
  <corpus:content>
    <corpus:title>La France sponsorise le piratage</
        corpus:title>
    <corpus:epigraph>Il l'a dit : Steve Jobs, Le PDG d'Apple
        propose désormais de vendre la musique sans protection
```

```
        .</corpus:epigraph>
    <corpus:bloc>
      <corpus:header><typo:b>Il a changé d'avis</typo:b></
          corpus:header>
      <corpus:paragraph>
        En mars 2006, Steve Jobs, le PDG d'Apple, ne cachait
            pas sa colère en dénonçant le projet de loi
            français sur les droits d'auteur. Le texte
            stipulait que la musique légalement téléchargée
            devait être écoutable sur tous les lecteurs. Une
            pierre lancée dans le jardin d'Apple et de
            FairPlay, son système de gestion numérique de
            droits (DRM) qui interdit d'écouter des morceaux
            achetés sur iTunes autrement qu'avec un iPod. La
            France est redevenue respectable depuis la semaine
             dernière. Dans une lettre publiée sur le web,
            Steve Jobs fait machine arrière et propose
            désormais de vendre la musique sans protection.
        <citation:discoursrepris source="1" type="direct">
        S'affranchir des DRM, assure-t-
        <citation:locuteur id="1">il</citation:locuteur>
        aujourd'hui, permettra de doper l'innovation tout en
            augmentant le nombre d'acheteurs et les recettes
            des vendeurs comme iTunes
        </citation:discoursrepris>.
        En réalité, le patron d'Apple revient à son idée
            originelle. En mars 2002, cinq mois après le
            lancement de l'iPod,
        <citation:locuteur id="2">Steve Jobs</
            citation:locuteur>
        estimait, dans le  Wall Street Journal , que «
        <citation:discoursrepris source="2" type="direct">
        la musique légalement acquise doit pouvoir être gérée
             sur n'importe quel type d'appareil
        </citation:discoursrepris>
        ». Quelques mois plus tard, les majors du disque l'
            avaient obligé à revoir son point de vue. Aujourd'
            hui, fort de ses 80 % de part de marché dans la
            musique numérique légale, il espère les faire
            plier en se posant en héraut des consommateurs.
            Son pari est risqué, mais il en a les moyens.
      </corpus:paragraph>
    </corpus:bloc>
  </corpus:content>
</corpus:article>
```

B.2 Extrait des corpus de dérivations

B.2.1 Extrait du corpus Piithie

B.2.1.1 Document source

Fichier de contenu (source-document00071.txt) :

```
Carrefour se lance dans la VOD
* Auteur: Maxime Gaillard
* Publié dans: Contenu
Samedi
19 avril 2008
Le groupe Carrefour va prochainement se lancer dans la VOD en
    France mais pas seulement. Fier de ses parts de marché
    dans la vente de DVD dans d'autres pays d'europe (autour
    de 13%), Carrefour va aussi ouvrir son service de Vidéo à
    la Demande en Espagne, Italie et Belgique.
Pour le moment aucune date de lancement et aucun prix n'a été
    annoncé mais Carrefour promet une interface simple et des
    téléchargements rapides (encore heureux !). Il est fait
    également mention d'une possibilité de streaming.
L'arrivée du mastodonte Carrefour dans la VOD ne va pas ravir
    tout le monde. En effet la force de frappe du second
    groupe de grande distribution du monde est impressionnante
    .
```

Fichier d'annotations (source-document00071.xml) :

```
<?xml version="1.0" encoding="utf-8"?>
<document reference="source-document00071.txt" xmlns:xsi="
    http://www.w3.org/2001/XMLSchema-instance"
    xsi:noNamespaceSchemaLocation="http://www.uni-weimar.de/
    medien/webis/research/corpora/pan-pc-09/document.xsd">
  <feature download_date="not available" journal="not
      available" name="piithie_metadata" piithie_reference="
      S00009/S00009.xml" publication_date="not available"
      source_used="" url=""/>
  <feature name="language" value="fr"/>
</document>
```

B.2.1.2 Document dérivé

Fichier de contenu (suspicious-document00848.txt) :

```
VOD Carrefour complément.... peut être    10:09 20/04/08
Le géant national de la grande distribution française lancera
    une offre de VOD, vidéo à la demande en France, Belgique,
    Italie et Espagne, où par-ailleurs il détient une part de
    marché de 13,3% dans la vente de DVD. Un terreau idéal
    pour y lancer une boutique de VoD !
Si aucune date de lancement, tarification ou matériel
    nécessaire n'a été précisé, Carrefour promet une interface
    simple et que les téléchargements seront rapides. Il est
    fait également mention d'une possibilité de streaming.
```

```
Dans cette partie de l'info ni Glowria ni NTG ne sont
    précisées. Néanmoins ds les docs recueillis à l'occasion
    de la dernière AG on peut lire p55 les principales
    activités de GLO " CLO opère aujourd'hui le service VOD
    pour NEUF, FNAC, DARTY, et prochainement pour ALLOCINE,
    SFR, CARREFOUR et P$ T Luxembourg."
Cette lecture permet qq espoirs puisque fin 07 ce partenariat
     était envisagé et si GLO est retenu NTG devrait
    probablement pouvoir placer sa Netbox ds la foulée.
    Gardons la tête froide et attendons l'annonce officielle.
```

Fichier d'annotations (suspicious-document00848.xml) :

```
<?xml version="1.0" ?>
<document reference="suspicious-document00848.txt" xmlns:xsi=
    "http://www.w3.org/2001/XMLSchema-instance"
    xsi:noNamespaceSchemaLocation="http://www.uni-weimar.de/
    medien/webis/research/corpora/pan-pc-09/document.xsd">
        <feature download_date="2008-04-23" journal="
            Boursorama" name="piithie_metadata"
            piithie_reference="S00009/C00001.xml"
            publication_date="2008-04-20" source_used="not
            available" url="not available"/>
        <feature name="language" value="fr"/>
        <!--Document level derivation information-->
        <feature derivation_origin="press" name="annotated-
            doclevel-derivation" source_mentionned="no"
            source_reference="source-document00071.txt"
            source_used="not available"/>
        <!--Local derivations-->
        <feature name="annotated-locallevel-derivation"
            obfuscation="mix" source_length="274"
            source_offset="102" source_reference="source-
            document00071.txt" this_length="268" this_offset="
            56" translation="false"/>
        <feature name="annotated-locallevel-derivation"
            obfuscation="mix" source_length="225"
            source_offset="377" source_reference="source-
            document00071.txt" this_length="224" this_offset="
            325" translation="false"/>
</document>
```

B.2.2 Extrait du corpus Wikinews

B.2.2.1 Document source

Fichier de contenu (source-document00054.txt) :

```
Des manifestations et des perturbations dans les transports
    rythment cette journée de mobilisation qui devrait selon
    les organisation syndicales rassemblée un million de
    personnes. Il s'agît de la 1er journée de contestation
    importante pour le gouvernement de Villepin. Elle est
    marquée par les revendications du 29 mai soit : la
```

```
    sauvegarde du service public, le retrait de la directive
    Bolkenstein

Sources
*
```

Fichier d'annotations (source-document00054.xml) :

```
<?xml version="1.0" encoding="utf-8"?>
<document reference="source-document00054.txt" xmlns:xsi="
    http://www.w3.org/2001/XMLSchema-instance"
    xsi:noNamespaceSchemaLocation="http://www.uni-weimar.de/
    medien/webis/research/corpora/pan-pc-09/document.xsd">
        <feature name="wikinews_metadata" page_id="3675"
            revision_depth="0" revision_id="108" source="
            Wikinews France" url="http://fr.wikinews.org/w/
            index.php?oldid=108"/>
        <feature name="language" value="fr"/>
</document>
```

B.2.2.2 Document dérivé

Fichier de contenu (suspicious-document00711.txt) :

```
Des manifestations et des perturbations dans les transports
    rythment cette journée de mobilisation qui devrait selon
    les organisations syndicales rassembler un million de
    personnes. Il s'agit de la première journée de
    contestation importante pour le gouvernement de Villepin.
    Elle est marquée par les revendications du 29 mai soit :
    la sauvegarde du service public, le retrait de la
    directive Bolkenstein...

La manifestation a réuni au moins 1 039 000 de manifestants
    dans toute la France selon les syndicats, et environ 470
    000 selon la police.

Le journal L'Humanité annonce que 74% des Français expriment
    leur soutien ou leur sympathie à cette mobilisation.

Source
*

Catégorie:Europe
Catégorie:France
Catégorie:Social
```

Fichier d'annotations (suspicious-document00711.xml) :

```
<?xml version="1.0" encoding="utf-8"?>
<document reference="suspicious-document00711.txt" xmlns:xsi=
    "http://www.w3.org/2001/XMLSchema-instance"
    xsi:noNamespaceSchemaLocation="http://www.uni-weimar.de/
    medien/webis/research/corpora/pan-pc-09/document.xsd">
```

```
        <feature name="wikinews_metadata" page_id="3675"
            revision_id="22906" source="Wikinews France" url="
            http://fr.wikinews.org/w/index.php?oldid=22906"/>
        <feature name="language" value="fr"/>
        <!--Document level derivation information-->
        <feature name="annotated-doclevel-revision"
            revision_depth="9" source_reference="source-
            document00054.txt"/>
        <!--Local level derivation (whole document as it is a
             revision)-->
        <feature name="annotated-locallevel-revision"
            source_length="409" source_offset="0"
            source_reference="source-document00054.txt"
            this_length="746" this_offset="0"/>
</document>
```

B.2.3 Extrait du corpus PANini

B.2.3.1 Document source

Fichier de contenu (source-document13444.txt) :

```
The Emperor had now lost the greater part of Bohemia, and the
     Saxons were advancing against
Austria, while the Swedish monarch was rapidly moving to the
    same point through Franconia,
Swabia, and Bavaria. A long war had exhausted the strength
    of the Austrian monarchy, wasted
the country, and diminished its armies. The renown of its
    victories was no more, as well as
the confidence inspired by constant success; its troops had
    lost the obedience and discipline
to which those of the Swedish monarch owed all their
    superiority in the field. The confederates
of the Emperor were disarmed, or their fidelity shaken by the
     danger which threatened themselves.
Even Maximilian of Bavaria, Austria's most powerful ally,
    seemed disposed to yield to the seductive
proposition of neutrality; while his suspicious alliance with
     France had long been a subject
of apprehension to the Emperor. The bishops of Wurtzburg and
     Bamberg, the Elector of Mentz,
and the Duke of Lorraine, were either expelled from their
    territories, or threatened with immediate
attack; Treves had placed itself under the protection of
    France. The bravery of the Hollanders
gave full employment to the Spanish arms in the Netherlands;
    while Gustavus had driven them
from the Rhine. Poland was still fettered by the truce which
     subsisted between that country
and Sweden. The Hungarian frontier was threatened by the
    Transylvanian Prince, Ragotsky, a
```

```
successor of Bethlen Gabor, and the inheritor of his restless
    mind; while the Porte was making
great preparation to profit by the favourable conjuncture for
    aggression. Most of the Protestant
states, encouraged by their protector's success, were openly
   and actively declaring against
the Emperor. All the resources which had been obtained by
   the violent and oppressive extortions
of Tilly and Wallenstein were exhausted; all these depots,
   magazines, and rallying-points,
were now lost to the Emperor; and the war could no longer be
   carried on as before at the cost
of others. To complete his embarrassment, a dangerous
   insurrection broke out in the territory
of the Ens, where the ill-timed religious zeal of the
   government had provoked the Protestants
to resistance; and thus fanaticism lit its torch within the
   empire, while a foreign enemy was
already on its frontier. After so long a continuance of good
    fortune, such brilliant victories
and extensive conquests, such fruitless effusion of blood,
   the Emperor saw himself a second
time on the brink of that abyss, into which he was so near
   falling at the commencement of his
reign. If Bavaria should embrace the neutrality; if Saxony
   should resist the tempting offers
he had held out; and France resolve to attack the Spanish
   power at the same time in the Netherlands,
in Italy and in Catalonia, the ruin of Austria would be
   complete; the allied powers would divide
its spoils, and the political system of Germany would undergo
    a total change.
```

Fichier d'annotations (source-document13444.xml) :

```
<?xml version="1.0" encoding="UTF-8"?>
<document xmlns:xsi="http://www.w3.org/2001/XMLSchema-
    instance" xsi:noNamespaceSchemaLocation="http://www.uni-
    weimar.de/medien/webis/research/corpora/pan-pc-09/document
    .xsd" reference="source-document13444.txt">
  <feature name="project-gutenberg" etext_number="6772" url="
      http://www.gutenberg.org/dirs/etext04/fs12w10.txt"/>
  <feature name="language" value="en" />
</document>
```

B.2.3.2 Document dérivé

Fichier de contenu[1] (suspicious-document13592.txt) :

```
RAVEN***
```

1. Une grande partie du fichier n'a pas été rapporté ici afin d'économiser quelques pages dépourvues de dérivations. Seule une fenêtre de texte autour de la partie dérivée est rapportée.

```
Transcribed from the 1913 Thomas J. Wise pamphlet by David
    Price, email ccx074@pglaf.org

                                THE NIGHTINGALE
                            THE VALKYRIE AND RAVEN
                            AND OTHER BALLADS

                                    BY
                                GEORGE BORROW

                                LONDON:
                        PRINTED FOR PRIVATE CIRCULATION

                                    1913

                    _Copyright in the United States of America_
                _by Houghton_, _Mifflin and Co. for Clement
                    Shorter_.

alas! for them! R.N.

    _Pleasure Gardens._—Has it never
occurred to any nurseryman that
    his garden might be made delightful and profitable
        promenades for
the public, at a low charge for admission?

[suppression de 135 lignes]

we A long war had the
    protection of France. The
exhausted the property of
    the bravery of the
Hollanders gave full
    employment embarrassment,
or their fidelity shaken
    by the danger which threatened themselves the Wurtzburg
        and Bamberg,,
time on the of the Elector of Hungarian frontier was'd
    threatened
brilliant slain victories and extensive conquests, such
    itself under
renown of its victories was Porte was making great For
    preparation
With to profit by the favourable conjuncture for aggression.
Most of the to the particular
    of neutrality; while his
suspicious alliance with France
    had long been a subject
apprehension'd to the second
    that abyss, into which
the Austrian monarchy,
```

```
   wasted country, and diminished try its armies.

The had driven them from
   the Rhine. was still fettered
by the, and the inheritor
   of his restless mind;
while, encouraged by depots
   by the Transylvanian
Prince well as the Spanish
   arms in the Netherlands; while Emperor.

After so confidence inspired
   by constant success; to
the long a continuance of good
   fortune, such,,
magazines, rallying-points,
   were now lost to the
Emperor; and the war could
   no longer foreign on was already on its frontier.

Even Halfdan Maximilian of Bavaria,
   Twixt Austria's most powerful
ally, seemed disposed to
   yield its sea-spray their,
protector's success,
   were openly and actively
declaring against" the.
   All the resources which had been obtained by the violent
      and
oppressive extortions of Tilly and Ragotsky, a speed
   dangerous
insurrection broke out in the territory the Ens, where the
   ill-timed
religious zeal of speed the government had provoked the
   Bloody
Protestants to From resistance,; and thus fanaticism lit its
   torch
empire, while a disarmed, a successor of Bethlen Gabor
   Wallenstein
Who were exhausted; all these had lost the obedience and
   discipline
to which those of the Swedish monarch all their superiority
   in the
field.
The confederates the Duke of
   Lorraine, Mentz, and territories,
or threatened with immediate attack;
   Treves had placed between
that country and were are either
   expelled war from their of the
Emperor be carried on as before
```

```
   at the cost of others. near falling at the commencement
      beautiful
of his complete his Protestant states The bishops of no more,
    as
which subsisted fruitless effusion of blood, the Emperor
   himself a
Gustavus he so reign

[suppression de 160 lignes]
```

Fichier d'annotations (suspicious-document13592.xml) :

```
<?xml version="1.0" encoding="UTF-8"?>
<document xmlns:xsi="http://www.w3.org/2001/XMLSchema-
    instance" xsi:noNamespaceSchemaLocation="http://www.uni-
    weimar.de/medien/webis/research/corpora/pan-pc-09/document
    .xsd" reference="suspicious-document13592.txt">
  <feature name="project-gutenberg" etext_number="26834" url=
      "http://www.gutenberg.org/files/26834/26834-0.txt"/>
  <feature name="language" value="en" />
  <feature name="artificial-plagiarism" translation="false"
      obfuscation="none" this_offset="585" this_length="3762"
      source_reference="source-document12990.txt"
      source_offset="6805" source_length="3650" />
  <feature name="artificial-plagiarism" translation="false"
      obfuscation="low" this_offset="5155" this_length="2420"
      source_reference="source-document13444.txt"
      source_offset="205" source_length="2352" />
  <feature name="artificial-plagiarism" translation="false"
      obfuscation="high" this_offset="8040" this_length="2620"
       source_reference="source-document12990.txt"
      source_offset="4029" source_length="2610" />
  <feature name="artificial-plagiarism" translation="false"
      obfuscation="high" this_offset="11325" this_length="279"
       source_reference="source-document12990.txt"
      source_offset="3325" source_length="255" />
  <feature name="artificial-plagiarism" translation="false"
      obfuscation="high" this_offset="11604" this_length="228"
       source_reference="source-document12990.txt"
      source_offset="0" source_length="261" />
  <feature name="artificial-plagiarism" translation="false"
      obfuscation="high" this_offset="11987" this_length="581"
       source_reference="source-document12990.txt"
      source_offset="1119" source_length="706" />
  <feature name="artificial-plagiarism" translation="false"
      obfuscation="high" this_offset="12714" this_length="253"
       source_reference="source-document09737.txt"
      source_offset="1917" source_length="262" />
  <feature name="artificial-plagiarism" translation="false"
      obfuscation="high" this_offset="12994" this_length="251"
       source_reference="source-document09221.txt"
      source_offset="46317" source_length="242" />
```

```
  <feature name="artificial-plagiarism" translation="false"
      obfuscation="high" this_offset="13400" this_length="256"
       source_reference="source-document09221.txt"
      source_offset="15569" source_length="259" />
  <feature name="artificial-plagiarism" translation="false"
      obfuscation="high" this_offset="13808" this_length="351"
       source_reference="source-document09221.txt"
      source_offset="64335" source_length="241" />
  <!-- Use tags like the one below to annotate plagiarism you
       detected. -->
  <!-- <feature name="detected-plagiarism" this_offset="50"
      this_length="1000" source_reference="source-documentX.
      txt" source_offset="750" source_length="1700"/> -->
</document>
```

Annexe C

Observation en corpus des objets citationnels

> *« Elle est à l'horizon, dit Fernando Birri. Je me rapproche de deux pas, elle s'éloigne de deux pas. Je chemine de dix pas et l'horizon s'enfuit dix pas plusloin. Pour autant que je chemine, jamais je ne l'atteindrai. A quoi sert l'utopie ? Elle sert à cela : cheminer. »*
>
> — Eduardo Galeano, Las Palabras andantes (1993)

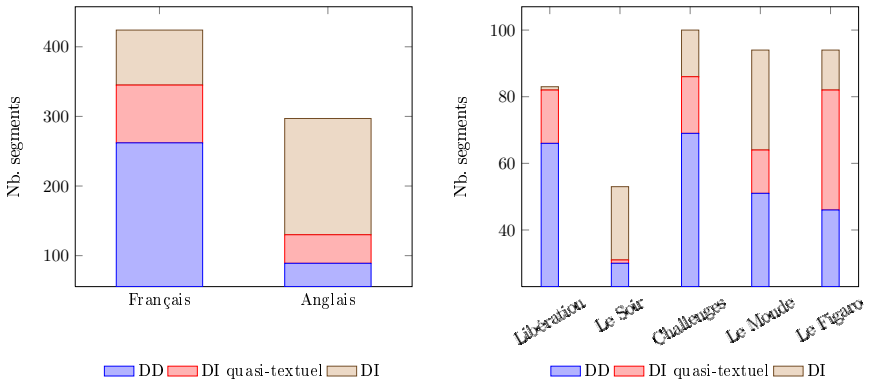
Nous avons cherché à vérifier l'adéquation des approches de Giguet et Lucas (2004) et Mourad et Desclés (2004) dans le cadre de l'article de presse en français et en anglais. Nous avons observé les différents objets citationnels annotés dans nos corpus (*cf. Section 3.2.2*) afin de vérifier la pertinence des indices proposés dans la littérature et éventuellement les compléter. Nous présentons dans cette section nos observations concernant les segments dérivés, les expressions locutrices et les quelques éléments relateurs.

C.1 Segments dérivés

La proportion des segments textuels dérivés dans les articles est assez variable. Ainsi, ils ne représentent en moyenne que 17 % du texte des articles du journal Challenges, contre près de 70 % des articles de Reuters [1]. Nous nous intéressons à la distribution des styles de discours employés et les constructions associées. Nous nous focalisons plus particulièrement sur le français.

Le graphique comparatif de la figure C.1(a) illustre la distribution des différents styles de discours utilisés dans chaque corpus. Pour le français, 80 % des segments textuels dérivés sont rapportés au style direct ou au style indirect quasi-textuel, préservant ainsi la forme originale. Les journalistes français emploient donc majoritairement des styles qui ne nécessitent pas de fortes modifications morpho-syntaxiques, c-à-d soit une intégration de type verbatim ou syntaxique (*cf. Section 1.4*). En revanche, la proportion s'inverse pour l'anglais où ces styles ne représentent plus que 44 % des annotations. Le graphique de la figure C.1(b) nous laisse penser que cette propension aux styles à guillemets est française puisque la proportion des DD et DI

1. Les articles issus de l'agence Reuters font partis d'un corpus construit en commun par le LINA et le LIA dans le cadre du projet Piithie. Ce corpus est détaillé dans Poulard et collab. (2008).



(a) Styles de discours par corpus (b) Styles de discours par titre

FIGURE C.1 – Répartition des styles de discours utilisé dans les citations pour chacun de nos corpus, et selon les titres de journaux pour le français.

quasi-textuel est majoritaire et stable pour chacun des titres français, mais que pour LE SOIR elle est plus proche de la distribution anglaise.

Les styles de discours employés pour rapporter les citations, en français tout du moins, font des guillemets les marques les plus utilisées pour encadrer tout ou partie des segments dérivés. Malheureusement, ces guillemets sont des marques ambiguës qui « servent soit à inclure en subordonnant, soit à exclure en isolant » (Mourad et Desclés, 2002). Le cas du DI quasi-textuel est le plus problématique étant donné que les segments placés entre guillemets sont potentiellement très courts. Il est alors difficile de les distinguer d'une emphase, comme l'illustre l'exemple 22. Nous considérerons par la suite l'approche consistant à considérer tous les segments entre guillemets comme des segments dérivés comme point de comparaison des résultats.

> Washington avance une estimation des réserves mondiales « **ultimes** » de pétrole à 2 275 milliards de barils.
>
> Elle était bien l'« **organisatrice** » du concert. Ce concert était une activité de « **service public** ». Les agents qui ont commis des fautes disposaient d'un « **pouvoir de représentation** » de la ville.

EXEMPLE 22: Ambiguïté des guillemets : le DI quasi-textuel produit de petit segments entre guillemets difficiles à distinguer des emphases.

C.1.1 Régularités du style direct

Nous avons observé deux constructions prédominantes pour le style direct. Le segment dérivé est délimité, à sa frontière droite, gauche ou même en son centre, par un syntagme prépositionnel d'une part (*cf. Exemple 23*) ou par une construction verbale d'autre part (*cf. Exemple 24*). Dans les deux cas, le segment dérivé constitue la majorité de la phrase et le reste se compose de l'expression locutrice et de la construction qui l'introduit. Il se peut également que cette construction soit embarquée au sein du segment entre guillemets comme l'illustre le premier passage de l'exemple 24.

Selon **eux**, « beaucoup [des saisonniers OMI] auraient bénéficié de CDI en d'autres temps. Relativement qualifiés, ils reviennent régulièrement dans les mêmes exploitations. On dit même de certains que ce sont les véritables chefs d'exploitation ».

© *www.liberation.fr – 22 Février 2007*

EXEMPLE 23: Introduction de la citation par le biais d'un syntagme prépositionnel.

"En 2003 , **explique -t-il,** j'ai fait effectuer douze tests sur des vols en France, dans onze cas sur douze, des armes et explosifs ont pu être introduits dans les avions ".

© *Le Figaro – 20 Février 2007*

"Avec la fin de la session, les parlementaires sont beaucoup plus disponibles. Cette réorganisation était donc nécessaire ", **renchérit -on dans l'entourage de Sarkozy**.

© *Le Figaro – 20 Février 2007*

EXEMPLE 24: Utilisation d'une construction verbale pour marquer la citation.

C.1.2 Régularités du style indirect

Nous avons également observé deux constructions majeures pour le style indirect. Le segment dérivé est délimité, à sa frontière droite, gauche ou même en son centre, par un syntagme prépositionnel d'une part (*cf. Exemple 25*) ou par une proposition complétive d'autre part (*cf. Exemple 26*). Le premier cas est similaire à ce que l'on a observé pour le DD. Dans le second cas, le segment textuel dérivé est introduit par une construction du type *verbe + que*. Nous avons également observé l'emploi du DI sans construction particulière, ce qui complexifie fortement le problème d'identification des frontières du segment textuel dérivé. Comme l'illustre l'exemple 27, la citation est amorcée par une phrase qui présente le contexte, le discours rapporté s'étend ensuite sur plusieurs phrases. Il se termine sans marque particulière.

D'après **sa mère,** Julien faisait une sieste dans l'appartement familial lorsqu'il a disparu .

© *Le Figaro – 20 Février 2007*

EXEMPLE 25: Le segment textuel dérivé est juxtaposé à un syntagme prépositionnel qui l'introduit.

C.2 Expressions locutrices

Les expressions locutrices sont moins nombreuses que les segments dérivés. On ne trouve pas d'expression locutrice sans la présence de segment dérivé mais on peut observer la construction inverse. Les expressions locutrices sont construites autour de trois formes linguistiques : les groupes nominaux, les noms propres capitalisés et les pronoms. Les groupes nominaux et les noms propres capitalisés correspondent à des entités nommées. La distribution de ces formes varie selon le corpus, comme le montre la figure C.2. Les groupes nominaux et les noms propres capitalisés sont autant utilisés que les pronoms pour le français, alors que ces derniers sont en minorité face aux groupes nominaux en anglais.

Les précédents travaux ont observé que les expressions locutrices moins précises (réductions lexicales et anaphores) apparaîssent généralement après une forme com-

Arnaud Montebourg, le porte-parole de Ségolène Royal, **promet** **ainsi** **que** , si la candidate de la gauche est élue, la construction de l'EPR ne serait pas interrompue .

© *Libération – 22 Février 2007*

EXEMPLE 26: Le segment textuel dérivé se présente comme une proposition complétive introduite par *verbe + que*.

Edward Lu, physicien et ancien astronaute au centre Johnson de la NASA, a exposé les moyens envisagés pour repousser ces envahisseurs! Première méthode : envoyer un petit vaisseau spatial de 1 000 kg pour aller impacter à la vitesse de 5 km/sec l'astéroïde menaçant. L'énergie du contact [...] sur la même trajectoire .

© *Le Figaro – 20 Février 2007*

EXEMPLE 27: Une fois la citation amorcée par la première phrase, elle peut s'étendre au-delà de la phrase.

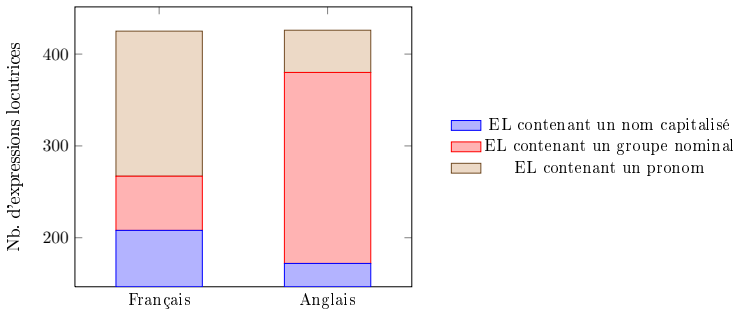


FIGURE C.2 – Répartition des formes des expressions locutrices au sein de nos corpus.

plète (nom capitalisé et groupe nominal). Nous l'observons également dans nos corpus (*cf. Exemple 28*). Cette observation semble expliquer la distribution que nous avons observé : les formes précises sont aussi ou plus nombreuses que les formes imprécises puisque ces dernières ne sont utilisées que pour une deuxième référence, au moins, à la source.

> Brigitte Liberman, la directrice générale de Cosmétique Active (La Roche-Posay, Vichy) , rétorque : « Les grandes innovations ne peuvent pas naître chaque année. »[...] Mais elle admet que « Jean-Paul nous poussant, nous pouvons faire encore mieux ».

EXEMPLE 28: La source est tout d'abord présentée par une forme très précise (nom + titre), puis rappelée par une forme réduite (pronom *elle*).

L'absence d'expression locutrice dans le contexte phrastique du segment textuel dérivé représente 10 % des cas dans le corpus français. Deux scénarios expliquent cette absence : le plus fréquemment, le locuteur a été précédemment introduit et est suffisamment saillant pour permettre au lecteur de faire le lien (*cf. Exemple 29*) ou plus rarement, le contexte permet d'identifier la source (*cf. Exemple 30*).

> Il a insufflé une culture d'entreprise à la Kennedy. "Aujourd'hui, on pense d'abord à Philips, ensuite à son business, et enfin à soi ."

EXEMPLE 29: La dernière expression locutrice (**Il**) est suffisamment saillante pour qu'on lui rattache le segment textuel dérivé de la phrase suivante.

> "Nul ne peut être condamné à la peine de mort " : cet article unique du projet de loi constitutionnelle modifiera le titre VIII de la Constitution, consacré à l'autorité judiciaire.

EXEMPLE 30: Sans expression locutrice, le contexte permet parfois de déduire la source.

C.3 Relateurs

Le composant citationnel *relateur* est certainement la plus complexe à caractériser. Nous n'avons d'ailleurs pas été en mesure de l'annoter dans nos corpus tant sa définition est floue, son interprétation subjective et surtout sa réalisation textuelle fluctuante.

Nous avons toutefois observé les relateurs les plus « évidents » : les prépositions et les regroupements de formes verbales et de compléments circonstanciels ou d'adverbes. Premièrement, les prépositions spécifiques (*selon*, *pour*, *d'après*...) sont constamment suivies d'une expression locutrice et forment avec cette dernière des syntagmes prépositionnels assez aisément identifiables (*cf. exemples 23 et 25*). Deuxièmement, les regroupements de formes verbales et de compléments circonstanciels ou d'adverbes. Ces formes se construisent souvent autour des verbes de communication (Mourad et Desclés, 2004), et offrent une contextualisation très précise de l'énonciation. Les compléments du verbe précisent le contexte d'énonciation, comme le montre l'exemple 31.

Les relateurs sont des marques ambiguës de la présence d'une citation. Outre les constructions susmentionnées, les relateurs peuvent se manifester sous la forme de structures syntaxiques particulières beaucoup plus diffuses dans le texte. Ils peuvent

> Le premier ministre, Tony Blair, a annoncé dimanche à la BBC vouloir durcir la loi sur les armes à feu afin de lutter contre leur circulation parmi les jeunes.
>
> © *Le Monde – 20 Février 2007*

EXEMPLE 31: Mise en relation d'une expression locutrice à un segment textuel dérivé par une construction à base de verbe de communication et de compléments circonstanciels.

notamment être confondus aux segments de textes dérivés ou aux expressions locutrices. Nous ne les considérons pas comme des composants citationnels à identifier, mais nous considérons les constructions particulières offertes par les formes verbales et les syntagmes prépositionnels décrites précédemment comme des indices pertinents de la présence des expressions locutrices et des segments textuels dérivés.

C.4 Conclusion

Nous avons repéré dans notre corpus une vingtaine de motifs, tels que proposés par Giguet et Lucas (2004), impliquant des segments textuels assimilables aux source, relateur et discours rapporté. Dans certains cas, ces composants sont découpés en segments discontinus intercalés de segments rattachés à d'autres composants. Dans d'autres cas, plusieurs sources sont rattachées à un discours rapporté. Dans d'autres cas encore, on ne retrouve qu'un sous-ensemble de ces composants. Les motifs « source + relateur + discours rapporté »et « discours rapporté + relateur + source »sont les plus présents avec respectivement 104 (28 %) et 148 (40 %) occurences. Plusieurs motifs (9 %) sont dépourvus de relateur ou de source (*cf. Exemple 32*). Les autres combinaisons impliquant l'ensemble des composants, « relateur + source + discours »(*cf. Exemple 25*) et « discours + relateur + source + discours »(*cf. Exemple 24*), représentent 15 % des citations.

> Alors Baloua accepte tout, les mois sans jour de repos, les heures sup pas payées, les salaires en dessous des minima. « On est juste une main d'oeuvre moins chère. Ils t'exploitent. Les patrons disent qu'ils ont trop de charges, des dettes. Alors ils piquent à nous, les plus pauvres. On est des victimes. »
>
> © *www.liberation.fr – 22 Février 2007*

EXEMPLE 32: Segment textuel dérivé non explicitement rattaché à une expression locutrice par un relateur.

L'apparition de nouveaux motifs réduit l'intérêt d'une approche positionnelle, même si elle reste envisageable pour 70 % des cas. Il nous semble donc nécessaire de chercher à identifier au mieux les deux composants citationnels que nous avons proposé : les expressions locutrices et les segments textuels dérivés. L'observation en corpus nous a permis de mieux cerner ceux-ci. Les premières se présentent tout d'abord sous des formes très précises (noms capitalisés, groupes nominaux...), puis éventuellement comme des références à cette première forme (anaphores pronominales ou nominales...). Les seconds sont majoritairement composés de segments entre guillemets, du moins pour le français, et sont introduits par des syntagmes nominaux ou des constructions verbales à base de verbes d'énonciation. Finalement, d'après nos observations et pour les articles de presse en ligne, la taille d'un segment citationnel est bornée *a minima* par le mot et *a maxima* par le paragraphe.

Annexe D

Mesures pour l'évaluation des classifieurs

L'évaluation consiste en la confrontation des prédictions des modèles par rapport aux connaissances d'un oracle. L'oracle connaît, habituellement sur la base d'annotations en corpus, la classe à laquelle appartient un candidat. Le tableau ci-dessous met en correspondance les prédictions d'un modèle et de l'oracle et la classification de la prédiction en VP (vrai positif), FP (faux positif), VN (vrai négatif) et FN (faux négatif), pour le cas particulier des composants sources :

	Oracle	
Modèle	composant source	¬ composant source
composant source	VP	FP
¬ composant source	FN	VN

Dans le cadre particulier d'une validation croisée nous obtenons des taux de VP, FP, VN, FN pour chaque partition évaluée. Nous calculons les métriques de précision, de rappel et de *fall out* à partir de la moyenne des taux de VP, FN... obtenus sur les différentes partitions, d'après les formules suivantes (Manning et Schütze, 1999, p.268) :

$$\text{précision} = \frac{VP}{VP + FP} \tag{D.1}$$

$$\text{rappel} = \frac{VP}{VP + FN} \tag{D.2}$$

$$\text{fallout} = \frac{FP}{FP + VN} \tag{D.3}$$

La *précision* (équation D.1) permet de mesurer le taux d'erreur de classification par rapport aux candidats classés comme VP. Plus la précision est élevée, meilleure est cette classification.

Le *rappel* (équation D.2) est une mesure qui complète la précision. En effet, si la précision mesure le nombre d'erreurs de classification parmi les candidats classés comme avérés, le rappel mesure le nombre d'avérés qui n'ont pas été classés comme tels.

Le *fall out* (équation D.3) est une mesure beaucoup moins utilisée, mais pertinente dans notre cas. Elle permet de mesurer le taux de candidats non avérés qui sont classés comme tels. Par exemple, une grande majorité des expressions entre guillemets sont des segments dérivés, il est donc important de classer correctement les candidats non avérés comme tels. En d'autres termes, plus le *fall out* est bas, mieux sont classés les vrais négatifs.

Dans la plupart des tâches de classification, il est possible de gagner en précision aux dépens du rappel et inversement. La F-mesure combine les scores de rappel et de précision afin d'en donner une vision conjointe. Elle s'exprime comme l'inverse de la somme pondérée des deux scores :

$$\text{f-mesure} = \frac{1}{\alpha \frac{1}{\text{précision}} + (1-\alpha)\frac{1}{\text{rappel}}} \tag{D.4}$$

Classiquement, la précision et le rappel sont considérés avec autant d'importance ce qui s'exprime par le choix d'un coefficient de pondération $\alpha = 0.5$. La F-mesure vaut alors :

$$\text{f-mesure} = \frac{2 \cdot \text{précision} \cdot \text{rappel}}{\text{précision} + \text{rappel}} \tag{D.5}$$

Annexe E

Calibrage des algorithmes d'apprentissage

La configuration des algorithmes d'apprentissage utilisés dans le chapitre 3 pour la construction des classifieurs des composants constitutifs de la citation a été choisie suite aux expérimentations de calibrage décrites dans cette annexe.

E.1 Calibrage pour la catégorisation des composants discours rapporté

Nous cherchons empiriquement la configuration des algorithmes d'apprentissage qui donne les meilleurs résultats pour l'ensemble des candidats. Nous explorons pour cette tâche le calibrage des algorithmes *AD Tree* et *C-SVC*.

En ce qui concerne *AD Tree*, algorithme de type arbre de décision avec *boosting*, nous avons joué sur le nombre d'itérations. La figure E.1 montre l'évolution de la F-mesure et du *fall out* en fonction de ce nombre d'itérations. Nous pouvons observer que si la F-mesure (compromis entre précision et rappel) est stable, le *fall out* chute brutalement dès la deuxième itération pour le français, et chute plus doucement à partir de la troisième itération pour l'anglais, or plus le *fall out* est faible, mieux sont reconnus les candidats qui ne sont pas avérés. Nous observons que le *fall out* se stabilise pour le français et l'anglais entre 10^1 et 10^2 itérations, ce qui correspond au paramétrage le plus performant. Au final, nous avons retenu la valeur de 14 itérations pour le français et de 100 pour l'anglais.

En ce qui concerne *C-SVC*, algorithme de classification à vastes marges, nous avons joué sur le paramètre de coût. La figure E.2 montre l'évolution de la F-mesure et du *fall out* en fonction de ce paramètre de coût. Nous observons que les valeurs de coût entre 0 et 15 rendent le *fall out* instable, notamment pour l'anglais, au dessus de 15 ce dernier croît lentement. La F-mesure quant à elle est assez stable, chutant légèrement à partir d'un coût de 30. Il est difficile de décider du paramétrage le plus performant compte tenu de l'instabilité du *fall out*, mais ce dernier semble se situer entre 5 et 10. Au final, nous avons retenu la valeur de 1 pour le français et de 9 pour l'anglais.

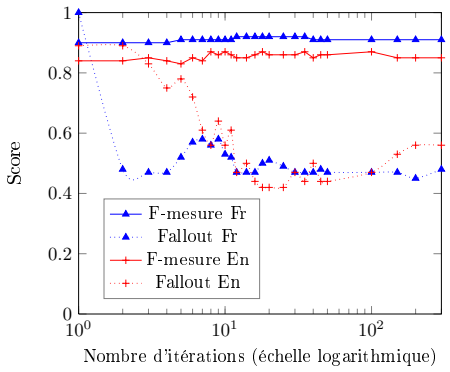


FIGURE E.1 – Résultats du modèle généré par *AD Tree* en fonction du nombre d'itérations

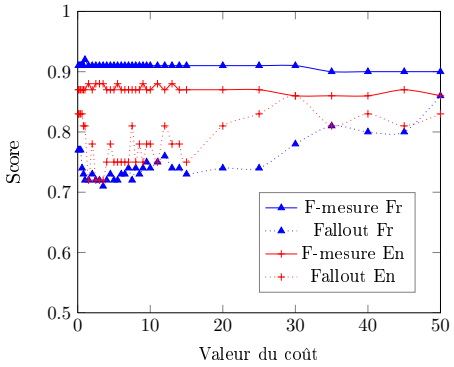


FIGURE E.2 – Résultats du modèle généré par *C-SVC* en fonction du paramètre de coût

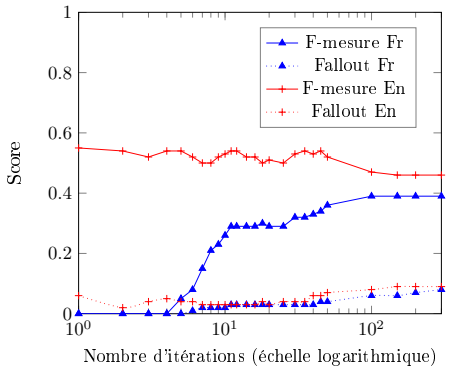


FIGURE E.3 – Résultats du modèle généré par *AD Tree* en fonction du nombre d'itérations

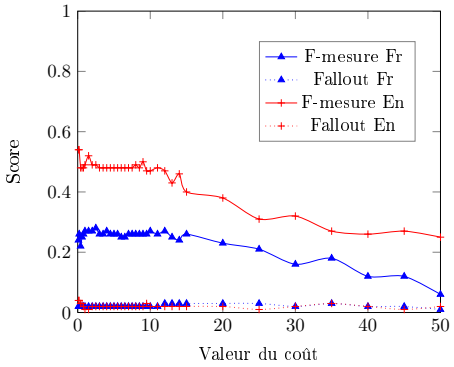


FIGURE E.4 – Résultats du modèle généré par *C-SVC* en fonction du paramètre de coût

E.2 Calibrage pour la catégorisation des composants source

De façon analogue à la section précédente, nous avons cherché, de manière empirique, le paramétrage des algorithmes d'apprentissage qui donnait les meilleurs résultats sur nos données. Nous avons expérimenté les mêmes algorithmes d'apprentissage supervisé (*AD Tree* et *C-SVC*) en faisant varier leur paramétrage de manière identique.

La figure E.3 montre l'évolution de la F-mesure et du *fall out* en fonction du nombre d'itérations pour l'algorithme *AD Tree*. Nous pouvons observer que les modèles pour le français et l'anglais se comportent très différemment, contrairement à ce que l'on avait pu observer pour les segments dérivés. Alors que la courbe de la F-mesure pour l'anglais est légèrement décroissante avec l'augmentation du nombre d'itérations, celle pour le français croît brutalement à partir de la cinquième itération puis est légèrement croissante à partir de la dixième. Les courbes de *fall out* restent quant à elles très basses, ce qui signifie que les candidats qui ne sont pas des expressions locutrices sont correctement classés. Le paramétrage pour l'anglais est le meilleur entre 0 et 500 itérations environ, tandis que pour le français les meilleurs

résultats sont obtenus à partir de 100 itérations. Au final, nous avons retenu la valeur de 100 itérations pour le français et d'une seule itération pour l'anglais.

La figure E.4 montre l'évolution de la F-mesure et du *fall out* en fonction du paramètre de coût pour l'algorithme *C-SVC*. Contrairement à ce que l'on pouvait observer pour l'algorithme *AD Tree*, les courbes pour le français et l'anglais se comportent de manière identique. Le *fall out* reste encore ici très bas, par contre la F-mesure se dégrade assez rapidement lorsque le coût est supérieur à 10. Au final, nous avons retenu un coût de 2,5 pour le français et de 0,2 pour l'anglais.

Annexe F

Exploration *in extenso* du paramétrage de la signature complète

Nos travaux sont à notre connaissance les premiers à s'intéresser à la langue française. Par conséquent nous n'avons pas de résultats de référence auxquels comparer ceux de nos propositions. La littérature cite régulièrement la signature complète à base de n-grammes mots comme approche de référence pour l'anglais. Ainsi, Yang (2006a) regroupe des textes dérivés à l'aide de cette méthode avec une précision et un rappel oscillant entre 90 % et 100 %. Nous proposons d'utiliser une telle approche comme point de comparaison.

Nous explorons par la suite l'impact de la variation des différents paramètres de cette approche sur les résultats : tailles des n-grammes, mesures de similarité, structures de données et normalisation des éléments, afin de sélectionner les configurations donnant les meilleurs résultats pour chacun des types dérivation représentés par nos corpus.

F.1 Taille des n-grammes

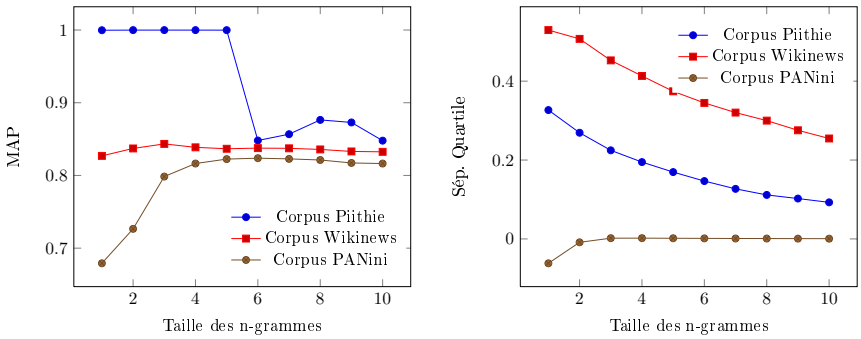
Nous évaluons les différentes configurations de la taille des n-grammes, de 1 à 10 mots, pour chaque corpus. Nous appliquons dans ce but les métriques de l'évaluation décrites en section 4.3.2. Nous fixons pour le moment la mesure de similarité, la *resemblance* (*cf. Équation 2.1*), et la structure de données, un ensemble.

F.1.1 Qualité de la classification

Nous employons la MAP pour mesurer la qualité de l'identification des liens de dérivation (*cf. Section 4.3.2*). Plus la MAP est élevée, meilleure est la classification. Le graphique F.1(a) montre l'évolution de la MAP selon la taille des n-grammes pour nos trois corpus.

Notre première observation est que les courbes des trois corpus se comportent différemment. Nous en concluons que la détection des différentes formes de dérivation que ces corpus illustrent nécessitent des approches différentes, au moins dans leur paramétrage.

La courbe du corpus Piithie avoisine 1 pour les plus petites tailles puis chute brutalement à environ 0,85. Nous pensons que cet excellent score pour les n-grammes de



(a) MAP (b) Sép. Quartile

FIGURE F.1 – Variation de la détection de dérivation pour l'approche par signature complète selon la taille des n-grammes sur nos trois corpus.

petite taille s'explique par la stabilité du contenu des articles de presse : les journalistes reprennent essentiellement les informations de la dépêche et la complètent peu. Toutefois, ils adaptent le texte à leur style (ou à la charte éditoriale) ce qui entraîne de légères réécritures qui font diverger les n-grammes de plus grande taille.

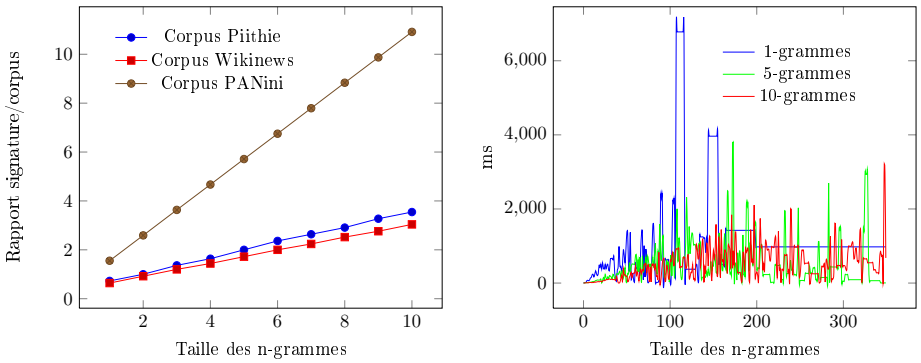
La courbe du corpus Wikinews est globalement stable, oscillant entre un minimum de 0,826 pour les unigrammes et de 0,843 pour les trigrammes. La constance de la qualité de classification quelque soit la taille des n-grammes utilisée nous paraît concordante avec la forme de dérivation du corpus : les révisions. La dérivation entre la version originale et les différentes révisions ne porte bien souvent que sur des parties localisées, la majorité de la forme textuelle restant inchangées. Contrairement à ce que nous attendions, ces faibles variations n'en font pas des liens de dérivation aussi facilement détectables qu'il n'y paraît. Nous supposons que ce score assez bas s'explique par un effet de dilatation du texte entre la première version et les dernières : les parties communes entre ces révisions se recouvrent largement mais les éventuels ajout ou retraits de texte sont sanctionnés par la mesure de *resemblance*.

Enfin, la courbe du corpus PANini a le comportement le plus classique : la MAP croît avec la taille des n-grammes jusqu'à un pallier stable à partir des 5-grammes. La plus mauvaise performance est obtenue avec les unigrammes (0,679) tandis que le pallier oscille entre 0,817 pour les 9-grammes et 0,823 pour les 6-grammes. L'approche par 6-grammes semble donc la plus intéressante au seul regard de la MAP.

F.1.2 Capacité de discrimination

Nous employons la séparation des quartiles pour mesurer la capacité de discrimination entre les dérivés et les non-dérivés (*cf. Section 4.3.2*). Plus elle est élevée, mieux sont discriminés les dérivés des non-dérivés. La graphique F.1(b) montre l'évolution de la séparation des quartiles selon la taille des n-grammes pour nos trois corpus. Nous pouvons observer que les courbes des corpus Piithie et Wikinews se comportent de manière similaire, alors que la courbe du corpus PANini à une allure différente.

Les courbes des corpus Piithie et Wikinews sont comparables : leur maximum est atteint pour les unigrammes puis elle décroissent avec la taille des n-grammes. Ainsi, les maximums sont respectivement de 0,326 et 0,529 pour les unigrammes et de 0,092 et 0,254 pour les 10-grammes. En d'autres termes, le recouvrement de simples mots permet mieux de séparer les dérivés des non-dérivés que des séquences de mots. Nous l'expliquons par le fait que les textes de ces deux corpus se recouvrent énormément. Les scores de similarité avec les dérivés sont donc très élevés tandis que les scores des



(a) Taille des signatures (b) Temps de traitement

FIGURE F.2 – Évolution du coût de la méthode pour les différentes tailles de n-grammes en fonction de la taille des signatures (a) et du temps de traitement (b).

non-dérivés sont beaucoup plus bas. L'utilisation de n-grammes réduit mécaniquement le taux de recouvrement à cause des quelques réécritures ce qui fait baisser les scores des dérivés tout en maintenant ceux des non-dérivés, réduisant l'espace entre ces deux classes.

La courbe du corpus PANini est beaucoup plus basse, pointant les limites de la signature complète pour des dérivations plus difficiles. Le minimum de la courbe, obtenu pour les unigrammes, est négatif (−0,061), les valeurs oscillant autour de 0 entre −0,008 pour les bigrammes et 0,002 pour les 4-grammes. Ces scores indiquent que les dérivés et non-dérivés ont tendance à se mélanger au sein du classement (séparation négative). Le niveau plus élevé de difficulté du corpus PANini s'illustre dans ces chiffres, alors que les résultats étaient comparables pour la MAP. Cette observation montre la nécessité d'une évaluation multicritère des méthodes de détection de dérivation.

F.1.3 Coût de la méthode

Le graphique F.2(a) montre l'évolution du rapport entre la taille de la signature et celle du corpus. Les tailles utilisées pour le calcul correspondent à l'espace disque occupé en Mo [1]. Les valeurs tiennent compte des méta-informations associées à chaque texte dans le cas des corpus. Bien qu'elles soient d'une taille négligeable au regard de celle des textes. Les résultats observés sont ceux attendus. Plus la taille des n-grammes est élevée, plus les signatures sont volumineuses (le ratio est supérieur à 1). Nous associons la différence de valeur entre les courbes de Piithie et de Wikinews et la courbe de PANini à la façon dont les tailles sont mesurées. En effet, nous utilisons l'utilitaire *du* qui calcule l'espace disque occupé et non la taille des fichiers. La taille supérieure de PANini s'explique par son nombre supérieur de fichiers qui entraîne un arrondi plus important.

Le graphique F.2(b) montre le temps de traitement nécessaire à la comparaison des signatures selon leur taille. La taille en abscisse correspond au nombre de caractères traités par tranche de 20 000, les nombres en ordonnées sont obtenus en faisant la moyenne des temps de de comparaison pour les lots dont la taille en caractère correspond à la taille en abscisse et pour les trois corpus confondus. Ces temps de

1. À titre informatif, les tailles des corpus Piithie, Wikinews et PANini sont respectivement de 11 Mo, 25 Mo et 188 Mo

traitement sont une approximation puisqu'ils sont calculés comme les temps écoulés sur le système et ne correspondent donc pas au temps réel. Ceci explique le flottement des courbes que l'on peut observer qui ne montrent pas de différence nette pour les différentes tailles de n-grammes.

Synthèse

Le corpus Piithie représente les dérivations telles qu'opérées par les journalistes depuis les dépêches d'agence ou les articles de leurs collègues. Nous avons pu observer que les petites tailles de n-grammes (1 à 5) donnent indifféremment la même qualité de classement. Toutefois, la discrimination entre les dérivés et non-dérivés est meilleure pour les unigrammes. Ce dernier choix nous semble donc le mieux approprié pour identifier les dérivations de ce type.

Le corpus Wikinews représente les dérivations de type révision. Nous avons pu observer que la taille des n-grammes n'a pas d'impact réel sur la qualité du classement. De même que pour Piithie, les unigrammes permettent toutefois d'obtenir la meilleure discrimination entre les dérivés et non-dérivés ce qui en fait le meilleur choix pour ce type de dérivation.

Le corpus PANini présente des dérivations beaucoup plus variées et difficiles à identifier que les précédentes. Les n-grammes de grande taille donnent les meilleures classifications, le maximum étant atteint avec les 6-grammes. Du point de vue de la discrimination, tous les n-grammes d'une taille supérieure à un sont virtuellement équivalents. Les 6-grammes semblent donc les meilleurs descripteurs à ce stade pour le corpus PANini.

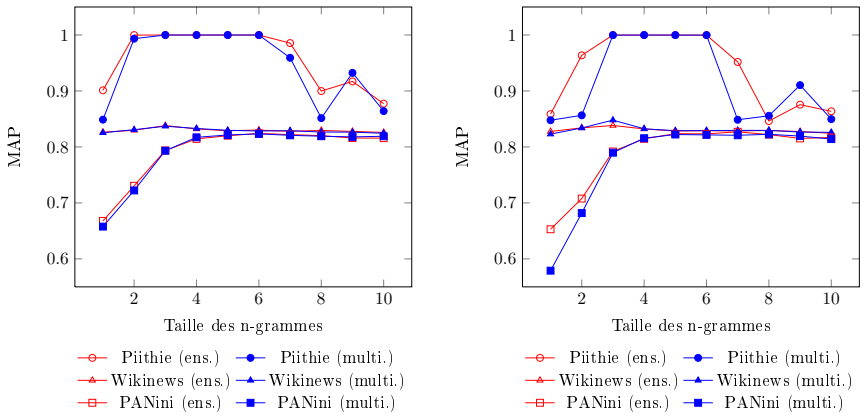
Nous avons observé que les tailles des n-grammes provoquent une augmentation de la taille des signatures, ce que nous attendions. Notre méthode de mesure ne permet de rendre compte de différences dans le temps de traitement pour ces différentes tailles.

F.2 Mesures de similarité et modèles

Nous explorons dans cette section l'impact du choix de la mesure de similarité utilisée pour comparer les signatures et du modèle utilisé pour représenter les signatures. Nous avons expérimenté chacune des cinq mesures de similarité ($c(ci, ca)$, $c(ca, ci)$, $c_{max}(ci, ca)$, $c_{min}(ci, ca)$, (*cf. Section 4.4.1*)) sur nos trois corpus et pour les différentes tailles de n-grammes. Nous y avons combiné les deux modèles de données : l'ensemble et le multiensemble.

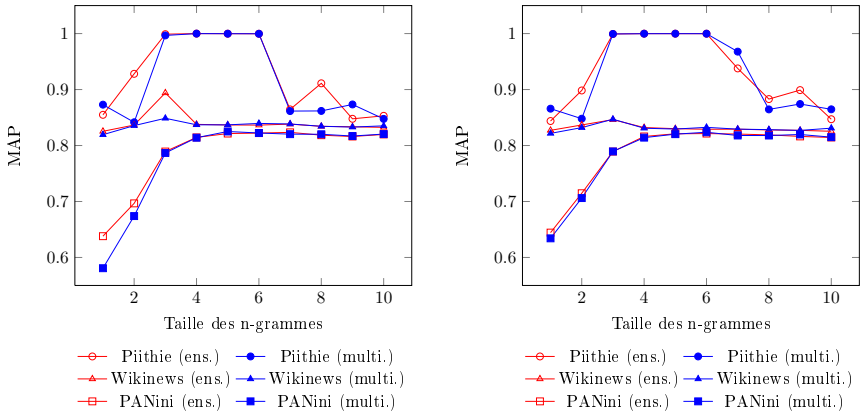
F.2.1 Mesures de similarité

Les graphiques de la figure F.3 montrent l'évolution de la MAP selon la taille des n-grammes pour les différentes mesures de similarité expérimentées. Nous observons que les différentes mesures de similarité ont un impact minime sur la qualité de la classification. Nous notons tout de même quelques légères différences. Pour le corpus Piithie, le pallier maximal démarre dès les bigrammes avec la mesure c_{max} et s'étend jusqu'au 7-grammes pour $c(ca, ci)$ et la *resemblance*. Le choix de la mesure de similarité fait donc glisser le pallier selon la taille des n-grammes. Pour le corpus Wikinews, les mesures de similarité n'ont virtuellement aucun impact à l'exception de $c(ci, ca)$ qui permet de faire bondir la MAP à presque 0,9 pour les trigrammes, soit une augmentation de 0,08 points. Pour le corpus PANini, le choix des mesures de similarité n'influent que pour les n-grammes de petites tailles pour lesquels c_{max} permet d'augmenter significativement la MAP.



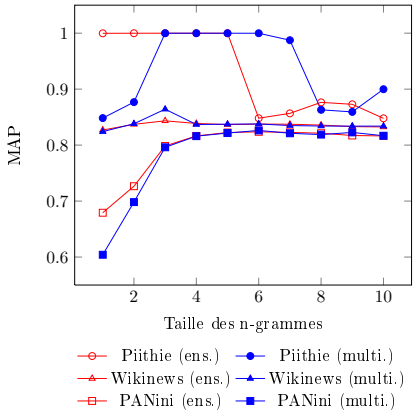
(a) Containment c_{max}

(b) Containment c_{min}



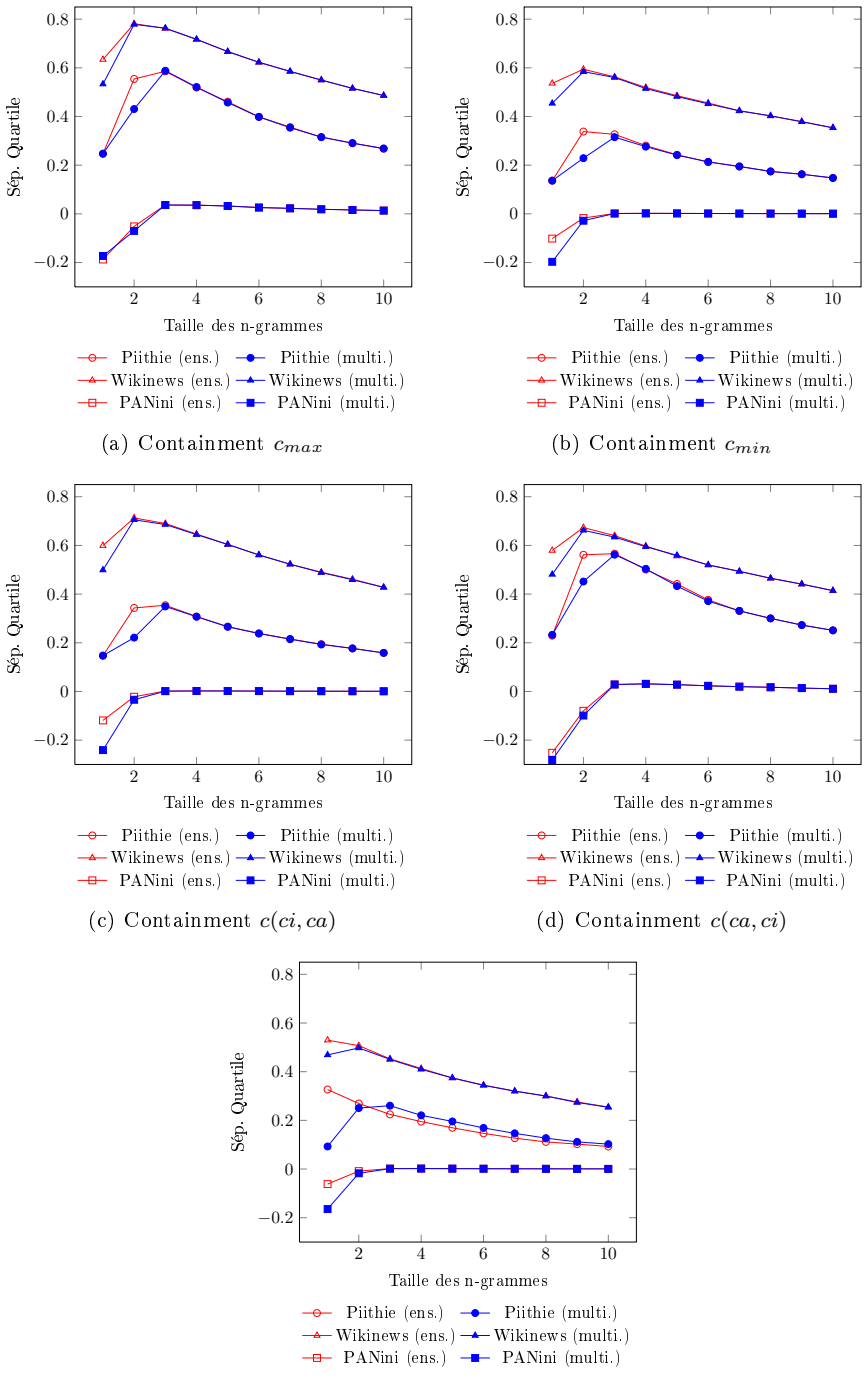
(c) Containment $c(ci, ca)$

(d) Containment $c(ca, ci)$



(e) Resemblance

FIGURE F.3 – Comparaison de l'évolution de la MAP pour les différents corpus et pour les différentes mesures de similarité selon la considération ensembliste ou multiensembliste. Afin de faciliter la lecture des courbes, les graphiques sont à la même échelle pour chacune des mesures d'évaluation.



(a) Containment c_{max}

(b) Containment c_{min}

(c) Containment $c(ci, ca)$

(d) Containment $c(ca, ci)$

(e) Resemblance

FIGURE F.4 – Comparaison de l'évolution de la séparation des quartiles pour les différents corpus et pour les différentes mesures de similarité selon la considération ensembliste ou multiensembliste. Afin de faciliter la lecture des courbes, les graphiques sont à la même échelle pour chacune des mesures d'évaluation.

Les graphiques de la figure F.4 montre l'évolution de la séparation des quartiles selon la taille des n-grammes pour les différentes mesures de similarité expérimentées. Nous observons que, contrairement à la MAP, les mesures de similarité ont un impact réel sur la capacité de discrimination. La mesure de *resemblance*, utilisée jusqu'à présent, donne les plus mauvais résultats tandis que la mesure c_{max} se détache nettement en offrant la meilleure MAP pour les trois corpus. La mesure $c(ci, ca)$ donne des résultats presque aussi bons pour Piithie et Wikinews ce qui est logique puisqu'elle mesure exactement la configuration « source–candidat »que nous cherchons à caractériser. Cependant, la mesure $c(ca, ci)$ est la plus proche de c_{max} pour le corpus PANini. Nous l'expliquons par la granularité partielle de ce corpus qui fait que les éléments en communs entre source et candidat ne correspondent bien souvent qu'à une petite partie de la source et il est donc préférable de calculer la proportion par rapport aux éléments du candidat plutôt que ceux de la source.

En résumé, toutes les mesures de similarité offrent une qualité de classification similaire mais c_{max} améliore nettement la capacité de discrimination de la méthode.

F.2.2 Modèles

Les graphiques de la figure F.3 illustrent également l'évolution de la MAP pour les deux choix de modèles dont nous avons discuté. Nous observons que les deux modèles obtiennent des résultats virtuellement identiques pour les n-grammes de taille suffisante (à partir des 4-grammes) ce qui s'explique par l'absence de redondance des n-grammes suffisamment grands. Cependant, cette observation est fausse dans le cas de Piithie pour lequel les résultats varient pour les n-grammes de grande taille (à partir des 6-grammes). Le modèle ensembliste est globalement meilleur que le multiensembliste justifiant le choix en faveur du premier dans la littérature. Les variations sont les plus appuyées pour le corpus Piithie. L'approche ensembliste se détache légèrement pour les n-grammes de petite taille excepté pour les mesures $c(ci, ca)$ et $c(ca, ci)$ pour lesquels les unigrammes sont plus performants rassemblés dans un multiensemble que dans un ensemble. La différence pour ces mêmes n-grammes et toutefois clairement en faveur de l'approche ensembliste pour la *resemblance*. Les variations observées sur Wikinews sont extrêmement légères à l'exception des trigrammes pour $c(ci, ca)$ et la *resemblance*. L'approche ensembliste est meilleure dans le premier cas mais moins bonne dans le second. Enfin, seules les n-grammes de petites tailles (unigrammes et bigrammes) sont impactées pour le corpus PANini. Le choix de la modélisation ensembliste l'emportant dans ces configurations.

Les graphiques de la figure F.4 illustrent l'évolution de la séparation des quartiles pour les deux modèles. À l'exception de la *resemblance* pour Wikinews, le choix du modèle n'a virtuellement aucun impact sur la capacité de discrimination pour les n-grammes suffisamment grands (trigrammes). Pour les plus petits n-grammes, les résultats sont en faveur de l'approche ensembliste. Nous noterons que pour les mesures de type *containment*, le maximum est obtenu pour les bigrammes ou trigrammes alors qu'il est atteint pour les unigrammes pour la *resemblance*, PANini excepté.

En résumé, si les deux modèles donnent globalement des résultats comparables, le modèle ensembliste dépasse le modèle multiensembliste pour les n-grammes de petite taille.

Synthèse

Nos expérimentations ont montré que les différentes mesures de similarité impactent peu la qualité de classification des liens de dérivations représentés par nos corpus. Toutefois, les mesures de *containment*, et plus particulièrement c_{max}, offrent

une meilleure capacité de discrimination dans la majorité des cas. Le choix d'une modélisation ensembliste ou multiensembliste a peu d'incidence sur les résultats de la signature complète pour des n-grammes de tailles suffisantes mais l'approche ensembliste semble préférable pour les n-grammes de petites tailles (notamment pour Piithie et PANini). Nous observons également que ces paramètres peuvent entraîner un glissement du meilleur résultat vers une autre taille de n-gramme.

En conclusion, il nous semble qu'une modélisation ensembliste couplée à une mesure de similarité de type c_{max} soit la combinaison la plus pertinente dans tous les cas. C'est la configuration que nous retenons pour la suite de nos expérimentations où nous influons sur le contenu des n-grammes.

F.3 Normalisation des éléments de la signature

Le dernier paramètre contrôlable qui nous semble pouvoir influer sur les résultats est celui de la normalisation des éléments de la signature. Nous nous intéressons à deux formes de normalisation : le filtrage des mots outils et la racinisation.

F.3.1 Filtrage des mots outils

Nous avons utilisé les listes de mots outils créés par Savoy (1999) et librement disponibles [2]. Les recherches en attribution d'auteur montrent que les mots outils sont très caractéristiques du style (Stamatatos, 2009b), un filtrage trop appuyé pourrait faire disparaître des indices pertinents. Par conséquent, nous avons regroupé les mots outils des listes en quatre catégories : (i) les articles et prépositions, (ii) les pronoms et adjectifs, (iii) les adverbes et auxiliaires et (iv) les autres mots outils non grammaticaux.

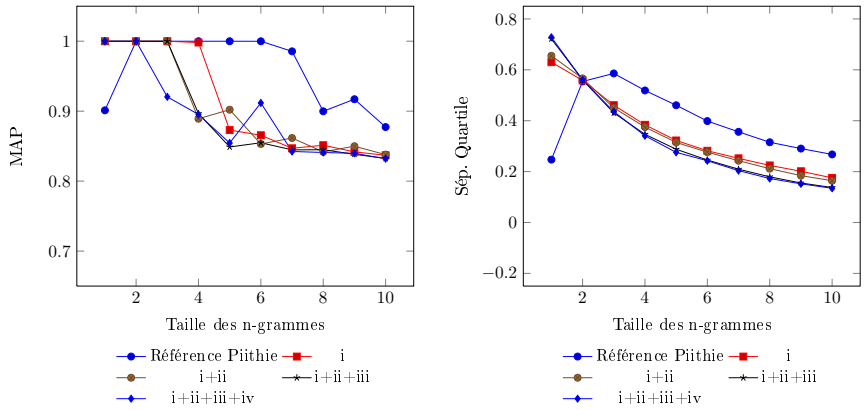
La figure F.5 montre l'impact sur la MAP et la séparation des quartiles du filtrage de ces différentes listes pour nos corpus. Nous pouvons constater tout d'abord que le filtrage des mots outils n'a aucun impact pour le corpus PANini, que ce soit en termes de qualité de la classification F.5(e) ou de capacité de discrimination F.5(f). Plus généralement, la variation de la taille des listes a un impact a peine notable sur l'ensemble des résultats, la plus petite liste donnant toutefois les meilleurs résultats. Le filtrage a un impact positif léger sur la qualité de classification pour les n-grammes de petite taille sur le corpus Wikinews et Piithie. Pour Piithie, cet avantage évolue rapidement vers un désavantage à partir des trigrammes. Le filtrage apporte un gain notable en terme de capacité de discrimination pour les unigrammes ce qui est logique puisque, comme présenté plus haut, les mots outils de par leur distribution uniforme n'ont aucun pouvoir discriminatif. Leur retrait pour les unigrammes réduit l'effet de nivellement qu'ils peuvent avoir. L'impact est cependant clairement négatif pour les autres tailles de n-grammes que ce soit pour Piithie ou Wikinews.

En résumé le filtrage des mots outils a un intérêt mitigé en termes de qualité de classification et de capacité de discrimination mais se justifie pour les unigrammes où il réhausse la capacité de discrimination.

F.3.2 Racinisation

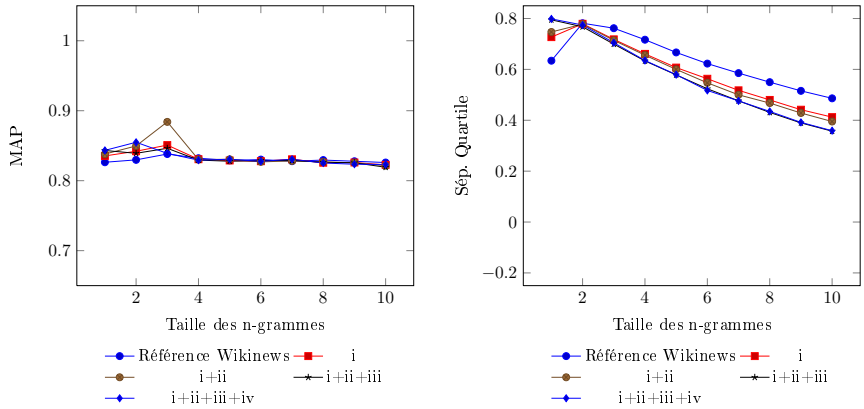
Nous utilisons le racinisateur *Snowball* qui a l'avantage de supporter plusieurs langues dont le français et l'anglais. Il offre également la possibilité de gérer l'agressivité de la racinisation, c-à-d le nombre maximum de transformations morphologiques opérées. Nous la faisons varier de 1 à 5 opérations morphologiques maximum.

2. `http://members.unine.ch/jacques.savoy/clef/`



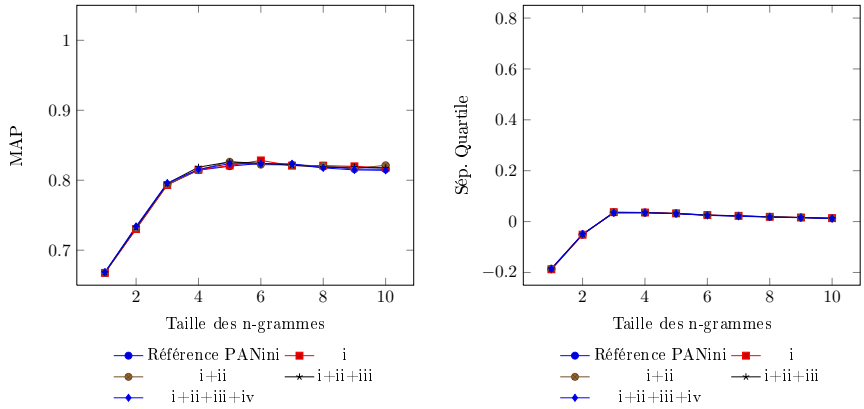
(a) MAP - Piithie

(b) Sép. Quartile - Piithie



(c) MAP - Wikinews

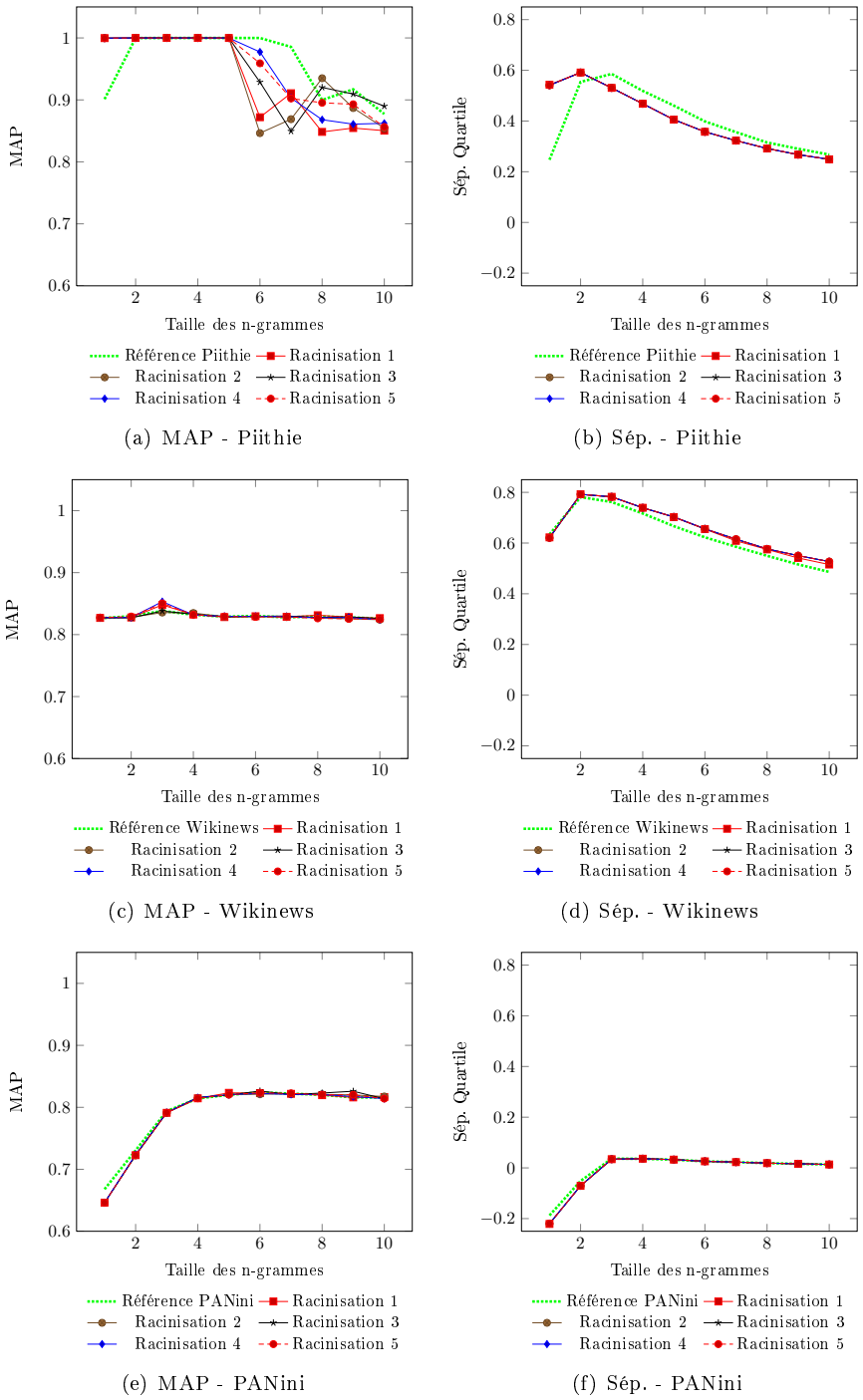
(d) Sép. Quartile - Wikinews



(e) MAP - PANini

(f) Sép. Quartile - PANini

FIGURE F.5 – Impact du filtrage des mots outils en terme de qualité de classification ((a), (c), (e)) et de capacité de discrimination ((b), (d), (f)).



(a) MAP - Piithie

(b) Sép. - Piithie

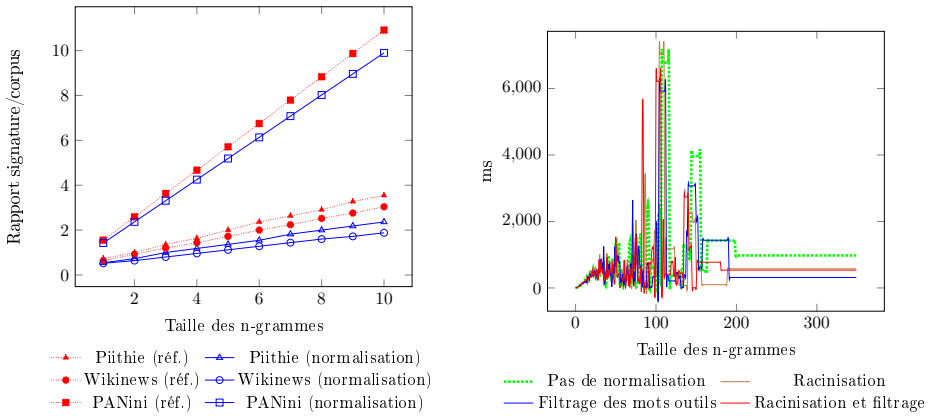
(c) MAP - Wikinews

(d) Sép. - Wikinews

(e) MAP - PANini

(f) Sép. - PANini

FIGURE F.6 – Comparaison de l'impact de différents niveaux d'agressivité de la racinisation sur l'évolution des métriques MAP ((a), (c), (e)) et Séparation Quartile ((b), (d), (f)) sur nos corpus.



(a) Coût de stockage (b) Coût de comparaison

FIGURE F.7 – Évolution du coût de stockage (a) et de comparaison (b).

Nos expérimentations rapportées dans la figure F.6 montrent que l'aggressivité de la racinisation a peu d'impact sur les résultats. Ainsi, la plupart des courbes de résultat de la figure F.6 sont confondues pour les différents niveaux de racinisation.

Pour le corpus PANini, la racinisation dégrade très légèrement les résultats pour les n-grammes de petites tailles mais n'a aucune incidence pour les autres. Pour le corpus Wikinews, nous pouvons noter un léger pic de la MAP pour les trigrammes. Nous pouvons également remarquer que la racinisation permet d'augmenter quelque peu la capacité de discrimination, toujours pour le corpus Wikinews, pour les bigrammes et plus. Enfin, pour le corpus Piithie, nous pouvons noter une nette amélioration pour les n-grammes de petite taille (unigrammes et bigrammes) pour la qualité de la classification et la capacité de discrimination. Au delà, la racinisation dégrade les résultats.

En résumé, la racinisation se justifie sur les n-grammes de petites tailles pour Piithie et pour toutes les tailles de n-grammes pour le corpus Wikinews. Elle permet alors d'améliorer la qualité de classification ainsi que la capacité de discrimination.

Synthèse

La normalisation (filtrage des mots outils et racinisation) a un impact mitigé sur les résultats de la détection en termes de MAP et de séparation des quartiles. Cependant, comme le montre la figure F.7, elle permet de réduire la taille des signatures ce qui se répercute positivement sur le coût de calcul des comparaisons.

Les dérivations de type révisions représentées par le corpus Wikinews semblent les plus sensibles à la normalisation. Nous pouvons notamment noter une légère amélioration de la qualité de classification pour les trigrammes avec le filtrage de certains mots outils ainsi qu'avec le procédé de racinisation. Il semble qu'au final ce pic soit représentatif et que les trigrammes soient véritablement la meilleure configuration pour le corpus Wikinews, notamment pour un certain niveau de normalisation (filtrage des mots outils i+ii et racinisation).

Pour les autres formes de dérivation représentées par les corpus Piithie et PANini, la normalisation a un impact sur l'utilisation des n-grammes de petite taille uniquement. Pour le corpus PANini cette amélioration ne permet pas d'obtenir de meilleurs résultats que ceux obtenus pour les n-grammes de grande taille. La normalisation se justifie pour le gain en termes de coût de stockage et d'exécution puisqu'elle n'entraîne

Corpus	Éléments	Filtrage	Racinisation	MAP	Sép. Quartile
Piithie	unigrammes	i+ii+iii+iv	1	0,999	0,708
Wikinews	bigrammes	i+ii	5	0,872	0,800
PANini	6-grammes	i+ii+iii+iv	1	0,823	0,024

TABLE F.1 – Approches de références pour nos différents corpus.

pas de dégradation. Pour le corpus Piithie, le filtrage de tous les mots outils et la racinisation apporte une sensible amélioration. Celle-ci décale la meilleure configuration vers des n-grammes de plus petite taille.

En ce qui concerne le coût de traitement, nous n'avons considéré dans la figure F.7(b) que le coût de comparaison. Il faut ajouter à celui-ci le coût de l'extraction des signatures qui augmente du fait des traitements nécessaires à la normalisation. Toutefois, l'extraction n'est réalisée qu'une seule fois tandis que la comparaison est opérée à chaque fois qu'un nouveau candidat se présente. Au final, le surcoût lors de l'extraction est compensé par le gain pour la comparaison, pour un nombre suffisant de comparaisons.

F.4 Conclusion

Le tableau F.1 fait la synthèse du paramétrage des approches de référence retenues et qui nous servirons de point de comparaison pour nos propositions.

Clough (2003a) s'est intéressé au paramétrage des signatures complètes sur le corpus METER, un corpus comparable à Piithie mais en anglais, notamment le filtrage des mots outils et la racinisation. Ces travaux concluaient que le *containment* entre les unigrammes des textes était l'approche la plus performante et que les différentes normalisations n'amélioraient pas les résultats. Nos conclusions sont similaires pour le français (Piithie uniquement). Toutefois, si la normalisation n'a pas d'impact sur les résultats en termes de classification, elle en a un sur le coût de stockage et d'exécution.

La principale conclusion que nous pouvons tirer de ces expérimentations est que, comme nous le pensions, la dérivation est un phénomène hectoplasmique et par conséquent chaque forme de dérivation réagit différemment à différentes approches de détection.

Annexe G

Détail des résultats de nos propositions

G.1 Exploitation des n-grammes rares

G.1.1 Corpus Piithie

Les courbes de la figure G.1 montrent que les signatures composées d'hapax majorent globalement l'approche de référence. Pour rappel, celle-ci consiste en une signature complète à base d'unigrammes construit sur un corpus normalisé (racinisation et filtrage des mots outils). Les courbes de la MAP (G.1(a)) de l'approche de référence et de l'approche hapax avec normalisation se confondent parfois mais l'approche sans normalisation se détache nettement. Ainsi la qualité de la classification est supérieure ou égale pour toutes les tailles de n-grammes à l'exception des unigrammes.

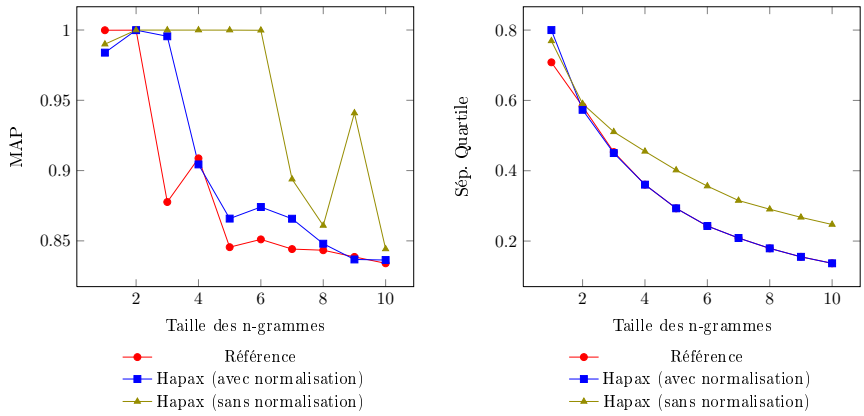
À l'opposé, la capacité de discrimination (G.1(b)) des unigrammes hapax est la plus élevée. Pour les autres tailles de n-grammes, l'approche de référence et les hapax avec normalisation sont équivalentes et majorées par les hapax sans normalisation.

Comme le rappelle le tableau G.1, les unigrammes de l'approche de référence permettent d'obtenir une MAP de 0,999 associée à une Sép.Q de 0,708. Les deux approches par hapax permettent d'égaler ou de dépasser, chacun de ces scores individuellement mais pas conjointement.

Nous observons deux principaux phénomènes liés à l'utilisation des n-grammes hapax pour la détection des dérivations sur le corpus Piithie : (i) l'approche sans normalisation donne de meilleurs résultats et (ii) les résultats avec les unigrammes sont inférieurs à ceux de l'approche de référence. En ce qui concerne les meilleurs résultats

		Éléments	MAP	Sép.Q.
Approche de référence		unigrammes	0,999	0,708
Hapax avec normalisation	max(MAP)	bigrammes	0,999	0,573
	max(Sép.Q)	unigrammes	0,983	0,8
Hapax sans normalisation	max(MAP)	bigrammes	0,999	0,590
	max(Sép.Q)	unigrammes	0,989	0,769

TABLE G.1 – Comparaison des meilleurs résultats pour l'approche par Hapax par rapport à l'approche de référence pour Piithie.



(a) MAP (b) Sép. Q

FIGURE G.1 – Évaluation de la sélection des éléments rares sur le corpus Piithie.

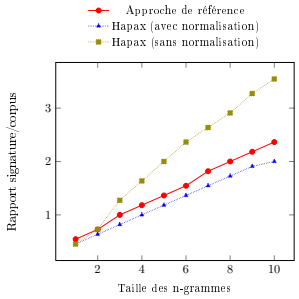


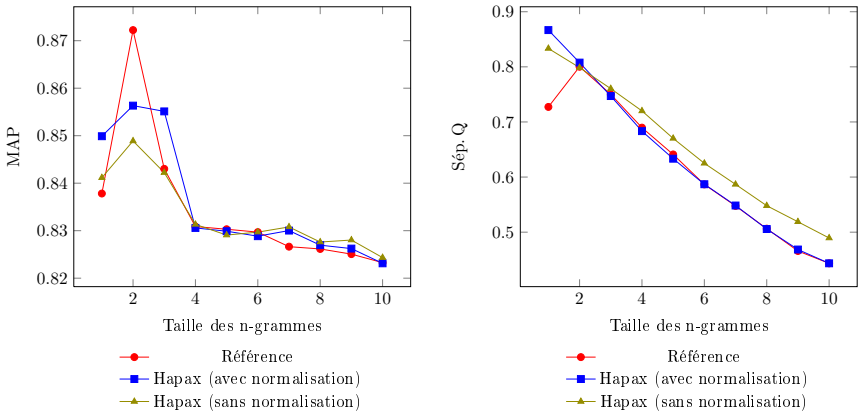
FIGURE G.2 – Évolution du coût de stockage pour la méthode hapax.

de l'approche sans normalisation, la taille des signatures nous semble en cause. La normalisation réduit le nombre de variations et donc d'hapax. En ce qui concerne les moins bonnes performances des unigrammes, nous pensons que deux causes y contribuent. Premièrement, certains passages de la source qui sont repris dans un dérivés ne contiennent pas d'hapax. Par conséquent, il n'y a pas de correspondance entre les signatures et la similarité est donc nulle. Deuxièmement, on retrouve parmi les unigrammes hapax des mots communs tels que « assure-t-il » et « ajoute-t-il » qui augmentent artificiellement la similarité entre des textes non dérivés. Si le découpage en mot est partiellement fautif, ces constructions ne correspondent pas à l'idée d'élément spécifiques au texte duquel ils proviennent. Le corpus de référence est ici en cause.

En résumé, la sélection des hapax ne permet pas d'obtenir de meilleurs résultats que l'approche de référence pour le corpus Piithie. Toutefois, si la MAP et la Sép.Q sont conjointement inférieures, elles restent comparables alors que le coût de la méthode est nettement moins élevé. La sélection des hapax ramène notamment le coût de stockage des signatures sans normalisation à celui des signatures avec normalisation invitant à l'économie des traitements de normalisation.

G.1.2 Corpus Wikinews

Les courbes de la figure G.3 montrent que les résultats obtenus par l'utilisation des hapax sur le corpus Wikinews, avec et sans normalisation, ne se détachent pas



(a) MAP (b) Sép. Q

FIGURE G.3 – Évaluation de l'approche par hapax sur le corpus Wikinews.

		Éléments	MAP	Sép. Q.
Approche de référence		bigrammes	0,872	0,800
Hapax avec normalisation	max(MAP)	bigrammes	0,856	0,807
	max(Sép. Q)	unigrammes	0,849	0,866
Hapax sans normalisation	max(MAP)	bigrammes	0,848	0,798
	max(Sép. Q)	unigrammes	0,841	0,833

TABLE G.2 – Comparaison des meilleurs résultats pour l'approche par Hapax par rapport à l'approche de référence pour Wikinews.

réellement de l'approche de référence. Les résultats en termes de qualité de classification (G.3(a)) sont d'ailleurs confondus pour les n-grammes de taille supérieure ou égale à quatre. Cette approche ne permet pas de faire mieux que le maximum obtenu par les bigrammes de l'approche de référence même si la différence se joue à quelques centièmes comme le montre le tableau G.2. La baisse de la qualité de la classification entre l'approche de référence et l'approche hapax avec la même normalisation est très légère. Les résultats des unigrammes et des trigrammes sont quant à eux très légèrement supérieurs à l'approche de référence.

Les résultats en termes de capacité de discrimination (G.3(b)) sont meilleurs en l'absence de normalisation et pour les unigrammes avec normalisation.

La sélection des hapax a un intérêt limité pour les révisions. Le filtrage statistique réduit très peu la taille des signatures : la plupart des éléments de la signature de l'approche de référence sont des hapax. L'allure de la courbe de la figure G.4 correspondant aux hapax avec normalisation le confirme, celle-ci rejoint la courbe de référence très rapidement. La différence entre les résultats s'explique par l'augmentation des scores nuls dûs à la suppression des quelques éléments en communs qui ne sont pas des hapax. Les quelques éléments des signatures en commun qui permettent de passer d'un score nul à un score non nul (aussi infime soit-il) suffisent à gagner les quelques centièmes manquants.

En résumé, l'utilisation des hapax pour le corpus Wikinews n'a que très peu d'apport par rapport à l'approche de référence. Cette dernière est très légèrement meilleure en termes de MAP et les deux approches sont équivalentes en termes de Sép. Q pour

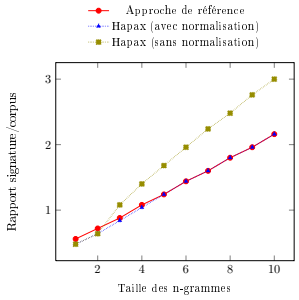
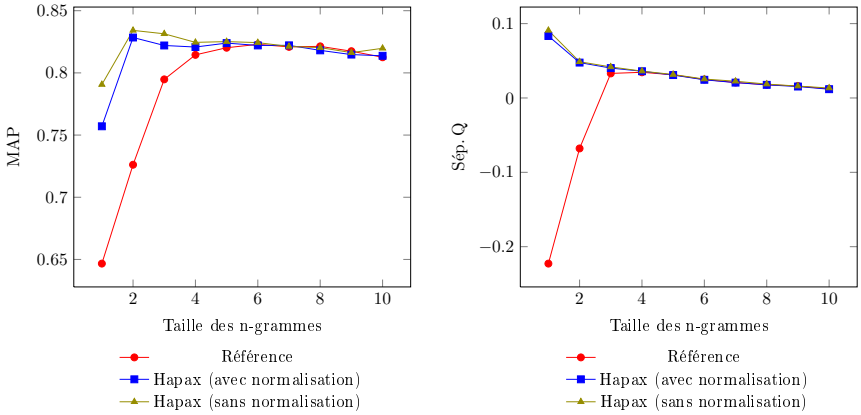


FIGURE G.4 – Évolution du coût de stockage pour la méthode hapax et pour le corpus Wikinews.



(a) MAP (b) Sép. Q

FIGURE G.5 – Évaluation de l'approche par hapax sur le corpus PANini.

les bigrammes.

G.1.3 Corpus PANini

Les résultats obtenus avec les hapax sur PANini se distinguent réellement de ceux de l'approche de référence, contrairement aux corpus précédents où les tendances de l'approche de référence et par sélection des hapax étaient globalement similaires. Les courbes de la figure G.5 montrent deux phénomènes. En premier lieu, la MAP (G.5(a)) a clairement augmenté pour les n-grammes de petites tailles (unigrammes à 4-grammes). Ainsi, la MAP progresse d'environ 0,1 points de sorte que la valeur pour les bigrammes (0,834) soit supérieure à la performance mesurée pour l'approche de référence (0,823). Elle reste similaire pour les n-grammes de plus grandes tailles. La normalisation dégrade légèrement les résultats de sorte que les résultats sans normalisation sont légèrement meilleurs. En second lieu, la courbe de la séparation des quartiles (G.5(b)) s'est inversée puisque la meilleure discrimination est obtenue pour les unigrammes alors que pour l'approche de référence il s'agit du plus mauvais score. Les unigrammes obtiennent ainsi une Sép.Q supérieure de 0,06 points à l'approche de référence. Au final, comme le confirme le tableau G.3, l'approche par sélection des hapax obtient de meilleurs résultats que l'approche de référence : les unigrammes offrent la meilleure capacité de discrimination et les bigrames la meilleure qualité de classification. L'abscence de normalisation est généralement préférable.

		Éléments	MAP	Sép.Q.
Approche de référence		6-grammes	0,823	0,024
Hapax avec normalisation	max(MAP)	bigrammes	0,834	0,047
	max(Sép.Q)	unigrammes	0,757	0,083
Hapax sans normalisation	max(MAP)	bigrammes	0,834	0,048
	max(Sép.Q)	unigrammes	0,790	0,090

TABLE G.3 – Comparaison des meilleurs résultats pour l'approche par Hapax par rapport à l'approche de référence pour PANini.

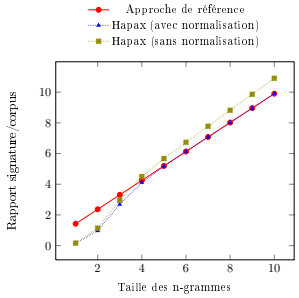
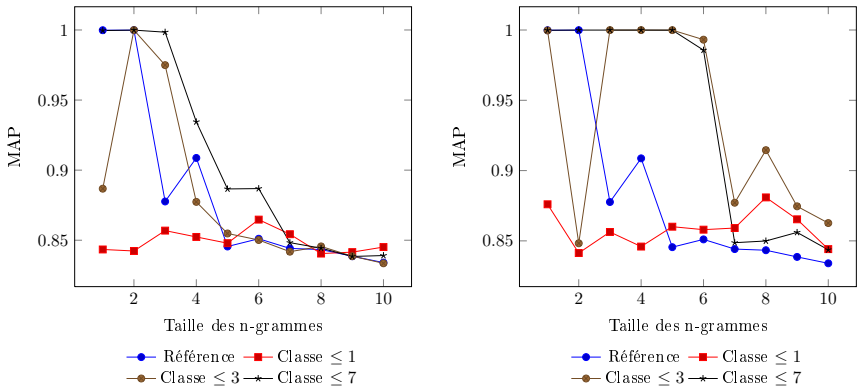


FIGURE G.6 – Évolution du coût de stockage pour la méthode hapax en fonction de la taille des n-grammes.

Les documents de PANini sont les plus volumineux et les dérivations qu'ils contiennent sont de granularité partielle contrairement aux deux autres corpus (*cf. Tableau 4.8 p. 120*). La combinaison de ces deux paramètres engendre des signatures très bruitées qui dégradent le score de similarité et donc mécaniquement les résultats de notre évaluation. Les n-grammes communs sont les premières causes de ces mauvais scores car ils trouvent des correspondances dans des textes non-dérivés. La sélection des hapax réduit le bruit occasionné en supprimant les n-grammes communs des signatures. Les courbes de la MAP (*cf. Figure G.5(a)*) soutiennent cette hypothèse. Les plus mauvais résultats des signatures complètes sont obtenus avec les n-grammes de petite taille car ceux-ci sont les plus à même d'être présents dans des textes non dérivés. À l'opposé, les n-grammes de grandes tailles sont très spécifiques aux textes et il est peu probable de les retrouver dans les textes candidats sans qu'il y ait de lien de dérivation. Les courbes de la figure G.6 se rejoignent pour les n-grammes de grande taille, ce qui confirme que ces derniers sont majoritairement des hapax et donc très spécifiques aux textes.

Les excellents scores obtenus pour les n-grammes de petites tailles par l'approche par sélection des hapax confirme que la suppression des éléments parasites améliore les performances. Il est intéressant de noter que la meilleure MAP est obtenue pour les bigrammes qui est également la meilleure configuration de l'approche de référence sur le corpus Wikinews. Cette taille de n-gramme semble particulièrement adaptée à la détection de dérivation. Les unigrammes ne permettent pas de capturer des éléments spécifiques et les n-grammes de plus grande taille ne sont pas conservés par les processus de dérivation impliquant beaucoup de réécritures. Les bigrammes semblent le meilleur compromis entre la spécificité par rapport au texte modélisé et l'invariance par rapport au processus de dérivation.

En résumé, l'approche par sélection des hapax a un réel intérêt pour le corpus



(a) MAP – avec normalisation (b) MAP – sans normalisation

FIGURE G.7 – Évaluation, en termes de MAP, de l'exploitation des n-grammes de plus fort poids informatif, par classes de scores de tf · idf, sur le corpus Piithie.

PANini. Elle obtient un score maximum pour la MAP et la Sép.Q supérieurs aux scores de l'approche de référence. De plus, elle provoque un glissement du maximum des 6-grammes pour l'approche de référence aux bigrammes, ce qui se traduit par un gain important du coût de stockage.

G.2 Exploitation des n-grammes de fort poids informatif

G.2.1 Corpus Piithie

Nous analysons dans cette section les résultats obtenus pour l'approche par sélection des n-grammes de plus fort poids informatif sur le corpus Piithie. Nous détaillons tout d'abord les résultats en termes de qualité de classification, mesurée par la MAP, et représentés par la figure G.7. Nous détaillons ensuite les résultats en termes de capacité de discrimination, mesurée par la Sép. Q, et représentés par la figure G.9. Finalement nous comparons les résultats des meilleurs configurations avec ceux de l'approche de référence.

G.2.1.1 Qualité de classification : évolution de la MAP

La figure G.7 montre l'évolution de la MAP en ordonnées selon les différentes tailles de n-grammes en abscisses, avec (G.7(a)) et sans normalisation du texte (G.7(b)). Nous observons trois comportements : (i) une stabilité basse de la MAP pour les classes ≤ 0, ≤ 1 et ≤ 2 ; (ii) une variation en dents de scie pour les classes ≤ 3 et ≤ 4 ; (iii) une évolution de la MAP similaire à celle de l'approche de référence (*cf. Annexe F.4*) pour les classes supérieures. Nous ne rapportons dans la figure G.7 qu'un seul jeu de résultat pour chacun de ces comportements : (i) ≤ 1, (ii) ≤ 3 et (iii) ≤ 7.

Le premier type de comportement, représenté par les courbes décrivant la classe ≤ 1 , est la stabilité basse de la MAP. Les résultats sont stables et oscillent, indépendamment de la taille des n-grammes et de la normalisation, aux alentours de 0,85. Nous supposons que cette performance s'explique par le trop petit nombre d'éléments

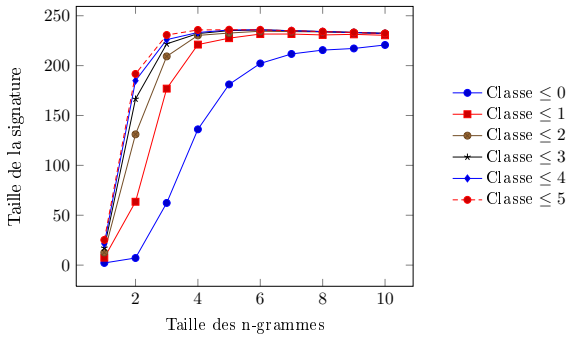


FIGURE G.8 – Nombre moyen d'éléments par signature pour les différents rangs expérimentés et pour le corpus Piithie.

composant les signatures (figure G.8). Le nombre moyen d'éléments par signatures illustré par la figure G.8 va dans le sens de cette hypothèse.

Le second type de comportement, représenté par les courbes décrivant la classe ≤ 3, est la variation de la MAP en dents de scie. La MAP varie fortement pour certaines tailles de n-grammes. Le pic maximum est obtenu pour les bigrammes avec normalisation. Le pic d'augmentation supporte notre hypothèse de l'adéquation de l'utilisation des bigrammes pour détecter les liens de dérivation. Nous ne pouvons toutefois pas expliquer l'effondrement constaté lors de l'absence de normalisation [1]. Il est d'autant plus étonnant que les n-grammes de tailles proches obtiennent d'excellents résultats.

Le troisième et dernier type de comportement, représenté par les courbes ≤ 7, est similaire à celui de l'approche de référence : un pallier maximum suivi d'une chute assez rapide. Avec ou sans normalisation, la MAP est meilleure qu'avec l'approche de référence, pour la plupart des n-grammes et la rejoint pour les 7-grammes et suivants. Comme le montre la figure G.8, la taille moyenne des signatures en fonction du rang converge assez rapidement. Les trois premières classes regroupent une grande partie des n-grammes car les courbes sont pratiquement confondues dès la classe ≤ 3. Nous pensons que pour les rangs supérieurs à 5, tous les éléments de la signature de référence qui sont pertinents y sont présents, ce qui explique ces bons résultats.

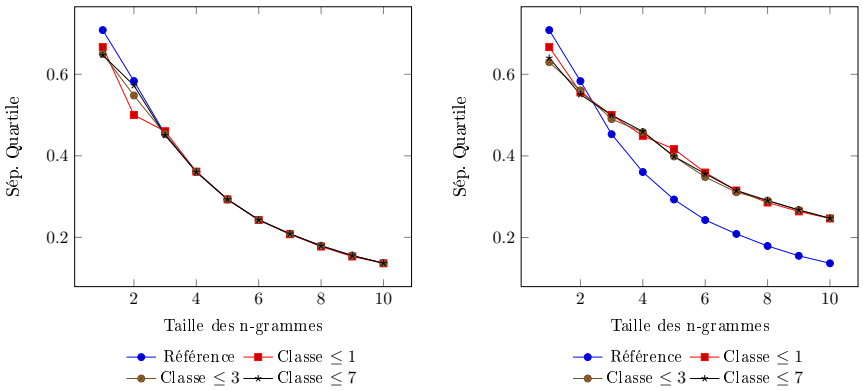
G.2.1.2 Capacité de discrimination : évolution de la Sép.Q

La capacité de discrimination de la signature exploitant les n-grammes de plus fort poids informatif est illustrée par les courbes de la figure G.9. L'évolution de la Sép. Q pour les signatures avec normalisation (G.9(a)) est similaire à celle de la signature complète. Toutefois, pour la signature sans normalisation (G.9(b)), les résultats sont légèrement inférieurs pour les n-grammes de taille inférieure ou égale à 3 et nettement supérieurs pour les autres.

G.2.1.3 Synthèse

Le tableau G.4 compare les meilleurs résultats de cette approche exploitant les n-grammes de plus fort poids informatif par rapport à l'approche de référence (*cf. Annexe F.4*). Il montre que la première ne permet pas d'obtenir des résultats aussi bon que ceux de l'approche de référence. Les configurations permettant d'obtenir la

1. Après vérification, nous pouvons confirmer que la chaîne de traitement a fonctionné correctement lors du calcul de ce résultat. Ce dernier ne semble donc pas résulter d'un bogue.



(a) Sép. Q – avec normalisation

(b) Sép. Q – sans normalisation

FIGURE G.9 – Évaluation, en termes de Sép. Q, de l'exploitation des n-grammes de plus fort poids informatif, par classes de scores de tf · idf, sur le corpus Piithie.

		Classes	Éléments	MAP	Sép.Q.
Approche de référence		NA	unigrammes	0,999	0,708
tf · idf avec	max(MAP)	≤ 7	bigrammes	0,999	0,572
normalisation	max(Sép.Q)	≤ 1	unigrammes	0,843	0,666
tf · idf sans	max(MAP)	≤ 7	bigrammes	0,999	0,550
normalisation	max(Sép.Q)	≤ 1	unigrammes	0,876	0,666

TABLE G.4 – Comparaison des meilleurs résultats pour l'exploitation des n-grammes de plus fort poids informatif par rapport à l'approche de référence pour Piithie.

meilleure qualité de classification, une MAP de valeur similaire à celle de l'approche de référence, offrent une moins bonne capacité de discrimination. Il est bien entendu nécessaire de relativiser ces résultats au regard à la fois de la faible différence des performances par rapport à l'approche de référence, mais également de la différence en termes de coût de stockage des signatures qui est favorable à notre approche, comme l'illustre la figure G.10. Les signatures exploitant les n-grammes de plus fort poids informatif ont un coût de stockage moindre que l'approche de référence lorsque le texte est normalisé. En l'absence de normalisation, la méthode est plus coûteuse à partir d'une certaine taille de n-grammes définie par la classe retenue (4-grammes pour ≤ 1, bigrammes pour ≤ 10).

En conclusion, l'approche par signature composée des n-grammes les plus représentatifs permet d'obtenir des performances équivalentes à l'approche de référence en termes de qualité de classification, mais en deça en termes de capacité de discrimination. Ces performances comparables sont toutefois obtenus pour des signatures de plus petite taille.

G.2.2 Corpus Wikinews

Nous analysons dans cette section les résultats obtenus pour l'approche par sélection des n-grammes de plus fort poids informatif sur le corpus Wikinews. Nous détaillons tout d'abord les résultats en termes de qualité de classification, mesurée par la MAP, et représentés par la figure G.11. Nous détaillons ensuite les résultats

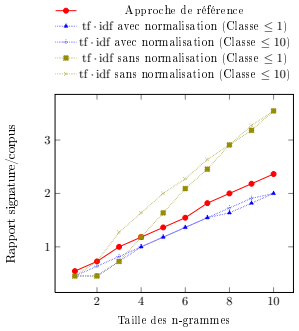
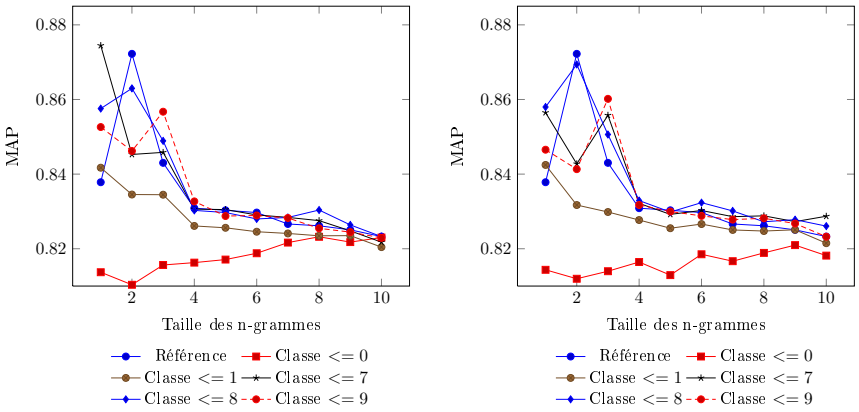


FIGURE G.10 – Évolution du coût de stockage de l'approche exploitant les n-grammes de plus fort poids informatif pour Piithie.



(a) MAP – avec normalisation

(b) MAP – sans normalisation

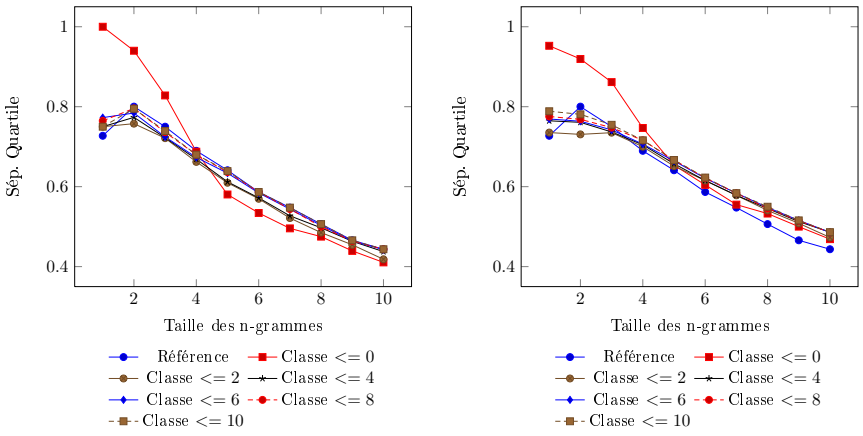
FIGURE G.11 – Évaluation, en termes de MAP, de l'exploitation des n-grammes de plus fort poids informatif, par classes de scores de tf · idf, sur le corpus Wikinews.

en termes de capacité de discrimination, mesurée par la Sép. Q, et représentés par la figure G.12. Finalement nous comparons les résultats des meilleurs configurations avec ceux de l'approche de référence.

G.2.2.1 Qualité de classification : évolution de la MAP

La figure G.11 montre l'évolution de la MAP en ordonnées selon les différentes tailles de n-grammes en abscisses, avec (G.11(a)) et sans normalisation du texte (G.11(b)). Nous observons cinq comportements : (i) stables et bas pour ≤ 0, (ii) décroissants monotones pour ≤ 1, ≤ 2, ≤ 3 et globalement décroissants à l'exception d'un pic avec (iii) les unigrammes pour les rangs ≤ 5 et ≤ 7, (iv) les bigrammes pour les rangs ≤ 6, ≤ 8 et ≤ 10, ou encore (v) les trigrammes pour ≤ 9. Afin de clarifier les graphiques de résultats, la figure G.11 ne montre qu'un seul jeu de résultat par comportement, soient : (i) ≤ 0, (ii) ≤ 1, (iii) ≤ 7, (iv) ≤ 8 et (v) ≤ 9.

Le premier type de comportement, représenté par la courbe décrivant la classe ≤ 0, est la stabilité basse de la MAP. Celle-ci est stable et oscille, indépendamment de la taille des n-grammes et de la normalisation, aux alentours de 0,82. Nous supposons que cette performance s'explique, comme pour le corpus Piithie, par le trop petit nombre d'éléments correspondant à ces classes, ce qui entraîne une signature trop



(a) Sép. Q – avec normalisation

(b) Sép. Q – sans normalisation

FIGURE G.12 – Évaluation, en termes de Sép. Q, de l'exploitation des n-grammes de plus fort poids informatif, par classes de scores de tf · idf, sur le corpus Wikinews.

peu représentative du texte.

Le deuxième type de comportement, représenté par la courbe décrivant la classe ≤ 1, est la décroissance monotone de la MAP. Le maximum est atteint pour les unigrammes et décroit avec l'augmentation de la taille des n-grammes. Le score obtenu par les unigrammes est le seul supérieur à celui de l'approche de référence pour ces classes. Nous pensons que cette performance s'explique également par le trop petit nombre d'éléments dans les signatures. Les unigrammes font exception car, contrairement aux autres n-grammes, ils offrent moins de variation et par conséquent la probabilité de trouver des correspondances est plus élevée.

Les trois derniers types de comportement, représentés par les courbes décrivant les classes ≤ 7, ≤ 8 et ≤ 9, correspondent à une décroissance ponctuée de pics correspondant aux maximums. La différence entre ces courbes réside dans la taille des n-grammes pour laquelle le pic a lieu : unigrammes, bigrammes ou trigrammes. On pourrait considérer comme similaires les courbes, telles que ≤ 7, où le pic se produit pour les unigrammes avec le deuxième type de comportement représenté par la courbe décrivant la classe ≤ 1. Cependant, nous pouvons observer sur la figure G.11(a) que le maximum est très nettement supérieur au reste de la courbe. Il dépasse d'ailleurs le maximum obtenu par l'approche de référence pour les bigrammes. Les maximums des autres courbes ne permettent pas d'atteindre un score aussi élevé. La valeur pour les bigrammes de la courbe décrivant la classe ≤ 8 sans normalisation frôle le maximum de l'approche de référence. On retrouve dans ces comportements, celui de l'approche pas signature complète où les résultats sont nettement meilleurs pour une taille de n-grammes particulière.

G.2.2.2 Capacité de discrimination : évolution de la Sép.Q

À l'exception des rangs ≤ 0, les résultats en termes de capacité de discrimination sont similaires à ceux de l'approche de référence. Nous ne traçons dans la figure G.12 qu'une courbe sur deux à des fins de lisibilité. Comme pour le corpus Piithie, nous pouvons observer un comportement légèrement différent de la courbe décrivant la classe ≤ 0, que nous expliquons ici encore par le nombre réduit d'éléments dans la signature. Les autres courbes suivent globalement celle de l'approche de référence.

		Classes	Éléments	MAP	Sép.Q.
Approche de référence		NA	bigrammes	0,872	0,800
tf · idf avec normalisation	max(MAP)	≤ 10	bigrammes	0,880	0,794
	max(Sép.Q)	≤ 0	unigrammes	0,813	1,0
tf · idf sans normalisation	max(MAP)	≤ 8	bigrammes	0,869	0,768
	max(Sép.Q)	≤ 0	unigrammes	0,814	0,952

TABLE G.5 – Comparaison des meilleurs résultats pour l'exploitation des n-grammes de plus fort poids informatif par rapport à l'approche de référence pour Wikinews.

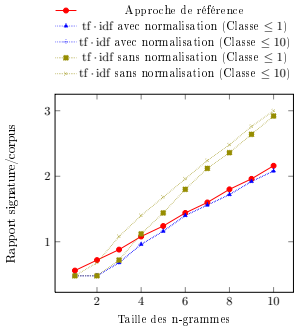


FIGURE G.13 – Évolution du coût de stockage de la méthode par représentativité pour Wikinews.

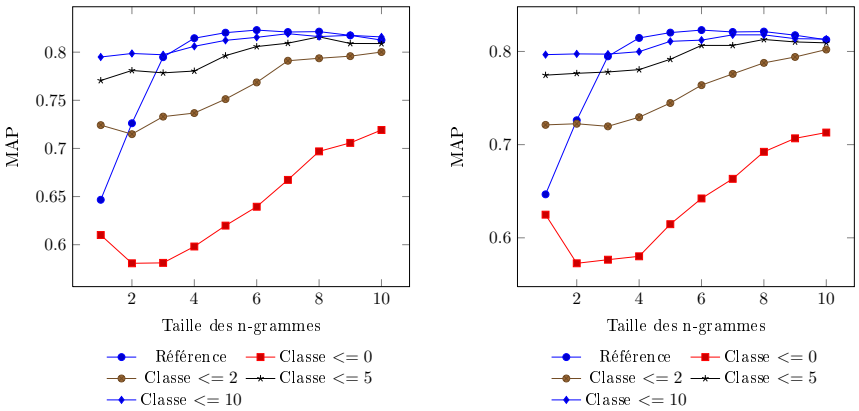
G.2.2.3 Synthèse

Le tableau G.5 reprend les meilleurs résultats obtenus pour cette approche sur le corpus wikinews et les compare à ceux de l'approche de référence. Il montre que la première permet d'obtenir de meilleurs résultats que la seconde en termes de qualité de classification ou de capacité de discrimination, mais pas les deux à la fois. Nous pouvons toutefois noter que les résultats sont très proches et que la meilleurs qualité de classification est encore une fois obtenue avec des bigrammes. Les bigrammes correspondent d'ailleurs à une des configurations les plus intéressantes en termes de gain de l'espace de stockage. La figure G.13 montre l'exploitation des n-grammes de plus fort poids informatif est plus efficace en termes de coût de traitement que l'approche de référence. Le gain est le plus important pour les n-grammes de taille inférieure à 4. Tout comme pour Piithie, lorsque le texte n'est pas normalisé, les coûts sont plus conséquents et dépassent ceux de l'approche de référence.

En conclusion, l'approche par sélection des éléments de la signature complète selon leur représentativité permet d'obtenir des performances potentiellement supérieures à l'approche de référence en termes de qualité de classification ou de capacité de discrimination, tout en réduisant le coût de stockage et donc de traitement.

G.2.3 Corpus PANini

Nous analysons dans cette section les résultats obtenus pour l'approche par sélection des n-grammes de plus fort poids informatif sur le corpus PANini. Nous détaillons tout d'abord les résultats en termes de qualité de classification, mesurée par la MAP, et représentés par la figure G.14. Nous détaillons ensuite les résultats en termes de capacité de discrimination, mesurée par la Sép. Q, et représentés par la figure G.15.



(a) MAP – avec normalisation (b) MAP – sans normalisation

FIGURE G.14 – Évaluation, en termes de MAP, de l'exploitation des n-grammes de plus fort poids informatif, par classes de scores de tf · idf, sur le corpus PANini.

Finalement nous comparons les résultats des meilleurs configurations avec ceux de l'approche de référence.

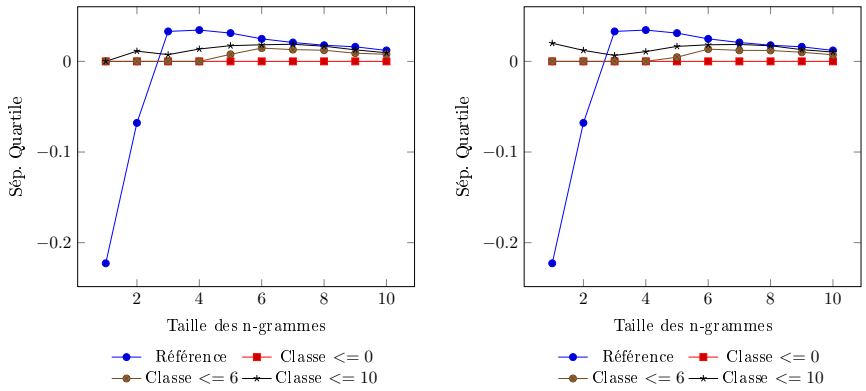
G.2.3.1 Qualité de classification : évolution de la MAP

La figure G.14 montre l'évolution de la MAP en ordonnées selon les différentes tailles de n-grammes en abscisses, avec (G.14(a)) et sans normalisation du texte (G.14(b)). Contrairement aux deux autres corpus, les résultats en termes de qualité de classification pour le corpus PANini se comportent de la même façon quelque soit le rang sélectionné : ils croient lentement avec la taille des n-grammes. Afin de clarifier le graphique des résultats, nous ne montrons dans la figure G.14 que certaines courbes : ≤ 0 qui donne les plus mauvais résultats, ≤ 10 qui donne les meilleurs résultats et ≤ 2 et ≤ 5 qui donnent des résultats intermédiaires.

On retrouve dans les courbes de l'approche par sélection des éléments les plus représentatifs l'allure de la courbe de l'approche de référence : les résultats les plus bas sont obtenus pour les n-grammes de petites tailles, les plus hauts pour les n-grammes de tailles plus importantes. Les différences résident dans la valeur du minimum et la vitesse de croissance des résultats. Pour l'approche de référence les résultats croient très rapidement avec la taille des n-grammes pour se stabiliser dès les 4-grammes. Pour l'approche exploitant les n-grammes de plus fort poids informatif, la croissance est plus lente pour les premières classes alors que la courbe se stabilise dès les unigrammes pour les autres classes. Quelque soit le rang sélectionné, les résultats rejoignent asymptotiquement ceux de l'approche de référence. Des expérimentations complémentaires non rapportés dans ces graphiques montrent que l'augmentation du rang pousse à la convergence des résultats avec ceux de l'approche de référence sans les atteindre.

G.2.3.2 Capacité de discrimination : évolution de la Sép.Q

À l'opposé, les résultats en termes de capacité de discrimination se comportent différemment de l'approche de référence. Ils sont stables, entre 0 et 0,2, indépendamment de la taille des n-grammes. L'augmentation du rang les fait faiblement décoller mais sans réellement atteindre le haut de la courbe de l'approche de référence. Nous ne traçons dans la figure G.15 que les courbes ≤ 0, ≤ 6 et ≤ 10 à des fins de lisibilité. Les autres courbes peuvent être déduites de celles-ci.



(a) Sép. Q – avec normalisation (b) Sép. Q – sans normalisation

FIGURE G.15 – Évaluation, en termes de Sép. Q, de l'exploitation des n-grammes de plus fort poids informatif, par classes de scores de tf · idf, sur le corpus PANini.

		Classes	Éléments	MAP	Sép.Q.
Approche de référence		NA	6-grammes	0,823	0,024
tf · idf avec	max(MAP)	≤ 10	7-grammes	0,819	0,018
normalisation	max(Sép.Q)	≤ 10	7-grammes	0,819	0,018
tf · idf sans	max(MAP)	≤ 10	8-grammes	0,817	0,017
normalisation	max(Sép.Q)	≤ 10	unigrammes	0,796	0,02

TABLE G.6 – Comparaison des meilleurs résultats pour l'exploitation des n-grammes de plus fort poids informatif par rapport à l'approche de référence pour PANini.

G.2.3.3 Synthèse

Le tableau G.6 compare les meilleurs résultats de l'approche exploitant les n-grammes de plus fort poids informatif par rapport à l'approche de référence sur le corpus PANini. Nous pouvons observer que l'approche par sélection des éléments les plus représentatifs ne permet pas d'obtenir d'aussi bons résultats que l'approche de référence, que ce soit en termes de qualité de classification ou de capacité de discrimination. Cependant, les résultats sont assez proches alors que la figure G.16 montre que le gain en termes de coût de stockage est particulièrement important, plus encore que pour les corpus Piithie et Wikinews, notamment pour les n-grammes de petites tailles qui atteignent des performances similaires aux plus grands. Nous pouvons noter que, pour l'approche avec normalisation, le meilleur résultat en termes de qualité de la classification MAP est également le meilleur résultat en termes de capacité de discrimination. Les résultats de cette configuration majorent tous les autres. Les courbes des expérimentations précédentes semblaient montrer que, tout comme la précision et le rappel, l'augmentation de l'un se faisait au détriment de l'autre. Ce résultat confirme, comme nous le pensions, que les deux mesures ne sont pas corrélées et qu'elles mesurent des phénomènes ni similaires, ni opposables, juste distincts.

En conclusion, l'approche par sélection des éléments de la signature complète selon leur représentativité permet de maintenir des résultats similaires à ceux de l'approche de référence en termes de qualité de classification ou de capacité de discrimination, tout en réduisant nettement le coût de stockage et donc de traitement.

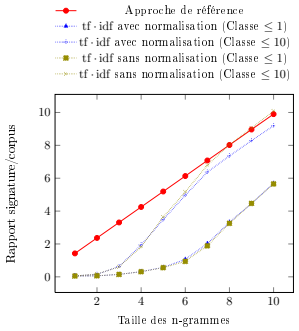


FIGURE G.16 – Évolution du coût de stockage de la méthode par représentativité pour PANini.

Corpus	Négatifs les plus hauts			Positifs les plus bas		
	rang	sim.	comparaison	rang	sim.	comparaison
Pithie	1059	0,72	src00049–cand00582	1518	0,17	src00026–cand00317
	1295	0,5	src00011–cand00871	1516	0,18	src00082–cand00980
	1350	0,4	src00051–cand00425	1508	0,19	src00026–cand00315
	1351	0,4	src00024–cand00583	1462	0,22	src00026–cand00313
	1355	0,4	src00056–cand00911	1456	0,25	src00025–cand00295
Wikinews	987	1,0	src00204–cand01629	3070	0,2	src00088–cand01084
	2708	0,66	src00018–cand01305	3067	0,2	src00088–cand01086
	2790	0,5	src00159–cand02279	3051	0,2	src00088–cand01087
	2816	0,5	src00167–cand00505	3045	0,2	src00088–cand01098
	2821	0,5	src00015–cand01312	3036	0,25	src00088–cand01093
PANini	7	1,0	src10358–cand06069	1371	0,01	src10126–cand07923
	8	1,0	src14098–cand00417	1368	0,01	src08486–cand13086
	98	0,66	src09082–cand04406	1366	0,01	src07393–cand12062
	219	0,5	src01028–cand13419	1360	0,01	src11196–cand11393
	227	0,5	src01418–cand01209	1359	0,01	src06236–cand06899

TABLE G.7 – Paires de textes sélectionnées pour l'analyse des erreurs de l'approche basée sur les entités nommées.

G.3 Exploitation des entités nommées

G.4 Exploitation des composés nominaux

Corpus	Négatifs les plus hauts			Positifs les plus bas		
	rang	sim.	comparaison	rang	sim.	comparaison
Pïthie	655	0,31	src00049–cand00582	746	0,02	src00080–cand00954
	699	0,17	src00035–cand00383	729	0,07	src00007–cand00079
	711	0,12	src00028–cand00934	728	0,07	src00002–cand00015
	713	0,11	src00065–cand00848	718	0,09	src00057–cand00684
	714	0,11	src00034–cand00567	715	0,1	src00057–cand00678
Wikinews	791	1,0	src00212–cand00094	3773	0,02	src00127–cand01570
	2369	0,5	src00058–cand00034	3725	0,03	src00039–cand00472
	2372	0,5	src00064–cand01366	3678	0,03	src00151–cand01849
	2373	0,5	src00207–cand01853	3677	0,03	src00151–cand01848
	2374	0,5	src00047–cand00093	3676	0,03	src00039–cand00471
PANini	328	0,25	src04312–cand02524	1179	0,01	src04981–cand03574
	358	0,23	src04830–cand00208	1178	0,01	src02490–cand06388
	378	0,21	src04574–cand07248	1175	0,01	src08182–cand08331
	423	0,18	src07380–cand03516	1172	0,01	src00786–cand04032
	436	0,17	src04689–cand02070	1169	0,01	src08084–cand10434

TABLE G.8 – Paires de textes sélectionnées pour l'analyse des erreurs de l'approche basée sur les composés nominaux.

Annexe H

Apache UIMA : une brève introduction

Early optimization is the root of much evil

— Donald Knuth

Le projet UIMA (*Unstructured Information Management Architecture*) désigne à la fois un standard pour les outils de structuration de l'information [1], et une implémentation de ce standard initiée par IBM et désormais portée par la fondation Apache [2].

Nous présentons tout d'abord les intérêts du développement d'un standard et d'un cadriciel tel qu'UIMA (*cf. Section H.1*), puis nous présentons les concepts principaux de CAS (*cf. Section H.2*) et de chaîne de traitement (*cf. Section H.3*) qui le définissent. Nous passons sur les autres facilités offertes par la plateforme telles que la gestion des ressources ou des erreurs. Finalement, nous présentons brièvement la communauté UIMA-Fr (*cf. Section H.4*).

H.1 De l'intérêt d'UIMA

La production et la diffusion de l'information croît considérablement depuis l'avènement de l'imprimerie (apparue en Chine au XI^{e} siècle et en Europe au XV^{e} siècle), et s'est de nouveau accélerée depuis les débuts de l'informatique, de la numérisation et des réseaux. Pour autant, la majorité de cette information en circulation reste le produit direct des communications humaines (documents en langues naturelles, courriels, discours, images et vidéos), autrement appelée *information non-structurée*, par opposition aux données structurées telles que les bases de données. Cette masse d'information produite par les humains, et à destination des humains, n'est pas directement et aisément exploitable par les machines. Elle se résume, lorsqu'elle est numérisée, à un flux de bits correspondant à des chaînes de caractères, des matrices de pixels ou des fréquences audio. Nous appelerons *artefact* ces contenus non-structurés. Afin de rendre cette information accessible aux applications informatiques, il est nécessaire de réaliser des traitements de structuration (décodage, projection en base de données...) permettant d'associer à l'information brute une sémantique rendant possible un traitement automatisé. Nous appelerons la tâche d'assignation d'une telle structuration (ou sémantique) à tout ou partie d'un artefact une *analyse*.

INFORMATION NON-STRUCTURÉE

ARTEFACT

ANALYSE

1. `http://docs.oasis-open.org/uima/v1.0/uima-v1.0.html`
2. `http://uima.apache.org`

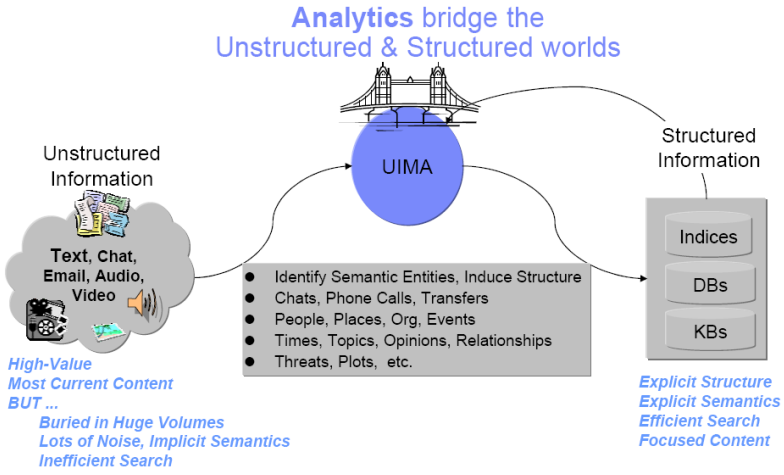


FIGURE H.1 – Apache UIMA est une proposition de pont entre les informations non-structurées et leur structuration. Schéma tiré de la documentation Apache.

L'émergence de l'intelligence artificielle et ses domaines adjacents (recherche d'information, extraction de connaissances, reconnaissance de formes...) a permis de mettre au point des algorithmes de structuration de l'information non-structurée. Ce genre d'algorithmes se développe notamment dans le cadre du traitement automatique des langues naturelles. En parallèle, des normes concernant le stockage et l'interrogation de ces structurations ont été développées : SGML, SQL, HTML, XML, Unicode... Les structures de données et l'hétérogénéité des analyses rend difficile leur combinaison d'où le besoin de formaliser le processus ainsi que les échanges entre les analyses. UIMA offre une façon de matérialiser ce pont entre l'information non-structurée et l'information structurée (*cf. Figure H.1*).

UIMA est avant tout **une spécification**, validée par l'OASIS [3], garantissant l'interopérabilité des briques logicielles entre les différentes plateformes, les différents cadriciels et les différentes modalités (texte, audio, vidéo...). Elle encourage la réutilisation des composants d'analyse ce qui limite la duplication des développements, et repose sur quatre axes : (i) la représentation des données, (ii) l'échange et la modélisation des données, (iii) la découverte, la réutilisation et la composition des briques logicielles et (iv) l'interopérabilité au niveau des services.

Les diverses briques logicielles d'analyse ont une API normalisée. Les résultats des analyses sont transmis de brique en brique à l'aide d'une structure normalisée : le *CAS* (*Common Analysis Structure*). La combinaison d'une API et d'une structure de données commune permet de contrôler aisément l'orchestration du traitement, c'est le rôle du *CPE* (*Collection Processing Engine*).

H.2 Le CAS : structure d'échange entre composants d'analyse

L'un des objectifs d'UIMA est de permettre l'échange des données de structuration de manière unifiée entre les différents composants d'analyse. C'est le rôle du couple

3. L'OASIS est un consortium qui conduit le développement et l'adoption de normes ouvertes dédiées à la société de l'information. La norme UIMA v1.0 est disponible sur le site de l'OASIS : `http://docs.oasis-open.org/uima/v1.0/uima-v1.0.html`

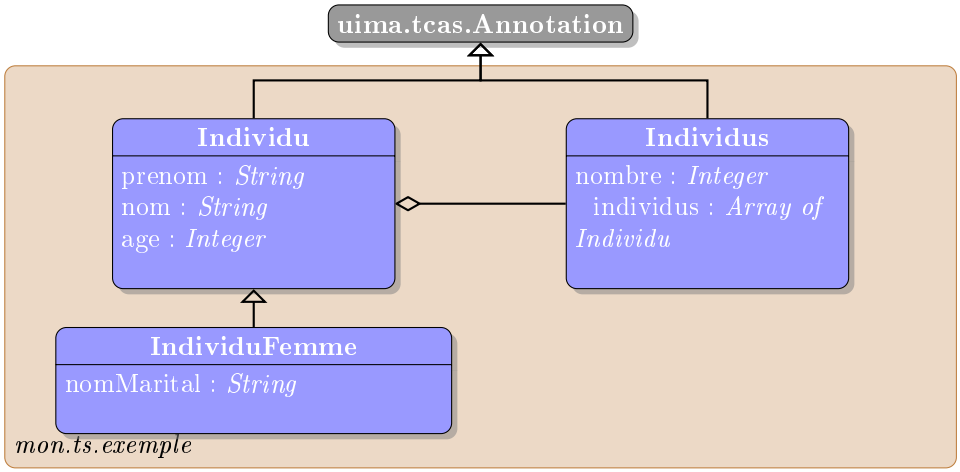


FIGURE H.2 – Exemple de type system modélisant des entités individus.

CAS (*Common Analysis Structure*) et *Type System* qui permettent à un composant d'analyse de connaître les résultats d'analyses précédentes et de faire connaître aux composants suivants le résultat de ses propres analyses. Le *CAS* contient l'artefact à analyser — le *SOFA* (*Subject of Analysis*) — ainsi que les informations de structuration — les annotations — issues des analyses et organisées selon les indications données par le *Type System*. CAS

H.2.1 Type System

Le *Type System* (TS) permet au développeur de définir un modèle pour structurer les données issues de l'analyse. Ce mécanisme permet à l'utilisateur de définir un schéma de données qui correspond exactement à ses besoins d'analyse. Cette approche est plus souple que la définition d'un modèle de données commun global qui ne serait pas adapté à toutes les tâches qui peuvent être appréhendées par UIMA. TYPE SYSTEM

Un TS est un ensemble de types munis d'une relation d'ordre (l'héritage). Un type est constitué d'un nom et d'un ensemble de traits. Un trait est une paire formée d'un nom et d'un type. UIMA introduit des types primitifs sans trait (Integer, String, Array...) qui sont nécessaires à la définition des nouveaux types. Chaque type défini par l'utilisateur doit hériter d'un type pré-existant (défini par l'utilisateur ou par le système).

Le *type system* illustré par la figure H.2 définit trois types d'annotation : *Individu*, *IndividuFemme* et *Individus*. Chacun de ces types hérite du type *uima.tcas.Annotation*. Ce dernier est type standard défini au sein de l'implémentation Apache d'UIMA. Comme l'illustre la figure H.2, il est possible de définir des types qui héritent d'autres. C'est le cas d'*IndividuFemme* qui hérite d'*Individu*. Il hérite ainsi de tous les attributs définis dans ce dernier type, auxquels vient s'ajouter son attribut propre (*nomMarital*). Il est également possible de composer les types, c'est le cas pour *Individus* qui a pour attribut un tableau d'instances de type *Individu*.

La norme UIMA ne propose pas de *type system* unifié pour quelque tâche que ce soit, laissant la liberté au développeur de le définir selon ses besoins. La documentation du projet invite toutefois la communauté à développer des *types systems* communs pour différents domaines. De telles initiatives peuvent grandement réduire les efforts nécessaires au développement de composants d'analyse. Nous pouvons souligner en ce sens l'initiative de Hahn et collab. (2007) sur la création d'un *type system* généraliste

pour les tâches de traitement automatique des langues orientées données. Toutefois, l'étendue des traitements qu'il couvre et les nombreuses ramifications entre les types le rendent difficile d'utilisation et difficile à maintenir. Il est ainsi complexe de l'adapter à des nouvelles tâches non anticipées et va à l'encontre de la mise en place d'outils simples à utiliser pour la communauté. Nous avons donc choisi de ne pas l'utiliser.

H.2.2 SOFA et Annotations

Si UIMA ne fournit pas de *type system* dédié, il définit plusieurs types de base sur lesquels se construisent les *types systems* utilisateurs. Les deux types fondamentaux fournis par UIMA sont le Sofa (*Subject of Analysis*) et l'Annotation.

SOFA

Le Sofa est le type permettant de stocker au sein du CAS l'artefact sur lequel portent les analyses. Il n'est pas directement accessible aux utilisateurs. Le type Annotation est le parent de tous les types définis par les utilisateurs. Il est spécialisé pour les artefacts textuels par le type *uima.tcas.Annotation* qui en hérite. Il est ainsi doté d'attributs *begin* et *end* précisant l'index de début et de fin de la couverture de l'annotation.

ANNOTATION

VIEW

Le type View permet de regrouper une collection spécifique d'instances d'annotations. Un CAS peut contenir plusieurs instances de View, et en contient au moins une nommée *_InitialView*. Finalement, plusieurs types primitifs utilisés pour la construction des types utilisateurs sont également disponibles. Ils sont tirés du méta-modèle Ecore.

H.3 Les chaînes de traitement

La mise à disposition d'un système de structuration unifié des données est le premier avantage de la plateforme UIMA. Le second est la portabilité et la compatibilité des composants d'analyse, ainsi que leur ordonnancement en chaînes de traitement. C'est le rôle des *collection processing engine* (CPE).

COLLECTION PROCESSING ENGINE

H.3.1 Le CPE

Un CPE (*Collection Processing Engine*) correspond à ce que l'on appelle communément une chaîne de traitement, c-à-d un ensemble de composants logiciels mis bout à bout afin d'opérer un traitement particulier à partir d'un ensemble de ressources. La figure H.3 est un exemple d'une telle chaîne dans UIMA.

Nous pouvons tout d'abord observer que certains éléments dans UIMA sont extérieurs au CPE, il s'agit des artefacts à analyser réprésentés dans la figure H.3 par le nuage à gauche) ainsi que des ressources extérieures telles que les dictionnaires, les bases de données... (réprésentées dans la figure H.3 par les cylindres à droite). Ensuite, un CPE se compose d'un certain nombre d'éléments :

- un *collection reader* qui charge les artefacts dans la chaîne de traitement ;
- des *annotateurs* qui opèrent des tâches atomiques sur les artefacts ;
- des *flow controllers* qui contrôlent l'échange des CAS entre les différents annotateurs ;
- des *CAS consumers* qui, en fin de chaîne, exportent le résultat du traitement.

Depuis la version 2.2 d'Apache UIMA, les *CAS consumers* ne sont plus distingués des classiques annotateurs, ces derniers étant également en mesure d'exporter le résultat du traitement ainsi que des résultats intermédiaires. Les autres éléments sont décrits par la suite dans cette section.

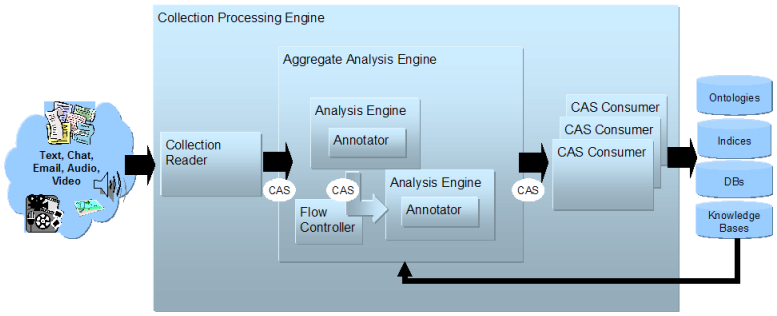


FIGURE H.3 – Un CPE se compose d'un *collection reader* et d'un ou plusieurs *analysis engines*. Depuis la version 2.2, les composants de type CAS Consumer ne sont plus différenciés des *analysis engines*. Schéma tiré de la documentation Apache.

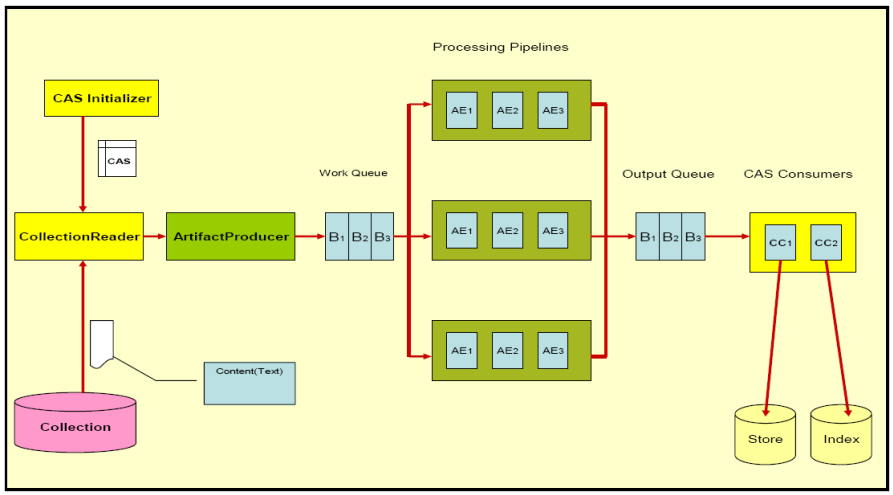


FIGURE H.4 – Schéma détaillant la mise en œuvre d'un CPE au sein d'UIMA. Schéma tiré de la documentation Apache.

Un CPE est une chaîne assez complexe, décrite par un fichier XML à des fins d'interopérabilité, où s'enchevêtrent, au sein de différents processus, plusieurs briques logicielles. La figure H.4 schématise cette mise en œuvre.

Le *collection reader* tout d'abord, comme le montre la figure H.4, est exécuté au sein du thread principal et ne peut-être parallelisé. Pour les traitements peu lourds, le collection reader est le goulot d'étranglement des données. Un découpage de la collection d'artefacts à traiter est une solution de parallelisation.

Les *processing pipelines*, ensuite, sont les chaînes de traitement à proprement parler. Elles se constituent de séquences d'annotateurs et sont dupliquées dans des threads distincts afin de paralléliser le traitement. Le CPE se charge éventuellement, lorsqu'une exception est levée au sein d'un des annotateurs, de retirer le CAS incriminé de la chaîne. Comme l'illustre la figure H.4, le CPE maintient un tampon entre le *collection reader* et les annotateurs. Ce tampon est le *cas pool*, il permet de préparer plus de CAS que les *processing pipelines* ne peuvent en traiter en parallèle afin de palier le monolithisme du *collection reader*.

H.3.2 Les composants d'analyse

ANALYSIS ENGINE

DESCRIPTEUR

Les composants d'analyse, appelés *analysis engines* (AE) au sein d'UIMA, sont des briques logicielles implémentant une API particulière et décrits comme le CPE par un fichier XML : le descripteur. Les descripteurs XML des AE définissent les informations :

- d'identification : un nom unique, une description et un numéro de version ;
- de configuration : noms des paramètres du composant et valeurs par défaut ;
- de comportement : dépendances en entrée et opérations fournies en sortie ;
- sur le type system utilisé.

Les principaux AE utilisés dans UIMA sont ceux directement responsables du traitement : les annotateurs et les *collection reader*.

H.3.2.1 Les annotateurs

Les annotateurs sont les composants chargés des tâches métiers. Ils opèrent des analyses sur les CAS et mettent éventuellement leurs contenus à jour par l'ajout d'instances d'annotations. Ils peuvent être de type primitif, auquel cas il s'agit de briques logicielles autonomes, ou bien agrégés auquel cas il s'agit d'une chaîne d'annotateurs ordonnancée par un *flow controler*. Les annotateurs, d'un point de vue logiciel, sont représentés par une classe contenant au moins une méthode *process* (ou *processCas* selon l'implémentation) qui lance l'exécution de l'analyse sur le CAS passé en paramètre.

H.3.2.2 Les chargeurs de collections

Les *collection reader* sont les composants en entrée de chaîne. Ils chargent les artefacts sous forme de CAS afin qu'ils soient transférés aux différents composants d'analyse. Ils sont représentés, d'un point de vue logiciel, par une classe contenant au moins les méthodes *hasNext* et *getNext*. La première permet d'interroger le composant sur la disponibilité d'autres CAS en préparation et la seconde permet d'obtenir un nouveau CAS à traiter.

H.3.2.3 Les autres types de composants d'analyse

Il existe d'autres types d'AE qui ont des rôles utilitaires : les *CAS Multiplier* et les *flow controler*. Un *CAS Multiplier* effectue les mêmes tâches qu'un annotateur, mais peut créer de nouveaux CAS. Ces composants sont classiquement utilisés pour segmenter un CAS en plusieurs morceaux, chaque morceau correspondant à un nouveau CAS en sortie du composant. Ils peuvent également être utilisés à l'inverse pour fusionner plusieurs CAS en un seul. Un *flow controller* permet de router les CAS au travers des différents annotateurs disponibles.

H.4 La communauté UIMA-FR

L'intérêt principal d'UIMA est d'offrir une base technique commune et fiable. Le passage récent du projet Apache UIMA de l'incubateur à un projet officiel Apache est une preuve de la qualité du développement et de la maturité de la communauté associée. L'architecture à composants choisie pour UIMA ouvre la voie à un partage facilité des développements des différents organismes, de recherche notamment, qui travaillent sur le traitement des données textuelles. Cette collaboration, si elle est désormais techniquement possible, nécessite de se concrétiser humainement. Nous avons initié la communauté UIMA-Fr dans ce but.

H.4.1 La nécessité d'une communauté

Le traitement automatique des langues repose aujourd'hui sur un mille-feuilles applicatif :

- outils de chargement des différents corpus ;
- découpage des textes, des sons ou des images en unités mots, morphèmes, graphèmes ;
- analyses morphologiques, syntaxiques... ;
- application de modèles de langage ;
- ...

Ces différentes briques logicielles, lorsqu'elles sont disponibles, collaborent difficilement les uns avec les autres. Les chercheurs en traitement automatique des langues doivent donc réussir à mettre en place une chaîne de prétraitement à partir de laquelle ils pourront envisager mettre en œuvre leurs expérimentations. À ce premier obstacle fastidieux s'ajoute souvent un protectionnisme injustifié des équipes envers leurs propres développements. Ces obstacles sont un frein à l'utilisation des outils les plus performants et à la recherche en général.

La mise en place d'une communauté a pour objectif, grâce à la plateforme technique UIMA, de partager simplement ces différentes briques logicielles entre tous les acteurs. Nous pensons qu'en levant ces deux obstacles principaux et en encourageant une mutualisation des efforts par l'échange de briques logicielles performantes, nous ne pouvons que faciliter la recherche des membres de la communauté. Au delà du logiciel, la communauté sert également de vivier de connaissances et de compétences autour d'UIMA et du TAL en général. Elle offre une porte d'entrée unique sur des ressources et des services tels que l'actualité en cours, un annuaire des acteurs francophones, des tutoriels et de la documentation, un dépôt de composants, le déploiement de composants sous forme de services webs...

Les enjeux sont nombreux : faciliter les échanges, capitaliser les savoir-faire et les produits développés au cours du temps, se donner la capacité de dépasser le cadre du prototypage pour offrir des solutions de force industrielle, viser la production d'applications de plus en plus complexes, gagner en réactivité face aux appels à projets nationaux et internationaux...

H.4.2 Actions et infrastructure

L'action que nous avons menée dans le cadre de la construction de la communauté UIMA Fr s'est développée sur deux axes : la mise en place d'une infrastructure de communication et l'organisation d'une réunion de travail des acteurs intéressés.

Nous avons beaucoup travaillé pour nous approprier l'architecture UIMA et nous avons réalisé plusieurs développements :

- composants encapsulateurs pour plusieurs outils (Brill, Tree Tagger, Flemm...) ;
- portage d'outils existants au sein de l'équipe ;
- création de paquets Debian pour le déploiement.

Ces tâches nous ont permis de monter en compétence et être ainsi en mesure d'écrire des tutoriaux, d'organiser des formations internes pour les membres de l'équipe et d'utiliser UIMA comme outil pédagogique dans les cours de Master et projets étudiants. Nous avons voulu partager ces compétences avec d'autres afin de faciliter leur prise en main d'UIMA.

Dans ce but, nous avons lancé le portail `uima-fr.org` qui sert de vitrine à la construction de la communauté francophone d'utilisateurs et de développeurs, industriels ou académiques, autour de l'architecture UIMA. Ce portail donne accès à une liste de discussion destinée à toutes les personnes désireuses de s'essayer à UIMA.

Plusieurs personnes compétentes sur UIMA y sont abonnées et répondent aux questions qui y sont posées. Il se compose également d'un *planet*, un site de syndication des billets de blogs qui traitent d'UIMA. À l'heure actuelle, le *planet* regroupe les publications de cinq blogs, soit une trentaine d'articles dédiés à différentes thématiques autour d'UIMA.

La technique ne suffisant pas à la construction d'une communauté, nous avons organisé en juin 2009 un workshop sur UIMA. Nous avons profité de la tenue à Nantes des Rencontres Mondiales du Logiciel Libre [4] (RMLL) pour organiser cet évènement de deux jours composé d'une journée d'atelier et d'une journée de conférence. L'atelier a permis de faire découvrir UIMA à travers un tutoriel de création d'une chaîne de traitement et d'un composant configurable. Nous avions dans ce but préparé un environnement de développement complet dans une salle machine. La journée de conférence a vu plusieurs intervenants présenter leurs travaux autour de la plateforme UIMA. Elle s'est conclue par une table ronde qui a abordé quelques considérations techniques ainsi que les perspectives de la naissante communauté UIMA Fr.

H.4.3 Perspectives

La mise en place d'une communauté est un processus long. L'infrastructure actuelle, qui devrait prochainement se compléter par une plateforme de développement collaboratif, devrait suffire à l'émergence de la communauté. Il est aujourd'hui nécessaire de faire un nouveau pas en avant en initiant des collaborations.

Plusieurs laboratoires, dont le LINA, ont développé des composants basés sur UIMA. Pour en faire profiter la communauté, il nous semble nécessaire de faire un bilan des développements réalisés, de les rendre disponibles et de les faire connaître. Ceci passe également par la définition d'un cadre de distribution : licences, qualité du code, documentation, dépendances...

Le workshop organisé durant les RMLL a permis aux personnes intéressées par UIMA en France de se rencontrer. Il est primordial de pérenniser ce type de rencontre afin de suivre l'évolution de la communauté et permettre à chacun de profiter du retour d'expérience des autres. Ce type de rendez-vous pourrait également permettre de définir des objectifs à court ou moyen terme et éventuellement tenter de répondre aux besoins les plus urgents par le biais de sessions de développements en communauté.

Finalement, il est plus facile de développer une communauté en se reposant sur une communauté existante. Le croisement de la communauté scientifique avec la communauté des logiciels libres nous semble la plus prometteuse. Le projet *Debian Science* [5] a pour objectif de faciliter l'accès aux logiciels et ressources à visée scientifique. Il se donne ainsi pour tâches de classer et de distribuer les solutions logicielles pertinentes pour la recherche et de s'assurer de leur qualité. Il nous semble que UIMA Fr gagnerait à rejoindre et participer à ce projet.

4. `http://2009.rmll.info`
5. `http://wiki.debian.org/DebianScience`

Annexe I

Classe de fréquence moyenne des mots

Eissen et Stein (2006) ont utilisé la classe de fréquence moyenne des mots pour caractériser la complexité du style d'un auteur et la richesse de son vocabulaire.

La classe de fréquence d'un mot w dans une collection $\mathcal{C}$ est calculée à l'aide de la formule I.1. Elle correspond grossièrement à la position de la fréquence du mot considéré par rapport à la fréquence du mot le plus fréquent.

$$\begin{aligned} \forall w \in \mathcal{C}, c(w) \quad &= \quad \lfloor \log_2(f(w^*)/f(w)) \rfloor \\ &\quad\; \text{avec } f(w) \text{ fréquence du mot } w \text{ dans } \mathcal{C} \\ &\quad\; \text{et } w^* \text{ mot le plus fréquent dans } \mathcal{C} \end{aligned} \tag{I.1}$$

Étant donnée la phrase suivante :

> « L'apathéisme est un mot-valise formé des mots apathie d'une part et de athéisme d'autre part, qui désigne le manque d'intérêt envers la croyance, ou l'absence de croyance en une divinité. »

la classe de fréquence des mots la composant serait celle donnée dans la colonne $c(w)$ du tableau suivant :

Mot	Nb. Occ.	c(w)	Mot	Nb. Occ.	c(w)
le/la	4	0	athéisme	1	2
apathéisme	1	2	autre	1	2
est	1	2	qui	1	2
un(e)	3	0	désigne	1	2
mot-valise	1	1	manque	1	2
formé	1	2	envers	1	2
de(s)	5	0	croyance	2	1
mots	1	2	ou	1	2
apathie	1	2	absence	1	2
de	1	2	en	1	2
part	2	1	divinité	1	2
et	1	2			

Le mot w^* le plus fréquent dans cet extrait textuel est « de(s) », avec $f(w^*) = \frac{5}{34} \simeq 0,147$.

Plus la classe est proche de 0, plus le mot est fréquent. Ainsi, les mots « le/la », « un(e) » sont de la même classe que « de(s) » alors que les mots « divinité », « part » ou encore « apathéisme » apartiennent à la classe des mots les moins fréquents.

La comparaison des couples mots/classe de fréquence est, selon Eissen et Stein (2006), un trait caractéristique du style de l'auteur.

Publications

Conférences internationales :

Poulard, F., N. Hernandez B. Daille. 2011, Detecting derivatives using specific and invariant descriptors, *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2010)*, Tokyo, Japon

Hernandez, N., F. Poulard, M. Vernier J. Rocheteau. 2010, Building a French-speaking community around UIMA, gathering research, education and industrial partners, mainly in Natural Language Processing and Speech Recognizing domains, *Workshop Abstracts LREC 2010 Workshop 'New Challenges for NLP Frameworks'*, La Valleta Malte, p64. `http://hal.archives-ouvertes.fr/hal-00481459/en/`.

Revues nationales :

Poulard, F., N. Hernandez, S. D. Afantenos B. Daille. 2010, Evaluation de descripteurs statistiques et linguistiques pour la détection de dérivation de texte, *Document numérique*, 13, 3/2010, 69–93. `http://hal.archives-ouvertes.fr/hal-00554351/en/`.

Conférences nationales :

Dejean, C., M. Fortun, C. Massot, V. Pottier, F. Poulard M. Vernier. 2010, Un étiqueteur de rôles grammaticaux libre pour le français intégré à Apache UIMA, *Actes de la 17e Conférence sur le Traitement Automatique des Langues Naturelles 17e Conférence sur le Traitement Automatique des Langues Naturelles*, Montréal Canada, –. `http://hal.archives-ouvertes.fr/hal-00493847/en/`, Experimentation, Performance.

Poulard, F., S. D. Afantenos N. Hernandez. 2009, Nouvelles considérations pour la détection de réutilisation de texte, *Actes de la 16ème conférence sur le Traitement Automatique des Langues Naturelles Conférence sur le Traitement Automatique des Langues Naturelles*, Senlis France, 67. `http://hal.archives-ouvertes.fr/hal-00401072/en/`.

Hernandez, N., F. Poulard, S. Afantenos, M. Vernier J. Rocheteau. 2009, Apache UIMA pour le Traitement Automatique des Langues, `http://hal.archives-ouvertes.fr/hal-00423728/en/`, 16ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'09) - Session Démonstration.

Poulard, F., T. Waszak, N. Hernandez P. Bellot. 2008, Repérage de citations, classification des styles de discours rapporté et identification des constituants citationnels en écrits journalistiques, *Actes de la 15e Conférence sur le Traitement Automatique des Langues Naturelles Traitement Automatique des Langues Naturelles*, Avignon France, 450–459. `http://hal.archives-ouvertes.fr/hal-00401011/en/`.

Poulard, F. 2008, Analyse quantitative et qualitative de citations extraites d'un corpus journalistique, *Actes de la 12e édition de RECITAL Rencontre des Etudiants-Chercheurs en Informatique et en Traitement Automatique des Langues (RÉCITAL)*, Avignon France, 101–110. `http://hal.archives-ouvertes.fr/hal-00401001/en/`.

Bibliographie

1993, *Proceedings of the 5th Message Understanding Conference*, Association for Computational Linguistics (ACL).

1995, *Proceedings of the 6th Message Understanding Conference*, Morgan Kauffman.

1999, « Glatt plagiarism screening program », URL `http://www.plagiarism.com/screen.id.htm`.

2009. URL `\url{http://sites.google.com/site/whatplagiarismlookslike/}`.

Abeillé, A., L. Clément et F. Toussenel. 2003, *Building a treebank for French*, Kluwer Academic Publishers, p. 165–187.

Afantenos, S., V. Karkaletsis et P. Stamatopoulos. 2005, « Summarization from medical documents : a survey », *Artificial Intelligence in Medicine*, vol. 33, n° 2, p. 157–177.

Aho, A. V. et M. J. Corasick. 1975, « Efficient string matching : an aid to bibliographic search », *Communications of the ACM*, vol. 18, n° 6, p. 333–340.

Aizawa, A. 2003, « Analysis of source identified text corpora : exploring the statistics of the reused text and authorship », dans *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, Association for Computational Linguistics, Sapporo, Japan, p. 383–390.

Bao, J., C. Lyon, P. Lane, W. Ji et J. Malcolm. 2006, « Copy detection in chinese documents using the ferret : a report on experiments. », *Language Resources and Evaluation*, vol. 40, n° 3, p. 357–365.

Bao, J., C. Lyon, P. Lane, W. Ji et J. Malcolm. 2007, « Comparing different text similarity methods. », Technical Report 461, University of Hertfordshire. URL `http://hdl.handle.net/2299/1772`.

Barrón-Cedeño, A., P. Rosso, E. Agirre et G. Labaka. 2010, « Plagiarism detection across distant language pairs », dans *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, p. 37–45.

Beauchene, D., C. Million-Rousseau et C. Rieu. 2002, « Détection automatique de l'insatisfaction du client dans un contexte de commerce électronique », dans *Colloque international sur la fouille de texte (CIFT'02)*, Hammamet, Tunisie, p. 107–117.

Bendersky, M. et B. Croft. 2009, « Finding text reuse on the web », dans *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, ACM, Barcelona, Spain, ISBN 978-1-60558-390-7, p. 262–271. URL `http://portal.acm.org/citation.cfm?id=1498835`.

Bernstein, Y., M. Shokouhi et J. Zobel. 2006, « Compact features for detection of near-duplicates in distributed retrieval », dans *In Proceedings of String Processing and Information Retrieval Symposium*, doi :10.1.1.88.3243. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.88.3243`.

Bernstein, Y. et J. Zobel. 2004, « A scalable system for identifying Co-Derivative documents », *In Proceedings of the Symposium on String Processing and Information Retrieval*, doi :10.1.1.70.2306, p. 55–67. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.70.2306`.

Bernstein, Y. et J. Zobel. 2005, « Redundant documents and search effectiveness », dans *Proceedings of the 14th ACM international conference on Information and knowledge management*, ACM New York, p. 736–743.

Bethard, S., H. Yu, A. Thornton, V. Hativassiloglou et D. Jurafsky. 2004, « Automatic extraction of opinion propositions and their holders », dans *In Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text*, p. 22–24.

Bourdaillet, J. 2007, *Alignement textuel monolingue avec recherche de délocationments : algorithmique pour la critique génétique*, thèse de doctorat, Pierre et Marie Curie.

Bourigault, D. 1992, « Surface grammatical analysis for the extraction of terminological noun phrases », dans *Proceedings of the 14th conference on Computational linguistics-Volume 3*, Association for Computational Linguistics, p. 977–981.

Bowker, L. 2002, *Terminology-Management System*, University of Ottawa Press, p. 77–91.

Boyer, R. S. et J. S. Moore. 1977, « A fast string searching algorithm », *Communications of the ACM*, vol. 20, n° 1010, doi :10.1145/359842.359859, p. 762–772. URL `http://www.akira.ruc.dk/~keld/teaching/algoritmedesign_f05/Artikler/09/Boyer77.pdf`.

Brin, S., J. Davis et H. Garcia-molina. 1995, « Copy detection mechanisms for digital documents », *In Proceedings of the ACM SIGMOD Annual Conference*, vol. 24, doi :10.1.1.43.8485, p. 398–409. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.43.8485`.

Broder, A. Z. 1997, « On the resemblance and containment of documents », dans *In Compression and Complexity of Sequences (SEQUENCES'97)*, p. 21–29, doi : 10.1.1.24.779. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.24.779`.

Broder, A. Z., S. C. Glassman, M. S. Manasse et G. Zweig. 1997, « Syntactic clustering of the web », *Computer Networks and ISDN Systems*, vol. 29, n° 8-13, p. 1157–1166.

Brumfiel, G. 2007, « Turkish physicists face accusations of plagiarism », *Nature*, vol. 449, n° 71587158, doi :10.1038/449008b, p. 8. URL `http://dx.doi.org/10.1038/449008b`.

Bull, J., C. Collins, E. Coughlin et D. Sharp. 2001, « Technical review of plagiarism detection software report », cahier de recherche, University of Luton.

Cerbah, F. 2000, « Exogeneous and endogeneous approaches to semantic categorization of unknown technical terms », dans *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, p. 145–151.

Chandrasekar, R., C. Doran et B. Srinivas. 1996, « Motivation and methods for text simplification », dans *Proceedings of the 16th conference on Computational linguistics*, vol. 2, p. 1041–1044. URL `http://portal.acm.org/citation.cfm?id=993268.993361&coll=Portal&dl=ACM&CFID=78536803&CFTOKEN=82640970`.

Chang, W. I. et E. L. Lawler. 1994, « Sublinear approximate string matching and biological applications », *Algorithmica*, vol. 12, p. 327–344.

Charras, C. et T. Lecroq. 2004, *Handbook of exact string matching algorithms*, King's College London Publications. URL `http://www-igm.univ-mlv.fr/~lecroq/string/index.html`.

Chase, P. J. et S. Argamon. 2006, « Stylistic text segmentation », dans *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, p. 633–634.

Chaski, C. E. 2001, « Empirical evaluations of language-based author identification techniques », *Forensic Linguistics*, vol. 8, n° 1, p. 1–65.

Cho, J., N. Shivakumar et H. Garcia-molina. 2000, « Finding replicated web collections », dans *In ACM SIGMOD*, p. 355–366, doi :10.1.1.18.5269. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.18.5269`.

Choi, Y., C. Cardie, E. Riloff et S. Patwardhan. 2005, « Identifying sources of opinions with conditional random fields and extraction patterns », dans *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Vancouver, B.C., Canada, p. 355–362. URL `http://www.aclweb.org/anthology/H/H05/H05-1045`.

Chowdhury, A., O. Frieder, D. Grossman et M. C. Mccabe. 2002, « Collection statistics for fast duplicate document detection », *ACM Transactions on Information Systems*, vol. 20, doi :10.1.1.5.3673, p. 2002. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.5.3673`.

Church, K. W. 1988, « A stochastic parts program and noun phrase parser for unrestricted text », dans *Proceedings of the second conference on Applied natural language processing*, Association for Computational Linguistics, p. 136–143.

Church, K. W. et W. A. Gale. 1995, « Inverse document frequency (IDF) : A measure of deviations from poisson », dans *Proceedings of the Third Workshop on Very Large Corpora*, p. 121–130.

Clough, P. 2000, « Plagiarism in natural and programming languages : an overview of current tools and technologies », *Research Memoranda : CS-00-05, Department of Computer Science, University of Sheffield*.

Clough, P. 2003a, *Measuring text reuse*, thèse de doctorat, University of Sheffield.

Clough, P. 2003b, « Old and new challenges in automatic plagiarism detection », *National UK Plagiarism Advisory Service*.

Clough, P. et R. Gaizauskas. 2008, « Corpora and text re-use », dans *Corpus Linguistics. An International Handbook, Handbooks of Linguistics and Communication Science*, vol. 1, mouton de gruyter éd., Anke Lüdeling & Merja Kytö, Berlin, ISBN 978-3-11-018043-5, p. 1249–1272. URL `http://www.degruyter.de/cont/imp/mouton/detailEn.cfm?id=IS-9783110180435-1`.

Clough, P., R. Gaizauskas, S. S. Piao et Y. Wilks. 2002, « METER : MEasuring TExt reuse », dans *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, p. 152–159. URL http://www.aclweb.org/anthology/P02-1020.pdf.

Conrad, J. G. 2003, « Online duplicate document detection : Signature reliability in a dynamic retrieval environment », *Proceedings of the twelfth international conference on Information and knowledge management, Pages : 443 - 452*, doi :10.1.1.100.6457, p. 443–452. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.100.6457.

Covington, M. A. 1996, « An algorithm to align words for historical comparison », *Computational linguistics*, vol. 22, n° 4, p. 496. URL http://portal.acm.org/citation.cfm?id=256329.256333.

Crochemore, M., C. Hancart et T. Lecroq. 2007, *Algorithms on strings*, Cambridge University Press, ISBN 0521848997.

Cunningham, H., D. Maynard, K. Bontcheva et V. Tablan. 2002, « GATE : A framework and graphical development environment for robust NLP tools and applications », dans *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL 2002)*.

Daille, B. 2007, « Variations and application-oriented terminology engineering », *Application-Driven Terminology Engineering*, vol. 2, p. 163–177, ISSN 1874-0081.

Daille, B., N. Fourour et E. Morin. 2000, « Catégorisation des noms propres : une étude en corpus », *Cahiers de grammaire*, vol. 25, p. 115–129.

Dave, K., S. Lawrence et D. M. Pennock. 2003, « Mining the peanut gallery : Opinion extraction and semantic classification of product reviews », dans *Twelfth International World Wide Web Conference (WWW'03)*.

Deguy, M. 2009, « Traduire », *Que veut dire traduire ?*, vol. 7. URL http://www.reseau-terra.eu/article891.html.

Charles Dejean, Manoel Fortun, Clotilde Massot, Vincent Pottier, Fabien Poulard et Matthieu Vernier. 2010, rançais« Un étiqueteur de rôles grammaticaux libre pour le français intégré à Apache UIMA », dans *Actes de la 17e Conférence sur le Traitement Automatique des Langues Naturelles 17e Conférence sur le Traitement Automatique des Langues Naturelles*, Montréal Canada. URL http://hal.archives-ouvertes.fr/hal-00493847/PDF/article-taln-2010.pdf, Experimentation, Performance.

Dice, L. R. 1945, « Measures of the amount of ecologic association between species », *Ecology*, vol. 26, n° 3, doi :10.2307/1932409, p. 297. URL http://www.jstor.org/stable/1932409?origin=crossref.

Doermann, D., H. Li, O. Kia et K. Kilic. 1997, « The detection of duplicates », dans *In Document Image Databases, Proceedings of the International Conference on Document Analysis and Recognition*, p. 314–318, doi :10.1.1.34.1246. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.34.1246.

Ducrot, O. et J. Schaeffer. 1995, *Nouveau dictionnaire encyclopédique des sciences du langage*, ISSN 0768-0481, Éd. du Seuil, [Paris], ISBN 2-02-014437-9, 668 p..

Duggan, F. 2006, « Plagiarism : prevention, practice and policy », *Assessment & Evaluation in Higher Education*, vol. 31, n° 2, p. 151–154.

zu Eissen, S. M., B. Stein et M. Kulig. 2007, *Plagiarism Detection without Reference Collections*, Springer, p. 359–366.

Eissen, S. M. Z. et B. Stein. 2006, « Intrinsic plagiarism detection », *Proceedings of the European Conference on Information Retrieval (ECIR-06)*, doi :10.1.1.110.5366. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.110.5366`.

EL-Manzalawy, Y. et V. Honavar. 2005, *WLSVM : Integrating LibSVM into Weka Environment*. Software available at `http://www.cs.iastate.edu/~yasser/wlsvm`.

Erny-Newton, E. 2010, `http://owni.fr/2010/07/28/le-plagiat-dans-la-culture-du-partage/`.

Fetterly, D. 2005, « Detecting phrase-level duplication on the world wide web », dans *In Proceedings of the 28th Annual International ACM SIGIR Conference on Research & Development in Information Retrieval*, p. 170–177, doi :10.1.1.69.3965. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.69.3965`.

Fetterly, D., M. Manasse et M. Najork. 2003, « On the evolution of clusters of near-duplicate web pages », dans *1st Latin American Web Congress (LA-WEB 2003)*, IEEE Computer Society, p. 37–45, doi :10.1.1.1.9275. URL `http://www.informatik.uni-trier.de/~ley/db/conf/la-web/la-web2003.html`.

Filatova, E., V. Hatzivassiloglou et K. McKeown. 2006, « Automatic creation of domain templates », dans *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, Association for Computational Linguistics Morristown, NJ, USA. URL `http://www.informatik.uni-trier.de/~ley/db/conf/acl/acl2006.html`.

Finkel, R. A., A. Zaslavsky, K. Monostori et H. Schmidt. 2002, « Signature extraction for overlap detection in documents », *Australian Computer Science Communications*, vol. 24, n° 1, doi :563857.563809, p. 59–64. URL `http://portal.acm.org/citation.cfm?id=563857.563809`.

Flesch, R. 1948, « A new readability yardstick », *Journal of Applied Psychology*, vol. 32, p. 221–233.

Fourour, N. 2004, *Identification et catégorisation des entités nommées dans les textes français*, thèse de doctorat, Université de Nantes.

Fourour, N., E. Morin et B. Daille. 2002, « Incremental recognition and referential categorization of french proper names », dans *Proceedings of the Third International Conference on Language Ressources and Evaluation (LREC 2002)*, vol. 3, p. 1068–1074.

Francis, W. N. et H. Kucera. 1979, « Brown corpus manual », *Brown University*.

Friburger, N. et D. Maurel. 2002, « Textual similarity based on proper names », dans *Proceedings of the workshop Mathematical/Formal Methods in Information Retrieval (MFIR'2002) at the 25 thACM SIGIR Conference*, p. 155–167.

Fullam, K. et J. Park. 2000, « Improvements for scalable and accurate plagiarism detection in digital documents », dans *Proceedings of the 8th International Conference on Parallel and Distributed Systems*, vol. 1, p. 8–23.

Fung, P. 1998, « A statistical vieux on bilingual lexicon extraction : from parallel corpora to non-parallel corpora », dans *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA'98)*, édité par D. Farwell, L. Gerber et E. H. Hovy, p. 1–17.

Gaizauskas, R., J. Foster, Y. Wilks, J. Arundel, P. Clough et S. S. L. Piao. 2001, « The METER corpus : a corpus for analysing journalistic text reuse », dans *Proceedings of the Corpus Linguistics 2001 Conference*, p. 214–223. URL `http://nlp.shef.ac.uk/meter/`.

Genette, G. 1982, *Palimpsestes. La littérature au second degré*, Seuil, Paris, ISBN 2-02-006116-3, 470 p..

Giampiccolo, D., B. Magnini, I. Dagan et B. Dolan. 2007, « The third pascal recognizing textual entailment challenge », dans *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, p. 1–9.

Giguet, E. et N. Lucas. 2004, « La détection automatique des citations et des locuteurs dans les textes informatifs », dans *Le discours rapporté dans tous ses états : question de frontiéres*, édité par J.-M. L.-M. noz, S. Marnette et L. Rosier, L'Harmattan, Paris, p. 410–418.

Giraudoux, J. 1928, « Siegfried », .

Glover, A. et G. Hirst. 1996, *Detecting stylistic inconsistencies in collaborative writing*, Springer-Verlag.

Grieve, J. 2007, « Quantitative authorship attribution : An evaluation of techniques », *Literary and linguistic computing*, vol. 22, n° 3, p. 251–270. URL `http://llc.oxfordjournals.org/cgi/content/abstract/fqm020v1`.

Habert, B. 2000, « Des corpus représentatifs : de quoi, pour quoi, comment », dans *Cahiers de l'Université de Perpignan*, vol. 31, édité par M. Bilger, Presses Universitaires de Perpignan, p. 11–58.

Hahn, U., E. Buyko, K. Tomanek, S. Piao, J. McNaught, Y. Tsuruoka et S. Ananiadou. 2007, « An annotation type system for a data-driven NLP pipeline », dans *The Linguistic Annotation Workshop (LAW) of ACL 2007*.

Hahn, U., N. Chater et L. B. Richardson. 2003, « Similarity as transformation », *Cognition*, vol. 87, n° 1, p. 1–32. URL `http://linkinghub.elsevier.com/retrieve/pii/S0010027702001841`.

van Halteren, H., R. H. Baayen, F. Tweedie, M. Haverkort et A. Neijt. 2005, « New machine learning methods demonstrate the existence of a human stylome », *Journal of Quantitative Linguistics*, vol. 12, p. 65–77. URL `http://www.informaworld.com/index/J78751280421V845.pdf`.

Hamming, R. W. 1950, « Error detecting and error correcting codes », *Bell System Technical Journal*, vol. 29, n° 2, p. 147–160.

Hannabuss, S. 2001, « Contested texts : issues of plagiarism », *Library Management*, vol. 22, n° 6/7, p. 311–318.

Hatzivassiloglou, V., J. L. Klavans et E. Eskin. 1999, « Detecting text similarity over short passages : exploring linguistic feature combinations via machine learning », dans *In Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods*

in Natural Language Processing and Very Large Corpora, p. 203–212, doi :10.1.1.14.7387.

Hearst, M. A. 1997, « Textiling : Segmenting text info multi-paragraph subtopic passages », *Computational Linguistics*, vol. 23, p. 33–64.

Heintze, N. 1996, « Scalable document fingerprinting (extended abstract) », http ://www.cs.cmu.edu/afs/cs/user/nch/www/koala/main.html. URL `http://www.cs.cmu.edu/afs/cs/user/nch/www/koala/main.html`.

Hirsch, J. E. 2005, « An index to quantify an individual's scientific research output », *Proceedings of the National Academy of Sciences*, vol. 102, n° 46, p. 16 569–16 572.

Hoad, T. C. et J. Zobel. 2002, « Methods for identifying versioned and plagiarised documents », *Journal of the American Society for Information Science and Technology*, vol. 54, doi :10.1.1.18.2680, p. 203–215. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.18.2680`.

Hose, R. 2003, *Investigation of Sentence Level Text Reuse Algorithms*, Master's thesis, Cornell University.

Ibekwe-SanJuan, F. 2007, *Fouille de textes : méthodes, outils et applications*, Hermès Science Publications.

Irving, R. W. 2004, « Plagiarism and collusion detection using the smith-waterman algorithm », URL `http://www.dcs.gla.ac.uk/publications/PAPERS/7444/TR-2004-164.pdf`.

Ivekovic, R. 2009, « Que veut dire traduire? les enjeux sociaux et culturels de la traduction. », *Que veut dire traduire ?*, , n° 7. URL `http://www.reseau-terra.eu/article889.html`.

Jaccard, P. 1912, « The distribution of the flora in the alpine zone », *New Phytologist*, p. 37–50.

Jackiewicz, A. 2006, « Relations intersubjectives dans les discours rapportés », *Revue TAL*, vol. 47, n° 2, p. 65–87, ISSN 1965-0906. URL `http://atala.org/Relations-intersubjectives-dans`.

Jeh, G. 2002, « Simrank : A measure of structural-context similarity », *KDD*, doi :10.1.1.12.4975, p. 538–543. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.12.4975`.

Jones, K. S. 2004, « A statistical interpretation of term specificity and its application in retrieval », *Journal of Documentation*, vol. 60, n° 5, doi :10.1108/00220410410560573, p. 493–502. URL `http://www.emeraldinsight.com/10.1108/00220410410560573`.

Joy, M. et M. Luck. 1999, « Plagiarism in programming assignments », *IEEE Transactions on Education*, vol. 42, n° 2, p. 129–133.

Karp, R. M. et M. O. Rabin. 1987a, « Efficient randomized pattern-matching algorithms », *IBM Journal of Research and Development*, vol. 31, n° 2, p. 249–260. URL `http://portal.acm.org/citation.cfm?id=1012171`.

Karp, R. M. et M. O. Rabin. 1987b, « Efficient randomized pattern-matching algorithms », *IBM Journal of Research and Development*, vol. 31, n° 2, p. 249–260.

Kim, J., H. C. Mi et I. H. Mei. 2006, « Opinmind, outil en ligne collecte de l'opinion des blogs sur de mulitples sujets », URL `www.opinmind.com`.

Kleppe, A., D. Braunsdorf, C. Loessnitz et S. M. zu Eissen. 2005, « On web-based plagiarism analysis », dans *International Workshop on Text-Based Information Retrieval*, édité par B. Stein et S. M. zu Eissen, Universität Koblenz-Landau, p. 77–86.

Knuth, D. E., J. Morris et V. R. Pratt. 1977, « Fast pattern matching in strings », *SIAM Journal on Computing*, vol. 6, n° 2, doi :10.1137/0206024, p. 323–350. URL `http://link.aip.org/link/SMJCAT/v6/i2/p323/s1&Agg=doi`.

Kołcz, A. 2004, « Improved robustness of Signature-Based Near-Replica detection via lexicon randomization », dans *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, p. 605–610, doi :10.1.1.68.606. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.68.606`.

Kołcz, A. et A. Chowdhury. 2008, « Lexicon randomization for near-duplicate detection with I-Match », *The Journal of Supercomputing*, vol. 45, n° 3, doi :10.1007/s11227-007-0171-z, p. 255–276. URL `http://dx.doi.org/10.1007/s11227-007-0171-z`.

Kulkarni, P., F. Douglis, J. Lavoie et J. M. Tracey. 2004, « Redundancy elimination within large collections of files », dans *In USENIX Annual Technical Conference, General Track*, p. 59–72, doi :10.1.1.85.7036. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.85.7036`.

Labbé, C. et D. Labbé. 2006, « A tool for literary studies : intertextual distance and tree classification », *Literary and Linguistic Computing*, vol. 21, n° 3, p. 311–326. URL `http://llc.oxfordjournals.org/cgi/content/abstract/21/3/311`.

Lee, M. D., B. Pincombe et M. Welsh. 2005, « An empirical evaluation of models of text document similarity », dans *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, p. 1254–1259.

Levenshtein, V. I. 1966, « Binary codes capable of correcting deletions, insertions and reversals », dans *Soviet Physics Doklady*, vol. 10, p. 707–710.

Levitan, S. et S. Argamon. 2006, « Fixing the federalist : correcting results and evaluating editions for automated attribution », *Digital humanities*.

Locard, E. 1940, « L'enquête criminelle », *Traité de criminalistique*.

Lyon, C., R. Barrett et J. Malcolm. 2004, « A theoretical basis to the automated detection of copying between texts, and its practical implementation in the ferret plagiarism and collusion detector », dans *JISC (UK) Conference on Plagiarism : Prevention, Practice and Policies Conference*.

Lyon, C., R. Barrett et J. Malcolm. 2006, « Plagiarism is easy, but also easy to detect », *Plagiary*, vol. 1, p. 1–10.

Lyon, C., J. Malcolm et B. Dickerson. 2001, « Detecting short passages of similar text in large document collections », *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, doi :10.1.1.7.2630, p. 118–125. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.7.2630`.

Manber, U. 1994, « Finding similar files in a large file system », dans *Proceedings of the USENIX Winter 1994 Technical Conference*, p. 1–10.

Manku, G. S., A. Jain et A. D. Sarma. 2007, « Detecting near-duplicates for web crawling », *International World Wide Web Conference*.

Manning, C. D. et H. Schütze. 1999, *Foundations of Statistical Natural Language Processing*, The MIT Press, 680 p..

Martin, B. 1994, « Plagiarism : a mislocationd emphasis », *Journal of Information Ethics*, vol. 3, n° 2, p. 36–47.

Martins, B. E. D. G. 2004, « Inter-document similarity in web searches », doi :10.1.1.80.6273. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.80.6273`.

Maurer, H., F. Kappe et B. Zaka. 2006, « Plagiarism - A survey », *Journal of Universal Computer Science*, vol. 12, n° 8, p. 1050–1084.

Metzler, D., Y. Bernstein, W. B. Croft, A. Moffat et J. Zobel. 2005, « Similarity measures for tracking information flow », dans *Proceedings of the 14th ACM international conference on Information and knowledge management*, ACM New York, NY, USA, p. 517–524.

Metzler, D., S. Dumais et C. Meek. 2007, « Similarity measures for short segments of text », *Advances in Information Retrieval*.

Molino, J. 1982, « Le nom propre dans la langue », *Langages*, , n° 66, p. 5–20.

Monostori, K., A. Z. Alej et R. Bia. 2001, « Using the MatchDetectReveal system for comparative analysis of texts », dans *Proceedings of the 6th Australian Document Computing Symposium*, Coffs Harbour, Australia, doi :10.1.1.17.563. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.17.563`.

Monostori, K., R. A. Finkel, A. Zaslavsky, G. Hodász et M. Pataki. 2002, « Comparison of overlap detection techniques », dans *The 2002 International Conference on Computational Science*, p. 51–60.

Monostori, K., A. Zaslavsky et H. Schmidt. 2000, « Parallel and distributed overlap detection on the web », dans *In Proceedings of the Workshop on Applied Parallel Computing*, doi :10.1.1.23.5205. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.23.5205`.

Mosteller, F. et D. Wallace. 1964, *Inference and disputed authorship : The Federalist*, CSLI Publications, Standford.

Mourad, G. 2001, *Analyse informatique des signes typographiques pour la segmentation de textes et l'extraction automatique des citations*, thèse de doctorat, Université Paris-Sorbonne.

Mourad, G. et J. P. Desclés. 2002, « Citation textuelle : identification automatique par exploration contextuelle », dans *Faits de langues*, édité par L. Danon-Boileau et M.-A. Morel, 19, Ophrys, p. 179–188.

Mourad, G. et J.-P. Desclés. 2004, « Identification et extraction automatique des informations citationnelles dans un texte », dans *Le discours rapporté dans tous ses états : question de frontiéres*, édité par J.-M. L.-M. noz, S. Marnette et L. Rosier, L'Harmattan, Paris.

Mourad, G. et J.-L. Minel. 2000, « Filtrage sémantique du texte, le cas de la citation », dans *3e Colloque International sur le Document Électronique*, édité par G. M. et T. E., Lavoisier, p. 41–56.

Muñoz, J. M. L., S. Marnette et L. Rosier. 2004, *Le discours rapporté dans tous ses états*, sémantiques, Paris : L'Harmattan, ISBN 2-7475-6445-2.

Needleman, S. et C. Wunsch. 1970, « A general method applicable to the search for similarities in the amino acid sequence of two proteins », *Journal of Molecular Biology*, vol. 48, n° 3, doi :10.1016/0022-2836(70)90057-4, p. 443–453. URL `http://linkinghub.elsevier.com/retrieve/pii/0022283670900574`.

O'Shea, J. D., Z. Bandar, K. Crockett et D. McLean. 2008, *A Comparative Study of Two Short Text Semantic Similarity Measures*, vol. 4953, Springer Berlin / Heidelberg, p. 172–181, doi :10.1007/978-3-540-78582-8. URL `http://www.springerlink.com/index/v0867641u342pm28.pdf`.

Pereira, A. R. J. et N. Ziviani. 2003, « Syntactic similarity of web documents », dans *Proceedings of the First Latin American Web Congress*, p. 194–200.

Piao, S. S. L. 2001, « Detecting and measuring text reuse via aligning texts », Research Memorandum CS-01-15, Department of Computer Science, University of Sheffield. URL `http://www.dcs.shef.ac.uk/intranet/research/resmes/CS0115.pdf`.

Piao, S. S. L. et T. McEnery. 2003, « A tool for text comparison », dans *Proceedings of the Corpus Linguistics*, p. 637–646.

Piao, S. S. L. et T. Mcenery. 2004, « A tool for text comparison », *In Archer, Rayson, Wilson and McEnery (eds)*, doi :10.1.1.105.4386. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.105.4386`.

Pingoud, L. et J. Fabre. 2010, « Étudiant puni pour avoir pompé sur wikipédia », `http://www.24heures.ch/vaud-regions/actu/eleve-pompe-travail-matu-internet-sanction-echec-2010-08-11`.

Pirró, G. 2009, « A semantic similarity metric combining features and intrinsic information content », *Data & Knowledge Engineering*, ISSN 0169023X. URL `http://dx.doi.org/10.1016/j.datak.2009.06.008`.

Popescu-Belis, A. 2008, « Le rôle des métriques d'évaluation dans le processus de recherche en TAL », *TAL (Traitement Automatique de la Langue)*, vol. 47, n° 2, p. 25. URL `http://atala.org/IMG/pdf/TAL-2007-48-1-03-Popescu-Belis.pdf`.

Potthast, M., A. Barrón-Cedeño, B. Stein et P. Rosso. 2010a, « Cross-language plagiarism detection », *Language Resources and Evaluation*, doi :10.1007/s10579-009-9114-z. URL `http://www.springerlink.com/index/10.1007/s10579-009-9114-z`.

Potthast, M. et B. Stein. 2007, « New issues in near-duplicate detection », dans *Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e.V.*, vol. 200, Springer, Heidelberg, p. 601–609, doi :10.1007/978-3-540-78246-9. URL `http://www.springerlink.com/index/u80u2v66666t7701.pdf`.

Potthast, M., B. Stein, A. Eiselt, A. Barrón-Cedeño et P. Rosso. 2009, « Overview of the 1st international competition on plagiarism detection », dans *Proceedings of the 3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social software misuse*, p. 1–9.

Potthast, M., B. Stein et P. Rosso. 2010b, « An evaluation framework for plagiarism detection », dans *Proceedings of the 23rd International Conference on Computational Linguistics, COLING 2010*.

Poulard, F., T. Waszak, N. Hernandez et P. Bellot. 2008, « Repérage de citations, classification des styles de discours rapporté et identification des constituants citationnels en écrits journalistiques », dans *Actes de la 15e Conférence sur le Traitement Automatique des Langues Naturelles Traitement Automatique des Langues Naturelles*, Avignon France, p. 450–459. URL `http://hal.archives-ouvertes.fr/hal-00401011/en/`.

Prasad, R., N. Dinesh, A. Lee, A. Joshi et B. Webber. 2006, « Attribution and its annotation in the penn discourse treebank », *Revue TAL*, vol. 47, n° 2, p. 43–64.

Prochasson, E. 2009, *Alignement multilingue en corpus comparables spécialisés*, thèse de doctorat, Université de Nantes.

Rabatel, A. 2001, « Les verbes de perception, entre point de vue représenté et discours représentés », dans *Le discours rapporté dans tous ses états : question de frontières*, L'Harmattan.

Radev, D. R., E. Hovy et K. McKeown. 2002, « Introduction to the special issue on summarization », *Computational Linguistics*, vol. 28, n° 4, p. 399–408.

Ratinov, L. et D. Roth. 2009, « Design challenges and misconceptions in named entity recognition », dans *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, Association for Computational Linguistics, p. 147–155.

Ritchie, A., S. Teufel et S. Robertson. 2006, « How to find better index terms through citations », dans *Can Computational Linguistics Improve Information Retrieval? Workshop at ACL/COLING*, Sydney, Australia.

Robertson, S., S. Walker et M. Beaulieu. 1999, « Okapi at TREC-7 : automatic ad hoc, filtering, VLC and interactive track », dans *Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, NIST Special Publication 500-242, p. 253–264. URL `http://trec.nist.gov/pubs/trec7/papers/okapi_proc.pdf.gz`.

Rosier. 1999, *Le discours rapporté : Histoire, théories, pratiques*, Duculot.

Sag, I., T. Baldwin, F. Bond, A. Copestake et D. Flickinger. 2002, « Multiword expressions : A pain in the neck for NLP », dans *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, p. 1–15.

Sagot, B., L. Danlos et R. Stern. 2010, nglais« A Lexicon of French Quotation Verbs for Automatic Quotation Extraction », dans *7th international conference on Language Resources and Evaluation - LREC 2010*, Valetta Malte. URL `http://hal.archives-ouvertes.fr/inria-00515461/en/`.

Salton, G. et M. J. McGill. 1986, *Introduction to modern information retrieval*, McGraw-Hill Inc., New York, NY, USA, ISBN 0-07-054484-0, 464 p..

Salton, G. et C. Yang. 1973, « On the specification of term values in automatic indexing », *Journal of documentation*, vol. 29, p. 351–372.

Sattath, S. et A. Tversky. 1977, « Additive similarity trees », *Psychometrika*, vol. 42, n° 3, p. 319–345.

Savoy, J. 1999, « A stemming procedure and stopword list for general french corpora », *Journal of the American Society for Information Science*, vol. 50, n° 10, p. 944–952.

Schleimer, S. 2003, « Winnowing : Local algorithms for document fingerprinting », *null*, doi :10.1.1.112.9320, p. 76–85. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.112.9320.

Seaward, L. et S. Matwin. 2009, « Intrinsic plagiarism detection using complexity analysis », dans *Proceedings of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse*, édité par B. Stein, P. Rosso, E. Stamatatos, M. Koppel et E. Agirre, p. 56–61. URL http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-502/paper10.pdf.

Sebastiani, F. 2002, « Machine learning in automated text categorization », *ACM Comput. Surv.*, vol. 34, doi :http://doi.acm.org/10.1145/505282.505283, p. 1–47, ISSN 0360-0300. URL http://doi.acm.org/10.1145/505282.505283.

Seo, J. et W. B. Croft. 2008, « Local text reuse detection », dans *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, ACM New York, NY, USA, p. 571–578.

Seretan, V. et E. Wehrli. 2008, « Multilingual collocation extraction with a syntactic parser », *Language Resources & Evaluation*.

Shannon, C. 1948, « A mathematical theory of communication », *Bell System Technical Journal*, vol. 27, p. 379–423.

Shivakumar, N. et H. Garcia-Molina. 1995, « SCAM : A copy detection mechanism for digital documents », dans *Proceedings of the Second Annual Conference on the Theory and Practice of Digital Libraries*.

Shivakumar, N. et H. Garcia-Molina. 1996, « Building a scalable and accurate copy detection mechanism », dans *In Proceedings of 1st ACM Conference on Digital Libraries (DL'96)*, p. 160–168, doi :10.1.1.51.6064. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.51.6064.

Shivakumar, N. et H. Garcia-Molina. 1999, « Finding near-replicas of documents on the web », *Lecture notes in computer science*, doi :10.1.1.51.4870, p. 204–212. URL http://www.springerlink.com/index/l6p53843711v1751.pdf.

Si, A., H. Va, L. Rynson et W. H. Lau. 1997, « CHECK : A document plagiarism detection system », *In Proceedings of ACM Symposium for Applied Computing*, doi :10.1.1.47.9726, p. 70–77. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.47.9726.

Siddharthan, A. et S. Teufel. 2007, « Whose idea was this, and why does it matter ? attributing scientific work to citations », *Proceedings of NAACL/HLT-07*.

Smith, T. F. et M. S. Waterman. 1981, « Identification of common molecular subsequences », *Journal of Molecular Biology*, vol. 147, doi :10.1016/0022-2836(81)90087-5, p. 195–197. URL http://gel.ym.edu.tw/~chc/AB_papers/03.pdf.

Somasundaran, S., T. Wilson, J. Wiebe et V. Stoyanov. 2007, « QA with attitude : Exploiting opinion type analysis for improving question answering in on-line discussions and the news », dans *International Conference on Weblogs and Social Media (ICWSM'07)*.

Sorokina, D., J. Gehrke, S. Warner et P. Ginsparg. 2006, « Plagiarism detection in arxiv », dans *Proceedings of the Sixth International Conference on Data Mining*, p. 1070–1075, doi :10.1.1.121.7883.

Stamatatos, E. 2009a, « Intrinsic plagiarism detection using character n-gram profiles », dans *Proceedings of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse*, vol. 2, édité par B. Stein, P. Rosso, E. Stamatatos, M. Koppel et E. Agirre, p. 38–46. URL `http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-502/paper8.pdf`.

Stamatatos, E. 2009b, « A survey of modern authorship attribution methods », *Journal of the American Society for Information Science and Technology*, vol. 60, n° 3, p. 538–556.

Stein, B. 2005, « Fuzzy-fingerprints for text-based information retrieval », dans *Proceedings of the 5th International Conference on Knowledge Management (I-KNOW 05), Graz, Journal of Universal Computer Science*, p. 572–579.

Stein, B. et S. M. zu Eissen. 2005, « Near similarity search and plagiarism analysis », dans *Proceedings of the 29th Annual Conference of the Gesellschaft für Klassifikation e.V.*, p. 430–437.

Stein, B. et S. M. zu Eissen. 2007, « Intrinsic plagiarism analysis with meta learning », dans *SIGIR Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN 07)*, édité par B. Stein, M. Koppel et E. Stamatatos, CEUR-WS.org, p. 45–50. URL `http://ceur-ws.org/Vol-276`.

Stein, B. et S. M. Z. Eissen. 2006, « Near similarity search and plagiarism analysis », *From Data and Information Analysis to Knowledge Engineering*, p. 430–437.

Stein, B., N. Lipka et P. Prettenhofer. 2010, « Intrinsic plagiarism analysis », *Language Resources and Evaluation*, doi :10.1007/s10579-010-9115-y. URL `http://www.springerlink.com/index/10.1007/s10579-010-9115-y`.

Teufel, S., A. Siddharthan et D. Tidhar. 2006, « Automatic classification of citation function », dans *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, p. 103–110. URL `http://portal.acm.org/citation.cfm?id=1610091`.

Thomas, L. 2009, « Broadcast plagiarism », URL `http://blog.seattlepi.com/thenewschick/archives/169587.asp`.

Todorov, T. 1987, *La notion de littérature et autres essais*, Seuil, Paris.

Tversky, A. et collab.. 1977, « Features of similarity », *Psychological review*, vol. 84, n° 4, p. 327–352.

Özlem Uzuner et A. Davis. 2003, « Content and Expression-Based copy recognition for intellectual property protection », *In the Proceedings of the 3rd ACM Workshop on Digital Rights Management (DRM'03)*, vol. 2003, doi :10.1.1.60.2566, p. 103–110. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.60.2566`.

Uzuner, O., R. Davis et B. Katz. 2004, « Using empirical methods for evaluating expression and content similarity », dans *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on*, p. 8.

Uzuner, O., B. Katz et T. Nahnsen. 2005, « Using syntactic information to identify plagiarism », dans *Proceedings of the 2nd Workshop on Building Educational Applications Using NLP*, p. 37–44.

Vandendorpe, C. 1992, « Le plagiat », URL `http://www.uottawa.ca/academic/arts/lettres/vanden/plagiat.htm`.

Veron, E. 1988, « Presse écrite et théorie des discours sociaux : production, réception, régulation », *La press, produit, production, réception. Paris : Didier.*

Wiebe, J., T. Wilson et C. Cardie. 2005, « Annotating expressions of opinions and emotions in language », *Language Resources and Evaluation*, vol. 39, n° 2-3, p. 165–210.

Wilson, T., P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff et S. Patwardhan. 2005, « Opinionfinder : A system for subjectivity analysis », dans *HLT–EMNLP 2005.*

Winkler, W. 1999, « The state of record linkage and current research problems », *Statistical Research Division, US Bureau of the Census, Wachington, DC.* URL `http://eprints.kfupm.edu.sa/71379/`.

Wise, M. J. 1996, « YAP3 : improved detection of similarities in computer program and other texts », *ACM SIGCSE Bulletin*, vol. 28, n° 1, p. 130–134, ISSN 0097-8418.

Witten, I. H. et E. Frank. 2005, *Data Mining : Practical machine learning tools and techniques*, 2e éd., Morgan Kaufmann.

Yang, H. 2006a, « Near-duplicate detection by instance-level constrained clustering », dans *In Proceedings of the 29th ACM Conference on Research and Development in Information Retrieval (SIGIR-06). 2006*, p. 421–428, doi :10.1.1.116.4413. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.116.4413`.

Yang, H. 2006b, « Next steps in near-duplicate detection for erulemaking », dans *In Proceedings of the 7th National Conference on Digital Government Research*, p. 21–24, doi :10.1.1.111.3732. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.111.3732`.

Yang, H. et J. Callan. 2005, « Near-Duplicate detection for eRulemaking », *In Proceedings of the 5th National Conference on Digital Government Research (DG.O2005)*, doi :10.1.1.60.4235, p. 15–18. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.60.4235`.

Yilmaz, I. 2007, « Plagiarism ? no, we're just borrowing better english. », *Nature*, vol. 449, n° 71637163, doi :10.1038/449658a, p. 658. URL `http://www.ncbi.nlm.nih.gov/pubmed/17928839`.

Zechner, M., M. Muhr, R. Kern et M. Granitzer. 2009, « External and interinsic plagiarism detection using vector space models », dans *Proceedings of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse*, édité par B. Stein, P. Rosso, E. Stamatatos, M. Koppel et E. Agirre, p. 47–55.

Zobel, J. et Y. Bernstein. 2006, « The case of the duplicate documents : Measurement, search, and science », dans *Proceedings of the APWeb Asia Pacific Web Conference*, p. 26–39.

parindent

Détection de dérivation de texte

L'Internet permet la production et la diffusion de contenu sans effort et à grande vitesse. Cela pose la question du contrôle de leur origine. Ce travail s'intéresse à la détection des liens de dérivation entre des textes. Un lien de dérivation unit un texte dérivé et les textes préexistants à partir desquels il a été écrit. Nous nous sommes concentré sur la tâche d'identification des textes dérivés étant donné un texte source, et ce pour différentes formes de dérivation. Notre première contribution consiste en la définition d'un cadre théorique posant les concepts de la dérivation ainsi qu'un modèle mulitidimensionnel cadrant les différentes formes de dérivation. Nous avons ensuite mis en place un cadre expérimental constitué d'une infrastructure logicielle libre, de corpus d'évaluation et d'un protocole expérimental inspiré de la RI. Les corpus Piithie et Wikinews que nous avons développé sont à notre connaissance les seuls corpus en français pour la détection de dérivation. Finalement, nous avons exploré différentes méthodes de détection fondées sur l'approche par signature. Nous avons notamment introduit les notions de singularité et d'invariance afin de guider le choix des descripteurs utilisés pour la modélisation des textes en vue de leur comparaison. Nos résultats montrent que le choix motivé des descripteurs, linguistiques notamment, permet de réduire la taille de la modélisation des textes, et par conséquence des coûts de la méthode, tout en offrant des performances comparables à l'approche état de l'art beaucoup plus volumineuse.
Mots-clés : *détection de dérivation, révisions, plagiat, approche par signature, mesures de similarité, recherche d'information*

Detecting textual derivatives

Thanks to the Internet, the production and publication of content is possible with ease and speed. This possibility raises the issue of controling the origins of this content. This work focuses on detecting derivation links between texts. A derivation link associates a derivative text and the pre-existing texts from which it was written. We focused on the task of identifying derivative texts given a source text for various forms of derivation. Our first contribution is the definition of a theoretical framework defines the concept of derivation as well as a model framing the different forms of derivation. Then, we set up an experimental framework consisting of free software tools, evaluation corpora and evaluation metrics based on IR. The Piithie and Wikinews corpora we have developed are to our knowledge the only ones in French for the evaluation of the detection of derivation links. Finally, we explored different methods of detection based on the signature-based approach. In particular, we have introduced the notions of specificity and invariance to guide the choice of descriptors used to modelize the texts in the expectation of their comparison. Our results show that the choice of motivated descriptors, including linguistically motivated ones, can reduce the size of the modelization of texts, and therefore the cost of the method, while offering performances comparable to the much more voluminous state of the art approach.
Keywords: *detection of derivation, revisions, plagiarism, signature approach, similarity metrics, information retrieval*

Discipline : Informatique
Spécialité : Traitement Automatique des Langues
Laboratoire : LINA – CNRS UMR 6241
2, rue de la Houssinière, BP 92208 – 44322 Nantes cx 3