Similarités Textuelles Sémantiques Translingues : vers la Détection Automatique du Plagiat par Traduction

Jérémy Ferrero

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Informatique**

Arrêté ministériel : 25 mai 2016

Présentée par

Jérémy FERRERO

Thèse dirigée par **Laurent BESACIER, Professeur des Universités, Université Grenoble Alpes**, et
codirigée par **Didier SCHWAB, Maître de Conférences, Université Grenoble Alpes**

préparée au sein du **Laboratoire d'Informatique de Grenoble**
dans l'**École Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique**

Similarités Textuelles Sémantiques Translingues : vers la Détection Automatique du Plagiat par Traduction

Thèse soutenue publiquement le **8 décembre 2017**,
devant le jury composé de :

Mme Isabelle TELLIER
Professeur des Universités, Université Paris 3 - Sorbonne Nouvelle, Présidente
M. Emmanuel MORIN
Professeur des Universités, Université de Nantes, Rapporteur
M. Juan-Manuel TORRES-MORENO
Maître de Conférences, HDR, Université d'Avignon et des Pays de Vaucluse, École Polytechnique de Montréal - DGIGL, Rapporteur
M. Frédéric AGNÈS
Ingénieur R&D, Compilatio, Membre
M. Laurent BESACIER
Professeur des Universités, Université Grenoble Alpes, Membre
M. Didier SCHWAB
Maître de Conférences, Université Grenoble Alpes, Membre

Remerciements

Je tiens tout d'abord à remercier grandement Frédéric Agnès et Alain Simac-Lejeune pour avoir eu confiance en moi et m'avoir donné l'opportunité d'effectuer cette thèse.

Je tiens tout particulièrement à remercier une seconde fois Frédéric pour l'intérêt qu'il a eu envers mon travail et surtout pour m'avoir guidé avec enthousiasme et patience tout au long de cette aventure.

J'adresse évidemment un immense merci à Laurent Besacier et Didier Schwab, mes directeurs de thèse, pour avoir également contribué à rendre tout cela possible. Merci à Didier pour sa grande disponibilité, la qualité et la pertinence de ses remarques et questionnements, et merci à lui d'avoir partagé son bureau (la pièce et non le meuble) avec moi pendant la première moitié de ma thèse. Merci à Laurent pour son implication dans mon travail et pour son aiguillage ambitieux tout au long de cette thèse. Je les remercie tous les deux pour leurs précieux conseils, pour le temps qu'ils ont consacré à relire et corriger mes travaux et pour leurs critiques constructives à la source de la rigueur de ces derniers. Au-delà de leurs qualités professionnelles et de leurs apports scientifiques, je tiens à dire que ce fut un plaisir de travailler sous leur supervision.

Je remercie également tous les collègues du laboratoire pour l'ambiance toujours conviviale ainsi que les nombreuses discussions, autour ou non d'un café, qui ont souvent contribué à l'avancée de mes recherches. Je tiens à remercier plus particulièrement Élodie, Alexis et Loïc pour leur aide ainsi que pour les repas, les cafés et tous les autres moments partagés. Je garderais un très bon souvenir de cette expérience parmi eux.

Je remercie tout naturellement l'ensemble de l'équipe Compilatio de ces trois dernières années, notamment Jeanine, Lucile, Aurélie, Laure, Perle, Benoît, Julien, Gabriel, Joffrey, Maxime, Clément, François, Thierry et Lingxiao, pour leur travail, leur aide et leur soutien qu'il fut technique ou humain. Merci pour leur bonne humeur et le temps qu'ils ont passé à me supporter. Je tiens à remercier plus particulièrement Gabriel et Thierry pour les nombreux fous rires qui ont fait passer les journées plus vite, Clément pour avoir apporté sa culture gastronomique au sein de l'entreprise ainsi que Lingxiao pour son aide précieuse dans les derniers mois. Je remercie aussi Valérie pour son rapide passage parmi nous mais sa néanmoins grande contribution apportée à ma thèse.

Merci tout simplement à tous les membres de Compilatio, pour les restaurants, les sorties, les parties de billard enflammées, les célèbres balades du midi et tout le reste. Il est sûr que sans eux, cette expérience n'aurait pas été la même.

Enfin, je remercie toute ma famille, en particulier mes parents et mes grands parents, pour leur soutien durant cette thèse tout comme ce fut le cas durant le reste de ma scolarité, pour leurs encouragements, pour l'intérêt constant porté à mes travaux et pour l'éducation qu'ils m'ont apporté qui a sans doute fait de moi ce que je suis aujourd'hui.

Résumé

La mise à disposition massive de documents via Internet (pages Web, entrepôts de données, documents numériques, numérisés ou retranscrits, *etc.*) rend de plus en plus aisée la récupération d'idées. Malheureusement, ce phénomène s'accompagne d'une augmentation des cas de plagiat. En effet, s'approprier du contenu, peu importe sa forme, sans le consentement de son auteur (ou de ses ayants droit) et sans citer ses sources, dans le but de le présenter comme sa propre œuvre ou création est considéré comme plagiat. De plus, ces dernières années, l'expansion d'Internet a également facilité l'accès à des documents du monde entier (écrits dans des langues étrangères) et à des outils de traduction automatique de plus en plus performants, accentuant ainsi la progression d'un nouveau type de plagiat : le plagiat translingue. Ce plagiat implique l'emprunt d'un texte tout en le traduisant (manuellement ou automatiquement) de sa langue originale vers la langue du document dans lequel le plagiaire veut l'inclure. De nos jours, la prévention du plagiat commence à porter ses fruits, grâce notamment à des logiciels anti-plagiat performants qui reposent sur des techniques de comparaison monolingue déjà bien éprouvées. Néanmoins, ces derniers ne traitent pas encore de manière efficace les cas translingues. Cette thèse est née du besoin de Compilatio, une société d'édition de l'un de ces logiciels anti-plagiat, de mesurer des similarités textuelles sémantiques translingues (sous-tâche de la détection du plagiat).

Après avoir défini le plagiat et les différents concepts abordés au cours de cette thèse, nous établissons un état de l'art des différentes approches de détection du plagiat translingue. Nous présentons également les différents corpus déjà existants pour la détection du plagiat translingue et exposons les limites qu'ils peuvent rencontrer lors d'une évaluation de méthodes de détection du plagiat translingue. Nous présentons ensuite le corpus que nous avons constitué et qui ne possède pas la plupart des limites rencontrées par les différents corpus déjà existants. Nous menons, à l'aide de ce nouveau corpus, une évaluation de plusieurs méthodes de l'état de l'art et découvrons que ces dernières se comportent différemment en fonction de certaines caractéristiques des textes sur lesquelles elles opèrent. Ensuite, nous présentons des nouvelles méthodes de mesure de similarités textuelles sémantiques translingues basées sur des représentations continues de mots (*word embeddings*). Nous proposons également une notion de pondération morphosyntaxique et fréquentielle de mots, qui peut aussi bien être utilisée au sein d'un vecteur qu'au sein d'un sac de mots, et nous montrons que son introduction dans ces nouvelles méthodes augmente leurs performances respectives. Nous testons ensuite différents systèmes de fusion et combinaison entre différentes méthodes et étudions les performances, sur notre corpus, de ces méthodes et fusions en les comparant à celles des méthodes de l'état de l'art. Nous obtenons ainsi de meilleurs résultats que l'état de l'art dans la totalité des sous-corpus étudiés. Nous terminons en présentant et discutant les résultats de ces méthodes lors de notre participation à la tâche de similarité textuelle sémantique (STS) translingue de la campagne d'évaluation SemEval 2017, où nous nous sommes classés 1er à la sous-tâche correspondant le plus au scénario industriel de Compilatio.

Abstract

The massive amount of documents through the Internet (*e.g.* web pages, data warehouses and digital or transcribed texts) makes easier the recycling of ideas. Unfortunately, this phenomenon is accompanied by an increase of plagiarism cases. Indeed, claim ownership of content, without the consent of its author and without crediting its source, and present it as new and original, is considered as plagiarism. In addition, the expansion of the Internet, which facilitates access to documents throughout the world (written in foreign languages) as well as increasingly efficient (and freely available) machine translation tools, contribute to spread a new kind of plagiarism: cross-language plagiarism. Cross-language plagiarism means plagiarism by translation, *i.e.* a text has been plagiarized while being translated (manually or automatically) from its original language into the language of the document in which the plagiarist wishes to include it. While prevention of plagiarism is an active field of research and development, it covers mostly monolingual comparison techniques. This thesis is a joint work between an academic laboratory (LIG) and Compilatio (a software publishing company of solutions for plagiarism detection), and proposes cross-lingual semantic textual similarity measures, which is an important sub-task of cross-language plagiarism detection.

After defining the plagiarism and the different concepts discussed during this thesis, we present a state-of-the-art of the different cross-language plagiarism detection approaches. We also present the preexisting corpora for cross-language plagiarism detection and show their limits. Then we describe how we have gathered and built a new dataset, which does not contain most of the limits encountered by the preexisting corpora. Using this new dataset, we conduct a rigorous evaluation of several state-of-the-art methods and discover that they behave differently according to certain characteristics of the texts on which they operate. We next present new methods for measuring cross-lingual semantic textual similarities based on word embeddings. We also propose a notion of morphosyntactic and frequency weighting of words, which can be used both within a vector and within a bag-of-words, and we show that its introduction in the new methods increases their respective performance. Then we test different fusion systems (mostly based on linear regression). Our experiments show that we obtain better results than the state-of-the-art in all the sub-corpora studied. We conclude by presenting and discussing the results of these methods obtained during our participation to the cross-lingual Semantic Textual Similarity (STS) task of SemEval-2017, where we ranked 1st on the sub-task that best corresponds to Compilatio's use-case scenario.

Table des matières

Liste des figures

Liste des tableaux

Introduction

« Our best thoughts come from others. » [1]

— Ralph Waldo Emerson (1803-1882)

Motivations et Objectifs

Le plagiat consiste à s'approprier du contenu, peu importe sa forme, sans le consentement de son auteur (ou de ses ayants droit) et sans citer ses sources, dans le but de le présenter comme sa propre œuvre ou création.

La mise à disposition massive de documents via Internet (pages Web, entrepôts de données, documents numériques, numérisés ou retranscrits, *etc.*) rend de plus en plus aisée la récupération d'idées. Dès lors, les auteurs, les ayants droit, les éditeurs, ou bien encore les établissements délivrant des diplômes ou des certificats après validation d'un rendu textuel, doivent garantir l'originalité de ces contenus. C'est pour répondre à ce besoin que les logiciels anti-plagiat ont vu le jour. Ils sont de plus en plus nombreux à être utilisés au sein des établissements d'études supérieures et autres institutions académiques. Ces logiciels se basent sur des techniques de comparaison déjà bien éprouvées.

Cependant, ces dernières années, l'expansion d'Internet a également facilité l'accès à des documents du monde entier écrits dans des langues étrangères et à des outils de traduction automatique de plus en plus performants, accentuant ainsi la progression d'un nouveau type de plagiat : le plagiat translingue. Ce plagiat implique l'emprunt d'un texte tout en le traduisant (manuellement ou automatiquement) de sa langue originale vers la langue du document dans lequel le plagiaire veut l'inclure, comme l'illustre la Figure 0.1. L'expansion d'Internet a donc facilité la possibilité de plagier mais a également permis de répandre un nouveau type de plagiat plus difficile à prévenir.

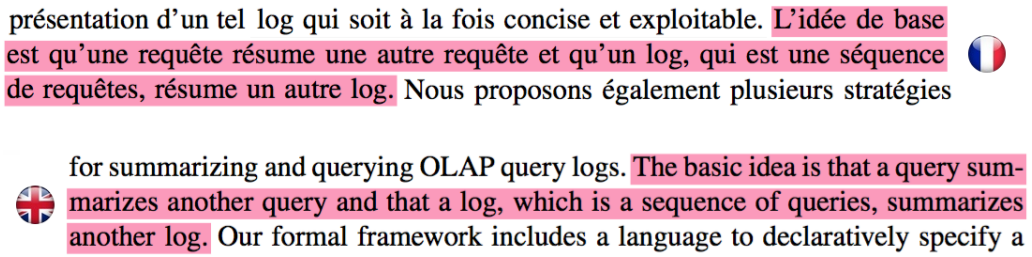


FIGURE 0.1 – Exemple de cas de plagiat translingue d'une phrase.

On peut définir la détection automatique du plagiat par un système composé de deux tâches successives. La première tâche est la collecte de documents sources candidats (des documents suspects à comparer par la suite) et la seconde est la comparaison (la recherche d'alignements de passages similaires) de documents deux à deux, entre le document suspect en cours d'analyse et chacune des sources renvoyées par la première tâche. Le sujet de ma thèse se focalise uniquement

1. *« Nos meilleures idées viennent des autres. »*

sur la seconde tâche : la comparaison de textes suspects avec des textes sources candidats, mais dans un contexte translingue, c'est-à-dire que les textes comparés ne sont pas écrits dans la même langue.

Il va de soit que détecter de façon automatique une similarité textuelle, ne revient pas à détecter un cas de plagiat. Le plagiat est le fait de copier du texte sans attribution, or dans le cas d'une similarité textuelle, il n'y a aucun moyen de connaître pourquoi les textes sont similaires et donc d'assimiler cette similitude à du plagiat. Il reviendra ensuite à un humain d'identifier si des éventuelles similarités détectées relèvent ou non d'un cas de plagiat. En effet, elles peuvent être, par exemple, issues de coïncidences ou de citations proprement référencées. Dans ces travaux, nous ne prenons aucune décision et ne portons aucun jugement, nous nous concentrons uniquement à retrouver des passages similaires entre deux textes.

L'objectif de cette thèse consiste donc à détecter des passages communs suffisamment importants pour être considérés comme suspect entre deux textes écrits dans deux langues différentes. Les passages ne sont pas identiques, ni même similaires, mais signifient la même chose dans deux langues différentes, ils sont donc plus complexes à détecter qu'un « copier/coller ».

Partenariat industriel

La thèse s'est faite dans le cadre d'une convention de type CIFRE [2] entre le Laboratoire d'Informatique de Grenoble (LIG) et l'entreprise Compilatio [3].

Compilatio s'engage depuis 2005 dans la prévention du plagiat et le respect de la propriété intellectuelle et du droit d'auteur en mettant à disposition un service de détection de similitudes entre documents sur le Web. Compilatio s'est imposée comme leader sur le marché francophone et propose des outils d'aide à la détection du plagiat, à l'usage des enseignants, chercheurs, étudiants, et maisons d'éditions, pensés pour contribuer à améliorer l'originalité et l'authenticité des travaux universitaires.

Le projet initial de cette thèse résulte de la nécessité de Compilatio de répondre à une nouvelle problématique : la détection du plagiat trans-lingue. Le besoin est exprimé par le souhait de réaliser un outil intégrable dans la future architecture de la nouvelle application Compilatio, qui sera capable de détecter des paraphrases, des reformulations et du plagiat par traduction.

Le partenaire académique est l'équipe GETALP [4] du Laboratoire d'Informatique de Grenoble. Elle a été choisie pour son expérience sur le traitement automatique de données textuelles et pour son expertise en traduction automatique.

Dans un contexte applicatif où le service développé est destiné à des clients usagers réguliers, comme c'est ici le cas, il y a certaines contraintes à respecter. Dans un premier temps, il est primordial d'éviter les faux positifs, quitte même à privilégier la non complétude des résultats du moment que ceux-ci sont tous exacts. En effet, la confiance que peuvent porter les utilisateurs dans le logiciel final est une priorité.

De plus, les documents traités proviennent soit de travaux d'élèves (parfois truffés de fautes d'orthographe) ne respectant aucune normalisation, utilisant souvent des modèles, des feuilles de style de traitements de textes ou des encodages PDF trompant les extracteurs de texte, soit de pages Web non normalisées avec un important bruit (tableaux, images, balises, codes, caractères non UTF-8, *etc.*). Il sera donc primordial que la solution développée soit robuste aux problèmes éventuels d'encodage et de lecture de flux de texte.

Enfin, il est important de noter que la récente tâche de détection de plagiat translingue ne possède aucune évaluation officielle dans une conférence ou revue. De ce fait, elle ne dispose donc ni de définition, ni de contraintes clairement explicitées, ni de corpus conséquents et de protocole d'évaluation ayant donné lieu à un benchmarking rigoureux.

2. Conventions Industrielles de Formation par la REcherche.
3. www.compilatio.net
4. Groupe d'Études pour la Traduction Automatique du Langage et de la Parole.

Plan du Manuscrit

Dans le chapitre 1, nous définirons, dans un premier temps, la notion de *plagiat*, avant de nous attarder plus particulièrement sur le *plagiat textuel*. Quelques exemples de cas récents de plagiat mettront en évidence le fait que ce phénomène est un problème d'actualité et plus particulièrement dans le milieu académique et l'enseignement supérieur. Ensuite, nous introduirons un phénomène moins traité dans la littérature, le *plagiat translingue*. C'est ce type de plagiat en particulier, plus difficile à prévenir, qui fait l'objet de cette thèse.

Afin de répondre aux objectifs et aux contraintes posées par cette thèse, nous étudierons ensuite l'état de l'art relatif aux méthodes de détection du plagiat dans le chapitre 2. Nous établirons un rapide état de l'art des techniques de détection du plagiat monolingue extrinsèque et intrinsèque, avant de fournir un état de l'art plus complet ainsi qu'une étude comparative de l'ensemble des méthodes de détection du plagiat translingue. Malgré l'assimilation du problème de détection du plagiat avec celui de détection de similarités textuelles sémantiques, nous traiterons dans cette thèse uniquement les méthodes éprouvées dans la littérature de détection du plagiat et non chaque méthode de détection de similarités textuelles sémantiques existantes.

Le chapitre 3 présentera ensuite les différents corpus existants déjà pour la détection du plagiat translingue ainsi que les campagnes d'évaluation dans lesquelles ils s'inscrivent. Nous montrerons que, pris séparément, ces corpus ne sont pas assez diversifiés pour servir à eux seuls à une évaluation rigoureuse. Ils couvrent, pour la plupart, un domaine spécifique et sont principalement issus de corpus comparables. De plus, ils sont traduits manuellement, écrits par la même catégorie d'auteurs et présentent la même granularité de passages plagiés. Il est donc difficile de tirer, à partir des évaluations effectuées sur ces corpus, des conclusions scientifiques pouvant être prises comme cas général et pouvant être ainsi exploitées pour implémenter un outil commercial à destination d'une clientèle exigeante. Idéalement, un corpus permettant une évaluation rigoureuse des méthodes de détection du plagiat translingue ne devra pas contenir ces limites et devra être aussi diversifié que possible.

Nous passerons ensuite aux contributions scientifiques apportées durant cette thèse.

Dans le chapitre 4, nous introduirons le corpus que nous avons construit et qui ne présente pas la plupart des limites rencontrées par les autres corpus existants évoqués dans le chapitre 3. Ce corpus est multilingue (français, anglais et espagnol), présente des textes à trois granularités différentes (syntagme nominal, phrase et document), comporte des textes écrits et traduits (manuellement ou automatiquement) par différentes catégories d'auteurs et qui couvrent différents domaines. Nous présenterons la méthodologie de collecte et de construction de ce jeu de données ainsi qu'un protocole d'évaluation reproductible afin de permettre une évaluation rigoureuse des méthodes de détection du plagiat translingue. Nous mènerons ensuite, à l'aide de ce nouveau corpus, une évaluation de plusieurs méthodes de l'état de l'art et découvrirons que ces dernières se comportent différemment en fonction des caractéristiques des textes comparés mais aussi que la taille de ces textes impacte leurs performances. Finalement, nous montrerons que ces méthodes se comportent différemment en matière de classification, même si elles semblent similaires en termes de performances. Cela encourage une fusion de méthodes, qui pourrait conduire à des résultats bien meilleurs que l'état de l'art.

Dans le chapitre 5, nous présenterons dans un premier temps une nouvelle méthode de détection de similarités textuelles sémantiques translingues basée exclusivement sur des représentations distributionnelles distribuées continues et nous proposerons une augmentation d'une méthode de l'état de l'art en y introduisant également de telles représentations. Ensuite, nous proposerons une notion de pondération morphosyntaxique et fréquentielle de mots qui peut aussi bien être utilisée au sein d'un vecteur qu'au sein d'un sac de mots et nous montrerons que son introduction dans les deux nouvelles méthodes présentées augmente leurs performances respectives. Nous essayerons ensuite de tirer parti des observations faites au cours du chapitre 4, pour

fusionner les nouvelles méthodes introduites au cours de ce chapitre avec les méthodes de l'état de l'art afin d'améliorer leurs performances globales. Nous étudierons les performances, sur notre corpus, de ces méthodes et fusions et les comparerons à celles des méthodes de l'état de l'art. Nous terminerons en présentant et discutant les résultats de ces méthodes lors de notre participation à la tâche de détection de similarité textuelle sémantique (STS) translingue espagnol-anglais de la campagne d'évaluation SemEval 2017.

Première partie

État de l'Art

1 Le plagiat

> *« An idea's birth is legitimate only if one has the feeling that one is catching oneself plagiarizing oneself. »* [1]
>
> Half-truths & One-and-a-half Truths: Selected Aphorisms (Engendra Press, 1976, p. 59)
> — Karl Kraus (1874-1936)

Ce chapitre définit dans un premier temps la notion de *plagiat* et s'attarde plus particulièrement sur le *plagiat textuel*. Par la suite, quelques exemples de cas récents de plagiat mettront en évidence le fait que ce phénomène est un problème d'actualité et plus particulièrement dans le milieu académique et l'enseignement supérieur. Le chapitre introduit ensuite un phénomène moins traité dans la littérature, le *plagiat translingue*. C'est sur ce type de plagiat en particulier, plus difficile à prévenir, que nos travaux se porteront.

1.1 Définitions générales

1.1.1 Définition du plagiat

En France, le CNRTL [2] (*Centre National de Ressources Textuelles et Lexicales*) définit le plagiat comme « *l'action de plagier une œuvre (et donc par métonymie son auteur)* » [3]. Toujours d'après le CNRTL, le terme *plagier* signifie « *emprunter à un ouvrage (ici ce terme désigne tout produit issu d'un travail) original et donc par métonymie à son auteur, des éléments, des fragments, 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* » [4].

Le dictionnaire en ligne des éditions Larousse définit le plagiat comme « *l'acte de quelqu'un qui, dans le domaine artistique ou littéraire, donne pour sien ce qu'il a pris à l'œuvre d'un autre* » [5], tandis qu'Outre-Manche, le dictionnaire en ligne d'Oxford qualifie le plagiat comme « *la pratique de prendre le travail ou les idées d'un autre et de les faire passer comme étant les siennes* » [6]. Il ajoute que son origine est plutôt récente et provient du 17e siècle du latin *plagiarius*, issu de *plagium* (signifiant *voler*, issu lui même du grec ancien *plagion*) et qui signifie avec l'ajout du suffixe *-arius* : *vol d'homme*.

Le site *Dictionary.com*, qui décrit le plagiat comme étant « *l'acte d'utiliser ou d'imiter de façon très proche les mots ou les pensées d'un autre auteur sans autorisation et de présenter l'œuvre de cet auteur comme étant la sienne* », positionne plus précisément son origine entre les années 1615 et 1625 [7].

En effet, même si l'on pourrait penser que les premiers cas de plagiat remontent à l'Antiquité (Aragione, 2010), à cette époque, il ne s'agissait pas réellement de plagiat puisque ceci était toléré,

1. *« Une pensée n'est légitime que si on a le sentiment de se surprendre en flagrant délit de plagiat de soi. »*
2. Créé en 2005 par le CNRS (*Centre National de la Recherche Scientifique*), le CNRTL fédère, au sein d'un portail unique, un ensemble de ressources linguistiques informatisées et d'outils de traitement de la langue.
3. http://www.cnrtl.fr/definition/plagiat (consulté le 31/03/2017 à 10h)
4. http://www.cnrtl.fr/definition/plagier (consulté le 31/03/2017 à 10h)
5. http://www.larousse.fr/dictionnaires/francais/plagiat/61301 (consulté le 31/03/2017 à 10h)
6. https://en.oxforddictionaries.com/definition/plagiarism (consulté le 31/03/2017 à 10h)
7. http://www.dictionary.com/browse/plagiarism (consulté le 31/03/2017 à 10h)

autorisé et parfois même encouragé. Par exemple, avant le 18e siècle et le Romantisme en Europe, les auteurs et artistes étaient encouragés à copier les œuvres de leurs pères et maîtres. Plus concrètement, jusqu'à récemment en France, au début des années 1900, les articles scientifiques appartenaient *de facto* au domaine public, de même la notion d'auteur était totalement absente de la presse généraliste. À partir de 1852 toutefois, « *les traités internationaux de protection littéraire prévoient une licence libre par défaut (tous les articles de revues peuvent être reproduits, sans l'accord de l'auteur, sous réserve d'être cités)* » [8]. Il faudra attendre après 1908 pour que la conférence de Berlin (BIALA, 1910) [9] mette un terme définitif à la licence libre et qu'un début d'application de droit d'auteur voit le jour.

Outre-Atlantique, le dictionnaire en ligne Merriam-Webster définit le terme *plagier* comme le fait de « *voler les idées ou les mots d'un autre et les faire passer comme étant les siennes, cela sans citer ses sources* » et « *présenter comme nouveau et orignal une idée ou un produit dérivé de la source existante* » [10]. En d'autres mots, il conclut par « *plagier est un acte de fraude délibéré qui implique deux méfaits, le vol du travail de quelqu'un d'autre et le mensonge à propos de sa provenance et de son appartenance* ».

Par extension, on peut donc affirmer qu'utiliser une image, une vidéo ou une musique dans une production sans avoir reçu l'autorisation ou avoir cité de façon appropriée l'ayant droit est un acte de plagiat. Et donc par métonymie, copier un film d'un format propriétaire à un format non propriétaire, utiliser une musique en fond dans un montage vidéo, ou reproduire une peinture afin de la faire passer pour l'œuvre originale, sont des plagiats.

Il est important de préciser que, d'après les lois Européennes et Américaines, une idée elle-même ne peut pas être protégée et n'appartient à personne, seulement son expression matérielle ou sa représentation sous toute forme est sujette à la propriété intellectuelle et au droit d'auteur. En France, l'INPI (l'Institut National de la Propriété Industrielle) met d'ailleurs en garde sur son site [11] : « *Attention : le droit d'auteur ne protège pas les idées ou les concepts* ». Le vol d'idées n'est donc pas du plagiat.

Pour conclure, le plagiat consiste donc à s'approprier du contenu, peu importe sa forme, sans le consentement de son auteur (ou de ses ayants droit) et sans citer ses sources, ceci dans le but de le présenter comme sa propre œuvre ou création originale.

1.1.2 Définition du plagiat textuel

Le plagiat textuel, plagiat d'un écrit, est un plagiat impliquant le vol d'une œuvre écrite. Le *copier/coller* de tout ou partie d'un texte, sans citer sa source afin de faire passer ce texte comme étant sa propre production est la forme de plagiat textuel la plus triviale et répandue, mais un plagiat textuel peut se présenter sous diverses formes. En effet, il a été dit en section 1.1.1 que le vol d'idées était également du plagiat à partir du moment où elles avaient donné lieu à un support, comme une œuvre écrite en l'occurrence. C'est pourquoi, s'attribuer les idées originales d'une autre personne, peu importe que ce soit sous la forme d'une théorie, d'une opinion, d'une d'interprétation de données, d'une méthode, d'un algorithme ou d'une formule, sans citer cette personne même si ces idées sont dites avec ses propres mots, est un plagiat du moment où ces idées avaient déjà été mises à l'écrit. Il n'est donc pas nécessaire de faire du mot à mot (*copier/coller*) pour plagier à l'écrit. Il est également considéré comme plagiat le fait de paraphraser ou reformuler un texte.

8. http://scoms.hypotheses.org/409 (consulté le 31/03/2017 à 10h)
9. ftp://ftp.wipo.int/pub/library/ebooks/wipopublications/Berlin_1908.pdf (consulté le 31/03/2017 à 10h)
10. https://www.merriam-webster.com/dictionary/plagiarizing (consulté le 31/03/2017 à 10h)
11. https://www.inpi.fr/fr/comprendre-la-propriete-intellectuelle/les-autres-modes-de-protection/le-droit-dauteur (consulté le 31/03/2017 à 10h)

On distingue donc les différentes cas suivants :

La copie à l'identique, qui consiste à copier mot à mot tout ou partie d'un texte dans un autre. Pour exemple, considérons la phrase suivante présente dans un texte :

> « En cinquante ans, grâce à des efforts considérables dans la recherche et l'élaboration de la fusion, la performance des plasmas a été multipliée par 10 000. »

Elle sera recopiée à l'identique dans un autre texte.

La paraphrase aussi appelée reformulation paraphrastique, consiste à reprendre une phrase d'un texte pour la détailler ou l'expliciter [12] [13]. C'est une opération qui conserve donc le sens et le plus souvent l'ordre des idées évoquées dans le texte, mais qui autorise le changement de vocabulaire, l'ajout, la suppression ou la substitution de certains mots afin de compléter ou d'ajouter de l'information.

Par exemple, la phrase suivante :

> « En cinquante ans, grâce à des efforts considérables dans la recherche et l'élaboration de la fusion, la performance des plasmas a été multipliée par 10 000. »

pourrait être paraphrasée de la façon suivante :

> « En une cinquantaine d'années, grâce à un immense effort de recherche, la performance des plasmas produits par les machines de fusion a été multipliée par 10 000. »

La reformulation consiste à produire une nouvelle formulation qui reproduit autrement ce qui a déjà été exprimé [14]. D'un point de vue linguistique, elle autorise toutes modifications textuelles à condition que le sens de la phrase soit conservé. Cela peut donc donner lieu à un changement d'ordre des idées évoquées dans le texte.

Par exemple, la phrase suivante :

> « En cinquante ans, grâce à des efforts considérables dans la recherche et l'élaboration de la fusion, la performance des plasmas a été multipliée par 10 000. »

pourrait être reformulée de la façon suivante :

> « La performance des plasmas produits par les machines de fusion a été multipliée par 10 000 grâce à un immense effort de la recherche bien que cela ait pris une cinquantaine d'années. »

Copier, paraphraser ou reformuler du texte, sans citer la source de ce dernier et en omettant les marques de citations, revient à commettre un plagiat. De manière plus générale, dans tous les cas, oublier des guillemets ou oublier de créditer la source d'origine de son emprunt, peu importe sa forme ou les changements qu'on y a opéré, est un acte de plagiat.

1.1.3 Limite d'interprétation et aspect éthique

Il est important de différencier la notion de similarité textuelle, d'un cas de plagiat. Le plagiat est le fait de copier du texte délibérément ou non, sans attribution. Dans le cas d'une similarité textuelle, il n'y a aucun moyen de savoir pour quelle raison les textes sont identiques ou similaires. En effet, il est possible qu'il s'agisse par exemple d'une citation, ce qui ne donnera alors pas lieu à un plagiat. Une copie d'un texte, même d'un extrait volumineux, depuis un ouvrage vers un autre, n'est pas un plagiat si cet extrait est correctement marqué et référencé (Wiwanitkit, 2011). En effet, cela n'entraine pas de problème pénal dû à un non respect du droit d'auteur, cependant cela peut poser des problèmes éthiques dans de nombreux cas et risque

12. http://www.larousse.fr/dictionnaires/francais/paraphrase/57993 (consulté le 31/03/2017 à 10h)
13. http://www.cnrtl.fr/definition/paraphrase (consulté le 31/03/2017 à 10h)
14. http://www.cnrtl.fr/definition/reformulation (consulté le 31/03/2017 à 10h)

d'enfreindre les règles d'acceptation et d'édition assujetties à l'ouvrage. Il est donc important de différencier une simple similarité textuelle, d'un cas de plagiat. Néanmoins, il est nécessaire de pouvoir détecter la similarité textuelle afin de pouvoir détecter le plagiat. On peut alors se poser la question suivante : À partir de quand une similarité textuelle sans citation peut être considérée comme étant du plagiat, autrement dit, à partir de combien de mots consécutifs un emprunt peut il être suffisamment suspect pour être considéré comme étant volontaire et non fortuit ? En 2012, la COPE (*Committee on Publication Ethics*) conjointement avec la OASPA (*Open Access Scholarly Publishers Association*), la DOAJ (*Directory of Open Access Journals*), et la WAME (*World Association of Medical Editors*) caractérisait un plagiat à partir de plus de six mots consécutifs copiés sans citation (Masic, 2012). Leurs règles d'éditions [15] [16] sont beaucoup plus nuancées maintenant et font part de plusieurs critères à prendre en compte, comme par exemple la section de l'article dans laquelle l'emprunt a eu lieu (celui-ci n'aura pas la même gravité selon s'il se trouve dans la section d'introduction, de présentation de nouvelles méthodes ou de résultats). Ainsi, il n'est plus fait état nulle part d'un nombre de mots consécutifs minimal caractérisant une similarité textuelle devant obligatoirement donner lieu à une citation sous peine d'être considérée comme plagiat.

Un auto-plagiat est, par définition, un emprunt à soi-même (de ses travaux antérieurs) sans y faire référence. Cette définition est sujette à débat. En effet, selon la définition du plagiat, est-ce que l'auto-plagiat n'est pas un oxymore ? Un oxymore est une figure de style qui réunit plusieurs mots contradictoires et qui rend ainsi cette figure ambiguë ou incompréhensible. Selon la définition du Larousse, vue en section 1.1.1, le plagiaire « *donne pour sien ce qu'il a pris à l'œuvre d'un autre* », ici un auto-plagiat serait donc un oxymore, mais selon la définition plus rigoureuse et nuancée du CNRTL, également vue en section 1.1.1, un plagiat est le fait d' « *emprunter à un ouvrage original [...] des éléments [...] que l'on présente comme personnelle* », ce qui rend dans ce cas-ci le terme auto-plagiat tout a fait pertinent et correct d'usage car le plagiaire tente de faire passer son emprunt comme personnel (ce qui est bien le cas) mais il tente également de le faire passer comme nouveau et original (ce qui n'est pas le cas).

Quoiqu'il en soit cette pratique est beaucoup moins encadrée (Collberg *et* Kobourov, 2005). La réutilisation de ses propres recherches est par exemple autorisée dans les publications d'actes de nombreux domaines scientifiques, comme c'est le cas pour les publications en informatique, sous réserve une fois de plus de citer ses travaux précédents bien entendu. Ce qui implique que dans le cas contraire, où il n'y aura aucune citation alors que cela aurait été nécessaire, cette réutilisation de travaux antérieurs serait bien considérée comme étant du plagiat. Cependant, pour des raisons d'éthique, les règles d'acceptation des grandes conférences du domaine limitent la couverture textuelle entre deux articles à généralement moins de 25% (Collberg *et* Kobourov, 2005; Bruton, 2014), comme c'est par exemple le cas pour les conférences COLING [17] et EACL [18]. Cette mesure permet ainsi la réutilisation de ses anciens travaux, afin d'éviter de représenter trop en profondeur un état de l'art ou un jeu de données, tout en préservant l'anonymat pour les soumissions en double aveugle et évitant l'ambiguïté de faire passer comme nouveau un travail déjà publié en grande partie.

Dans tous les cas, les outils et méthodes d'aide à la détection du plagiat ne prennent aucune décision. Leur objectif est de détecter les similarités textuelles, une relecture puis une prise de décision humaine sont nécessaires pour déterminer si cette similarité donne lieu à un plagiat.

15. http://publicationethics.org/resources/guidelines (consulté le 31/03/2017 à 10h)
16. http://publicationethics.org/files/Web_A29298_COPE_Text_Recycling.pdf (consulté le 31/03/2017 à 10h)
17. http://coling2016.anlp.jp/cfp/ (consulté le 31/03/2017 à 10h)
18. http://eacl2017.org/index.php/calls/call-for-papers (consulté le 31/03/2017 à 10h)

1.2 Le plagiat, un phénomène préoccupant

1.2.1 Le plagiat en pleine expansion

De plus en plus de gens ont accès à Internet. Ce phénomène est illustré dans la Figure 1.1 issue d'un rapport [19] sur la santé d'Internet rendu par la société Mozilla. On peut, par exemple, constater que fin 2016, plus de 50% de la population mondiale avait accès à Internet, contre 30% début 2011 [20] [21].

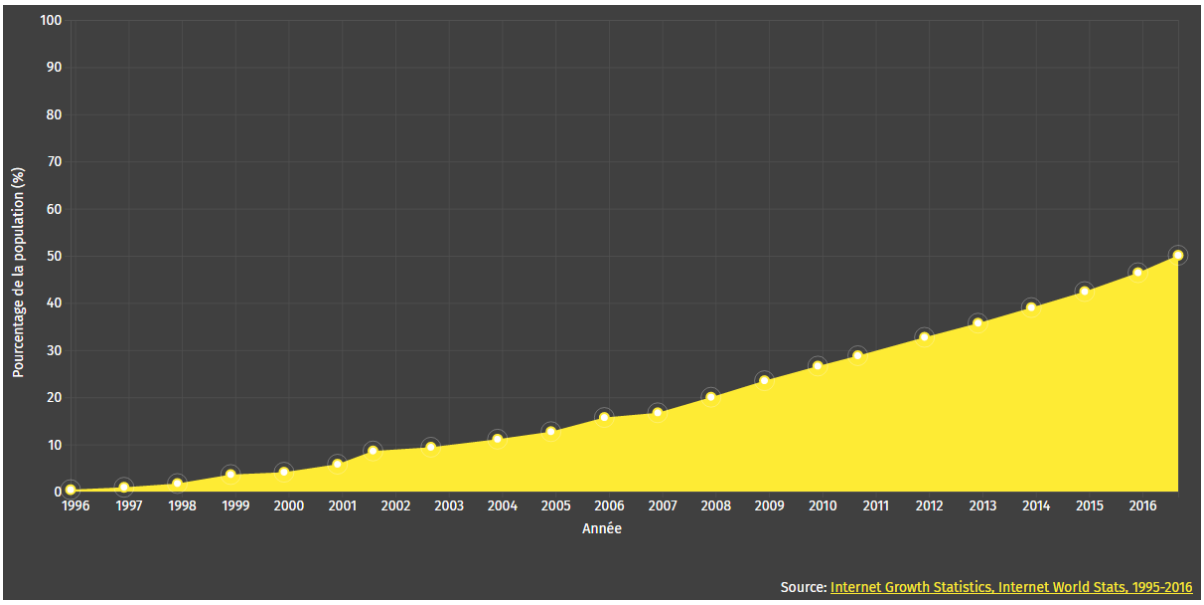


FIGURE 1.1 – Pourcentage de la population mondiale ayant accès à Internet en fonction de l'année.
source : https://internethealthreport.org/v01/fr/digital-inclusion/

La mise à disposition massive de plusieurs milliards de documents via Internet et la digitalisation de textes (pages Web, base de données, documents numériques, numérisés ou retranscrits) rend de plus en plus aisé le *copier/coller* et la récupération d'idées. Cette nouvelle ère numérique fluidifie les échanges de données et diminue les scrupules des gens à s'approprier le travail des autres. Ce phénomène est accentué par le fait que la plupart des données se trouvant sur Internet ne sont pas protégées par un dispositif ou par des lois les empêchant de se faire "voler". En effet, bien que plus de la moitié des pays du monde disposent de lois veillant à assurer la propriété intellectuelle [22] [23], on peut voir sur la Figure 1.2, qui illustre les différents types de mesures mises en place dans les différents pays à travers le monde pour protéger le droit d'auteur, que deux tiers des pays ne disposent pas de clauses d' « utilisation équitable » ou d' « usage acceptable » concernant le droit d'auteur sur Internet [24].

Le vide juridique sur Internet contribue fortement à accentuer le vol d'idées. Le plagiat n'est certes pas un problème datant de l'ère du numérique mais avec la mondialisation et l'évolution des médias et plus particulièrement d'Internet, de nombreuses informations se diffusent bien plus rapidement dans le monde et sont rendues accessibles plus facilement à plus de personnes, et il est certain que cette expansion accentue le plagiat. De plus, Internet facilite également l'accès à des

19. https://internethealthreport.org/v01/about/ (consulté le 31/03/2017 à 10h)
20. https://internethealthreport.org/v01/fr/digital-inclusion/ (consulté le 31/03/2017 à 10h)
21. http://www.internetworldstats.com/emarketing.htm (consulté le 31/03/2017 à 10h)
22. http://www.cncpi.fr/iaa145-45-propriete-intellectuelle-monde.htm (consulté le 31/03/2017 à 10h)
23. http://www.wipo.int/ipstats/en/statistics/country_profile/index.html (consulté le 31/03/2017 à 10h)
24. http://infojustice.org/archives/29136 (consulté le 31/03/2017 à 10h)

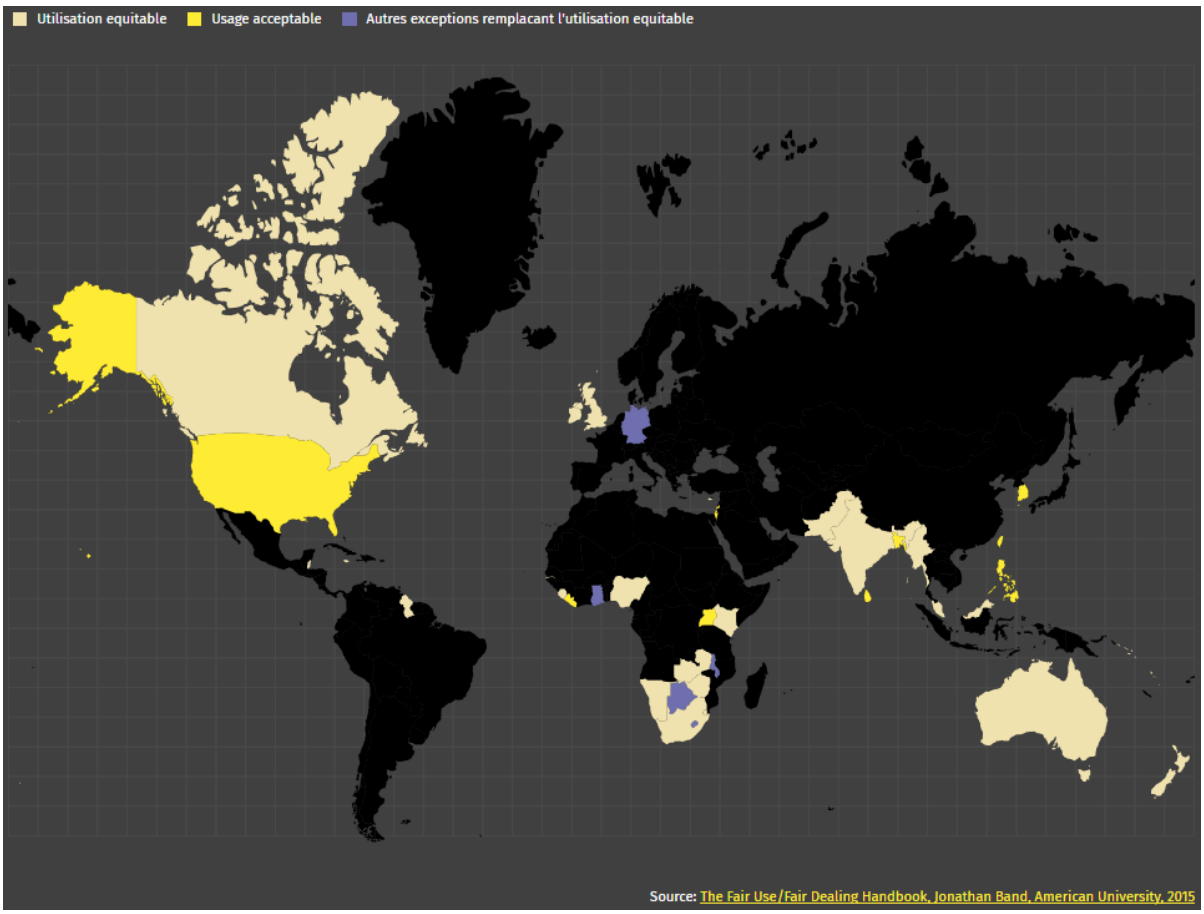


FIGURE 1.2 – Pays disposant de lois favorisant le respect des droits d'auteurs sur Internet.
source : https://internethealthreport.org/v01/fr/open-innovation/

outils aidant au plagiat comme des outils de *paraphrasing*, outils paraphrasant automatiquement du texte (Rogerson *et* McCarthy, 2017), ou des outils de *back-translation*, traduisant une première fois le texte dans une langue étrangère avant de le retraduire dans sa langue d'origine afin de le reformuler de façon implicite (Jones *et* Sheridan, 2015; Mallinson *et al.*, 2017).

Le plagiat étant devenu un phénomène de plus en plus préoccupant, les grands noms du numérique, tels que Twitter, se sont sentis obligés de mettre en place des dispositions pour freiner sa progression sur la toile. Le célèbre réseau social, utilisé par plus de 300 millions de personnes [25] et limitant les interventions de sa communauté à des messages de 140 caractères, s'est rendu compte récemment que ce qui faisait son principal succès était justement ce petit format d'expression. En effet, il encourage ainsi les citations et blagues, d'ailleurs majoritairement reprises sur les autres réseaux sociaux. Il a donc décidé de se pencher sérieusement sur la protection du droit d'auteur et il sera bientôt impossible de copier un tweet sans en mentionner l'auteur [26].

1.2.2 Le plagiat, un problème toujours autant d'actualité

Le plagiat textuel, malgré sa spécificité, touche de nombreux milieux (politique, littéraire, artistique, *etc.*), même s'il est plus difficile à rencontrer dans le milieu musical. Les paroles d'une chanson étant plutôt personnelles et facilement identifiables, les cas reconnus de plagiat gravitent autour des mélodies, à l'instar de l'affaire Calogero de fin 2016. En effet, « *Si seulement je pouvais*

25. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (consulté le 31/03/2017 à 10h)
26. http://www.20minutes.fr/web/1659023-20150728-twitter-part-guerre-contre-plagiat-blagues (consulté le 31/03/2017 à 10h)

lui manquer », le succès de 2004 du chanteur, présenterait plus de 15% de notes identiques (63% pour le refrain) avec « *Les Chansons D'Artistes* » de La Troupe Des Années Boum [27].

Le plagiat textuel, quant à lui, touche de nombreux domaines, aussi bien dans le milieu purement littéraire, que dans la politique ou les écrits académiques. Côté littérature, on peut citer les cas de *copier/coller* d'Étienne Klein, de Patrick Poivre d'Arvor ou bien encore de Joseph Macé-Scaron.

Fin 2016, le célèbre physicien et auteur Étienne Klein est accusé de plagiat dans plusieurs de ses ouvrages (notamment dans « *Le Pays qu'habitait Albert Einstein* », paru en septembre 2016) par la presse et par un rapport commandé par le ministère de la Recherche [28] [29]. Une autre affaire de plagiat littéraire, qui avait fait parler d'elle en France plus tôt dans la décennie, est celle de la biographie d'Ernest Hemingway écrite par Patrick Poivre d'Arvor. En effet, le journal L'Express avait reçu une version de pré-édition pour pouvoir écrire une critique de la biographie, dans laquelle se trouvait plus d'une dizaine de paragraphes entiers plagiés. L'Express révèle « *que l'ancien présentateur du 20 Heures a plagié une biographie signée Peter Griffin, parue aux Etats-Unis, en 1985, aux éditions Oxford University Press. Traduite en France, chez Gallimard, en 1989, elle est aujourd'hui quasiment introuvable en librairie* » [30]. Ces passages seront ensuite retravaillés dans la version finale sortie en librairie quelques jours plus tard. À la même période, le journaliste Joseph Macé-Scaron a quant à lui laissé jusqu'à la sortie officielle de son roman « *Ticket d'entrée* », des emprunts pris mot pour mot du livre de Bill Bryson, « *American rigolos : chroniques d'un grand* », sans les retravailler et sans les référencer, donnant donc lieu à un cas de réel plagiat [31].

Les affaires de plagiat n'épargnent pas le monde de la politique. On peut par exemple citer le cas de Annette Schavan, la ministre allemande de l'éducation et de la recherche qui, début 2013, s'est vue contrainte de présenter sa démission à Angela Merkel suite à l'invalidité de son doctorat en philosophie pour constat de plagiat [32]. Plus récemment, en 2016, lors du passage du RNC (Republican National Committee) à Cleveland, Melania Trump, la femme de Donal Trump, le président des États-Unis d'Amérique, a plagié mot pour mot un discours de Michelle Obama, femme de l'ex-président des États-Unis, Barack Obama [33]. La même année, le président Nigérian, Muhammadu Buhari, avoue avoir lui aussi plagié lors d'un discours prononcé au lancement d'une campagne de civisme de nombreuses phrases provenant du discours de Barack Obama après sa victoire à l'élection présidentielle américaine en 2008 [34].

Toujours en politique, début 2017, la maison blanche *copie/colle* mot pour mot dans l'un de ses communiqués de presse un paragraphe d'une annonce faite quelques jours plus tôt par la compagnie pétrolière et gazière ExxonMobil [35]. Côté français, Marine Le Pen, alors candidate en lice au second tour de l'élection présidentielle, plagie lors de son discours du 1^{er} mai, un discours de François Fillon qu'il avait tenu deux semaines plus tôt [36].

27. www.lefigaro.fr/musique/2016/11/24/03006-20161124ARTFIG00028-calogero-condamne-definitivement-pour-plagiat.php (consulté le 31/03/2017 à 10h)
28. www.lexpress.fr/culture/livre/plagiat-les-copier-coller-du-physicien-etienne-klein_1855198.html (consulté le 31/03/2017 à 10h)
29. www.lexpress.fr/actualite/societe/accuse-de-plagiat-etienne-klein-clame-qu-il-ne-demissionnera-pas_1894025.html (consulté le 31/03/2017 à 10h)
30. www.lexpress.fr/culture/livre/le-plagiat-de-ppda_949676.html (consulté le 31/03/2017 à 10h)
31. www.lexpress.fr/culture/livre/joseph-mace-scaron-reconnait-un-plagiat-une-connerie_1022984.html (consulté le 31/03/2017 à 10h)
32. www.lemonde.fr/europe/article/2013/02/09/accusee-de-plagiat-de-these-la-ministre-allemande-de-l-education-demissionne_1829628_3214.html (consulté le 31/03/2017 à 10h)
33. www.lemonde.fr/big-browser/article/2016/07/19/quand-melania-trump-copie-mot-pour-mot-un-discours-de-michelle-obama_4971564_4832693.html (consulté le 31/03/2017 à 10h)
34. www.20minutes.fr/monde/1926587-20160918-nigeria-presidence-reconnait-plagiat-discours-obama (consulté le 31/03/2017 à 10h)
35. www.leparisien.fr/international/quand-la-maison-blanche-copie-colle-un-communique-d-exxonmobil-08-03-2017-6742932.php (consulté le 31/03/2017 à 10h)
36. www.leparisien.fr/elections/presidentielle/candidats-et-programmes/video-marine-le-pen-a-visiblement-plagie-le-discours-de-francois-fillon-01-05-2017-6907671 (consulté le 04/05/2017 à 14h)

Dernier exemple, en avril 2017, Free, l'opérateur téléphonique français, fête alors ses 5 ans et décide pour cela de commencer une nouvelle campagne publicitaire vantant les mérites de son forfait tout illimité. « *L'une des affiches en question s'est révélée n'être que la copie conforme d'une page du livre « Avec Maman », écrit par Alban Orsini et paru en 2014 chez Chiflet & Cie. Conforme ? Pas tout à fait, puisque Free a remplacé le nom du personnage de « Maman » par « Mamie ». Un changement tout de même bien maigre pour que le copié-collé ne se remarque pas...* » [37].

1.2.3 Le plagiat dans le milieu académique et l'enseignement

Le *copier/coller* touche particulièrement les étudiants. En Europe, 34,5% d'entre eux auraient déjà recopié tout ou partie d'un document pour le présenter comme travail personnel (Guibert *et* Michaut, 2011). Cette fréquence rejoint celle d'études américano-canadiennes (Josephson Institute, 2011; McCabe, 2010) estimant à plus de 36% la proportion d'étudiants de premier cycle et à 24% la proportion d'étudiants du supérieur ayant déjà ré-utilisé des phrases provenant d'Internet sans en citer la source. Une étude européenne (Gibney, 2006) révèle que près d'un étudiant français sur deux (46%) a déjà fait usage du plagiat pendant son cursus, contre environ 33% des étudiants anglais et 10% des étudiants allemands. La plupart d'entre eux avouent ré-utiliser plusieurs sources pour produire un seul document, ce plagiat est appelé "mosaïque" ou bien *patchwork* en anglais. Ces résultats, qui paraissent déjà impressionnants, pourraient pourtant encore être sous-évalués. En effet, toujours selon la même étude, 40% des étudiants ne comprennent pas ce que signifie réellement le plagiat et n'assimilent pas le *copier/coller* à de la tricherie.

À noter également, l'apparition d'un nouveau phénomène de plus en plus répandu au sein des devoirs étudiants, le *ghostwriting*. Plus communément appelé le phénomène des écrivains fantômes, en français on appelle cela l'emploi d'un nègre littéraire, un écrivain sous-traitant anonyme. Ici, c'est donc lorsque l'on paie quelqu'un pour directement rédiger tout ou partie d'un ouvrage que l'on fera ensuite passer pour sien. Étant donné que le contenu produit est unique et nouveau, le plagiat peut ne pas paraître évident, mais en effet, « *faire appel à un ghostwriter pour rédiger son mémoire, de master ou de doctorat, relève bien du phénomène de plagiat puisque l'on s'attribue un écrit qui n'est pas le sien pour obtenir des diplômes. C'est bien en termes de conséquences, une fraude au système* » [38].

Le plagiat n'est pas seulement monnaie courante chez les étudiants, c'est aussi un phénomène existant chez leurs enseignants. À l'instar de l'exemple de la ministre allemande de l'éducation évoqué dans la section précédente, début 2009, la directrice de l'école de journalisme de Sciences Po, Agnès Chauveau, accusée de plagiat elle aussi sur sa thèse, s'est vu remerciée [39], alors que « *les étudiants de l'École de journalisme de Sciences Po doivent tous signer une charte déontologique qui stipule, entre autres, que tout étudiant ne commet aucun plagiat* » [40]. Autre cas, un professeur de la Faculté des sciences économiques de l'Université de Neuchâtel, qui avait été soupçonné de plagiat fin 2015, a lui préféré démissionner plutôt que d'attendre une décision juridique [41].

Une étude (Fang *et al.*, 2012), menée sur les 25 millions d'articles publiés sur PubMed depuis les années 40, comprenant plus de 2 000 rétractations, rend état du fait que 24% des rétractations d'articles scientifiques sont dus à des plagiats, contre 21,3% pour des erreurs ou toutes autres formes de mauvaises conduites expérimentales. On peut ajouter à cela le fait que ce taux est en constante évolution, moins de 100 cas de plagiat entre les années 2002 et 2006, contre plus

37. https://www.actualitte.com/article/monde-edition/un-editeur-va-porter-plainte-contre-free-pour-plagiat/70877 (consulté le 24/04/2017 à 11h)
38. https://responsable-academia.org/action/validite-des-diplomes/ghostwriters/ (consulté le 15/04/2017 à 18h)
39. http://www.liberation.fr/ecrans/2014/11/17/la-directrice-de-l-ecole-de-journalisme-de-sciences-po-suspendue-pour-plagiat_1144633 (consulté le 31/03/2017 à 10h)
40. http://www.francetvinfo.fr/societe/education/accusee-de-plagiat-la-directrice-de-l-ecole-de-journalisme-de-sciences-po-remerciee_803359.html (consulté le 31/03/2017 à 10h)
41. http://www.20min.ch/ro/news/vaud/story/Le-professeur-accuse-de-plagiat-a-demissionne-12235820 (consulté le 31/03/2017 à 10h)

de 400 cas entre 2007 et 2011 [42]. D'après une autre étude, celle du site américain Retraction Watch, « *plus de 650 études, sur 2 millions publiées, ont été retirées dans le monde en 2016. Depuis 10 ans, la proportion des études retirées par rapport à celles publiées est en augmentation. [...] L'ampleur de la fraude est méconnue en France, bien que ces 5 dernières années 46 cas de plagiat ont été rapportés par l'ensemble des universités françaises, de même que 2 fabrications de résultats et 22 falsifications.* » [43] [44].

Francopoulo *et al.* (2016) ont récemment étudié le taux de réutilisation d'articles de conférences dans le domaine du traitement automatique des langues au sein des actes de la conférence sur les ressources linguistiques et leur évaluation (LREC) [45]. Il apparaît lors de cette étude que 445 articles (environ 10% des articles de LREC publiés sur 9 éditions, entre 1998 et 2014) sur plus de 4550 analysés, ont été ré-utilisés par leurs propres auteurs en citant ou non leurs papiers sources. En revanche, seulement des cas d'auto-plagiats sont donc à remonter. Aucun cas de plagiat réel n'est reporté durant leur étude.

1.2.4 La prévention et la lutte contre le plagiat

La standardisation du plagiat oblige les auteurs, les ayants droit, les éditeurs, ou bien encore les établissements délivrant des diplômes ou des certificats après validation d'un rendu textuel, à garantir l'originalité de ces contenus. C'est pour répondre à ce besoin que les logiciels anti-plagiat ont vu le jour. Ils sont de plus en plus nombreux à être utilisés au sein des institutions publiques et d'enseignements. Cependant, ces logiciels sont surtout utilisés dans un but de prévention plutôt que de répression, en complément de signatures de chartes d'intégrité et de campagnes de sensibilisation auprès des étudiants.

Les différentes institutions d'État et administrations académiques prennent conscience de ce problème qui peut mettre en danger leur crédibilité et leur notoriété, que ce soit au niveau national ou international.

Ce phénomène devient tellement alarmant qu'en Suisse, l'Institut International de Recherche et d'Action sur la Fraude et le Plagiat Académique a même mis en place un label anti-plagiat [46] visant à certifier et à garantir l'originalité des travaux d'un établissement labellisé. Ce label aura pour mission de certifier qu'un établissement applique correctement un programme anti-plagiat (formations, sensibilisation, contrôles et sanctions) et ainsi de garantir l'originalité de ses productions et recherches. Il s'obtient suite à des audits vérifiant le respect du programme anti-plagiat appliqué au sein de l'établissement demandeur.

D'une initiative moins individuelle et plus officielle, l'Office Français de l'Intégrité Scientifique (OFIS) a officiellement vu le jour en début d'année 2017. Créé suite aux préconisations du rapport [47] du professeur Pierre Corvol, du Collège de France, cet office « *tiendra lieu d'observatoire et de référence pour toutes les questions relatives à l'intégrité, il s'assurera également que des formations sur l'intégrité soient mises en place [...] et sera en mesure de faire valoir au niveau européen et international les actions des acteurs nationaux de la recherche publique* » [48].

42. http://www.pnas.org/content/109/42/17028.full.pdf (consulté le 31/03/2017 à 10h)
43. http://www.lequotidiendumedecin.fr/actualites/article/2017/03/23/loffice-francais-de-lintegrite-scientifique-officiellement-cree_845931 (consulté le 31/03/2017 à 10h)
44. http://www.enseignementsup-recherche.gouv.fr/pid20536/bulletin-officiel.html?cid_bo=114318&cbo=1 (consulté le 31/03/2017 à 10h)
45. http://www.lrec-conf.org (consulté le 19/04/2017 à 11h)
46. https://responsable-academia.org/organisations/label-anti-plagiat-sgs/label-anti-plagiat-introduction/ (consulté le 31/03/2017 à 10h)
47. http://www.enseignementsup-recherche.gouv.fr/cid104249/remise-du-rapport-de-pierre-corvol-bilan-et-propositions-de-mise-en-oeuvre-de-la-charte-nationale-d-integrite-scientifique.html (consulté le 31/03/2017 à 10h)
48. http://www.lequotidiendumedecin.fr/actualites/article/2017/03/23/loffice-francais-de-lintegrite-scientifique-officiellement-cree_845931 (consulté le 31/03/2017 à 10h)

1.3 Un phénomène peu contrôlé : le plagiat translingue

L'expansion d'Internet a facilité l'accès à des outils de traduction automatique de plus en plus fiables et surtout à des documents en langues étrangères autrefois plus difficiles d'accès. De ce fait, un nouveau type de plagiat a pu devenir plus fréquent, le plagiat par traduction, aussi appelé *plagiat translingue*. Ce plagiat implique l'emprunt d'un texte tout en le traduisant (manuellement ou automatiquement) de sa langue originale vers la langue du document dans lequel le plagiaire veut l'inclure.

Le plagiat translingue a pu prendre de l'ampleur dû au fait que le flou, évoqué lors de l'étude de Gibney (2006) qui démontre que la plupart des gens ne savent pas ce qui est du plagiat et ce qui n'en est pas, persiste et se ressent d'autant plus lorsque l'on parle de la réutilisation d'un texte après sa traduction [49]. En effet, plus les gens ont effectué des efforts pour camoufler leur emprunt, plus ils ont la sensation que ce dernier leur appartient désormais et qu'il ne s'agit plus d'un plagiat.

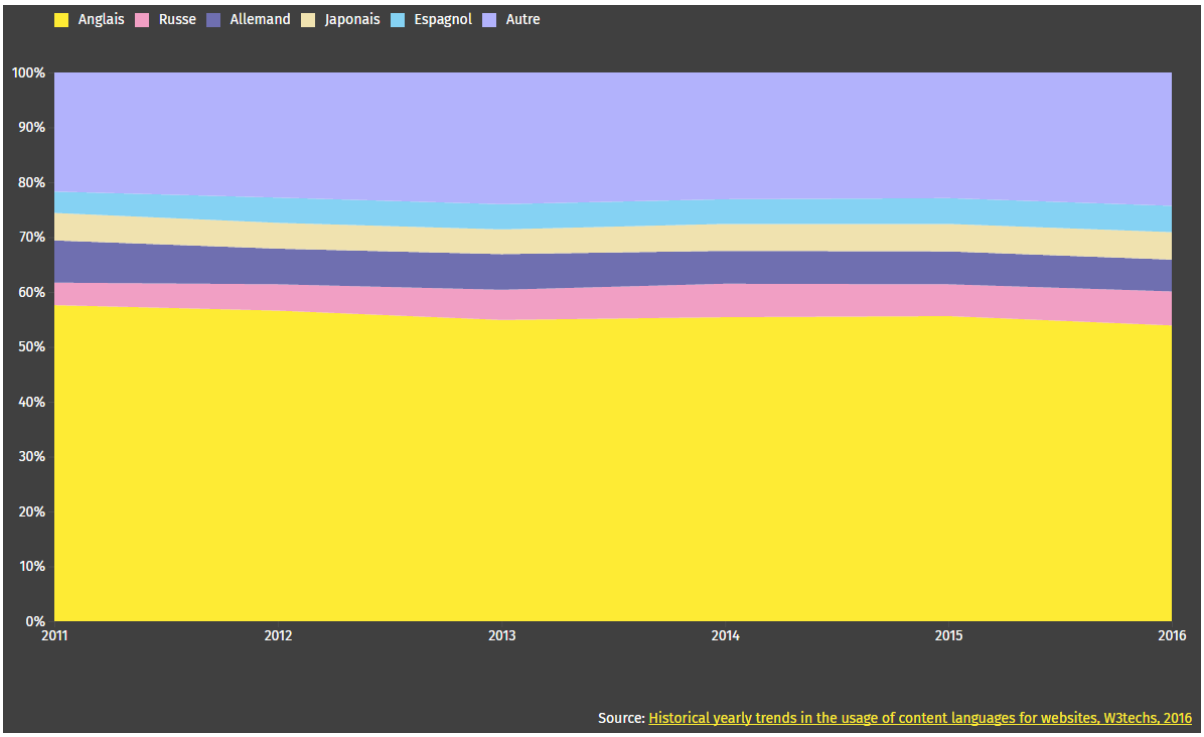


FIGURE 1.3 – Évolution au fil des années du pourcentage de couverture d'Internet par langue.
source : https://internethealthreport.org/v01/fr/digital-inclusion/

De plus, pour certains sujets ou domaines, il est plus aisé de trouver des sources dans une langue étrangère que dans sa propre langue. La Figure 1.3 détaille la répartition des langues majoritaires sur le Web et on peut y voir que plus de 50% des sites sont en anglais [50]. Ce phénomène est accentué selon les langues. En effet, ce dernier chiffre est contrasté par le fait que seulement 25% des habitants de la planète sont anglophones [51]. La Figure 1.4 représente le pourcentage d'internautes par rapport au pourcentage de contenu sur Internet dans une langue parlée par ces internautes. On peut y voir d'importantes disparités entre les demandeurs de contenu et les informations disponibles pour ces demandeurs (les gens pouvant lire et comprendre ce contenu). On peut par exemple y voir que le pourcentage de contenu Web en espagnol, en

49. http://writers.stackexchange.com/questions/20506/is-translating-to-other-language-plagiarism (consulté le 31/03/2017 à 10h)
50. https://w3techs.com/technologies/history_overview/content_language/ms/y (consulté le 31/03/2017 à 10h)
51. http://www.internetworldstats.com/stats7.htm (consulté le 31/03/2017 à 10h)

chinois ou en arabe est très nettement insuffisant par rapport au pourcentage de la population mondiale parlant ces langues. Certains utilisateurs peuvent donc se voir contraints d'aller chercher l'information dans d'autres langues et sachant que les contrôles anti-plagiat sont sans doute plus complexes à effectuer dans le cadre translingue, ils se laissent alors plus facilement tenter par le plagiat.

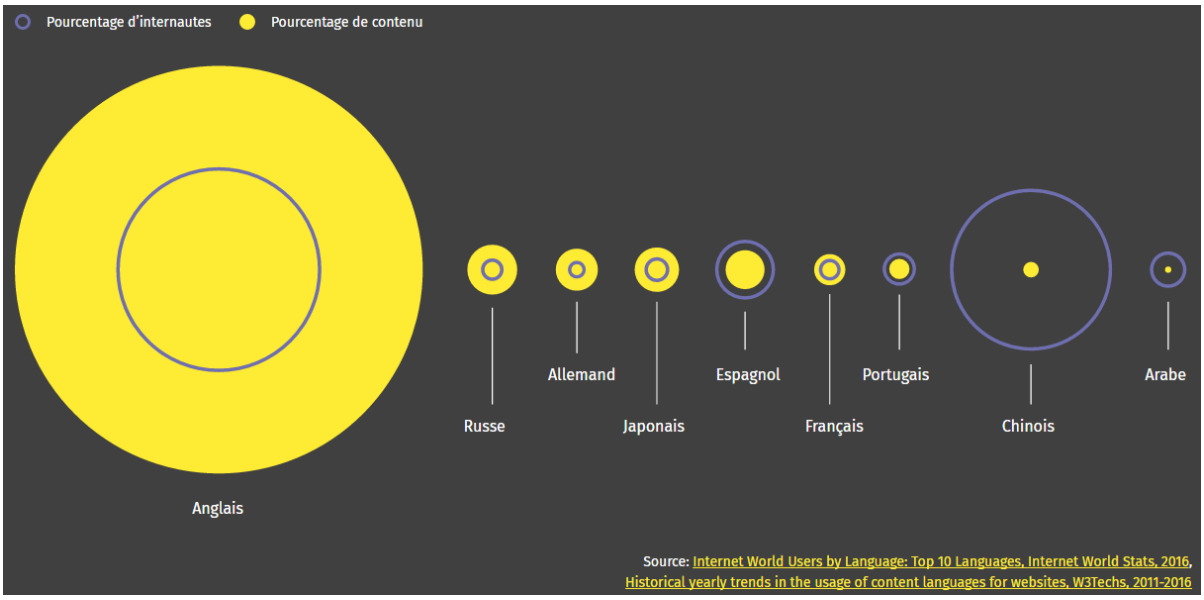


FIGURE 1.4 – Pourcentage d'internautes locuteurs d'une langue par rapport au pourcentage de contenu sur Internet dans cette langue.
source : https://internethealthreport.org/v01/fr/digital-inclusion/

Bien qu'il n'y a encore aucune étude traitant directement du plagiat translingue, l'étude de McCabe (2010), qui affirme que plus d'un tiers des étudiants ont déjà plagié des documents provenant du Web, comprend les cas de plagiat translingue. De plus, une entreprise de veille et de prévention du plagiat sur Internet comme Compilatio reçoit chaque année plusieurs demandes de ses clients pour répondre au besoin de détecter le plagiat translingue. On peut donc supposer que le plagiat translingue est un phénomène bien répandu.

La difficulté dans la détection de ce type de plagiat réside dans le fait que le document sur lequel pèsent les soupçons n'est pas écrit dans la même langue que sa prétendue source. La problématique peut revenir à détecter si un texte est bien la traduction d'un autre, mais s'en éloigne du fait que l'on ne cherche pas à évaluer la justesse d'une traduction mais seulement à évaluer si deux textes écrits dans deux langues différentes signifient ou non la même chose (expriment ou non les mêmes idées). Des recherches (Potthast *et al.*, 2011a; Eissen *et* Stein, 2006) soulignent que la traduction se trouve dans le même registre de modification textuelle que la reformulation. En effet, le sens principal de la phrase est conservé mais le vocabulaire employé et l'ordre des idées exprimées ont certainement changé.

Le chapitre suivant va s'efforcer d'établir un état de l'art des différentes techniques existantes pour détecter les différents types de plagiat.

2 La prévention du plagiat

> *« If you steal from one author, it's plagiarism; if you steal from many, it's research. »* [1]
>
> — Wilson Mizner (1876-1933)

Nous avons vu au cours de la section précédente que le plagiat est un problème touchant tous les milieux de création de contenu et qu'il se répand au fur et à mesure que ce contenu se digitalise. Il devient alors nécessaire, afin de garantir les droits et propriétés intellectuelles de chacun, de prévenir les cas de plagiat. Pour cela, le traitement automatique du langage naturel semble être l'une des meilleures solutions. La Figure 2.1 représente, selon les travaux de Eissen *et* Stein (2006), la taxonomie des différents types de plagiat ainsi que de leurs différentes méthodes de détection par traitement automatique du langage. Ce chapitre va d'abord établir un rapide état de l'art sur les techniques de détection du plagiat monolingue extrinsèque (1) avant de faire de même pour les techniques de détection intrinsèque (2). Ensuite un état de l'art plus complet ainsi qu'une étude sur l'ensemble des méthodes de détection du plagiat translingue (3) sera effectué.

Malgré l'assimilation du problème de la détection du plagiat avec la tâche d'extraction de phrases parallèles au sein d'un corpus comparable (voir section 3.1) ou la tâche de détection de similarités sémantiques textuelles (voir section 3.6), nous ne traiterons ici que les méthodes éprouvées dans la littérature de la détection du plagiat et non chaque méthode de détection de similarités textuelles existantes.

2.1 La prévention du plagiat monolingue

De nos jours de nombreuses recherches s'intéressent à la détection du plagiat. Certaines se concentrent plutôt sur l'alignement des passages similaires entre deux textes (Barrón-Cedeño *et* Rosso, 2009; Barrón-Cedeño *et al.*, 2013b). D'autres techniques, tentent d'analyser les changements de style d'écriture au sein d'un texte pour en détecter les zones suspectes (Oberreuter *et* Velásquez, 2013) et ainsi faire l'hypothèse que ces zones n'ont pas été écrites par le même auteur que le reste du texte et donc qu'elles ont été plagiées.

Le premier cas est ce que l'on appelle la détection de plagiat extrinsèque. Soit un jeu de documents suspects et un jeu de documents sources potentiels, la tâche est de trouver des passages plagiés dans les documents suspects et les passages sources correspondants dans les documents sources. Le second, est quant à lui appelé la détection de plagiat intrinsèque. Soit un jeu de documents suspects, la tâche est d'extraire tous les passages plagiés sans les comparer à des documents sources potentielles, donc sans avoir accès à un corpus de documents sources externes.

La campagne d'évaluation PAN [2] s'impose depuis plusieurs années comme la campagne d'évaluation de référence dans les tâches assimilées à la détection du plagiat. Elle a fortement contribué à l'avancée de l'état de l'art en matière de détection du plagiat extrinsèque et intrinsèque. Cette section va tenter de résumer les différentes avancées effectuées dans ces deux types de détection de plagiat.

1. *« Copier sur un seul, c'est du plagiat. Copier sur plusieurs, c'est de la recherche. »*
2. http://pan.webis.de (consulté le 15/04/2017 à 19h)

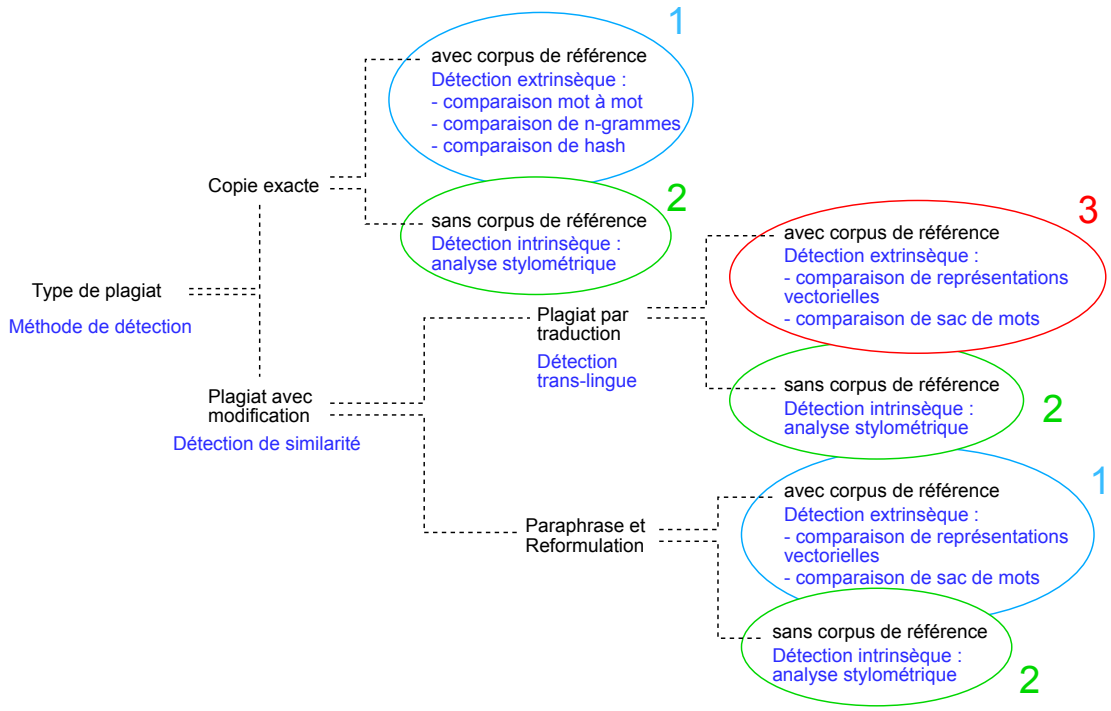


FIGURE 2.1 – Adaptation de la taxonomie de Eissen *et* Stein (2006) des différents types de plagiat et de leurs moyens de détection.

2.1.1 La détection extrinsèque

La tâche de détection de plagiat extrinsèque est souvent découpée en deux sous-tâches, la tâche de collecte de documents sources candidats et la tâche de comparaison (la recherche d'alignements de passages similaires) de documents deux à deux, entre le document suspect en cours d'analyse et chacune des sources renvoyées par la tâche de collecte. Le sujet de la thèse concernant seulement la seconde tâche, nous porterons notre attention uniquement sur celle-ci : la détection de similarités textuelles entre deux documents.

Les *copier/coller* sont en théorie les similitudes textuelles les plus facilement repérables et identifiables. En effet, la détection de ceux-ci équivaut à vérifier l'égalité entre deux textes. Pour effectuer naïvement cette recherche de façon automatique, on est obligé de procéder à une comparaison mot à mot des textes. Cette opération, étant beaucoup trop chronophage pour être intégrée dans des solutions à but commercial ou hébergées en ligne, comme c'est le cas des services anti-plagiat, des techniques alternatives ont dû être mises au point. Les méthodes les plus répandues restent à l'heure actuelle celles à base de n-grammes, consistant à représenter les textes sous forme de séquences de n éléments pouvant être des lettres, des syllabes, des mots, des entités nommées, *etc.* afin de les comparer plus facilement par la suite sur une granularité plus fine (Manning *et* Schütze, 1999).

La campagne d'évaluation PAN a arrêté en 2014 sa dernière tâche pouvant être apparentée à la détection du plagiat extrinsèque (les tâches officielles traitant ce sujet avaient déjà toutes été arrêtées en 2011). Depuis, cette campagne se focalise sur des tâches d'identification et de vérification d'auteurs. Lors de la dernière année de la tâche de détection du plagiat extrinsèque, Grman *et* Ravas (2011) remportèrent la compétition avec un système basé sur l'intersection de mots. La recherche de Barrón-Cedeño *et* Rosso (2009) avait prouvé qu'en prenant des n-grammes de mots (séquence de n mots se suivant dans un texte) de petites tailles, 2 ou 3 par exemple, les résultats sont bien meilleurs qu'en utilisant des longues séquences avec un n plus important. Grman *et* Ravas (2011) ont établi qu'à partir de 15 mots consécutifs communs, deux passages dans deux textes différents peuvent être considérés comme suffisamment suspects pour

être relevés. Toutefois, les n-grammes les plus pertinents à utiliser lors d'une comparaison ne sont pas toujours des séquences de mots, comme en atteste le travail de Shrestha *et* Solorio (2013). Des n-grammes de mots vides et d'entités nommées peuvent également être utilisés pour détecter des parties de textes similaires entre deux documents. Les mots vides sont des mots vides de sens n'apportant aucune valeur sémantique supplémentaire dans un texte, ce sont le plus souvent des articles, des prépositions, des pronoms *etc.* Les entités nommées sont des expressions, des groupes de mots faisant référence à des entités remarquables au sein d'un texte, comme un nom d'entreprise, un lieu ou une date par exemple.

D'autres méthodes assez répandues sont les méthodes à base d'empreinte textuelle, créant une empreinte du document pour la comparer avec celle d'autres documents. La plupart de ces méthodes, à l'instar des travaux de Kent *et* Salim (2010a), utilisent également des n-grammes pour construire l'empreinte des documents. Les méthodes à base d'empreinte textuelle divisent la plupart du temps les documents en n-grammes, ainsi les empreintes de deux documents peuvent être comparées. Certaines méthodes à base d'empreinte textuelle (Stein *et* Eissen, 2006, 2007a; Lyon *et al.*, 2001) vont au-delà de la recherche de similitudes exactes et introduisent la notion de *similarités proches* pouvant en théorie ainsi détecter les paraphrases.

Plus tard, Sanchez-Perez *et al.* (2014) proposent lors de la dernière tâche d'alignement textuel de la PAN, une mesure de similarité entre phrases basée sur un modèle vectoriel pondéré avec des fréquences de termes permettant de conserver les mots vides tout en évitant d'incrémenter le taux de faux positifs. Ils introduisent un algorithme récursif pour étendre la correspondance des phrases en construisant des passages de longueur maximale. Pour cela si deux fragments similaires dans les deux textes sont contigus avec deux autres fragments également similaires dans les deux textes, on les concatène pour former un fragment similaire plus large. Le but étant de trouver les plus longues séquences communes entre les deux textes.

Plus récemment, avec l'arrivée des réseaux de neurones, de nouvelles approches ont vu le jour. Yin *et* Schütze (2015), par exemple, utilisent des *word embeddings* (Mikolov *et al.*, 2013a) (voir section 2.2.6) avec une pondération pour les mots jugés plus importants dans le cadre de l'identification d'une paraphrase, pour constituer des représentations vectorielles de phrases plus pertinentes à employer lors d'une tâche de détection de paraphrases.

Cependant, la détection de plagiat extrinsèque devient inefficace lorsque l'on n'a pas accès aux documents potentiellement sources du plagiat ou lorsque l'on se confronte à un espace aussi vaste que le Web, ce qui est souvent le cas dans les logiciels anti-plagiat actuels. La détection de plagiat de façon intrinsèque (une analyse d'un document en interne) tente alors de pallier ce problème.

2.1.2 La détection intrinsèque

La grande majorité des techniques de détection du plagiat de façon intrinsèque utilisent des approches dites stylométriques. Ces dernières suggèrent qu'en analysant les caractéristiques d'un texte, on peut en reconnaître l'auteur, et ainsi, si un passage du document ne possède pas les mêmes caractéristiques que le reste du document, on peut en déduire que ce passage aura été emprunté à un autre auteur.

La stylométrie ou l'étude stylométrique d'un texte est une analyse à mi-chemin entre une analyse linguistique et statistique. Elle exploite des variables stylométriques, qui sont des caractéristiques linguistiques du texte, afin d'établir des statistiques propre à l'écriture du texte étudié. Effectuer l'analyse stylométrique d'un document consiste à surveiller les variations du style d'écriture du document en surveillant l'évolution des variables stylométriques au sein de celui-ci afin d'en détecter les irrégularités et ainsi pouvoir déterminer si certains passages, appelés phases stylistiques, sortent de la norme par rapport au reste du texte. Si de tels passages existent, dans le cadre d'une détection intrinsèque de plagiat, on pourra faire l'hypothèse que ces passages ont été écrits par un auteur différent (du reste du texte) et donc qu'ils ont été plagiés. Cette approche de détection du plagiat est utile lorsqu'aucune collection de référence n'est disponible

et que donc aucun algorithme de comparaison de documents deux à deux ne peut être appliqué pour retrouver les portions de textes similaires dans deux documents comparés.

Dès le 19e siècle, Mendenhall (1887) suggère qu'en analysant des caractéristiques internes d'un texte on peut en reconnaître l'auteur. Van Halteren (2004) a été le premier à introduire la notion de surveillance de plusieurs variables stylométriques, d'abord lexicales, puis syntaxiques, au fil d'un document pour en identifier les divers auteurs. Les auteurs utilisent des caractéristiques telles que la fréquence des mots, la taille des phrases ou bien encore la fréquence de divers n-grammes (séquences de n éléments contigus, pouvant être des caractères, des mots ou bien des groupes de ces deux derniers éléments). Eissen *et* Stein (2006) ont ensuite été les premiers à réutiliser ces approches pour la détection de plagiat intrinsèque. La tâche consiste à déterminer si des passages d'un document sont plagiés, en se fondant seulement sur le fait qu'ils sont suspects car ne concordant pas avec le reste du style d'écriture du document.

Stein *et al.* (2011) établissent un état de l'art très complet et décrivent de façon approfondie la tâche et ses limites. La plupart des méthodes de détection de plagiat intrinsèque fonctionnent de la même façon : elles consistent à découper le texte en blocs puis à représenter chacun de ces blocs par des ensembles de données stylistiques avant de tenter de détecter les blocs aberrants par rapport aux autres par le biais de différentes techniques de calcul de distance ou de classification.

La Figure 2.2, tirée du travail de Oberreuter *et* Velásquez (2013), illustre un exemple de cas de détection de plagiat intrinsèque par analyse stylométrique d'un document. Au fur et à mesure de la ligne de vie du document, une fenêtre d'analyse se déplace étudiant la stylométrie du document. Les passages ayant une stylométrie trop éloignée de la moyenne du document en fonction d'un seuil, sont jugés comme étant plagiés.

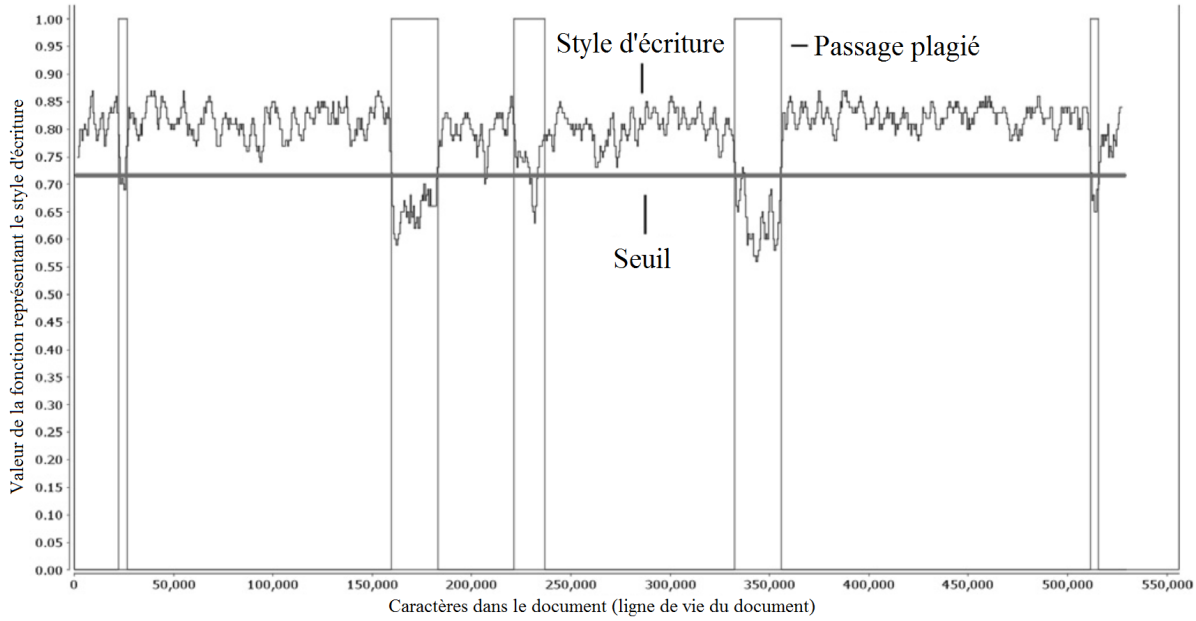


FIGURE 2.2 – Exemple de cas de détection de plagiat intrinsèque par analyse stylométrique d'un document (Oberreuter *et* Velásquez, 2013).

Certaines de ces recherches se concentrent sur l'extraction et la surveillance des données stylométriques les plus pertinentes. Stein *et* Eissen (2007b) ainsi que Zamani *et al.* (2014) et Ferrero *et* Simac-Lejeune (2015) surveillent les proportions d'utilisation des parties du discours (classes grammaticales) au sein de segments afin de discerner lesquelles sont les plus caractéristiques du style de l'auteur. Oberreuter *et* Velásquez (2013) privilégient, quant à eux, la fréquence des termes comme donnée stylométrique à surveiller. Cependant, les données stylistiques s'avérant être les plus efficaces se trouvent souvent être à base de n-grammes (Stamatatos, 2009; Jayapal *et* Goswami, 2013; Layton *et al.*, 2013). Jayapal *et* Goswami (2013) utilisent une mesure de

similarité d'intersection de vecteurs de n-grammes. Layton *et al.* (2013) utilisent un modèle à base de n-grammes pour créer des profils, consistant en une liste de n-grammes de caractères représentant le mieux le style d'écriture d'un auteur particulier. L'algorithme utilisé consiste à garder seulement les n-grammes de caractères les plus fréquemment utilisés par un auteur pour représenter cet auteur. Ensuite, ils calculent la distance entre un texte et la moyenne des distances des textes d'un corpus pour déterminer si ce texte s'apparente à ce corpus et donc à l'auteur ayant rédigé les textes de ce corpus. Kestemont *et al.* (2011), quant à eux, utilisent aussi des n-grammes mais ils font l'hypothèse que comparer les données stylistiques d'une fenêtre avec une autre fenêtre est plus pertinent que de comparer une fenêtre avec le reste du texte (ou un corpus de textes).

L'une des limites de toutes ces approches est que le style de l'auteur semble souvent être lié ou influencé par le sujet qu'il traite. Lorsque le domaine traité change, certaines variables stylistiques changent également, ce qui rend donc leur pertinence moindre et fait échouer certaines méthodes de détection. Pour limiter ce problème, Stamatatos (2017) met au point une technique robuste au changement de sujet en appliquant une phase de distorsion de texte avant la phase d'extraction des données stylométriques.

En 2016, la campagne d'évaluation PAN introduit une nouvelle tâche qu'elle assimile à la détection de plagiat intrinsèque, la tâche de reconnaissance d'auteur (*Author Diarization*). Étant donné un document, il s'agit d'identifier et grouper les passages de ce document qui ont été écrits par le même auteur. Cette tâche a permis de ré-actualiser la littérature sur la détection de plagiat intrinsèque. Kuznetsov *et al.* (2016) se basent sur la même architecture que les méthodes déjà éprouvées. Ils segmentent le texte en blocs, ils représentent ensuite chaque bloc par des vecteurs de fréquences de mots, de n-grammes, de ratio de ponctuation et de parties du discours et les classifient ensuite en utilisant un algorithme d'apprentissage par arbre de décision.

On peut également assimiler par extension, la tâche de détection de plagiat intrinsèque aux tâches d'identification, de vérification et de regroupement d'auteurs. En effet, les recherches visant à représenter au mieux le style d'écriture d'un auteur ou à différencier deux styles d'écriture tendent à faire avancer la détection de plagiat intrinsèque. De ce coté là, des propositions plus novatrices (Sari *et* Stevenson, 2016; Bagnall, 2016) sont apparues dernièrement avec l'arrivée des *word embeddings* (Mikolov *et al.*, 2013a) (voir section 2.2.6) ou des réseaux de neurones. Par exemple, en contraste avec les travaux précédemment évoqués, la méthode de Sari *et al.* (2017) utilise un modèle apprenant des représentations continues de n-grammes par le biais d'un réseau de neurones, plutôt que d'utiliser des représentations de données discrètes. On peut également citer les travaux de Shrestha *et al.* (2017), qui utilisent un réseau de neurones convolutionnel sur des séquences de caractères.

À noter que ces techniques visant à caractériser le style d'écriture d'un auteur et ainsi à différencier deux auteurs, peuvent aussi servir à déceler des cas de *ghostwriting*. Récemment, une société suisse, OrphAnalytics [3], a fait parler d'elle pour avoir analysée les ouvrages de Nicolas Sarkozy et François Bayrou, ou bien encore la saga des romans Millénium. La Figure 2.3 représente l'analyse par cette société de la célèbre série de romans suédois Millénium. « *Chaque point représente l'identité d'un segment de texte. En théorie, l'empreinte d'une personne devrait former un seul nuage de points. Les axes représentent des valeurs statistiques utilisées par le logiciel. Les trois premiers ouvrages ont été écrits par Stieg Larsson, décédé en 2004, alors que le quatrième, publié en 2015, est signé David Lagercrantz. L'analyse des livres confirme que les trois premiers textes (en rouge, orange et jaune) se recoupent dans la même empreinte stylistique, alors que celle du dernier est très distincte. Bien qu'ils traitent de thèmes totalement différents, en l'occurrence du mathématicien Alan Turing et de l'Everest, d'autres livres de David Lagercrantz (en bleu et violet) voient leur empreinte rejoindre celle du dernier Millénium* » [4].

3. http://www.orphanalytics.com/ (consulté le 15/04/2017 à 19h)
4. http://www.rts.ch/info/suisse/7517532-la-traque-aux-ghostwriters-pour-etudiants.html (consulté le 15/04/2017 à 19h)

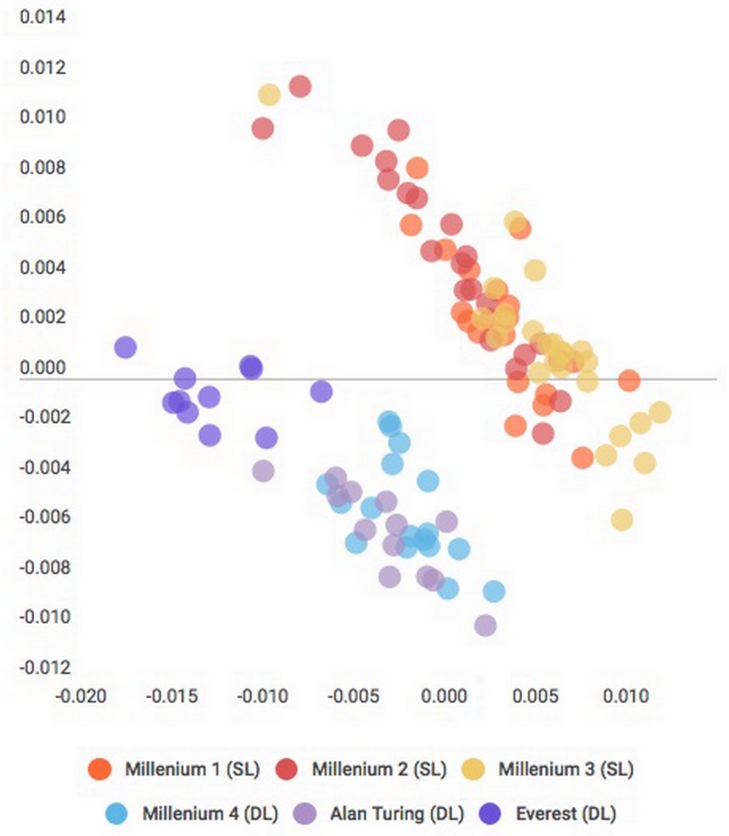


FIGURE 2.3 – Analyse stylométrique des romans de la saga Millénium.

2.2 La détection du plagiat translingue

La détection du plagiat translingue est la problématique qui nous intéresse particulièrement dans cette thèse. Pour le moment, il existe cinq familles d'approches de détection de plagiat translingue : les modèles basés sur une analyse lexicale ou syntaxique, ceux qui se servent de ressources lexicales ou sémantiques (bases de connaissances externes), ceux basés sur les corpus parallèles, ceux basés sur les corpus comparables et les modèles qui utilisent en premier lieu une phase de traduction automatique afin d'opérer par la suite une comparaison monolingue.

La Figure 2.4 représente la taxonomie des différentes approches de détection du plagiat translingue, basée sur les travaux de Potthast *et al.* (2011a) et Danilova (2013). La méthode marquée d'une * dans la taxonomie est une méthode classée dans la littérature dans deux types d'approches différents.

Cet état de l'art a pour but d'expliquer la méthodologie de chacune de ces familles d'approches et de répertorier les principales méthodes de chacune d'entre elles. En fin de section, nous proposons également une synthèse de toutes les études comparatives effectuées sur ces méthodes et verrons ainsi les avantages et inconvénients de chacune.

2.2.1 Modèles basés sur le lexique et la syntaxe

La première famille est l'ensemble des modèles basés sur la comparaison lexicale ou syntaxique. L'avantage de ces méthodes est qu'elles ne nécessitent aucune traduction des textes à comparer. Elles sont entièrement basées sur une analyse lexicale ou syntaxique et demandent seulement, dans la majorité des cas, une étape de nettoyage ou *pre-processing* (suppression des espaces, suppression des mots vides de sens, suppression des caractères diacritiques, *etc.*), ce qui justifie un temps de traitement assez rapide (aussi rapide que dans les cas monolingues).

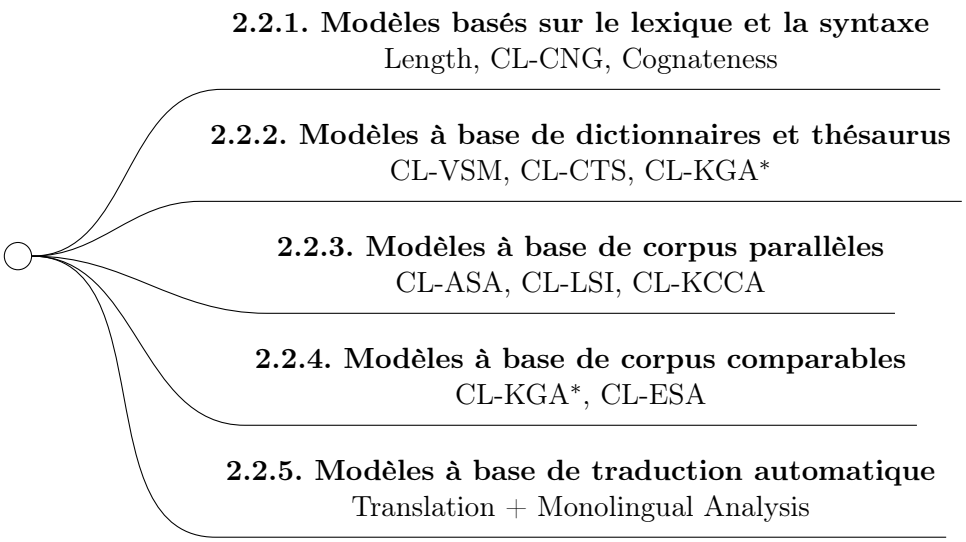


FIGURE 2.4 – Adaptation de la taxonomie des différentes approches de détection du plagiat translingue, basée sur les travaux de Potthast *et al.* (2011a) et Danilova (2013). La méthode CL-KGA, marquée d'une * dans la taxonomie, est une méthode classée dans la littérature dans deux types d'approches différents.

2.2.1.1 Vecteurs translingues de n-grammes de caractères (*Cross-Language Character n-Gram, CL-CnG*)

Ce modèle a pour objectif de représenter deux documents écrits dans deux langues différentes sous forme de vecteurs de n-grammes, les rendant ainsi comparables de façon triviale et indépendante de leur langue d'origine. Cette technique est basée sur les travaux de Mcnamee *et* Mayfield (2004) utilisés dans la recherche de documents. Cette méthode donne de bons résultats en recherche d'informations pour les langues avec la même origine et ce en raison des nombreux mots à racine commune au sein de ces langues.

Soit d et d' deux documents dans deux langues différentes (respectivement L et L'). Tout d'abord on normalise leur alphabet sur un espace $\sum = \{a-z, 0-9\}$, c'est-à-dire que l'on y conserve seulement les caractères alphanumériques. Tout autre caractère diacritique (accent, cédille, *etc.*), symbole (+, -, _, *etc.*) ou espace est alors supprimé et l'ensemble des lettres est passé en minuscule. Les textes résultant de ce filtrage sont ensuite segmentés en n-grammes (ici des séquences de n caractères contigus se suivant). Selon les études, on convient que $n = [3\,;\,5]$ – Mcnamee *et* Mayfield (2004) utilisent un modèle CL-C4G, tandis que Potthast *et al.* (2011a) préfèrent utiliser un modèle CL-C3G. Les textes sont ainsi codés sous forme de vecteurs de n-grammes dont les poids sont définis par le modèle standard *tf.idf* (*Term Frequency Inverse Document Frequency*) (Salton *et* Buckley, 1988). Cette méthode de pondération multiplie la fréquence d'apparition d'un terme par la fréquence inverse de document du même terme au sein d'un corpus. Elle prend donc en considération la fréquence d'apparition des n-grammes dans un texte vis à vis de leur fréquence d'apparition dans un corpus de textes censé représenter leur usage standard dans leur langue.

Soit un terme t dans un document d, on peut exprimer sa pondération *tf.idf* relative à un corpus D de la façon suivante :

$$tf.idf(t,d) = tf(t,d) \cdot idf(t,D) \tag{2.1}$$

où $tf(t,d)$ est défini comme la fonction retournant le nombre d'occurrences du terme t dans le document d et *idf* est défini comme suit :

$$idf(t, D) = \log\left(\frac{\mid D \mid}{n(t, D)}\right) \tag{2.2}$$

avec au numérateur le nombre de documents que comporte en tout le corpus D servant à calculer l'*idf* et $n(t, D)$ la fonction définie comme suit, qui retourne le nombre de documents du corpus D dans lesquelles apparaissent le terme t.

$$n(t, D) = \mid d \in D : t \in d \mid \tag{2.3}$$

On obtient donc deux vecteurs de n-grammes pondérés par *tf.idf*, v et v', représentant respectivement les documents d et d'. Pour finir, une mesure de similarité cosinus (Salton, 1989) entre ces deux vecteurs est opérée pour calculer la similarité entre les deux textes qu'ils représentent. La similarité cosinus entre deux vecteurs, notée φ, est définie par le rapport entre le produit scalaire de ces deux vecteurs et la multiplication de leur norme :

$$\varphi(v, v') = \frac{v \cdot v'}{\parallel v \parallel \cdot \parallel v' \parallel} = \frac{\sum\limits_{k \in (v \cap v')}^{|v \cap v'|} (v_k \cdot v'_k)}{\sqrt{\sum\limits_{i=1}^{|v|} (v_i)^2} \cdot \sqrt{\sum\limits_{j=1}^{|v'|} (v'_j)^2}} \tag{2.4}$$

où les ensembles des i et des j représentent respectivement l'ensemble des éléments contenus dans v et v', et k est l'ensemble des éléments communs aux vecteurs v et v'.

La Figure 2.5 illustre le fonctionnement de la méthode CL-C3G sur deux phrases exemples, une en français et l'autre en anglais.

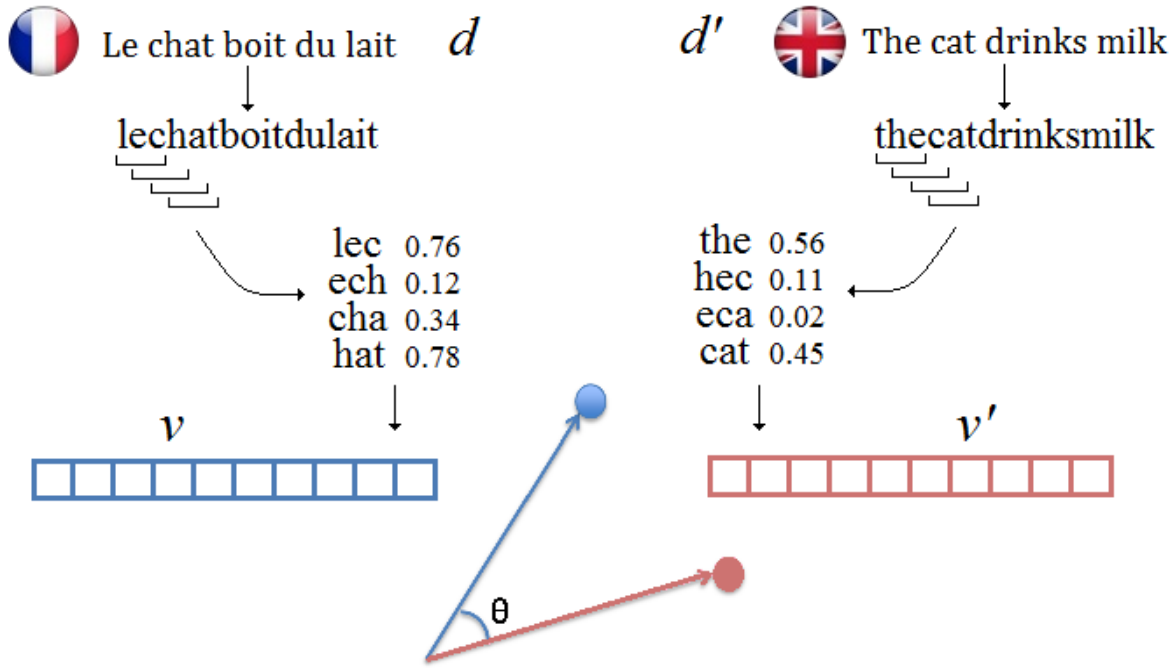


FIGURE 2.5 – Fonctionnement du modèle CL-C3G. Les deux phrases à comparer sont d'abord réduites sur un alphabet alphanumérique (caractères alphabétiques minuscules et caractères numériques). Elles sont ensuite représentées sous forme de vecteurs de 3-grammes pondérés avec du *tf.idf*. Ces deux vecteurs sont enfin comparés avec une similarité cosinus.

2.2.1.2 Correspondance de mots apparentés (*Cognateness*)

Ce modèle est essentiellement basé sur la comparaison de préfixes et de nombres. On le retrouve pour la première fois dans les travaux de Simard *et al.* (1993), où il est utilisé afin d'aligner des phrases dans un corpus parallèle. Depuis, il a inspiré de nombreux travaux, notamment des travaux d'alignement de corpus parallèles (Enright *et* Kondrak, 2007) ou de corpus comparables (Morin *et al.*, 2015), ainsi que des travaux dans la détection du plagiat translingue (Barrón-Cedeño *et al.*, 2014).

Dans leurs travaux, un terme t est un mot candidat pouvant être apparenté à un autre mot (ce qu'ils appellent un *cognate*), s'il rentre dans l'une de ces trois catégories :

— (a) : t contient au moins un chiffre ;
— (b) : t est uniquement constitué de lettres et sa taille est supérieure ou égale à quatre caractères ($| t | \geq 4$) ;
— (c) : t est un signe de ponctuation.

t et t' sont des mots apparentés (*pseudo-cognates*) si les deux appartiennent à la catégorie (a) ou (b) et sont identiques, ou si les deux appartiennent à la catégorie (b) et partagent les mêmes 4 premières lettres. Autrement dit, on peut synthétiser par les deux règles suivantes, les termes t et t' sont apparentés si :

— $t \in (a)$ OU $t \in (b)$ ET $t = t'$;
— $t \in (b)$ ET $t' \in (b)$ ET $head(t, 4) = head(t', 4)$;

avec $head(t, x)$ la fonction retournant les x premières lettres du terme t.

Par conséquent, un document est en fait découpé en termes et ramené à un alphabet $\sum = \{a-z, 0-9\}$ où tous les symboles et signes diacritiques (accents, cédilles, *etc.*) sont supprimés et les lettres sont passées en minuscule. Les termes qui représentent des mots (constitués de lettres) sont réduits à leur préfixe (leurs 4 premières lettres). Posant le postulat que les mots de moins de 4 lettres ne sont pas conservés car jugés comme étant des mots vides (vide de sens).

Le score de similarité entre les deux textes est ensuite calculé par un simple ratio entre le nombre de termes apparentés sur le nombre de termes totaux évalués.

2.2.1.3 Modèle de longueur (*Length*)

Ce modèle a pour but de comparer les tailles de deux textes afin de prédire s'ils expriment la même chose ou non. Bien qu'il soit peu probable que deux textes d et d' signifiant la même chose dans deux langues différentes (respectivement L et L') aient exactement la même longueur, tel que $| d | = | d' |$, on convient qu'il est probable que leur taille soit liée par un certain facteur. Pouliquen *et al.* (2003b) observent qu'il existe d'ailleurs un facteur différent pour chaque paire de langues. Ils observent par exemple que les traductions espagnole et française utilisent en moyenne respectivement 13,5% et 18% plus de caractères que le document original en anglais. Une fois ce nombre connu, ils ont isolé la formule suivante, reprise plus tard dans les travaux de Potthast *et al.* (2011a) :

$$\varrho(d, d') = exp\left(-0.5\left(\frac{\frac{| d' |}{| d |} - \mu}{\sigma}\right)^2\right) \tag{2.5}$$

où μ et σ sont respectivement la moyenne et l'écart-type du rapport des longueurs (en nombre de caractères) entre les textes originaux et leur traduction. Le Tableau 2.1 présente les valeurs de μ et σ pour les différentes paires de langues qui sont utilisées dans l'évaluation des travaux de Potthast *et al.* (2011a).

Paramètres	en → de	en → es	en → fr	en → nl	en → pl
μ	1,098	1,138	1,093	1,143	1,216
σ	0,268	0,631	0,157	1,885	6,399

Tableau 2.1 – Coefficients des variations des moyennes et écarts-types entre des paires de langues selon Potthast *et al.* (2011a). Chaque langue est notée par son code à deux lettres sous la norme ISO 639-1 (dit *alpha-2*).

2.2.2 Modèles à base de dictionnaires et thésaurus

Cette famille de méthodes repose sur l'utilisation de bases de connaissances lexicales ou conceptuelles externes, comme des dictionnaires, des thésaurus ou des réseaux sémantiques.

2.2.2.1 Ressources lexicales et conceptuelles

Le premier type de ressources qui peut être exploité est l'inventaire de sens, c'est-à-dire une ressource qui associe à chaque mot des sens possibles ou des définitions, comme par exemple un dictionnaire (Collins, 1988), un dictionnaire en ligne comme le Wiktionary [5], ou une encyclopédie en ligne comme Wikipédia [6]. Un dictionnaire est « *un recueil des mots d'une langue ou d'un domaine de l'activité humaine, réunis selon une nomenclature d'importance variable et présentés généralement par ordre alphabétique, fournissant sur chaque mot un certain nombre d'informations relatives à son sens et à son emploi et destiné à un public défini* » [7].

Le second type de ressources est l'inventaire de synonymes comme, par exemple, les thésaurus ou les dictionnaires de synonymes. Un thésaurus est « *une structuration hiérarchisée d'un ou plusieurs domaines de la connaissance et dans lequel les notions sont représentées par des termes d'une ou plusieurs langues naturelles et les relations entre notions par des signes conventionnels* » [8].

D'autre part, des ressources telles que les ontologies ou les réseaux sémantiques peuvent aussi être utilisées pour établir des liens entre les sens des différents mots. Un réseau sémantique est « *un graphe orienté pondéré et étiqueté où les nœuds représentent des concepts et des entités nommées multilingues et où les arêtes expriment les relations sémantiques entre ces nœuds* » (Navigli *et* Ponzetto, 2012; Franco-Salvador *et al.*, 2013c).

Nous allons présenter ici seulement les ressources sémantiques les plus couramment utilisées en détection de similitudes translingue.

La grande majorité de ces ressources est structurée de façon très similaire, à l'instar de WordNet (Miller, 1985; Miller *et al.*, 1988) qui joue le rôle d'inventaire de sens, mais donne également accès à une hiérarchie de ses sens (en quelque sorte un thésaurus structuré). WordNet est structuré autour de la notion de *synsets*, c'est-à-dire un ensemble de mots formant un concept. Un concept est une idée représentée par un ensemble de mots partageant le même sens, souvent des synonymes (Edmonds *et* Hirst, 2002). Les reformulations et paraphrases exploitent les propriétés paradigmatiques des mots (leur capacité à se substituer mutuellement) entraînant ainsi des changements de vocabulaire mais conservant les concepts et les idées qui y sont exprimés (Duclaye, 2003). Il est alors, dans le cadre de la détection de similitudes textuelles sémantiques, plus judicieux de représenter un mot par un concept. Par exemple, il est plus judicieux de représenter un mot par une liste de tous ses sens ou de tous les mots par lesquels il peut être substitué (tous ses synonymes, lui compris) plutôt que seulement par lui-même. Il est également important de noter que dans WordNet, les *synsets* sont reliés entre eux par des relations, soit lexicales (antonymie

5. https://fr.wiktionary.org (consulté le 02/06/2017 à 12h)
6. https://fr.wikipedia.org (consulté le 02/06/2017 à 12h)
7. http://www.cnrtl.fr/definition/dictionnaire (consulté le 25/04/2017 à 15h)
8. http://www.cnrtl.fr/definition/th%C3%A9saurus (consulté le 25/04/2017 à 15h)

par exemple), soit taxonomiques (hyperonymie, méronymie, *etc.*). La version 3.0 de WordNet contient plus de 117 000 *synsets*[9].

EuroVoc[10] (Eurovoc, 1995) a été élaboré par le Parlement Européen et l'Office des publications de l'Union européenne. Il existe dans les 24 langues officielles de l'Union européenne et c'est pour cette raison qu'il est la plupart du temps choisi comme thésaurus lors de l'extraction de connaissance pour la construction d'un graphe. EuroVoc se compose de plus de 6000 descripteurs organisés hiérarchiquement en plus de 20 domaines. Il dispose d'une large couverture et contient des descripteurs dans des domaines allant de la politique à la géographie, en passant par l'agriculture, les transports, l'économie, et plus encore.

BabelNet[11] (Navigli *et* Ponzetto, 2012) est un réseau sémantique multilingue représenté par un graphe orienté pondéré et étiqueté où les nœuds représentent des concepts et des entités nommées multilingues et où les arêtes expriment les relations sémantiques entre ces nœuds. Chacun des nœuds contient un ensemble de lexicalisations du concept qu'il représente dans différentes langues. Dans sa version 3.7, il couvre plus de 270 langues et est construit à partir de 7 sources externes dont WordNet, Wikipédia et Wiktionary. Il contient plus de 13 millions de *synsets* répartis sur plus de 6 millions de concepts.

La Figure 2.6 illustre un exemple de structure du réseau sémantique multilingue BabelNet, schéma issu des travaux de Franco-Salvador *et al.* (2013c). On peut voir que pour le mot *play* (*jouer* en français) dans le sens de jouer un rôle ou une pièce de théâtre et non jouer à des jeux, le réseau propose des traductions dans diverses langues en prenant en compte le sens du mot. On peut également voir des relations entre ce mot et des mots apparentés, ainsi que des exemples de phrases d'usage où il est mis en contexte.

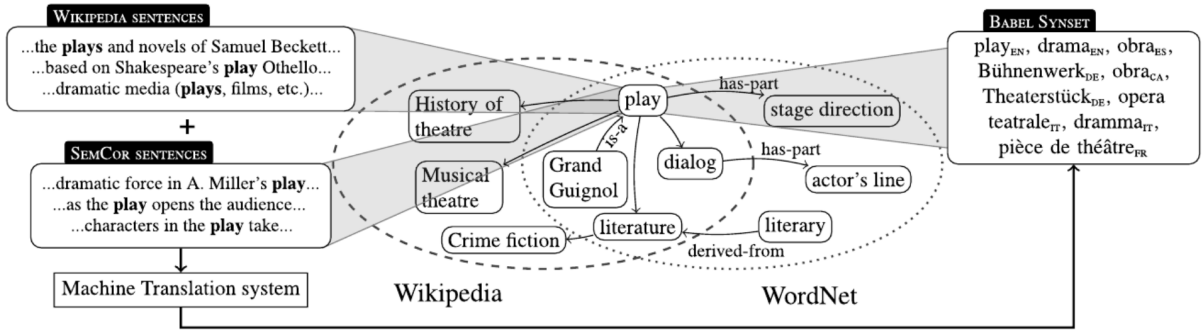


Figure 2.6 – Exemple de structure du réseau sémantique multilingue BabelNet. Schéma issu des travaux de Franco-Salvador *et al.* (2013c).

DBnary[12] (Sérasset, 2015) est une ressource multilingue en données lexicales liées ouvertes au format RDF (Klyne *et* Carroll, 2004) et dont les données sont représentées en utilisant le vocabulaire LEMON (McCrae *et al.*, 2011). Comme l'expliquent Servan *et al.* (2016), ces données lexicales sont automatiquement extraites à partir de Wiktionary, le dictionnaire de Wikipédia, pour 21 langues différentes. Il contient 2,9 millions d'entrées composées de catégories lexicales, de formes canoniques ou bien encore de formes fléchies. Ces entrées sont classées par sens et liées par des relations sémantiques du même ordre que celles de WordNet. Il contient également des définitions, des exemples d'utilisation et plus de 4,6 millions de traductions à partir des 21 langues extraites.

9. https://wordnet.princeton.edu/ (consulté le 15/04/2017 à 19h)
10. http://eurovoc.europa.eu/drupal/ (consulté le 15/04/2017 à 19h)
11. http://babelnet.org/ (consulté le 15/04/2017 à 19h)
12. http://kaiko.getalp.org/about-dbnary/ (consulté le 17/06/2017 à 11h)

2.2.2.2 Modèle vectoriel translingue (*Cross-Language Vector Space Model, CL-VSM*)

Bien que son nom peut prêter à confusion, cette approche est différente des *word embeddings* (Mikolov *et al.*, 2013a) que l'on présentera dans la section 2.2.6.

Ce modèle consiste à représenter le contenu des textes à comparer sous la forme de représentations vectorielles indépendantes de la langue des textes, ceci en utilisant des descripteurs issus de thésaurus multilingues indexés ou de toutes autres ressources conceptuelles. De cette manière, calculer une distance ou une similarité entre ces représentations est plus aisé que de le faire entre documents sous forme textuelle brute.

Steinberger *et al.* (2002), Pouliquen *et al.* (2003b) et Steinberger *et al.* (2004) utilisent les descripteurs issus du thésaurus multilingue EuroVoc pour construire les vecteurs. À la place des termes rencontrés dans les textes, ce sont les identifiants des concepts les représentant qui sont utilisés pour construire les vecteurs. L'avantage de ce type de modélisation est qu'elle représente les documents en vecteurs indépendants de la langue, elle ne nécessite donc aucune traduction explicite, ce sont les liaisons au sein du thésaurus qui vont établir des correspondances implicites entre deux concepts signifiant la même chose dans deux langues différentes. Les auteurs assignent automatiquement des descripteurs EuroVoc en utilisant une approche statistique, nécessitant un entrainement sur une collection de textes sur laquelle des descripteurs de termes ont été au préalablement assignés manuellement. Durant cette phase d'assignement, les lemmes des textes sont comparés avec la liste associative de descripteurs EuroVoc relative au langage du texte. Les descripteurs EuroVoc sont ensuite affectés automatiquement aux documents, en fonction de la similarité entre le contenu des documents et les descripteurs de la liste. Cette approche est assez restrictive et gourmande car elle exige un grand corpus annoté manuellement pour l'apprentissage. De plus, elle dépend également de la complétude des annotations et de la liste des descripteurs, ainsi que du domaine des documents considérés.

Au sein d'un même vecteur, les descripteurs peuvent être ordonnés par leur pertinence, représentée par le score utilisé lors de la phase d'assignement. Ce score est un méta-score calculé à partir de diverses mesures. La procédure de projection d'un texte vers les descripteurs EuroVoc et la fabrication de ce score sont décrites en détail dans les articles de Pouliquen *et al.* (2003a) et de Steinberger (2001). Les auteurs représentent donc chaque document par un vecteur des top 100 descripteurs EuroVoc assignés automatiquement avec comme pondération pour l'ordonnancement leur score d'assignement. Ensuite, pour calculer la similarité entre deux vecteurs, ils utilisent une similarité cosinus (Salton, 1989). Steinberger *et al.* (2002) montrent qu'en matière de calcul de similarité, la consistance des descripteurs EuroVoc est plus importante que leur précision (leur complétude est plus importante que leur exactitude). Pouliquen *et al.* (2003b) font intervenir la notion de longueur, vue dans la section 2.2.1.3 pour augmenter sensiblement les performances de leurs résultats.

2.2.2.3 Similarité translingue basée sur des thésaurus (*Cross-Language Conceptual Thesaurus-based Similarity, CL-CTS*)

Ce modèle, très proche du modèle précédent, a pour objectif de représenter les documents à comparer sous la forme de vecteurs de concepts pour ensuite mesurer leur similarité sémantique sous cette forme. Il n'implique également aucune traduction explicite, elle sera implicite à l'aide des correspondances internes au thésaurus utilisé.

Dans les travaux de Gupta *et al.* (2012), un modèle est proposé visant à mesurer la similarité entre deux textes écrits dans des langues différentes en projetant ces textes dans l'espace des domaines spécifiques présents dans EuroVoc. Pour cela, les auteurs filtrent au préalable les mots vides des textes et les mettent sous leur forme racine, car utiliser les termes tels qu'ils apparaissent dans les textes, sous leur forme lemmatisée, ne procure pas de résultats convaincants selon Pouliquen *et al.* (2003a). Ils construisent ensuite les vecteurs correspondants à ces textes en pondérant les termes selon leur fréquence d'apparition dans ces textes. La formule qu'ils utilisent

ensuite pour la comparaison de deux vecteurs fait intervenir une similarité cosinus (Salton, 1989), une correspondance d'entités nommées et une fois encore la notion de longueur (Pouliquen *et al.*, 2003b) vue dans la section 2.2.1.3. Ceska *et al.* (2008) procèdent de façon similaire avec EuroWordNet.

Levow *et al.* (2005) introduisent l'idée d'expansion de requêtes pour la recherche d'information translingue à l'aide de listes de traductions. Ils font l'hypothèse qu'une structure de données simple, comme une liste de termes multilingues ou un dictionnaire de traductions, peut être une bonne base pour créer un lien entre des dictionnaires ou des thésaurus de deux langues différentes qui n'étaient pas liés autrefois. Torrejón *et* Ramos (2011) présentent une combinaison de méthodes à base de n-grammes et d'approches avec dictionnaires de traductions. Ils utilisent ces dictionnaires, basés sur des extractions de Wiktionary [13] et Wikipédia [14], pour effectuer des transitions entre les mots et leur racine dans une langue cible, afin d'effectuer des liens translingues entre les concepts.

Pataki (2012) reprend cette idée d'utiliser des dictionnaires de traductions. Elle découpe le texte en phrases car, pour elle, c'est plus majoritairement à cette échelle que les traductions sont opérées et donc que les comparaisons doivent être effectuées. Elle y élimine les mots vides, pour deux raisons essentielles. Non seulement pour des raisons de rapidité d'indexation et de traitement, mais aussi pour des raisons linguistiques. La plupart des mots vides (prépositions, pronoms, *etc.*) ne disposent pas toujours d'une traduction claire et définie dans toutes les langues (inclusion de pronom en espagnol ou bien le cas des prépositions en hongrois). Ensuite, elle construit un sac de mots pour chaque phrase en traduisant les lemmes restants de chaque phrase. Le nombre optimal de traductions par lemme est estimé à 5. Pour récupérer les traductions d'un mot, basée sur les travaux précédemment évoqués de Levow *et al.* (2005) et Torrejón *et* Ramos (2011), Pataki (2012) préfère utiliser un dictionnaire de traductions plutôt qu'une ontologie ou un thésaurus, car l'utilisation d'une ontologie soulève deux problèmes. En premier lieu la limitation des données et dans un second temps l'asymétrie des données en fonction des langues. La distinction principale entre son approche et celles précédemment évoquées est la métrique qu'elle utilise pour comparer deux sacs de mots, une intersection d'ensemble minimale, ce qui permet d'être plus robuste à la différence de tailles des deux sacs de mots comparés.

Soit d un document de longueur n, les n mots du document sont représentés par w, avec w_i le $i^{\text{ème}}$ mot du document d, tel que :

$$d = \{w_1, w_2, w_3, ..., w_n\} \tag{2.6}$$

Soit d et d', deux documents dans deux langues différentes, respectivement L et L'. On construit le sac de mots S en filtrant les mots vides de d et en utilisant une fonction qui retourne pour chaque mot de d toutes ses traductions possibles vers L'.

Pour calculer la similarité entre d et d', la méthode la plus courante consiste alors à calculer l'intersection entre les sacs de mots S et S' :

$$Sim(d, d') = \mid S \cap S' \mid \tag{2.7}$$

Si une phrase possède suffisamment de concepts communs avec une autre, elle est alors jugée comme étant la traduction possible de cette phrase. Mais Pataki (2012) utilise un rapport plus complexe qui permet d'être plus robuste vis à vis de la différence des tailles des phrases comparées :

$$Sim(d, d') = min(\mid S \cap S' \mid - \mid S \setminus S' \mid , \mid S' \cap S \mid - \mid S' \setminus S \mid) \tag{2.8}$$

avec les relations entre les deux ensembles représentées dans la Figure 2.7.

À noter qu'une fois que les documents sont représentés sous forme de sacs de concepts, on peut également les comparer en appliquant des mesures de similarité sémantique plus triviales. On peut citer notamment comme exemple la similarité de Lesk (1986) ou plus généralement celle de Tversky (1977).

13. https://fr.wiktionary.org (consulté le 02/06/2017 à 12h)
14. https://fr.wikipedia.org (consulté le 02/06/2017 à 12h)

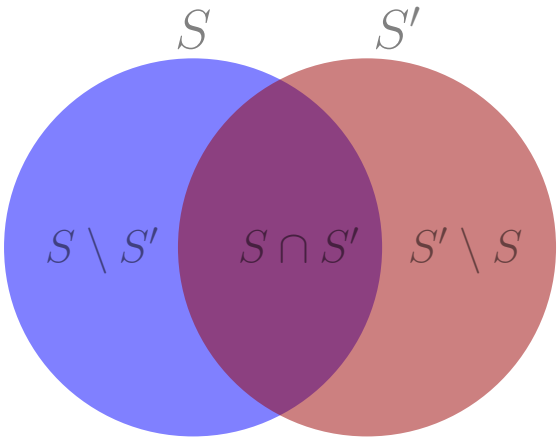


FIGURE 2.7 – L'intersection des sacs de mots S et S' est représentée dans ce diagramme de Venn par la zone violette centrale.

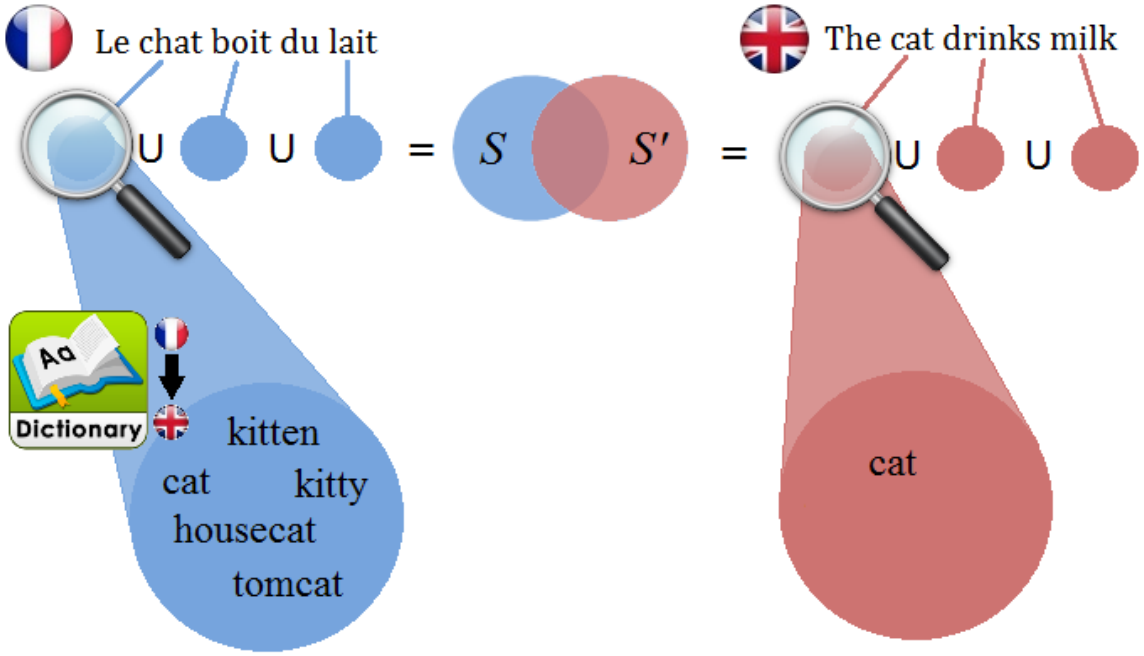


FIGURE 2.8 – Fonctionnement du modèle CL-CTS selon Pataki (2012). Un sac de mots est construit pour chaque phrase à partir d'un dictionnaire de traductions. Les deux sacs de mots qui en résultent sont des représentations dans la même langue des deux phrases d'origine. Un calcul basé sur une intersection ensembliste est alors utilisé pour mesurer la similarité entre ces deux représentations.

L'approche CL-CTS selon Pataki (2012) est schématisée sur la Figure 2.8. Dans un premier temps, on construit un sac de mots pour chaque mot porteur de sens de chacune des phrases. Pour la première phrase, ici en français, les 5 traductions les plus probables d'un mot vers l'anglais sont requêtées depuis un dictionnaire de traductions. Les sacs de mots de la seconde phrase contiennent uniquement le mot qui leur correspond. Enfin, on construit le sac de mots de chaque phrase en fusionnant tous les sacs de mots. Les deux sacs de mots qui en résultent sont des représentations dans la même langue des deux phrases d'origine. Pour calculer la similarité des deux phrases d'origine, on effectue un calcul basé sur l'intersection des deux sacs les représentant.

2.2.2.4 Analyse translingue de graphes de connaissances (*Cross-Language Knowledge Graph Analysis, CL-KGA*)

Cette approche, apparue pour la première fois dans Franco-Salvador *et al.* (2013a,c,b), utilise des graphes de connaissances générés à partir d'un réseau sémantique multilingue pour représenter des documents sous la forme de modèles contextuels translingues. La similarité entre deux graphes peut ensuite être mesurée dans un espace graphique sémantique. Dans la littérature, ce modèle est considéré comme étant basé à la fois sur des réseaux sémantiques (voir section 2.2.2.1) et à la fois sur des corpus comparables (voir section 2.2.4.1) (Danilova, 2013).

Dans l'état de l'art (Franco-Salvador *et al.*, 2013a,c,b, 2014), c'est le réseau sémantique multilingue BabelNet qui est utilisé. Dans BabelNet, des *synsets* WordNet et des pages de Wikipédia forment des concepts (des nœuds), tandis que les pointeurs sémantiques et les hyper-liens constituent des relations (arêtes), comme expliqué dans les travaux de Navigli *et* Ponzetto (2012). Cette façon de structurer les informations améliore la désambiguïsation des mots et la cartographie conceptuelle des documents qui y seront projetés, mais tout autre graphe de connaissances peut être utilisé lors de cette approche, comme le soulignent les auteurs.

Un graphe de connaissances est un graphe pondéré et étiqueté qui élargit et relie les concepts d'origine présents dans un texte. La création d'un graphe de connaissances à partir d'un réseau sémantique se fait comme décrit dans les articles de Navigli *et* Lapata (2010) et de Franco-Salvador *et al.* (2013a,c,b, 2014). Le graphe est créé en recherchant dans BabelNet, les chemins reliant les paires de *synsets* et en pondérant tous les concepts et les relations sémantiques du graphe. Pour les pondérations, les poids originaux de BabelNet sont utilisés. Ces poids fournissent le degré de parenté entre les points d'extrémité des *synsets* de chaque bord (Navigli, 2012; Navigli *et* Ponzetto, 2012; Franco-Salvador *et al.*, 2014).

De façon plus détaillée, pour créer un graphe de connaissances à partir d'un document d, la méthode consiste à :

— lemmatiser et étiqueter morphosyntaxiquement chaque mot du document d ;
— récupérer la liste des *synsets* s contenant chaque mot ;
— chercher les chemins entre toutes les paires des *synsets* s ;
— relier et fusionner ces chemins de façon à obtenir un graphe g ;
— pondérer les concepts (les nœuds) et les relations (les chemins) du graphe g obtenu.

La similarité entre deux documents peut ainsi être mesurée dans un espace graphique sémantique en calculant la similarité entre les deux graphes correspondants.

Soit deux documents d et d' dans deux langues différentes (respectivement L et L'), leurs graphes respectifs g et g' sont construits suivant la procédure expliquée ci-dessus. Ensuite, pour comparer ces deux graphes, une version adaptée au réseau sémantique multilingue de l'algorithme de comparaison flexible de graphes Sim_g, présenté dans les travaux de Montes-y-Gómez *et al.* (2001), est utilisée en suivant la formule :

$$Sim(d, d') = Sim_g(g, g') = Sim_c(g, g') \cdot (\alpha + \beta \cdot Sim_r(g, g')) \tag{2.9}$$

où Sim_c est le score des concepts, calculé par la Formule 2.10, Sim_r est le score des relations, calculé par la Formule 2.11, et α et β sont des paramètres permettant de gérer l'importance à donner aux relations par rapport aux concepts.

$$Sim_c(g, g') = \frac{2 \cdot \sum_{c \in g \cap g'}^{|g \cap g'|} W(c)}{\sum_{c \in g}^{|g|} W(c) + \sum_{c \in g'}^{|g'|} W(c)} \tag{2.10}$$

$$Sim_r(g, g') = \frac{2 \cdot \sum_{r \in N(c, g \cap g')}^{|N(c, g \cap g')|} W(r)}{\sum_{r \in N(c,g)}^{|N(c,g)|} W(r) + \sum_{r \in N(c,g')}^{|N(c,g')|} W(r)} \tag{2.11}$$

où c représente un concept, r représente une relation entre deux concepts, la fonction W retourne le poids dans le graphe pour un concept c ou une relation r donné et la fonction $N(c, g)$ retourne l'ensemble de toutes les relations connectées à un concept c au sein d'un graphe g.

2.2.3 Modèles à base de corpus parallèles

2.2.3.1 Corpus parallèles

On appelle corpus parallèle C, deux ensembles de documents D et D' dans deux langues différentes (respectivement L et L'), dans lesquels chaque document d_i du corpus D est la traduction de l'un des documents d'_i du corpus D' (voir Figure 2.9). Il peut être alors intéressant d'aligner ces corpus, c'est-à-dire de faire correspondre chaque unité textuelle (documents, paragraphes, phrases ou groupes de mots) de l'un des corpus avec l'unité textuelle lui correspondant dans l'autre corpus pour disposer d'un jeu de données bilingues comparables plus fin ou spécifique.

À titre d'exemple, citons les JRC-Acquis [15] (Steinberger *et al.*, 2006), qui sont des comptes rendus, disponibles dans plus de 20 langues, de débats du Parlement Européen.

Une présentation d'autres corpus parallèles couramment utilisés dans la détection de similarités textuelles sémantiques translingues sera donnée en section 4.1.1.

2.2.3.2 Similarité translingue basée sur l'alignement (*Cross-Language Alignment-based Similarity Analysis, CL-ASA*)

Cette méthode est évoquée pour la première fois pour la recherche de document par Pinto *et al.* (2007). Elle est introduite ensuite pour la détection de plagiat translingue par Barrón-Cedeño *et al.* (2008) et est développée par la suite dans Pinto *et al.* (2009). Elle implique la création d'un dictionnaire statistique bilingue contenant les probabilités de traductions de paires de mots déterminées à partir des alignements obtenus depuis un corpus parallèle en utilisant le modèle IBM-1 (Brown *et al.*, 1993). Dans ce modèle, l'objectif est d'estimer combien un document d écrit dans une langue L est potentiellement la traduction d'un document d' écrit dans une langue L'.

Le modèle est basé à l'origine sur une adaptation de la règle de Bayes (Bayes *et* Price, 1763) dans le cadre de la traduction automatique :

$$p(d \mid d') = \frac{p(d) \cdot p(d' \mid d)}{p(d')} \tag{2.12}$$

avec $p(d)$ la probabilité que l'événement d se produise, $p(d')$ celle que l'événement d' se produise et $p(d' \mid d)$ celle que l'événement d' survienne sachant d. Comme $p(d')$ ne dépend pas de d, elle peut être négligée si on cherche à maximiser $p(d \mid d')$. Dans un contexte de traduction automatique, la probabilité conditionnelle $p(d' \mid d)$ est la probabilité du modèle de traduction et est calculée sur la base d'un dictionnaire statistique bilingue. La probabilité $p(d)$, quant à elle, est la probabilité du modèle de langue, elle décrit la cible de langue L' dans le but d'obtenir des traductions qui seront grammaticalement acceptables (Brown *et al.*, 1993).

Traduire d en langue L' n'est pas le but de cette méthode. En effet, elle se focalise sur le fait de retrouver des textes écrits en langue L' qui pourraient être potentiellement une traduction de d. La Formule 2.12 est donc modifiée par Barrón-Cedeño *et al.* (2008) pour mieux répondre à cette

15. http://optima.jrc.it/Acquis/JRC-Acquis.3.0/doc/README_Acquis-Communautaire-corpus_JRC.html (consulté le 15/04/2017 à 19h)

tâche (Potthast *et al.*, 2011a). Le modèle de langue est ainsi remplacé par le modèle de longueur $\varrho(d, d')$, abordé par Pouliquen *et al.* (2003b) et vu dans la section 2.2.1.3, de façon à appliquer une comparaison robuste à la différence de taille des textes comparés. Ce modèle dépend donc de la taille des textes comparés plutôt que de leur structure. Le modèle de traduction est, quant à lui, adapté pour devenir une mesure simple, notée p' et exprimée dans la Formule 2.13.

Soit d et d', deux documents suivant le modèle de la Formule 2.6, tels que w_i représente le $i^{\text{ème}}$ mot du document d et w'_j, le $j^{\text{ème}}$ mot du document d'. On veut connaître la probabilité $p'(d, d')$ que d soit une traduction de d'.

$$p'(d, d') = \frac{1}{(\mid d' \mid +1)^{|d|}} \cdot \prod_{i=1}^{|d|} p(w_i \mid d') \tag{2.13}$$

où

$$p(w_i \mid d') = \sum_{j=1}^{|d'|} p(w_i \mid w'_j) \tag{2.14}$$

où $p(w_i \mid w'_j)$ est la probabilité que w'_j soit la traduction de w_i dans le dictionnaire statistique de probabilités bilingues utilisé.

La formule finale du calcul d'estimation CL-ASA devient donc :

$$Sim(d, d') = \varrho(d, d') \cdot p'(d, d') \tag{2.15}$$

Par la suite, des améliorations ont été proposées. Par exemple, Barrón-Cedeño *et al.* (2010) considèrent pour chaque mot, seulement les meilleures traductions, celles supérieures à une certaine probabilité minimale. Ils fixent ce seuil empiriquement à 0,4 durant leurs travaux.

2.2.3.3 Indexation sémantique latente translingue (*Cross-Language Latent Semantic Indexing, CL-LSI*)

L'analyse sémantique latente (Deerwester *et al.*, 1990; Berry *et* Young, 1995) extrait les mots relatifs à un sujet depuis un document lui-même et non pas depuis un corpus de documents externes comme c'est le cas pour l'analyse explicite CL-ESA (voir section 2.2.4.2). Elle induit la construction, à partir d'un corpus de documents, d'une matrice termes-documents, qui recense de façon structurée en deux dimensions l'importance de chaque terme dans chaque document du corpus. Puis, elle effectue la décomposition en valeurs singulières (Stewart, 1993; Businger *et* Golub, 1965; Golub *et* Reinsch, 1970) de cette matrice et permet ainsi de l'approximer en calculant une matrice de rang plus faible. La matrice obtenue définit ainsi un espace conceptuel commun à tous les documents du corpus.

Dumais *et al.* (1997), Littman *et al.* (1998), McCrae *et al.* (2013) et Saad *et al.* (2014) l'adaptent pour la recherche d'information translingue en utilisant un corpus parallèle pour la construction de la matrice. Dumais *et al.* (1997) et Littman *et al.* (1998) concatènent chaque document d_i d'une langue L avec son document relatif d'_i dans l'autre langue L', formant ainsi plus qu'un seul document au lieu de deux. Étant donné que l'indexation sémantique latente ne tient pas compte de l'ordre des mots mais considère plutôt chaque document comme un sac de mots, la concaténation de deux documents de deux langues différentes est traitée comme un seul document dans la matrice, alors constituée de plusieurs langues et représentant donc un espace multilingue. La décomposition en valeurs singulières peut être appliquée donnant alors lieu à un espace conceptuel translingue. McCrae *et al.* (2013) et Saad *et al.* (2014) quant à eux apprennent deux espaces monolingues (construisent donc deux matrices) à partir d'un corpus parallèle et identifient ensuite les représentations correspondantes à l'aide du parallélisme entre ces deux ensembles.

Une analyse sémantique latente consiste à représenter un corpus sous la forme d'une matrice qui décrit l'importance de chaque terme dans chacun des documents composant ce corpus. C'est

une matrice dont les lignes correspondent aux termes et dont les colonnes correspondent aux documents. Dans l'état de l'art les termes sont généralement ramenés à leur forme racine et la pondération utilisée pour construire les vecteurs est de type *tf.idf* (Salton *et* Buckley, 1988).

Soit X la matrice où l'élément $tf.idf(t_i, d_j)$ décrit la pondération *tf.idf* du terme t_i (le $i^{\text{ème}}$ terme du corpus) dans le document d_j (le $j^{\text{ème}}$ document du corpus), telle que :

$$X = \quad (\mathbf{t}_i) \rightarrow \quad \begin{matrix} \\ \\ \\ \\ \end{matrix} \begin{matrix} (\mathbf{d}_j) \\ \downarrow \\ \end{matrix} \begin{pmatrix} tf.idf(t_1, d_1) & tf.idf(t_1, d_2) & ... & tf.idf(t_1, d_n) \\ tf.idf(t_2, d_1) & tf.idf(t_2, d_2) & ... & tf.idf(t_2, d_n) \\ \vdots & \vdots & \vdots & \vdots \\ tf.idf(t_m, d_1) & tf.idf(t_m, d_2) & ... & tf.idf(t_m, d_n) \end{pmatrix} \tag{2.16}$$

où n est le nombre total de termes contenus dans le corpus considéré et m est le nombre total de documents qui forment le corpus.

Une ligne de cette matrice correspond à la représentation vectorielle d'un terme, où chaque dimension de cette représentation donne l'importance du terme dans chaque document du corpus. De même, une colonne de la matrice X correspond à la représentation vectorielle d'un document, où chaque dimension est l'importance d'un terme dans le contenu de ce document.

Une décomposition en valeurs singulières (Stewart, 1993; Businger *et* Golub, 1965; Golub *et* Reinsch, 1970) sur X est ensuite effectuée donnant deux matrices orthonormales U et V et une matrice diagonale Σ, telle que :

$$X = \quad (\mathbf{t}_i) \rightarrow \quad \begin{matrix} U \\ \\ \\ \\ \end{matrix} \begin{pmatrix} \begin{pmatrix} \\ \mathbf{u}_1 \\ \\ \end{pmatrix} \dots \begin{pmatrix} \\ \mathbf{u}_l \\ \\ \end{pmatrix} \end{pmatrix} \cdot \begin{matrix} \Sigma \\ \\ \\ \\ \end{matrix} \begin{pmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_l \end{pmatrix} \cdot \begin{matrix} V^T \\ (\mathbf{d}_j) \\ \downarrow \\ \end{matrix} \begin{pmatrix} (& \mathbf{v}_1 &) \\ & \vdots & \\ (& \mathbf{v}_l &) \end{pmatrix} \tag{2.17}$$

où U est formée des vecteurs $\mathbf{t}_i$ et V est formée des vecteurs $\mathbf{d}_j$. Σ est une matrice diagonale contenant les concepts les plus pertinents, indexés et pondérés par les valeurs singulières σ. Lorsque l'on conserve seulement les k plus grandes valeurs singulières, ainsi que les vecteurs singuliers correspondants dans U et V, on obtient une approximation de rang k de la matrice X. Cette approximation forme un espace conceptuel commun. À partir de là, les vecteurs $\mathbf{t}_i$ ne possèdent plus que k dimensions, chacune représentant l'importance du terme i pour l'un des k concepts, et les vecteurs $\mathbf{d}_j$ représentent l'importance des relations entre le document j et chacun des k concepts. On peut exprimer cette approximation de la façon suivante :

$$X_k \approx U_k \cdot \Sigma_k \cdot V_k^T \tag{2.18}$$

Par la suite, un nouveau document d peut être représenté sur l'espace conceptuel alors créé, grâce à la fonction Φ suivante :

$$\Phi(d) = U_k^T \cdot v_d \tag{2.19}$$

où v_d est le vecteur de termes pondérés *tf.idf* du document d.

Enfin, pour calculer dans quelle mesure un vecteur représentant un document est similaire d'un autre dans l'espace conceptuel latent, c'est la similarité cosinus (Salton, 1989) qui est plébiscitée dans l'état de l'art.

2.2.3.4 Analyse translingue par corrélation canonique de noyaux (*Cross-Language Kernel Canonical Correlation Analysis, CL-KCCA*)

Cette analyse n'étant que très peu présente dans la littérature et faisant appel à de nombreuses notions de mathématiques, nous ne présenterons ici que son approche générale. Pour une explication plus détaillée, nous invitons le lecteur à se reporter aux travaux de Vinokourov *et al.* (2002).

L'analyse par corrélation canonique de noyaux (Bach *et* Jordan, 2002) vise à créer un espace conceptuel translingue en identifiant des corrélations entre des termes à travers un corpus parallèle, c'est-à-dire deux espaces monolingues parallèles, et cela en se servant des noyaux de ces deux espaces comme de repère pour établir les corrélations. Elle identifie un ensemble de projections à partir de deux langues vers un espace sémantique commun. Cela procure donc un espace naturel de travail pour effectuer une recherche d'information translingue. Comme expliqué dans Vinokourov *et al.* (2002), cette analyse permet la détection de certaines similarités sémantiques, représentées par des ensembles de mots ayant les mêmes motifs d'occurrence au sein de paires de documents bilingues.

Considérons un corpus parallèle C contenant deux sous ensembles D et D' de textes alignés dans deux langues différentes (respectivement L et L'), tel que chaque texte $d_i \in D$ en langue L est la traduction d'un texte $d'_i \in D'$ en langue L'. Une fois une indexation sémantique latente (Deerwester *et al.*, 1990; Berry *et* Young, 1995) appliquée sur ce corpus, on obtient donc deux espaces F_D et $F_{D'}$, un pour chaque langue, le premier composé des $\Phi(d_i)$ et le second des $\Phi(d'_i)$. K_D et $K_{D'}$ sont respectivement les noyaux des deux espaces F_D et $F_{D'}$ (Cristianini *et* Shawe-Taylor, 2000; Vinokourov *et* Girolami, 2002), comme illustré sur la Figure 2.9. L'analyse par corrélation canonique de noyaux trouve alors des projections dans les deux espaces pour lesquels les valeurs qui y sont projetées sont hautement corrélées. L'approche exploite les combinaisons particulières de mots qui semblent avoir les mêmes motifs de co-occurence dans les deux langues. L'hypothèse est que trouver de telles corrélations à travers le corpus parallèle met en évidence des similarités sémantiques entre des concepts de deux langues différentes. Vinokourov *et al.* (2002) définissent ensuite une nouvelle notion, la F-corrélation canonique des noyaux, notée ρ. Ils sélectionnent ensuite le nombre n de dimensions sémantiques ayant les plus grandes valeurs de corrélation ρ.

Par la suite, un nouveau document d peut être représenté sur l'espace créé en projetant son vecteur v_d sur les n composants de la F-corrélation canonique. Enfin, pour calculer dans quelle mesure un vecteur représentant un document est similaire à un autre, c'est encore la similarité cosinus (Salton, 1989) qui est plébiscitée dans l'état de l'art.

2.2.4 Modèles à base de corpus comparables

2.2.4.1 Corpus comparables

Un corpus comparable est basé sur le même principe qu'un corpus parallèle. Il s'agit de deux ensembles de textes dans deux langues différentes, dans lesquels chaque document de l'un des ensembles est relatif à un document du second ensemble. Mais contrairement à un corpus parallèle, dans un corpus comparable, deux documents relatifs ne sont pas la traduction l'un de l'autre, ils traitent seulement du même sujet spécifique, c'est-à-dire que les deux documents partagent le même sujet dans le même registre tout en n'étant pas pour autant la traduction exacte l'un de l'autre.

Wikipédia est le corpus comparable le plus utilisé et le plus simple d'accès. En effet, deux articles équivalents dans deux langues différentes sont rarement les traductions exactes et parfaites l'un de l'autre mais plutôt deux articles indépendants, écrits par deux auteurs différents, dans deux langues différentes, traitant du même sujet.

Une présentation d'autres corpus comparables couramment utilisés dans la détection de similarités textuelles sémantiques translingues sera donnée dans la section 4.1.1.

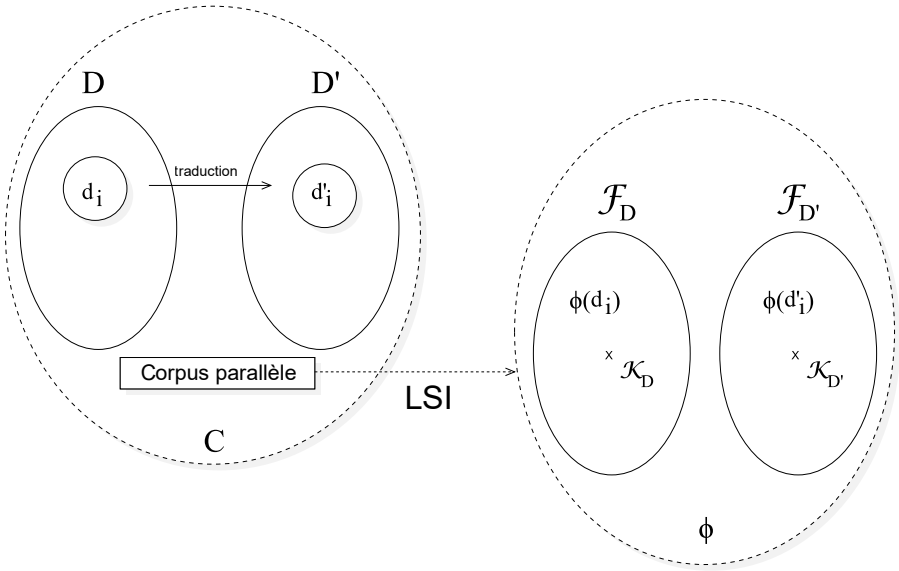


FIGURE 2.9 – Liaison entre un corpus parallèle et un espace conceptuel LSI.

À noter que certaines méthodes basées sur l'utilisation de corpus parallèles, peuvent être utilisées en se basant sur des corpus comparables, à l'image des travaux de Vulic *et* Moens (2014) qui utilisent une variante de CL-LSI où le modèle est construit à partir de données comparables.

2.2.4.2 Analyse sémantique explicite translingue (*Cross-Language Explicit Semantic Analysis, CL-ESA*)

Ce modèle est un modèle de recherche relatif à une collection, ce qui signifie qu'un document est représenté par sa similitude avec les documents d'une collection. Il fait l'hypothèse que l'on peut déterminer le sujet d'un document à partir de son vocabulaire, c'est-à-dire de la fréquence des termes qu'il contient. En calculant la similarité entre un document et chacun des documents d'un corpus, on peut obtenir un vecteur représentant au mieux le document en fonction de ce corpus. Dans un contexte translingue et en considérant l'utilisation d'un corpus comparable, on peut alors estimer la similarité entre deux documents en calculant la similarité entre leur représentation suivant ce modèle.

Il est basé sur le modèle d'analyse sémantique explicite introduit pour la première fois par Gabrilovich *et* Markovitch (2007) qui représente le sens d'un document par un vecteur basé sur des concepts dérivés de Wikipédia, afin de pouvoir retrouver un document au sein d'un corpus. Il est repris par Potthast *et al.* (2008) qui utilise un corpus comparable pour adapter ce modèle de recherche à un contexte translingue en se basant sur les travaux de Yang *et al.* (1998).

Soit d et d' deux documents dans deux langues différentes (respectivement L et L') et D et D' deux corpus comparables contenant un grand nombre de documents (respectivement dans les langues L et L'). On construit une représentation vectorielle explicite $\mathbf{d}$ pour d, où chaque dimension i dans $\mathbf{d}$ quantifie la similarité entre d et chaque document D_i du corpus D. Cette similarité est représentée par une représentation vectorielle dite secondaire, qui sera notée v. N'importe quelle représentation peut être utilisée mais dans l'état de l'art, c'est un vecteur de termes avec pondération *tf.idf* (Salton *et* Buckley, 1988) qui est adopté. Le document d est donc représenté sous la forme du vecteur $\mathbf{d}$ de n dimensions, tel que :

$$\mathbf{d} = (\varphi(v, v_1^*),\ \varphi(v, v_2^*),\ \varphi(v, v_3^*),\ .....,\ \varphi(v, v_n^*))^T \tag{2.20}$$

où v est le vecteur représentant d, v_i^* est le vecteur représentant le $i^{\text{ème}}$ document dans D et n est le nombre de documents que contient D. Dans l'état de l'art, la fonction φ est une similarité

cosinus (Salton, 1989) et si elle se trouve plus petite qu'un certain seuil, elle est réduite à zéro afin de minimiser le bruit et faciliter les calculs (Potthast *et al.*, 2008).

On procède de façon similaire pour le second document d' écrit en langue L', en construisant une représentation vectorielle explicite **d'** utilisant le corpus D'. La Figure 2.10 illustre le principe de création des vecteurs explicites **d** et **d'** et leurs interactions avec les vecteurs secondaires pour les créer.

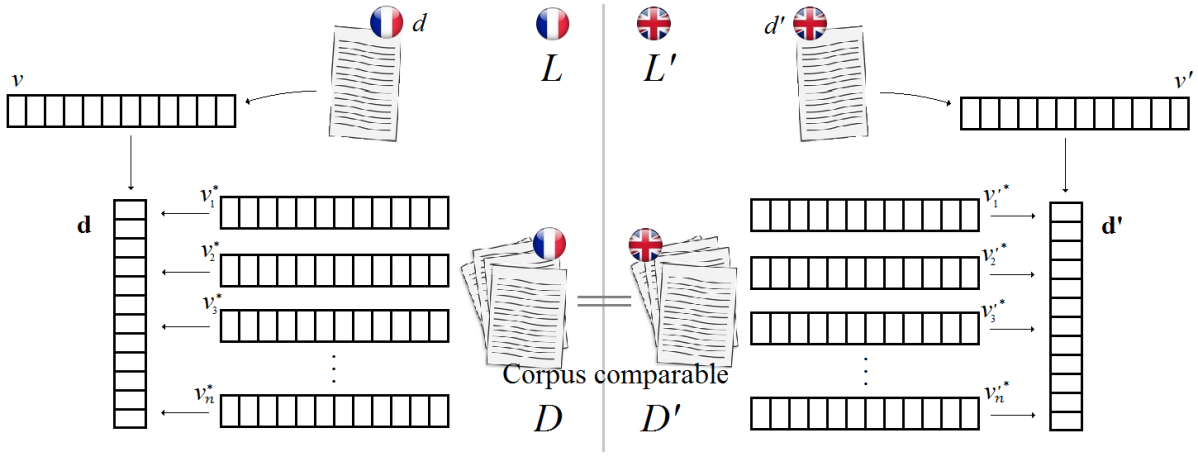


FIGURE 2.10 – Fonctionnement du modèle CL-ESA.

Si les documents à l'intérieur de D ont une correspondance un à un avec les documents à l'intérieur de D' (ce qui devrait être le cas dans un corpus comparable), alors les représentations **d** et **d'** sont comparables. Si la similarité cosinus est toujours notée φ, alors la similarité entre d et d' peut être exprimée de la façon suivante :

$$sim(d, d') = \varphi(\mathbf{d}, \mathbf{d'}) \tag{2.21}$$

2.2.5 Modèles à base de traduction suivie d'une analyse monolingue (*Translation + Monolingual Analysis, T+MA*)

Les méthodes présentées précédemment, utilisent des mécanismes et des ressources lexicales utilisés dans la traduction automatique, mais n'effectuent pas réellement de traduction à proprement parler. Les méthodes présentées dans cette section quant à elles, effectuent bel et bien une traduction automatique pour ramener les deux textes à comparer sous une forme textuelle de même langue afin de pouvoir les comparer avec des approches monolingues plus éprouvées dans la littérature.

Soit deux documents d et d' écrits dans deux langues différentes (respectivement L et L'). La première phase de ce type de méthode consiste à mettre les deux documents d et d' sous la même langue. Ceci entraîne alors plusieurs interrogations : traduire le document d en langue L', traduire le document d' en langue L, ou traduire les deux documents dans une troisième langue L'', que l'on appelle alors langue pivot ? Dans ce dernier cas, quelle langue pivot utiliser ? Les travaux de Linard *et al.* (2015), qui utilisent des vecteurs de contexte pour aligner des documents au sein de corpus comparables, peuvent donner des éléments de réponses. De leur côté, Pereira *et al.* (2010b) choisissent d'utiliser l'anglais comme langue pivot. Les raisons qu'ils exposent pour leur décision sont d'une part que la plupart des contenus disponibles sur Internet sont en anglais et d'autre part qu'une grande majorité d'outils de traduction automatique d'une langue quelconque vers l'anglais sont très facilement disponibles, ce qui n'est pas toujours le cas vers d'autres langues. À cela on pourrait ajouter une troisième raison qui est la disponibilité massive d'outils de pré- et post- traitement en anglais : listes de mots vides, lemmatiseurs, tokeniseurs, étiqueteurs, *etc.* Outils qui ne sont pas tout le temps disponibles dans des langues moins répandues.

Pour la traduction, Kent *et* Salim (2009, 2010b) proposent d'utiliser *Google Translate*. Une fois les deux textes dans la même langue, les mots vides sont retirés et les mots restants sont mis sous leur forme racine. Ensuite, les textes sont découpés en segments afin de constituer des empreintes. La méthode de comparaison qu'ils utilisent est une intersection des empreintes, composées à chaque fois des trois 4-grammes les moins fréquents de chaque segment.

Plutôt que d'utiliser une traduction, c'est-à-dire la sortie finale d'un traducteur, Muhr *et al.* (2010) essaient d'utiliser seulement une partie du processus de traduction automatique. Ils considèrent uniquement la sortie du modèle de traduction, c'est-à-dire du modèle se chargeant d'obtenir toutes les traductions possibles pour un mot donné. Leur modèle est construit avec l'outil BerkeleyAligner (Liang *et al.*, 2006; DeNero *et* Klein, 2007). Chaque token du texte est substitué par les 5 traductions candidates les plus plausibles (statistiquement parlant) ou par lui-même si aucune n'est retournée par le dictionnaire de traductions construit en utilisant le corpus Europarl (Koehn, 2005)

Une fois les deux textes mis sous la même langue, ils peuvent alors être comparés par des techniques de comparaison monolingue. Barrón-Cedeño *et al.* (2010) affirment que les traducteurs automatiques pouvant donner lieu à de multiples traductions (toutes justes mais étant sensiblement différentes et employant donc un vocabulaire différent), il n'est pas judicieux d'effectuer un alignement monolingue avec des méthodes basiques lexicales ou syntaxiques comme des méthodes à base d'intersections de n-grammes. Toujours d'après Barrón-Cedeño *et al.* (2010), il est donc préférable d'utiliser des méthodes telles que des vecteurs ou des sac de mots (avec pondération si possible) qui montrent de bien meilleurs résultats sur des comparaisons de textes monolingues (Barrón-Cedeño *et* Rosso, 2009). Ils construisent alors des vecteurs représentant d et d', qu'ils pondèrent avec *tf.idf* (Salton *et* Buckley, 1988). Leur similarité est ainsi calculée avec une similarité cosinus (Salton, 1989). Muhr *et al.* (2010) comparent déjà un mot du document d à un ensemble de mots du document d', leur méthode s'apparente donc déjà à des méthodes dites sac de mots.

Récemment, cette approche a encore une fois démontré ses performances en s'illustrant lors de la tâche 1, la tâche de détection de Similarité Textuelle Sémantique (STS), de l'édition 2016 de la campagne SemEval (Agirre *et al.*, 2016). Cette année là, pour la première fois, la tâche STS est étendue avec une sous-tâche translingue traitant la paire de langues espagnol-anglais et en effet, Brychcin *et* Svoboda (2016) ont fini 1^{er} en y soumettant une approche par traduction suivie d'une comparaison monolingue. Après une phase de traduction automatique où toutes les phrases espagnoles ont été traduites en anglais à l'aide de *Google Translate*, ils ont ré-utilisé l'aligneur monolingue [16] de Sultan *et al.* (2015), déjà vainqueur de la tâche 2015 en monolingue.

L'aligneur de Sultan *et al.* (2015) effectue un indice de Jaccard (Jaccard, 1912) entre les deux ensembles de mots alignés dans les deux documents comparés. Si d et d' sont deux documents dans la même langue, alors l'indice de Jaccard J peut s'exprimer avec la formule suivante :

$$J(d, d') = \frac{\mid A_d \mid + \mid A_{d'} \mid}{\mid d \mid + \mid d' \mid} \tag{2.22}$$

où d et d' sont les deux documents à comparer (représentés sous forme d'ensembles de mots) et A_d et $A_{d'}$ sont les ensembles de mots représentant respectivement l'intersection de d avec d' et l'intersection de d' et d. Bien que $A_d \neq A_{d'}$ car les mots s'alignant les uns aux autres ne sont pas identiques (synonymie), on peut noter toutefois que $\mid A_d \mid = \mid A_{d'} \mid$.

Brychcin *et* Svoboda (2016) ont ensuite amélioré cette mesure en y ajoutant des pondérations fréquentielles au sein de l'indice de Jaccard. Si d et d' sont deux documents dans la même langue, alors :

$$J(d, d') = \frac{\omega(A_d) + \omega(A_{d'})}{\omega(d) + \omega(d')} \tag{2.23}$$

où d et d' sont les deux documents à comparer (représentés sous forme d'ensembles de mots) et A_d et $A_{d'}$ sont les ensembles de mots représentant respectivement l'intersection de d avec d'

16. https://github.com/ma-sultan/monolingual-word-aligner (consulté le 10/08/2017 à 15h)

et l'intersection de d' et d. ω est une pondération fréquentielle d'un ensemble de mots, définie par la formule suivante :

$$\omega(A) = \sum_{i=1\,,\,w_i \in A}^{|A|} idf(w_i) \tag{2.24}$$

où w_i est le $i^{\text{ème}}$ mot de l'ensemble A et idf est la fonction qui retourne la fréquence inverse de document d'un mot, telle que définie dans la section 2.2.1.1. À noter qu'ici le fait que $A_d \neq A_{d'}$ dans la Formule 2.23 est un fait plus important que dans la Formule 2.22 car des mots différents auront une pondération fréquentielle différente.

Cette amélioration n'ayant pas été partagée par Brychcin *et* Svoboda (2016), nous décidons de partager notre ré-implémentation sur GitHub [17]. Nous nous en servirons dans le chapitre 5.

2.2.6 Travaux plus récents

Les méthodes de détection du plagiat peuvent être assimilées à des méthodes de détection de similarité textuelle sémantique, or la littérature pour cette tâche est en constante évolution. C'est pourquoi, nous allons passer ici en revue les derniers travaux en matière de détection de similarité textuelle sémantique, qui sont essentiellement à base de représentations distributionnelles distribuées continues de mots (*word embeddings*).

2.2.6.1 Modèles à base de représentations distributionnelles distribuées continues de mots (*word embeddings*)

Dans la littérature, plusieurs techniques existent pour construire des modèles vectoriels et ainsi représenter des textes sous forme de vecteurs (nous en avons vu en section 2.2.2.2 et section 2.2.1.1, par exemple), mais dernièrement, ces techniques se sont principalement appuyées sur les travaux de Mikolov *et al.* (2013a,b,c) sur les représentations distributionnelles distribuées continues de mots (*word embeddings*) qui ont montré des performances prometteuses dans diverses tâches du TAL que ce soit dans un contexte monolingue ou multilingue (Upadhyay *et al.*, 2016; Ammar *et al.*, 2016; Ghannay *et al.*, 2016). Un certain nombre de contributions ont étendu ces travaux à des séquences de mots (Mikolov *et al.*, 2013b; Le *et* Mikolov, 2014) et à des représentations translingues (Luong *et al.*, 2015). Ces représentations permettent de capturer implicitement les synonymies proches entre mots, sous-phrases ou phrases dans un contexte monolingue ou translingue. Dès lors, utiliser ces représentations pour la détection du plagiat semble être une idée attrayante puisque ces dernières permettent de calculer la similarité sémantique entre deux textes.

Les représentations distributionnelles distribuées continues de mots (*word embeddings*) ont dernièrement été rendues populaires par Mikolov *et al.* (2013a) qui s'appuient sur les architectures des réseaux neuronaux de Bengio *et al.* (2003) pour en offrir une version simplifiée tout en augmentant les performances d'apprentissage et de justesse de prédiction de ces réseaux. Ces représentations se fondent sur la construction d'un modèle qui projette les termes d'une langue dans un espace dans lequel certains liens sémantiques entre ces termes peuvent être observés et mesurés. Les mots sont projetés dans un espace continu multidimensionnel et ceux ayant un contexte similaire seront normalement les plus proches dans cet espace (Blacoe *et* Lapata, 2012; Mikolov *et al.*, 2013b). Ainsi, des similarités entre les mots peuvent être calculées, par exemple, grâce à une mesure de similarité cosinus (Salton, 1989) entre leurs représentations. Par ailleurs, les angles entre les projections (les vecteurs) des mots sont influencés par les diverses relations qui relient les mots. Grâce à cela, il est possible d'exploiter ces relations avec des opérations arithmétiques sur leurs vecteurs. Les exemples les plus connus sont, par exemple, le fait que le résultat de l'opération $vecteur(Madrid) - vecteur(Espagne) + vecteur(France)$

17. https://github.com/FerreroJeremy/monolingual-word-aligner (consulté le 10/08/2017 à 16h)

sera plus proche du $vecteur(Paris)$ que de n'importe quel autre vecteur. De même, le résultat de l'opération $vecteur(Roi) - vecteur(Homme) + vecteur(Femme)$ sera extrêmement proche du $vecteur(Reine)$.

Ces représentations sont qualifiées de *continues* car représentées dans un espace continu (composé de valeurs réelles non indexées), *distribuées* (sur un aspect technique et informatique) car dans le vecteur final la représentation d'un mot est distribuée dans chacune des composantes du vecteur (une seule dimension ne représente rien à proprement parler, c'est une caractéristique latente) et *distributionnelles* (sur un aspect mathématique et linguistique) car elles sont basées sur l'hypothèse distributionnelle où le contexte de chaque mot repose sur la distribution de ses mots voisins (Turian *et al.*, 2010). Dans la suite de ce manuscrit, nous utiliserons indifféremment plusieurs termes (variantes de ces qualificatifs) pour parler de ces mêmes représentations.

Parmi les modèles les plus utilisés on peut citer *word2vec*[18] (Mikolov *et al.*, 2013c) (dont nous allons expliquer le fonctionnement ci-dessous) ou bien encore *GloVe*[19] (Pennington *et al.*, 2014), qui se base lui sur une factorisation de matrice de co-occurences. Toutefois, *word2vec* est de loin le modèle le plus utilisé dans la littérature, c'est pourquoi nous nous concentrons majoritairement sur ce dernier pour illustrer cette section de notre état de l'art. Dans ce modèle, deux architectures sont proposées pour apprendre des *word embeddings*. La première est appelée sac de mots continu (*Continuous Bag-Of-Words (CBOW)*) (Mikolov *et al.*, 2013a) et la seconde, *skip-gram* (Mikolov *et al.*, 2013a,b). Ces deux architectures sont présentées ci-dessous et illustrées dans la Figure 2.11.

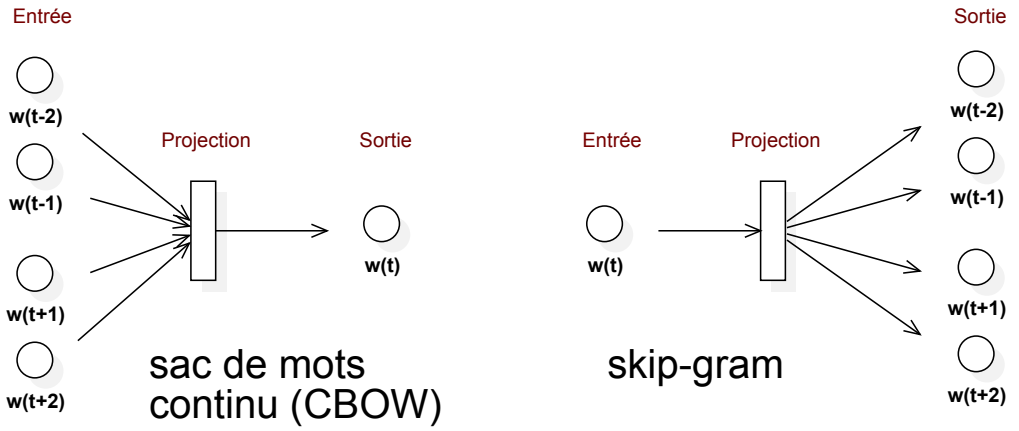


FIGURE 2.11 – Architecture des deux modèles, sac de mots continu (*Continuous Bag-Of-Words (CBOW)*) et *skip-gram* de *word2vec*.

La première architecture proposée est le sac de mots continu (*Continuous Bag-Of-Words (CBOW)*). Comme expliqué dans le papier de Mikolov *et al.* (2013a), elle est basée sur des modèles de langue neuronaux (Bengio *et al.*, 2003; Mikolov *et al.*, 2010), où la couche de projection est partagée pour tous les mots. Considérant une séquence de mots, l'objectif de cette approche est de prédire tous les mots depuis leur contexte. C'est-à-dire qu'un mot courant est prédit à partir des mots qui l'entourent. Cette architecture est représentée dans la partie gauche de la Figure 2.11. Elle est appelée sac de mots car l'ordre des mots dans l'entrée n'influence pas la projection et que des mots avant mais aussi après le mot courant sont utilisés.

La seconde architecture est appelée *skip-gram*. Elle est similaire à celle du *CBOW*, mais au lieu de prédire le mot courant à partir de son contexte (des mots qui l'entourent), elle essaie

18. https://code.google.com/archive/p/word2vec/ (consulté le 12/07/2017 à 10h)
19. https://nlp.stanford.edu/projects/glove/ (consulté le 12/07/2017 à 10h)

de maximiser la justesse de prédiction d'un mot à partir d'un autre mot de son contexte. Plus précisément, elle utilise chaque mot courant comme entrée d'un classifieur non linéaire avec une couche de projection continue qui essaie de prédire les mots dans une certaine fenêtre avant et après ce mot. C'est donc le fonctionnement inverse à celui de l'architecture *CBOW* qui est utilisé pour l'apprentissage de ce modèle : *skip-gram* utilise le mot courant pour prédire les mots de son contexte. Cette architecture est représentée dans la partie droite de la Figure 2.11.

Les représentations distributionnelles distribuées continues de séquences de mots peuvent se construire de diverses façons. La façon la plus simple et intuitive de construire la représentation d'une séquence de mots est de le faire à partir des représentations de ses mots, avec, par exemple, une somme comme le fait Blacoe *et* Lapata (2012). Ils génèrent le vecteur d'une séquence en faisant la somme des vecteurs des mots qui la composent.

Soit d un document de longueur n, les n mots du document sont représentés par w, avec w_i le $i^{\text{ème}}$ mot du document d, tel que défini dans la Formule 2.6. La représentation v du document d est la somme des vecteurs de ses mots :

$$v = \sum_{i=1}^{n} vector(w_i) \tag{2.25}$$

où w_i est le $i^{\text{ème}}$ mot du document d et *vector* est la fonction qui donne la représentation d'un mot.

Blacoe *et* Lapata (2012) effectuent une étude comparative de plusieurs méthodes de représentations de mots avec trois méthodes de composition (l'addition, le produit scalaire et l'auto-encodeur récursif de Socher *et al.* (2011)). Ils montrent que l'addition des *word embeddings* pour construire des *phrase embeddings* est compétitive, en dépit de sa naïveté et facilité d'implémentation.

En plus d'effectuer une mesure de similarité entre les vecteurs des séquences, qui sont eux-même la somme des vecteurs de leurs mots, comme c'est le cas dans les travaux de Blacoe *et* Lapata (2012), Bérard *et al.* (2016) proposent dans leur boîte à outils *MultiVec*[20], une seconde méthode qui consiste à moyenner les mesures de similarité des vecteurs de mots pris deux à deux dans chaque séquence (le vecteur du premier mot de la première séquence est comparé avec le vecteur du premier mot de la seconde séquence et ainsi de suite). Contrairement à la méthode de Blacoe *et* Lapata (2012), cette méthode est sensible à l'ordre des mots au sein des séquences comparées ainsi qu'à la taille des séquences (tous les mots de la séquence la plus longue ne sont pas comparés).

Soit deux documents, d et d', de même longueur n, avec w_i le $i^{\text{ème}}$ mot du document d et w'_i le $i^{\text{ème}}$ mot du document d', tels que définis dans la Formule 2.6. La similarité Sim entre ces deux documents peut s'exprimer de la façon suivante :

$$Sim(d, d') = \frac{\sum_{i=1}^{n} \varphi(vector(w_i), vector(w'_i))}{n} \tag{2.26}$$

où φ représente toujours la similarité cosinus et *vector* est la fonction qui donne la représentation d'un mot.

Les modèles à base de sac de mots offrent d'assez bons résultats mais présentent également certaines limites, en particulier la perte de toute information sur l'ordre des mots et le manque d'associations idiomatiques dans les mots composés (Mikolov *et al.*, 2013b). Afin de pallier ces limites, d'autres recherches apprennent directement un modèle pour des séquences de mots (Le *et* Mikolov, 2014; Pham *et al.*, 2015). Le modèle *paragraph vector*, introduit par Le *et* Mikolov

20. https://github.com/eske/multivec (consulté le 12/07/2017 à 10h)

(2014), possède une architecture similaire au modèle *CBOW* de *word2vec* à l'exception du fait qu'il ajoute, pour chaque séquence, un vecteur de pondération (de même taille que les vecteurs de mots) à la couche de projection. Cette méthode a l'avantage de ne nécessiter pour l'apprentissage du modèle que d'un simple corpus, contrairement à d'autres travaux qui requièrent des corpus annotés, avec, par exemple, des arbres syntaxiques (Socher *et al.*, 2013a,b).

Les approches les plus courantes apprennent un modèle de façon non supervisée, à la manière de *SkipThought*[21] (Kiros *et al.*, 2015) ou *FastSent*[22] (Hill *et al.*, 2016), alors que *InferSent*[23] (Conneau *et al.*, 2017) envisage plutôt un apprentissage supervisé. On peut également citer *FastText*[24] (Bojanowski *et al.*, 2017), qui est une version supervisée de *sent2vec*[25] (Pagliardini *et al.*, 2017).

Les représentations distributionnelles distribuées continues translingues se font plus rares en revanche. Il est vrai que les *word embeddings* peuvent être utilisés pour des tâches multilingues en entrainant indépendamment un modèle sur chaque langue. Pham *et al.* (2015), par exemple, étendent le *paragraph vector* proposé par Le *et* Mikolov (2014) au contexte bilingue. Néanmoins, les représentations qui en résultent sont dans un espace vectoriel différent (deux espaces monolingues et non un seul translingue). Toutefois, il existe plusieurs méthodes pour régler ce problème. Il est possible d'entrainer indépendamment deux modèles monolingues (dans deux langues différentes) et ensuite d'apprendre une équivalence de l'une des représentations sur l'autre (Vulić *et* Korhonen, 2016). Il est également possible d'apprendre directement un modèle translingue en substituant, à l'aide d'un dictionnaire d'équivalences, des mots du corpus d'apprentissage d'une langue vers une autre (Gouws *et al.*, 2015; Gouws *et* Søgaard, 2015). Ou bien encore, l'entrainement peut être fait conjointement en utilisant un corpus parallèle (Luong *et al.*, 2015).

BiVec (ou *BiSkip*)[26] (Luong *et al.*, 2015) appartient à la dernière catégorie et utilise un corpus parallèle en plus d'un dictionnaire d'équivalences. Pour chaque phrase du corpus parallèle, *BiVec* essaie de prédire les mots de la même façon que le *skip-gram* de *word2vec*, mais il utilise aussi les mots de la phrase source pour prédire les mots de la phrase cible dans l'autre langue et inversement. La boîte à outil *MultiVec* (Bérard *et al.*, 2016) procède de la même façon pour générer un modèle translingue. L'apprentissage dans *Bilbowa*[27] (Gouws *et al.*, 2015; Gouws *et* Søgaard, 2015) se base, quant à lui, sur l'architecture sac de mots de *word2vec* et ne nécessite aucun corpus parallèle. Il substitue, à l'aide d'un dictionnaire d'équivalences, des mots du corpus d'apprentissage depuis sa langue d'origine vers une autre langue (cible), afin que le modèle construit soit translingue. Certaines recherches utilisent des techniques de factorisation de matrice de co-occurences (Zou *et al.*, 2013; Shi *et al.*, 2015) comme le faisait *GloVe* (Pennington *et al.*, 2014) en monolingue.

Enfin, Duong *et al.* (2017) proposent divers algorithmes pour construire un modèle *word embeddings* multilingue dans un même espace vectoriel unifié. Ils proposent aussi bien des transformations linéaires après l'entraînement, qu'un apprentissage avec un lexique bilingue pendant l'entraînement.

Dès lors, utiliser les *word embeddings* pour la détection du plagiat translingue semble attrayant puisque, capturant implicitement la sémantique des mots, ils peuvent être utilisés pour calculer la similarité entre deux textes dans deux langues différentes.

21. https://github.com/ryankiros/skip-thoughts (consulté 12/07/2017 à 10h)
22. https://github.com/fh295/SentenceRepresentation (consulté 12/07/2017 à 11h)
23. https://github.com/facebookresearch/InferSent (consulté 12/07/2017 à 10h)
24. https://github.com/facebookresearch/fastText (consulté 12/07/2017 à 10h)
25. https://github.com/epfml/sent2vec (consulté 12/07/2017 à 10h)
26. https://github.com/lmthang/bivec (consulté 12/07/2017 à 10h)
27. https://github.com/gouwsmeister/bilbowa (consulté 12/07/2017 à 10h)

2.2.6.2 Les représentations distributionnelles distribuées continues dans la détection du plagiat

Les approches de détection du plagiat translingue à l'aide de représentations distributionnelles distribuées continues ne se basent pas uniquement sur ces représentations mais utilisent des systèmes hybrides ou des combinaisons. Stajner *et al.* (2017) emploient deux variantes plus ou moins strictes d'un algorithme qui, pour identifier des paraphrases, compare le score de similarité de toutes les paires de phrases possibles et aligne chacune des phrases avec sa phrase la plus proche en fonction d'une mesure de similarité calculée entre les deux représentations de ces phrases. Pour cela, ils utilisent trois représentations différentes. La première est simplement la représentation employée lors de la méthode CL-C3G présentée en section 2.2.1.1. La seconde, appelée WAVG, est une représentation de phrases construite à partir de la moyenne des représentations *word embeddings* de chaque mot de cette phrase (Adi *et al.*, 2017) (là où Blacoe *et* Lapata (2012) faisaient la somme). La dernière représentation, tirée des travaux de Franco-Salvadora *et al.* (2016), identifie l'alignement de mots optimal entre deux phrases en calculant la similarité cosinus (Salton, 1989) entre les représentations *word embeddings* des mots de toutes les paires de mots possibles entre les deux phrases et moyenne les scores des similarités cosinus les plus fortes ainsi obtenues pour chaque mot de la phrase la plus longue. Franco-Salvadora *et al.* (2016) appellent cette approche un calcul de similarité translingue basée sur l'alignement de représentations continues de mots (*Continuous Word Alignment-based Similarity Analysis (CWASA)*).

Certaines approches, au lieu de se concentrer sur la détection du plagiat à proprement parler, se concentrent sur la détection de similarité textuelle sémantique ou bien l'extraction de phrases parallèles. C'est par exemple le cas pour Grover *et* Mitra (2017), qui construisent une matrice de similarité entre deux phrases à l'aide de mesures de similarité entre les représentations *word embeddings* des mots de ces deux phrases. Parce que les phrases comparées peuvent avoir des tailles différentes, ils obtiennent une matrice à dimension variable. Ils réduisent ensuite cette matrice en une matrice de taille fixe et utilisent un réseau neuronal convolutif (*convolutional neural network (CNN)*) pour classer les phrases comme correspondantes ou non. España-Bonet *et al.* (2017), quant à eux, extraient des phrases parallèles au sein d'un corpus comparable à l'aide d'un réseau neuronal multilingue de traduction automatique (*Multilingual Neural Machine Translation (MNMT)*) basé sur *Nematus*[28] (Sennrich *et al.*, 2017).

Lors de la tâche de détection de similarité textuelle sémantique (STS) de la campagne d'évaluation SemEval 2017, la grande majorité des soumissions faisait appel à des réseaux neuronaux ou à des technologies basées directement sur des réseaux neuronaux comme des *word embeddings*, à l'instar de la soumission de l'équipe ECNU (Tian *et al.*, 2017) qui a remporté la tâche primaire (meilleur score sur la moyenne de toutes les sous-tâches). Cette équipe a proposé un score composite qui réalise la moyenne des scores de plusieurs systèmes, dont des systèmes de régression, mais aussi des systèmes basés sur des réseaux neuronaux comme un DAN (*deep averaging network*) (Iyyer *et al.*, 2015) et un LSTM (*long short-term memory*) (Hochreiter *et* Schmidhuber, 1997), en prenant comme traits d'entrée diverses mesures comme des distances d'édition, des alignements de mots, des n-grammes, des métriques de traduction automatique et de résumé automatique, des *word embeddings*, *etc.* (Cer *et al.*, 2017).

Certains travaux sont plus originaux, comme ceux de Gella *et al.* (2017), qui utilisent une architecture basée sur un réseau neuronal convolutif pour apprendre une représentation multilingue de deux phrases dans deux langues différentes en utilisant des images comme pivot entre ces deux langues. Leur modèle apprend la représentation de plusieurs images avec leurs descriptions (légendes) dans deux langues différentes et apprend également la correspondance entre deux images, ce qui permet ainsi d'utiliser les images comme pivot entre deux descriptions dans deux langues différentes.

28. https://github.com/EdinburghNLP/nematus (consulté 21/08/2017 à 12h)

2.2.7 Discussion sur les différentes approches

Étant donné la multiplicité des approches présentées, il est nécessaire d'étudier leur comportement selon différents facteurs. De telles études ont déjà été effectuées, mais pour la plupart, elles se contentent de comparer la méthode qu'elles visent à promouvoir avec des méthodes de l'état de l'art plus anciennes et déjà éprouvées. C'est par exemple le cas dans l'article de Barrón-Cedeño *et al.* (2010) ou de Franco-Salvador *et al.* (2016), où les auteurs introduisent respectivement la méthode T+MA et CL-KGA, puis comparent ces méthodes à l'état de l'art afin de prouver leur efficacité.

Nous allons tenter dans cette section de synthétiser les diverses études menées sur les méthodes de l'état de l'art présentées au cours de ce chapitre. Pour cela, on peut s'appuyer sur les travaux de Danilova (2013), qui recensent et résument quelques articles ayant effectué des évaluations de plusieurs de ces méthodes. Dans le même but, le Tableau 2.2 liste les méthodes comparées dans chaque étude précédemment menée, tandis que le Tableau 2.3 présente les différentes caractéristiques des corpus utilisés ainsi que les paires de langues traitées au cours de ces études.

Recherches	CL-C3G	CL-ASA	CL-KCCA	CL-ESA	CL-KGA	CL-LSI	T+MA
Vinokourov *et al.* (2002)			✓			✓	
Cimiano *et al.* (2009)				✓		✓	
Barrón-Cedeño *et al.* (2010)	✓	✓					✓
Potthast *et al.* (2011a)	✓	✓		✓			
Barrón-Cedeño (2012)	✓	✓					✓
Barrón-Cedeño *et al.* (2013a)	✓	✓					✓
Barrón-Cedeño *et al.* (2014)	✓						✓
Franco-Salvador *et al.* (2016)	✓	✓		✓	✓		

Tableau 2.2 – Méthodes déjà évaluées au cours des recherches précédemment menées.

Recherches	Corpus	Langues
Vinokourov *et al.* (2002)	Hansard (Germann, 2001)	en → fr
Cimiano *et al.* (2009)	Multext [29] ; JRC-Acquis [30] ; Wikipédia [31]	en ; fr ; de
Barrón-Cedeño *et al.* (2010)	Software ; Consumer (Alcázar, 2006)	en → eu ; es → eu
Potthast *et al.* (2011a)	Wikipédia ; JRC-Acquis	en → {es, de, fr, nl, pl}
Barrón-Cedeño (2012)	PAN-PC-11 [32] (Potthast *et al.*, 2011b)	en → eu ; es → eu
Barrón-Cedeño *et al.* (2013a)	PAN-PC-11	en → es
Barrón-Cedeño *et al.* (2014)	Wikipédia	el ; it ; lt ; es ; hr
Franco-Salvador *et al.* (2016)	PAN-PC-10 [33] (Potthast *et al.*, 2010b) ; PAN-PC-11	es → en ; de → en

Tableau 2.3 – Caractéristiques des corpus utilisés ainsi que les paires de langues traitées au cours des recherches précédemment menées. Chaque langue est notée par son code à deux lettres sous la norme ISO 639-1 (dit *alpha-2*).

Dans l'article de Potthast *et al.* (2011a), les performances des méthodes CL-C3G, CL-ESA et CL-ASA sont comparées. Ces méthodes sont évaluées sur des tâches de recherche de documents et de recherche de similarités entre documents, en utilisant des corpus comparables et parallèles. CL-C3G et CL-ESA montrent de meilleurs résultats quand les deux documents comparés partagent le même vocabulaire, tandis que CL-ASA obtient de meilleurs résultats dans les cas de traductions exactes. CL-C3G obtient de manière générale de bien meilleurs résultats. Toutefois,

29. http://www.issco.unige.ch/en/research/projects/MULTEXT.html (consulté le 07/06/2017 à 18h)
30. http://optima.jrc.it/Acquis/JRC-Acquis.3.0/doc/README_Acquis-Communautaire-corpus_JRC.html (consulté le 15/04/2017 à 19h)
31. http://download.wikimedia.org/backup-index.html (consulté le 07/06/2017 à 18h)
32. https://www.uni-weimar.de/de/medien/professuren/medieninformatik/webis/corpora/pan-pc-11/ (consulté le 07/06/2017 à 18h)
33. https://www.uni-weimar.de/en/media/chairs/computer-science-and-media/webis/corpora/corpus-pan-pc-10/ (consulté le 07/06/2017 à 18h)

CL-ASA et CL-ESA, contrairement à CL-C3G, peuvent être appliquées sur des paires de langues syntaxiquement éloignées, comme le pointe également Barrón-Cedeño (2012).

Les recherches de Barrón-Cedeño *et al.* (2010), Barrón-Cedeño (2012) et Barrón-Cedeño *et al.* (2013a) comparent les méthodes CL-C3G, CL-ASA et T+MA. CL-C3G obtient de moins bons résultats que ceux reportés par Potthast *et al.* (2011a). Cependant, lorsque les langues comparées possèdent une syntaxe semblable, CL-C3G s'aligne sur les performances de CL-ASA, alors que quand cette liaison syntaxique n'existe plus, il semble inenvisageable, d'après les auteurs, d'utiliser CL-C3G. T+MA ne semble pas, elle, dépendre de similarités lexicales ou syntaxiques entre les langues comparées. Néanmoins, elle souffre de deux défauts majeurs : elle est très couteuse en temps et nécessite un traducteur automatique, n'existant pas toujours pour toutes les langues. Les méthodes CL-C3G et T+MA obtiennent de meilleurs résultats sur des documents de grande taille. Sur la base des expériences menées, les auteurs concluent sur le fait que les méthodes T+MA et CL-C3G sont des systèmes privilégiant le rappel tandis que la méthode CL-ASA privilégie la précision.

Barrón-Cedeño *et al.* (2014) étudient les performances de l'approche T+MA par rapport à des méthodes plus naïves, basées sur des analyses lexicales ou syntaxiques (CL-CNG, comptage de mots, Cognateness). En complément des travaux précédents, ils étudient les performances de chaque méthode sur un groupe de langues peu dotées (comme le grec, le lituanien, l'estonien, ou bien le croate). Ils analysent la corrélation entre les résultats de ces modèles et les jugements humains. La méthode obtenant en moyenne les meilleures corrélations sur toutes les langues est une combinaison du comptage de mots et de la méthode CL-C3G, prouvant ainsi qu'une fusion de méthodes peut faire mieux que l'état de l'art.

D'après l'étude de Vinokourov *et al.* (2002), l'analyse par corrélation canonique de noyaux (CL-KCCA) obtient de meilleurs résultats, sur un même jeu de données, que la méthode LSI, malgré le fait qu'elle soit également basée sur une décomposition en valeurs singulières. Cependant, Potthast *et al.* (2011a) observent que pour les mêmes raisons de performance que la méthode CL-LSI justement, cette approche ne peut pas rivaliser avec les méthodes CL-C3G et CL-ASA.

Cimiano *et al.* (2009) comparent l'approche explicite CL-ESA avec des approches latentes (LSI et LDA) sur une tâche de recherche de documents translingues en utilisant les corpus Multext, JRC-Acquis et Wikipédia. L'approche CL-ESA obtient de meilleurs résultats que les deux modèles latents avec lesquels elle est comparée.

Plus récemment, Franco-Salvador *et al.* (2016) comparent différentes variantes de l'approche CL-KGA, avec diverses méthodes de l'état de l'art déjà éprouvées (CL-C3G, CL-ASA et CL-ESA). Il est montré dans leurs travaux que la variante de CL-KGA utilisant des pondérations basées sur des représentations distributionnelles distribuées continues de mots (*word embeddings*) (Mikolov *et al.*, 2013a) obtient de meilleurs résultats que la variante utilisant les poids issus du réseau sémantique BabelNet (Navigli *et* Ponzetto, 2012). Durant cette étude, il est également prouvé que, de manière générale, l'approche CL-KGA est plus performante que les autres méthodes évaluées.

De notre côté, nous avons brièvement comparé les implémentations de Sultan *et al.* (2015) et de Brychcin *et* Svoboda (2016) pour l'approche de traduction suivie d'un alignement monolingue (T+MA), comme décrites dans la section 2.2.5. Le Tableau 2.4 reporte les performances de ces deux implémentations sur les corpus d'évaluation 2016 et 2017 de la tâche 1 (détection de similarités textuelles sémantiques (STS) translingues sur le couple de langues anglais-espagnol) de la campagne SemEval (Agirre *et al.*, 2016; Cer *et al.*, 2017).

L'implémentation de Brychcin *et* Svoboda (2016) avec les pondérations fréquentielles intégrées dans l'indice de Jaccard (Jaccard, 1912) donne très nettement de meilleurs résultats lors de ces évaluations. Nous nous sommes nous-même inspirés de cette pondération pour améliorer les performances de certaines de nos méthodes (voir section 5.1).

Méthode	News	Multi-Src	Moyenne
Implémentation de Sultan *et al.* (2015)	0.8960	0.7185	0.8083
Implémentation de Brychcin *et* Svoboda (2016)	0.9060	0.8145	0.8608

(a) SemEval-2016 STS

Méthode	SNLI	WMT	Moyenne
Implémentation de Sultan *et al.* (2015)	0.6696	0.0825	0.3761
Implémentation de Brychcin *et* Svoboda (2016)	0.7601	0.1245	0.4423

(b) SemEval-2017 STS

Tableau 2.4 – Comparaison des performances des implémentations de Sultan *et al.* (2015) et de Brychcin *et* Svoboda (2016) de l'approche traduction suivie d'un alignement monolingue sur les corpus d'évaluation de la tâche 1 (détection de similarités textuelles sémantiques translingues sur le couple de langues anglais-espagnol) de SemEval 2016 (a) et 2017 (b).

La grande majorité des études menées se contentent donc d'évaluer et comparer seulement deux (Vinokourov *et al.*, 2002; Cimiano *et al.*, 2009) ou trois (Barrón-Cedeño *et al.*, 2010; Potthast *et al.*, 2011a; Barrón-Cedeño, 2012; Barrón-Cedeño *et al.*, 2013a) méthodes. Dans la majorité des cas, les auteurs de ces études se justifient par le fait que certaines méthodes nécessitent trop de ressources lexicales externes ou sont trop longues dans leur exécution et ne peuvent donc pas être comparées en toute impartialité avec les autres méthodes (Potthast *et al.*, 2011a). Bien sûr, des corollaires peuvent être établis comme, par exemple, si dans l'étude de Potthast *et al.* (2011a), il est montré que CL-C3G est meilleure que CL-ESA et que dans l'étude de Cimiano *et al.* (2009), il est montré que CL-ESA est meilleure que CL-LSI, alors on pourrait en déduire que CL-C3G est meilleure que CL-LSI. Cependant, ces évaluations ne prennent pas en compte les mêmes paramétrages des méthodes, sur les mêmes corpus (jeux de documents) sur la même tâche avec les mêmes protocoles et métriques d'évaluation. Il est donc difficile de s'en servir pour tirer des conclusions scientifiques rigoureuses.

Dans le chapitre suivant, nous nous proposons de présenter les différents corpus pouvant servir à l'évaluation de méthodes de détection de plagiat translingue, ainsi que leur protocole et leurs métriques d'évaluation respectifs, afin de mieux comprendre et identifier les limites de leur exploitation au cours d'une évaluation.

3 Corpus existants pouvant servir à évaluer la détection du plagiat translingue

> *« Le plagiat est la base de toutes les littératures, excepté de la première, qui d'ailleurs est inconnue. »*
>
> Siegfried (1928), I, 6, Robineau
> — Jean Giraudoux (1882-1944)

Comme on vient de le voir dans le chapitre 2, la littérature sur les méthodes de détection du plagiat translingue est assez récente et par conséquent ces méthodes de détection sont peu nombreuses. Il en est de même pour les corpus d'évaluation de la détection du plagiat translingue, qui se font plus rares que leurs homologues pour le plagiat monolingue. Cela peut s'expliquer en raison du fait que l'intérêt pour cette tâche est très récent et qu'un corpus avec des cas réels de plagiat ne peut pas être publié publiquement sans autorisation des auteurs des données de ce corpus, pour des raisons à la fois légales et éthiques. En effet, à cause des performances des algorithmes de recherche, l'anonymisation d'un tel corpus de cas de réel plagiat est pratiquement impossible à assurer (Potthast *et al.*, 2011a).

Clough (2003) explique que « *des exemples réels de plagiats d'étudiants sont difficiles à trouver en raison de restrictions devant garantir la confidentialité des étudiants* » et conclut donc que construire un corpus pour la détection du plagiat, même simulé, ne peut qu'offrir des avantages. Pour cette raison, la plupart du temps, des corpus parallèles ou comparables sont utilisés pour évaluer l'efficacité de la détection du plagiat translingue. Des corpus comme JRC-Acquis [1] (Steinberger *et al.*, 2006) peuvent, par exemple, servir à la détection du plagiat translingue au même titre qu'ils peuvent contribuer à l'évaluation de la traduction automatique. Potthast *et al.* (2011a) compilent un corpus à partir de JRC-Acquis et Wikipédia pour comparer les performances de différentes approches de détection du plagiat translingue. Un total de 23 000 documents de JRC-Acquis et de 45 000 documents de Wikipédia constituent ce corpus.

Toutefois, il existe certains corpus spécialement construits pour l'évaluation de la détection du plagiat translingue ou plus généralement pour la détection de similarités sémantiques textuelles translingues. Ce chapitre tente de présenter ces différents corpus tout en étudiant ce qu'ils permettent d'évaluer, afin de pouvoir déterminer au mieux les limites de l'usage de chacun d'eux.

3.1 Corpus de la tâche d'évaluation BUCC 2017

3.1.1 Le corpus

L'atelier BUCC sur la construction et l'utilisation de corpus parallèles (*Building and Using Comparable Corpora*), met à disposition, pour sa tâche d'évaluation de 2017 [2], un corpus pouvant être utile pour la détection du plagiat translingue. Le processus de construction de ce corpus est détaillé dans le papier de description de la tâche (Zweigenbaum *et al.*, 2016, 2017). Les détails de la construction du corpus ont quelque peu changé depuis la version présentée dans le papier

1. http://optima.jrc.it/Acquis/JRC-Acquis.3.0/doc/README_Acquis-Communautaire-corpus_JRC.html (consulté le 15/04/2017 à 19h)
2. https://comparable.limsi.fr/bucc2017/bucc2017-task.html (consulté le 10/05/2017 à 16h)

mais les principes fondamentaux restent les mêmes. Les organisateurs de la tâche décrivent eux-mêmes le corpus comme un soigneux mélange de phrases parallèles et non-parallèles. C'est-à-dire un corpus de documents comparables où sont insérées des phrases parallèles annotées. Les documents comparables proviennent de Wikipédia et les phrases parallèles à aligner proviennent du corpus *News Commentary*[3] (Tiedemann, 2012). *News Commentary* est un corpus parallèle constitué de commentaires politiques et économiques aspirés depuis le site *Project Syndicate*[4] et à la base produit comme données d'entraînement pour la traduction automatique statistique. Contrairement à ce qui est fait pour SemEval (voir section 3.6.1), le corpus n'est pas une liste de paires de phrases déjà alignées à valider ou non, mais est composé, pour chaque langue, de deux listes de phrases monolingues, représentant des textes où se trouvent des phrases à aligner. Plus concrètement, la tâche consiste alors à retourner toutes les paires de phrases alignées entre les deux textes. Le nombre de phrases à comparer peut donc potentiellement être le produit cartésien de ces deux listes. Ce qui oblige les systèmes proposés à être plus efficaces d'un point de vu temps de calcul que dans des tâches telles que la tâche de détection de similarité textuelle sémantique de SemEval (voir section 3.6).

Le Tableau 3.1 montre le nombre de phrases par sous-corpus et par couple de langues du jeu de données partagé pour la tâche d'évaluation de l'atelier BUCC 2017 (Zweigenbaum *et al.*, 2017). Les couples de langues que comporte ce corpus sont allemand-anglais, français-anglais, russe-anglais et chinois-anglais.

		Échantillon	Entrainement	Test
de-en	# de phrases alignées	1 038	9 580	NC
	# de phrases	32 593 - 40 354	413 869 - 399 337	413 884 - 396 534
	Taille (en Mo)	4,02 - 4,42	51,6 - 43,9	51,7 - 43,6
fr-en	# de phrases alignées	929	9 086	NC
	# de phrases	21 497 - 38 069	271 874 - 369 810	276 833 - 373 459
	Taille (en Mo)	2,43 - 4,18	30,9 - 40,6	31,4 - 41,0
ru-en	# de phrases alignées	2 374	14 435	NC
	# de phrases	45 459 - 72 766	460 853 - 558 401	457 327 - 566 356
	Taille (en Mo)	9,55 - 7,98	98,3 - 61,3	97,5 - 62,2
zh-en	# de phrases alignées	257	1 899	NC
	# de phrases	8 624 - 13 589	94 637 - 88 860	91 824 - 90 037
	Taille (en Mo)	0,95 - 1,49	10,2 - 9,83	9,89 - 9,93

Tableau 3.1 – Statistiques des sous-corpus de la tâche d'évaluation de l'atelier BUCC 2017. Chaque langue est notée par son code à deux lettres sous la norme ISO 639-1 (dit *alpha-2*). Le nombre de phrases alignées des corpus de tests est à ce jour non communiqué (NC) pour garantir le bon déroulement de la tâche d'évaluation.

3.1.2 Métriques d'évaluation

Considérant une méthode dont le but est de définir si deux phrases comparées écrites dans deux langues différentes sont équivalentes ou pas (c'est-à-dire pour un corpus tel que celui partagé durant l'atelier BUCC, de ne retourner que les paires de phrases alignées), on note *Positif* une comparaison que le système prédit comme équivalente et *Négatif* une comparaison qu'il prédit comme non équivalente. Le Tableau 3.2, emprunté aux travaux de Manning *et* Schütze (1999), recense les différents cas possibles qui peuvent être rencontrés après une comparaison et la prise de décision qui en découle.

3. http://www.casmacat.eu/corpus/news-commentary.html (consulté le 05/06/2017 à 10h)
4. https://www.project-syndicate.org/ (consulté le 15/06/2017 à 15h)

	Réalité	
Sortie du système	à retourner	à ne pas retourner
retourné (Positif)	Vrai Positif (VP)	Faux Positif (FP)
non retourné (Négatif)	Faux Négatif (FN)	Vrai Négatif (VN)

Tableau 3.2 – Tableau de contingence des cas possibles survenus après une recherche documentaire, emprunté aux travaux de Manning *et* Schütze (1999).

De façon plus détaillée, les cas du Tableau 3.2 signifient :
VP : Le système prédit à raison un texte comme similaire à un autre texte ;
FP : Le système prédit à tort un texte comme similaire à un autre texte ;
VN : Le système prédit à raison un texte comme non similaire à un autre texte ;
FN : Le système prédit à tort un texte comme non similaire à un autre texte.
À partir de ces observations, il est possible de calculer diverses métriques permettant de caractériser de diverses manières la performance du système considéré.

La précision est la proportion de solutions trouvées qui sont pertinentes. Cette mesure caractérise la capacité du système à refuser les solutions non-pertinentes. La précision P est définie par la proportion de solutions pertinentes retournées parmi toutes les solutions retournées :

$$P = \frac{VP}{(VP + FP)} \tag{3.1}$$

Le rappel est la proportion des solutions pertinentes qui sont trouvées. Cette mesure caractérise la capacité du système à donner toutes les solutions pertinentes. Le rappel R est la proportion de solutions pertinentes retrouvées parmi toutes les solutions pertinentes qu'il y avait à retrouver :

$$R = \frac{VP}{(VP + FN)} \tag{3.2}$$

La F-mesure est la moyenne harmonique de la précision P et du rappel R. Cette mesure caractérise la capacité du système à donner toutes les solutions pertinentes et à refuser les autres. Elle peut être exprimée par la formule :

$$F\text{-mesure} = \frac{2 \cdot P \cdot R}{(P + R)} \tag{3.3}$$

Ces mesures sont des métriques références dans les tâches de recherche d'informations. L'atelier BUCC 2017 prend la F-mesure comme métrique pour l'évaluation de sa tâche.

3.2 Corpus de la campagne d'évaluation PAN

3.2.1 Le corpus

Le corpus de la campagne d'évaluation PAN de 2009, le PAN-PC-09 [5] (Potthast *et al.*, 2009), est le premier corpus de taille importante pour l'évaluation de la détection du plagiat qui inclut des cas de plagiat translingue. La partie translingue de ce corpus couvre 10% de ces données et inclut des passages (fragments) plagiés traduits automatiquement depuis l'allemand et l'espagnol vers l'anglais. Ce corpus a subi quelques améliorations au fil des éditions de la PAN, donnant lieu

5. https://www.uni-weimar.de/en/media/chairs/computer-science-and-media/webis/corpora/corpus-pan-pc-09/ (consulté le 30/06/2017 à 15h)

à deux nouvelles versions. Les corpus PAN-PC-10[6] (Potthast *et al.*, 2010a,b) et PAN-PC-11[7] (Potthast *et al.*, 2011b) contiennent respectivement 14% et 11% de cas de plagiat translingue. De plus, pour augmenter la qualité des traductions de la partie translingue, 1% des fragments traduits automatiquement de la version PAN-PC-11 ont été corrigés manuellement. Depuis, le corpus n'ayant subi que quelques correctifs mineurs (Potthast *et al.*, 2014), c'est toujours la version PAN-PC-11 qui est utilisée lors des récentes campagnes PAN.

Dans les trois corpus, le processus de création des documents sources et cibles est identique. Un passage d'un document source est extrait et placé de façon aléatoire au sein d'un document suspect. Les textes utilisés sont des textes issus de la base de données libre de droit du projet Gutenberg[8], ainsi le corpus peut publiquement être ré-utilisé sans permission nécessaire. Une attention toute particulière est portée sur la conservation des phrases lors de l'extraction et sur la cohésion du thème des textes lors de l'insertion (depuis la version PAN-PC-10, 50% des cas de plagiat sont insérés dans des documents relatifs). De plus, ce corpus a la particularité de contenir des fragments de textes altérés automatiquement (opération sur le texte pour rendre la détection de plagiat plus compliquée).

Le corpus PAN-PC-11 contient 26 939 documents avec plus de 60 000 cas de plagiat, bien que seulement 11% de ce corpus couvre des cas de plagiat translingue. Pour générer ces cas de plagiat translingue, la plateforme de production participative (*crowdsourcing*) *Amazon Mechanical Turk*[9] a été utilisée.

Dans le but d'évaluer au mieux les systèmes de détection du plagiat, il est nécessaire que les passages (fragments) plagiés au sein des documents suspects soient explicitement marqués et étiquetés. C'est pourquoi les auteurs du corpus ont proposé d'accompagner chaque document suspect du corpus, d'un fichier XML comportant les méta-données ayant trait au document qui lui est relatif. Ainsi, ces fichiers XML contiennent, pour chaque fragment plagié, son caractère de commencement (*offset*) et sa taille au sein du document suspect, le nom du document source duquel il provient ainsi que l'offset de départ et la taille du passage originel auquel il correspond dans le dit document source.

Par exemple, le fichier compagnon XML du document suspect *suspicious-document00005.txt* contient les informations suivantes :

```
<?xml version="1.0" encoding="UTF-8"?>
<document reference="suspicious-document00005.txt">
  <feature name="about" authors="Saint-Simon, Louis de Rouvroy, duc de"
    title="Memoirs of Louis XIV and His Court and of the Regency - Volume 06"
    lang="en" />
  <feature name="md5Hash" value="ae971a081edc198d6a1132420419020f" />
  <feature name="plagiarism" type="artificial" obfuscation="low"
    this_language="en" this_offset="19254" this_length="1557"
    source_reference="source-document00178.txt" source_language="en"
    source_offset="3835" source_length="1560" />
</document>
```

Ces informations signifient que le document suspect *suspicious-document00005.txt* comporte un passage des *Memoirs of Louis XIV and His Court and of the Regency - Volume 06* écrit en anglais par Louis de Rouvroy, duc de Saint-Simon. Ces informations révèlent également qu'il comporte un passage plagié artificiellement avec une faible offuscation ajoutée, passage s'étalant du caractère 19 254 au caractère 20 811 (19 254 + 1 557) et qui est identique au passage

6. https://www.uni-weimar.de/en/media/chairs/computer-science-and-media/webis/corpora/corpus-pan-pc-10/ (consulté le 31/05/2017 à 11h
7. https://www.uni-weimar.de/en/media/chairs/computer-science-and-media/webis/corpora/corpus-pan-pc-11/ (consulté le 31/05/2017 à 11h)
8. https://www.gutenberg.org/ (consulté le 31/05/2017 à 16h)
9. https://www.mturk.com/ (consulté le 30/05/2017 à 12h)

s'étalant du caractère 3 835 au caractère 5 395 (3 835+1 560) dans le document source *source-document00178.txt*.

3.2.2 Métriques d'évaluation

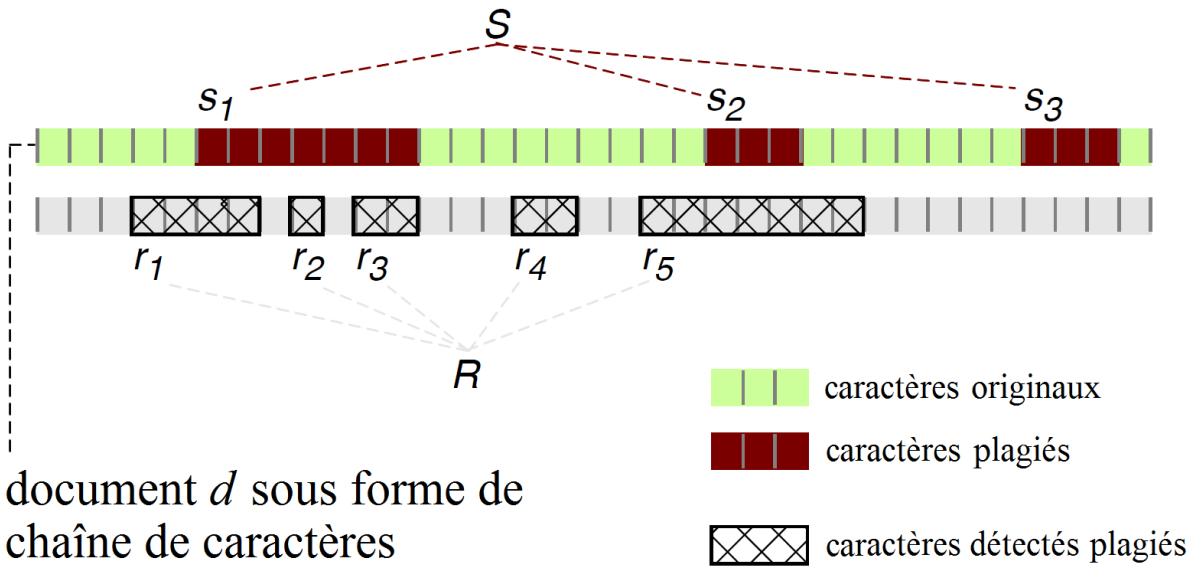


FIGURE 3.1 – Sortie schématique d'un système de détection du plagiat lancé sur un document d. Schéma issu du travail de Barrón-Cedeño (2012).

Dans cette sous-section, nous prendrons l'exemple de Barrón-Cedeño (2012) (voir Figure 3.1). Soit un document d contenant des passages plagiés S, tels que $S = \{s_1, s_2, s_3, ..., s_n\}$ et un système qui retourne des prétendus passages plagiés R tels que $R = \{r_1, r_2, r_3, ..., r_n\}$.

La précision orientée plagiat représente la fraction de fragments retrouvés qui sont réellement des cas de plagiat. Elle mesure le nombre de caractères correctement retournés comme plagiés sur le nombre total de caractères retournés. On la note ici P^*.

$$P^* = \frac{1}{|\,R\,|} \cdot \frac{|\,\cup_{s \in S}(s \cap r)\,|}{|\,r\,|} \tag{3.4}$$

Le rappel orienté plagiat représente la fraction des fragments plagiés qui ont bien été retrouvés. Il mesure le nombre de caractères correctement retournés comme plagiés sur le nombre total de caractères qu'il fallait retourner comme plagiés. On le note ici R^*.

$$R^* = \frac{1}{|\,S\,|} \cdot \frac{|\,\cup_{r \in R}(s \cap r)\,|}{|\,s\,|} \tag{3.5}$$

La F_{macro} est une macro F-mesure qui prend en compte la taille des passages plagiés au lieu de seulement considérer le nombre absolu de passages plagiés. C'est la moyenne harmonique entre P^* et R^* :

$$F_{macro} = \frac{2 \cdot P^* \cdot R^*}{(P^* + R^*)} \tag{3.6}$$

La granularité est une mesure introduite pour la première fois dans les travaux de Potthast *et al.* (2010b). Elle détermine si un fragment est détecté en totalité ou par morceaux. Cette mesure pénalise les cas où des passages retrouvés plagiés se chevauchent.

$$gran = \frac{1}{\mid S_R \mid} \cdot \sum_{s \in S_R} \mid C_S \mid \tag{3.7}$$

avec S_R l'ensemble des éléments $s \in S$ tel qu'un $r \in R$ existe pour lequel une intersection de s et r n'est pas vide et C_S l'ensemble des éléments $r \in R$ tel qu'une intersection entre s et r n'est pas vide.

Le Plagdet (*plagiarism detection*) est une mesure combinant la précision et le rappel orientés pour la détection du plagiat, ainsi que la granularité.

$$plagdet = \frac{F_{macro}}{\log_2(1 + gran)} \tag{3.8}$$

3.3 Corpus CL!TR 2011 de la campagne PAN@FIRE

3.3.1 Le corpus

Le corpus CL!TR 2011 (*Cross-Language !ndian Text Re-use*) [10] (Barrón-Cedeño *et al.*, 2011) est un autre corpus utilisé pour la détection du plagiat, plus particulièrement pour la ré-utilisation de texte en hindi. Il a été mis au point pour l'évaluation de la campagne PAN@FIRE de 2011. Ce corpus comprend 5 032 documents sources écrits en anglais issus de Wikipédia et 388 documents suspects écrits en hindi. Contrairement au corpus de la PAN, les cas de plagiat n'ont pas été ici générés automatiquement mais ont été écrits par des volontaires. Ces volontaires sont au nombre de quatorze et sont des locuteurs natifs hindi parlant couramment l'anglais. Le protocole utilisé pour construire le corpus est le même que celui décrit dans les travaux de Clough *et* Stevenson (2011) [11]. Des instructions spécifiques sont laissées aux volontaires afin de reproduire au mieux des cas de plagiats translingues s'approchant le plus de la réalité. Pour simuler ces cas de plagiat, il a été demandé aux participants de répondre à des questions en hindi sur un texte écrit en anglais et cela avec 4 niveaux différents de rédaction. Un texte en anglais est fourni, avec pour instructions de répondre à des questions en reprenant les phrases du texte et cela en :

Traduction automatique - le traduisant de façon automatique en hindi ;

Révision légère - le traduisant de façon automatique et en y apportant des corrections lexicales et grammaticales afin de corriger les erreurs du traducteur automatique ;

Traduction manuelle - le traduisant manuellement et en y faisant autant de modifications grammaticales et syntaxiques que possible tout en conservant le sens ;

Reformulation - comprenant le texte et en répondant aux questions avec leurs propres mots.

Le corpus est divisé en deux sous-corpus, un d'entraînement et un de test. Dans ces deux sous-corpus, les 5 032 documents sources écrits en anglais et issus de Wikipédia sont repris. Le Tableau 3.3 illustre la répartition des documents suspects en hindi dans les deux sous-corpus.

10. http://www.uni-weimar.de/medien/webis/events/panfire-11/panfire11-web/ (consulté le 30/06/2017 à 15h)
11. http://ir.shef.ac.uk/cloughie/resources/plagiarism_corpus.html (consulté le 30/06/2017 à 15h)

Sous-partie en anglais	
Texte original	5 032
Sous-partie en hindi	
Traduction automatique	45
Révision légère	30
Traduction manuelle	55
Reformulation	68
Total	198

(a) Corpus d'entraînement

Sous-partie en anglais	
Texte original	5 032
Sous-partie en hindi	
Traduction automatique	34
Révision légère	69
Traduction manuelle	43
Reformulation	44
Total	190

(b) Corpus de test

Tableau 3.3 – Statistiques du corpus CL!TR 2011.

3.3.2 Métriques d'évaluation

Lors de la compétition PAN@FIRE, l'efficacité d'un système à détecter la ré-utilisation de textes est mesurée avec la précision P, le rappel R et la F-mesure tels que décrits dans la section 3.1.2. Une détection est considérée comme correcte si le document ré-utilisé est identifié correctement et conjointement avec son document source correspondant.

3.4 Corpus ECLaPA

3.4.1 Le corpus

Le corpus ECLaPA [12] (Pereira *et al.*, 2010a; Pereira, 2010) ne fait pas partie d'une campagne d'évaluation comme tous les autres corpus présentés ici. Le corpus est composé de deux sous-corpus, un avec des cas de plagiat monolingue (contenant seulement des textes en anglais) et l'autre avec des cas de plagiat translingue. Nous allons nous intéresser ici seulement à la partie translingue. Dans cette partie, les documents suspects sont écrits en anglais et les sources sont écrites en portugais ou en français. Les textes sont issus des collections anglais-portugais et anglais-français d'Europarl (Koehn, 2005). De façon à simuler des cas de plagiat, des passages anglais correspondants à des passages portugais ou français (issus de documents sources portugais ou français) sont insérés au sein de documents suspects écrits entièrement en anglais. Plusieurs passages d'un même document suspect anglais peuvent être ainsi issus de documents sources portugais ou français. Cela donne 300 documents en anglais en tant que documents sources et 348 documents suspects (174 documents en portugais et 174 documents en français). Le Tableau 3.4 montre le nombre de documents du corpus par langue et catégorie.

Document	# de documents	Taille (en MB)	Langue		
			en	pt	fr
Suspect	300	89	300	0	0
Source	348	155	0	174	174

Tableau 3.4 – Statistiques du corpus ECLaPA. Chaque langue est notée par son code à deux lettres sous la norme ISO 639-1 (dit *alpha-2*).

Le même format que celui pour la PAN (Potthast *et al.*, 2009) est adopté pour annoter les passages plagiés (fichiers compagnons contenant des méta-données au format XML). Parmi les 300 documents suspects anglais, 100 ne contiennent aucun passage plagié et parmi les documents sources portugais et français, 100 ne sont pas utilisés comme source de plagiat (50 portugais et 50 français). Le corpus contient 2 196 cas de plagiat avec 30% de passages courts (de moins de

12. http://www.inf.ufrgs.br/~viviane/eclapa.html (consulté le 30/06/2017 à 15h)

1 500 caractères), 60% de passages de taille moyenne (entre 1 500 et 5 000 caractères) et 10% de passages plagiés de grande taille (entre 5 000 et 15 000 caractères). Chaque document suspect peut avoir jusqu'à 5 sources de plagiat différentes et pour chacune de ces sources, il peut y avoir jusqu'à 15 passages plagiés différents.

3.5 The Stanford Natural Language Inference (SNLI) Corpus

3.5.1 Le corpus

Le corpus SNLI [13] (*Stanford Natural Language Inference*) dans sa version 1, est une collection de 570 000 paires de phrases écrites en anglais et étiquetées pour servir à l'évaluation de la reconnaissance de correspondance sémantique textuelle (*inference*). Cette tâche d'inférence consiste à déterminer aux mieux la relation qui unit deux phrases. L'enjeu est de déterminer si elles sont sémantiquement équivalentes (*entailment*), contradictoires (*contradiction*) ou bien neutres (*neutral*) l'une par rapport à l'autre. Dans ce but, les paires du corpus SNLI sont étiquetées suivant ces trois catégories.

Ces paires de phrases ont été récoltées et étiquetées grâce à la plateforme de production participative (*crowdsourcing*) *Amazon Mechanical Turk* [14]. Les annotateurs ont été invités à produire pour une légende de photo donnée, et ce sans avoir accès à la photo elle-même, trois nouvelles légendes : une légende alternative *qui décrit également* la photo, une légende *qui peut décrire* la photo et une légende *qui ne décrit absolument pas* la photo. Les trois nouvelles légendes formulées sont ensuite respectivement étiquetées *entailment*, *neutral* et *contradiction* par rapport à la légende d'origine et forment un quadruplet avec cette dernière.

La version 1 du corpus (Bowman *et al.*, 2015) est monolingue puisque exclusivement en anglais. Toutefois, des recherches ultérieures ont étendu ce corpus au cas multi-genre (Williams *et al.*, 2017) et plus intéressant pour nous, au cas multilingue (Agić *et* Schluter, 2017). Dans les travaux de Agić *et* Schluter (2017), des experts ont traduit manuellement les 1 332 premières paires du corpus SNLI (Bowman *et al.*, 2015) de l'anglais à l'arabe, au français, au russe et à l'espagnol. Les étiquettes originales des relations de ces paires ont simplement été conservées. De cette façon, il est possible de comparer l'inférence translingue à travers ces 5 langues.

3.5.2 Métrique d'évaluation

Sur le site de présentation de ce corpus [15], un tableau de benchmarking recense tous les résultats des expérimentations réalisées sur ce corpus d'évaluation. La mesure d'évaluation utilisée pour évaluer ces approches est la performance de bonne classification, aussi appelée exactitude (*accuracy*) (Metz, 1978). En reprenant les notions vues dans la section 3.1.2, cette mesure s'exprime de la façon suivante :

$$ACC = \frac{(VP + VN)}{(VP + FP + VN + FN)} \tag{3.9}$$

13. https://nlp.stanford.edu/projects/snli/ (consulté le 30/05/2017 à 12h)
14. https://www.mturk.com/ (consulté le 30/05/2017 à 12h)
15. https://nlp.stanford.edu/projects/snli/ (consulté le 30/06/2017 à 15h)

3.6 Corpus d'évaluation de la tâche STS de la campagne SemEval 2017

3.6.1 Le corpus

La campagne d'évaluation SemEval propose depuis plusieurs années maintenant une tâche de détection de similarité textuelle sémantique (*Semantic Textual Similarity* en anglais ou STS) [16]. En 2016 [17], cette tâche est représentée par la tâche 1 et pour la première fois elle est étendue au cas translingue avec une sous-tâche de détection de similarité textuelle sémantique sur la paire de langues espagnol-anglais (Agirre *et al.*, 2016). En 2017 [18], cette sous-tâche translingue est renouvelée et une fois de plus étendue avec l'apparition de nouveaux couples de langues (Cer *et al.*, 2017). Contrairement au corpus BUCC, les corpus SemEval sont déjà sous forme de listes de paires de phrases.

Cette année, la tâche STS de SemEval, proposait donc les paires de langues suivantes :

— sous-tâche 1 : paires arabe-arabe
— **sous-tâche 2 : paires arabe-anglais**
— sous-tâche 3 : paires espagnol-espagnol
— **sous-tâche 4 (a et b) : paires espagnol-anglais**
— sous-tâche 5 : paires anglais-anglais
— **sous-tâche 6 : paires turc-anglais**

Les sous-tâches qui proposent une détection translingue sont donc les sous-tâches 2, 4a, 4b et 6.

Pour les sous-tâches 2, 4a et 6, les paires de phrases proviennent du corpus SNLI (Bowman *et al.*, 2015). Les paires utilisées ne sont pas directement les paires proposées dans le corpus SNLI d'origine. Elles sont ré-alignées avec un système développé pour l'occasion, à base d'une similarité cosinus entre vecteurs de mots. Ce système est détaillé dans le papier de présentation de la tâche (Cer *et al.*, 2017). Les phrases sources (première phrase de chaque paire) proviennent bien directement du corpus SNLI d'origine, qui dispose, pour rappel, seulement des paires de phrases anglais-anglais. Les phrases cibles sont, quant à elles, traduites manuellement, respectivement par des arabophones, des hispanophones et des turcophones, selon les sous-tâches.

Pour la sous-tâche 4b, les phrases sources (première phrase de chaque paire) proviennent du corpus de la tâche d'évaluation WMT13 [19] (Bojar *et al.*, 2013) tandis que les phrases cibles, les traductions en espagnol, sont produites avec le système de traduction baseline de la tâche d'évaluation WMT13. Une post-sélection des paires résultantes a été effectuée afin que les étiquettes de la tâche STS (voir section 3.6.2) puissent être comparées facilement avec les étiquettes de qualité de traduction de la tâche WMT13, afin que seulement les paires les plus pertinentes pour la tâche STS soient conservées.

Pour chaque sous-tâche, le corpus d'évaluation (*evaluation data*) est formé de 250 paires avec leur annotation de référence (*gold standard*) (rendues disponibles une fois l'évaluation terminée). Pour les données d'essai (*trial data*) et d'entraînement (*training data*), le nombre de paires disponibles est plus hétérogène : le Tableau 3.5 recense la répartition de ces données. Aucun corpus n'a été fourni avant la compétition pour la sous-tâche turc-anglais.

16. http://ixa2.si.ehu.es/stswiki/index.php/Main_Page (consulté le 30/06/2017 à 15h)
17. http://alt.qcri.org/semeval2016/task1/ (consulté le 30/06/2017 à 15h)
18. http://alt.qcri.org/semeval2017/task1/ (consulté le 30/06/2017 à 15h)
19. http://www.statmt.org/wmt13/translation-task.html (consulté le 24/07/2017 à 15h)
20. http://research.microsoft.com/en-us/downloads/607d14d9-20cd-47e3-85bc-a2f65cd28042/ (consulté le 24/07/2017 à 15h)
21. http://research.microsoft.com/en-us/downloads/38cf15fd-b8df-477e-a4e4-a4680caa75af/ (consulté le 24/07/2017 à 15h)
22. http://www.statmt.org/wmt08/shared-evaluation-task.html (consulté le 24/07/2017 à 15h)
23. http://ir.shef.ac.uk/cloughie/resources/plagiarism_corpus.html (consulté le 25/07/2017 à 15h)
24. https://archive.org/details/stackexchange (consulté le 25/07/2017 à 15h)

Année	Corpus	# de paires
2016	Trial	103
2016	News	301
2016	Multi-Source	294
2017	Trial	23
2017	WMT13 [19]	1 000

(a) Corpus pour le couple de langue es → en

Année	Corpus	# de paires
2017	Trial	23
2017	MSR-Paraphrase (*Microsoft Research Paraphrase Corpus*) [20]	1 020
2017	MSR-Video (*Microsoft Research Video Description Corpus*) [21]	736
2017	SMTeuroparl (*WMT2008 development dataset - europarl section*) [22]	406

(b) Corpus pour le couple de langue ar → en

Tableau 3.5 – Statistiques des corpus translingues des tâches STS correspondantes de la campagne d'évaluation SemEval, en fonction des paires de langues concernées, avec les langues notées sous la norme ISO 639-1. Les corpus News et Multi-Source pour le couple de langue es → en sont les données d'évaluation de l'édition 2016 et sont fournies comme données d'entraînement pour l'édition 2017. Elles sont principalement issues des corpus *Plagiarised Short Answers* [23] (Clough *et* Stevenson, 2011), *Europe Media Monitor* (Best *et al.*, 2005), des corpus de la tâche WMT12 (Callison-Burch *et al.*, 2012) et des archives *Stack Exchange Q&A Forums* [24].

3.6.2 Métrique d'évaluation

La mesure de similarité utilisée lors de la tâche est une étiquette qui représente une valeur numérique allant de 0 à 5, définie comme suit :

0 - Les deux phrases sont complètement différentes et n'ont strictement aucun rapport l'une avec l'autre ;

1 - Les deux phrases ne sont pas équivalentes mais traitent du même sujet ;

2 - Les deux phrases ne sont pas équivalentes mais partagent certains détails ;

3 - Les deux phrases sont à peu près équivalentes mais certaines informations importantes divergent ou sont manquantes ;

4 - Les deux phrases sont pratiquement équivalentes bien que quelques détails sans importance diffèrent ;

5 - Les deux phrases sont complètement équivalentes, elles signifient la même chose.

Pour les sous-tâches 4a et 6, l'annotation a été effectuée manuellement par cinq annotateurs par production participative (*crowdsourcing*) à l'aide de la plateforme *Amazon Mechanical Turk* [25]. Pour la tâche 2, une équipe d'experts de la Fondation du Qatar [26] s'est chargée de l'annotation. Pour chaque paire, la moyenne des annotations de tous les annotateurs est utilisée comme annotation de référence (*gold standard*). Pour la tâche 4b, un seul locuteur natif espagnol et parlant couramment anglais s'est chargé de l'annotation. Les instructions d'annotations ainsi que des exemples ont été fournis à chacun des annotateurs et peuvent être retrouvés dans l'article de Agirre *et al.* (2016).

La métrique d'évaluation de cette tâche est la corrélation de Pearson (Galton, 1886; Pearson, 1895) entre les sorties des méthodes soumises et les annotations humaines prises pour annotations de référence (*gold standard*).

25. https://www.mturk.com/ (consulté le 30/06/2017 à 15h)
26. https://www.qf.org.qa/ (consulté le 30/06/2017 à 15h)

3.7 Limites des corpus existants

Les corpus présentés ici peuvent tous servir à évaluer la détection de plagiat translingue et ils présentent certaines caractéristiques différentes, voire complémentaires. Certaines de ces différences sont illustrées dans le Tableau 3.6, qui recense les langues traitées, la structure et la provenance de chacun de ces corpus. En plus de leur volumétrie et des langues qu'ils comportent, l'élément qui différencie également ces corpus est leur structure et par conséquent leur protocole d'évaluation. Contrairement au corpus SNLI (Agić *et* Schluter, 2017) et SemEval (Cer *et al.*, 2017), les corpus de BUCC (Zweigenbaum *et al.*, 2017) et PAN (Potthast *et al.*, 2011b) ne sont pas formés d'une liste de paires de phrases (alignements de phrases qu'il faut affirmer ou infirmer), mais proposent des documents comparables comportant des phrases à aligner. L'ensemble des paires de phrases à comparer par les systèmes évalués est donc potentiellement le produit cartésien entre toutes les phrases d'un document source et toutes les phrases d'un document cible. Ceci, oblige les systèmes évalués à être très rapides ou à disposer d'un pré-traitement qui élague une grande majorité des paires de phrases candidates. Outre ces différences majeures, ils présentent tous des caractéristiques assez similaires et de ce fait comportent certaines limites dans leur exploitation pour une évaluation rigoureuse. Nous allons passer en revue ces limites au cours de cette section.

Corpus	Langues	Structure	Provenance
BUCC 2017 (Zweigenbaum *et al.*, 2017)	{de, fr, ru, zh} → en	Flux de texte	Wikipédia et News Commentary
PAN-PC-2011 (Potthast *et al.*, 2011b)	en → es	Flux de texte	Projet Gutenberg
CL!TR2011 (Barrón-Cedeño *et al.*, 2011)	en → hi	Flux de texte	Wikipédia
ECLaPA (Pereira *et al.*, 2010a)	{pt, fr} → en	Flux de texte	Europarl
SNLI (Agić *et* Schluter, 2017)	en ; ar ; fr ; ru ; es	Paire de phrases	SNLI
SemEval 2017 (Cer *et al.*, 2017)	{ar, en, tr} → en	Paire de phrases	SNLI

Tableau 3.6 – Caractéristiques des différents corpus existants qui peuvent servir à l'évaluation de la détection du plagiat translingue. Chaque langue est notée par son code à deux lettres sous la norme ISO 639-1 (dit *alpha-2*).

Pris séparément, ces corpus possèdent plusieurs limitations qui les empêchent d'être utilisés à eux seuls pour une évaluation rigoureuse.

La principale limite de ces corpus est qu'ils manquent de diversité. Le problème ne vient pas seulement du fait que les corpus soient tous comparables, mais plutôt du fait qu'ils traitent tous individuellement du même sujet avec le même style d'écriture, ou inversement avec des sujets et des styles d'écriture trop différents pour pouvoir établir des corrélations d'usages ou de résultats.

La deuxième limite à l'exploitation de ces corpus est le fait qu'ils sont presque tous construits à partir de traductions manuelles, que ce soit par le biais d'extraction depuis des corpus parallèles, comme c'est le cas pour BUCC et ECLaPA, ou que ce soit par appel à volontaires ou utilisation de plateformes participatives, comme c'est le cas pour le corpus PAN et SNLI. Seulement l'un des sous-corpus du jeu de données de SemEval (celui de la tâche 4b), créé à partir d'alignements de la campagne WMT13 (Bojar *et al.*, 2013), comporte des traductions automatiques. Cette préférence pour la traduction manuelle est appréciable car en effet, pour détecter au mieux des cas réels de plagiat translingue, il est important de s'approcher le plus possible de cas réels de reformulations et traductions. Cependant, les cas réels de plagiat translingue commis par des étudiants comportent également des cas avec traductions automatiques (avec ou sans relecture et post-correction). De plus, il serait bon d'avoir un corpus comportant des traductions manuelles et des traductions automatiques afin de comparer le comportement des méthodes évaluées à travers ces deux types de traductions.

Une autre limite de ces corpus s'exprime par un manque de diversité dans les tailles de leurs passages plagiés. Certains de ces corpus proposent des alignements phrastiques (SNLI et SemEval) tandis que d'autres proposent d'aligner des extraits de textes allant de une à plusieurs phrases au sein d'un texte plus conséquent (BUCC, PAN et ECLaPA). Ils ne proposent pas d'alignements plus fins que la phrase, alors que les cas de plagiat peuvent se produire à des gra-

nularités inférieures à la phrase. Utiliser plusieurs de ces corpus au cours d'une même évaluation afin de disposer ainsi d'un plus large panel de granularités ne donnerait pas lieu pour autant à une évaluation plus rigoureuse. Le fait d'utiliser un corpus documentaire et un corpus phrastique au cours de la même étude, ne contribue pas à produire une évaluation rigoureuse, car si les phrases ne sont pas directement tirées des documents (avec le même sujet, les mêmes auteurs, *etc.*), les résultats obtenus sur ces deux ensembles ne pourront pas être comparés objectivement. De ce fait, il ne sera pas possible de savoir si des différences significatives potentiellement observées seront dues à la différence de granularités ou bien à un autre facteur, comme la différence de lexique ou de style d'écriture.

Pour conclure, pris séparément, nous pensons que ces corpus ne sont pas assez diversifiés pour servir à eux seuls à une évaluation rigoureuse. Ils couvrent, pour la plupart, seulement un domaine spécifique (littéraire, législatif, légendes d'images) et sont principalement issus de corpus comparables. Ils sont traduits manuellement, écrits par la même catégorie d'auteurs et disposent de la même granularité de passages plagiés.

Il est donc difficile de tirer, à partir des évaluations effectuées sur ces corpus, des conclusions scientifiques pouvant être prises comme cas général et pouvant être ainsi exploitées pour implémenter un outil commercial à destination d'une clientèle exigeante. Idéalement, un corpus permettant une évaluation rigoureuse des méthodes de détection du plagiat translingue ne devra pas contenir ces limites et devra être aussi diversifié que possible.

C'est pourquoi, dans le chapitre suivant, nous présentons un nouveau jeu de données qui comporte des textes de tailles diverses, écrits en plusieurs langues, avec plusieurs styles d'écriture et traduits de diverses façons. Nous mènerons ainsi à l'aide de ce nouveau jeu de documents une évaluation de plusieurs méthodes de l'état de l'art, non pas pour déterminer laquelle est la plus performante de manière générale, mais plutôt dans le but d'étudier leur comportement sur différents types et tailles de textes ainsi que sur des couples de langues proches ou plus éloignées afin de déterminer s'il existe des liens entre leur comportement et les caractéristiques des textes traités.

Deuxième partie

Contributions

4 Un corpus multilingue, multi-genre et multi-granularité pour l'évaluation de la détection du plagiat trans-lingue

« Quand j'emprunte une idée à un autre, j'oublie souvent de la lui rendre. »

— Georges Wolinsky (1934-2015)

Dans le chapitre 3, nous avons vu l'existence de corpus utiles à l'évaluation de la détection du plagiat. Nous avons également vu que ces corpus connaissent certaines limites lors de leur utilisation dans cette tâche. En effet, pris séparément, ces corpus ne sont pas assez diversifiés pour servir, à eux seuls, à une évaluation rigoureuse. Ils couvrent, pour la plupart, seulement un domaine spécifique (littéraire, législatif, légendes d'images) et sont principalement issus de corpus comparables. De plus, ils sont traduits manuellement, écrits par la même catégorie d'auteurs et présentent la même granularité de passages plagiés.

Pour répondre à ces limites, nous avons construit un corpus multilingue, multi-genre et multi-granularité pour permettre une évaluation rigoureuse des méthodes de détection du plagiat trans-lingue. Pour le moment, notre corpus se focalise uniquement sur les langues anglaise, française et espagnole [1]. Plus précisément, ses caractéristiques sont les suivantes :

- il est multilingue : il contient des alignements en français, anglais et espagnol ;
- il propose des alignements à différentes granularités (taille de textes) : syntagme nominal, phrase et document ;
- il est basé à la fois sur des corpus parallèles et des corpus comparables ;
- il contient des textes traduits manuellement et automatiquement ;
- il contient des passages altérés automatiquement (c'est-à-dire ayant subi une opération volontaire pour rendre la détection de plagiat plus compliquée), des passages altérés involontairement (présence de balises HTML ou de fautes d'orthographe, par exemple) et d'autres passages sans bruit ;
- les documents sont écrits par différents types d'auteurs, allant du néophyte au professionnel, en passant par le contributeur du Web ;
- les documents traitent plusieurs sujets et thèmes différents (littérature, science, avis d'internautes, *etc.*)

Ce chapitre présente la méthodologie de collecte et de construction de ce jeu de données (Ferrero *et al.*, 2016) et son utilisation pour l'évaluation des méthodes de l'état de l'art (Ferrero *et al.*, 2016, 2017b), démontrant ainsi que ce corpus ne présente pas la plupart des limites rencontrées par les autres corpus existants.

1. Ces trois langues sont des langues majeures parlées en Europe et de ce fait intéressent tout particulièrement l'entreprise Compilatio.

4.1 Construction et propriétés du corpus

Les contributions de cette section ont fait l'objet d'une publication (Ferrero *et al.*, 2016).

4.1.1 Réutilisation de corpus parallèles et comparables existants

Nous décidons, en premier lieu, de réutiliser des collections de corpus parallèles et comparables déjà existantes afin de constituer une base déjà éprouvée pour notre corpus.

4.1.1.1 JRC-Acquis et Wikipédia

Le corpus CL-PL-09 [2], utilisé lors des recherches de Potthast *et al.* (2011a), regroupe des documents du corpus parallèle JRC-Acquis [3] [4] (Steinberger *et al.*, 2006) et du corpus comparable Wikipédia. Il est donc divisé en deux sous-corpus :

— une partie parallèle composée de textes extraits des JRC-Acquis, un corpus souvent utilisé dans les recherches en traduction et traitement automatique des langues. JRC-Acquis est un corpus parallèle (voir section 2.2.3.1) contenant des extraits de l'acquis communautaire (l'ensemble du corpus juridique communautaire, c'est-à-dire l'ensemble des droits et obligations juridiques qui lient les États-membres de l'Union Européenne), disponibles dans plus de 20 langues parlées au sein des pays de l'Union Européenne.

— une partie comparable composée de textes issus de Wikipédia. L'encyclopédie Wikipédia est considérée comme un corpus comparable (voir section 2.2.4.1) puisqu'elle comprend des documents dans plus de 200 langues qui sont alignés par sujet.

Ce corpus inclut des textes en néerlandais, anglais, français, allemand, polonais et espagnol. La collection JRC-Acquis et la collection Wikipédia contiennent chacun 10 000 documents alignés. Au total, le corpus CL-PL-09 contient 120 000 documents (10 000 documents pour les deux corpus dans les six langues à chaque fois). Pour le moment, notre corpus se focalise uniquement sur les langues : anglais, français et espagnol. Nous conservons donc uniquement les sous-collections de ces trois langues pour construire notre corpus.

4.1.1.2 Europarl

Europarl [5] (Koehn, 2005) est sans doute le corpus parallèle, avec JRC-Acquis, le plus couramment utilisé dans les tâches de comparaison de textes multilingues et de traduction automatique. Il est constitué des retranscriptions des échanges du Parlement Européen. Tiedemann (2012) rend disponible, à travers une archive [6], approximativement 10 000 documents parallèles issus d'Europarl, alignés simultanément dans 21 des langues parlées à travers l'Union Européenne.

Nous récupérons seulement les sous-collections anglaise, française et espagnole depuis le corpus de Tiedemann (2012), car notre corpus se focalise, pour l'heure, uniquement sur ces trois langues.

4.1.1.3 Revues de produits Amazon (Webis-CLS-10)

Le corpus *Cross-Lingual Sentiment* (*Webis-CLS-10*) [7] est constitué et utilisé pour la première fois par Prettenhofer *et* Stein (2010) pour l'évaluation de la tâche d'analyse translingue des sentiments. C'est une collection de revues *Amazon* pour trois catégories de produits (livres,

2. http://users.dsic.upv.es/grupos/nle/resources/abc/download-clpl09.html (consulté le 12/06/2017 à 14h)
3. http://optima.jrc.it/Acquis/JRC-Acquis.3.0/doc/README_Acquis-Communautaire-corpus_JRC.html (consulté le 15/04/2017 à 19h)
4. https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis (consulté le 13/06/2017 à 10h)
5. http://www.statmt.org/europarl/ (visité le 12/06/2017 à 16h)
6. http://opus.lingfil.uu.se/Europarl.php (consulté le 12/06/2017 à 16h)
7. https://www.uni-weimar.de/en/media/chairs/computer-science-and-media/webis/corpora/corpus-webis-cls-10/ (consulté le 12/06/2017 à 16h)

DVD et albums musicaux) et écrites dans quatre langues (français, anglais, allemand et japonais). Les revues allemandes, françaises et japonaises ont été aspirées respectivement depuis les sites *Amazon.de*, *Amazon.fr* et *Amazon.co.jp*, en 2009. Le corpus est étendu avec des revues de produits en anglais, récupérées depuis le jeu de données *Multi-Domain Sentiment Dataset* de Blitzer *et al.* (2007). Pour chacune des trois langues collectées depuis le Web – allemand, français et japonais – une traduction en anglais est effectuée en utilisant *Google Translate*. Cela donne lieu à 2 000 alignements par langue (français, anglais, allemand et japonais) et par catégorie de produits (livres, DVD et albums musicaux).

Nous conservons seulement les 6 000 alignements de la paire anglais-français (2 000 alignements pour chacune des trois catégories de produits) pour les inclure dans notre corpus.

4.1.2 Enrichissements

Outre le fait de réutiliser des corpus parallèles et comparables déjà existants, nous enrichissons également notre jeu de données en collectant et créant de nouveaux alignements.

4.1.2.1 PAN-PC-11

Gupta *et al.* (2012) font allusion dans leurs travaux à la construction d'un corpus translingue à partir du corpus multilingue PAN-PC-11. L'idée est d'utiliser les fichiers de méta-données qui accompagnent chaque texte du corpus de l'évaluation PAN pour reconstruire un corpus translingue. Le corpus PAN-PC-11 [8] (Potthast *et al.*, 2011b) a déjà été présenté au cours de la section 3.2. Pour rappel, les textes de ce corpus sont issus du projet Gutenberg [9] et les cas de plagiat translingue qui s'y trouvent ont été produits via la plateforme de production participative (*crowdsourcing*) *Amazon Mechanical Turk* [10].

Chaque document suspect du corpus est donc accompagné d'un fichier XML. Ce fichier XML contient, pour chaque passage plagié du document auquel il se réfère, les informations suivantes :

- *type* : le type de plagiat dont il s'agit ;
- *manual_obfuscation* : si le passage plagié comporte ou non une offuscation ;
- *this_language* : la langue dans laquelle est écrite le passage dans le document suspect ;
- *this_offset* : la position du caractère de départ du passage dans le document suspect ;
- *this_length* : la longueur en nombre de caractères du passage dans le document suspect ;
- *source_reference* : le nom du fichier du document source d'origine du passage ;
- *source_language* : la langue dans laquelle est écrite le passage dans le document source ;
- *source_offset* : la position du caractère de départ du passage dans le document source ;
- *source_length* : la longueur en nombre de caractères du passage dans le document source.

Pour exemple, un passage plagié est représenté au sein d'un fichier de méta-données par un élément *feature*, comme on peut le voir dans l'extrait suivant :

```
<feature name="plagiarism" type="translation" manual_obfuscation="false"
this_language="en" this_offset="1" this_length="1346"
source_reference="source-document01152.txt" source_language="es"
source_offset="33445" source_length="1449" />
```

Nous souhaitions ré-utiliser ces passages plagiés au sein de notre corpus afin de disposer d'alignements de type littéraire altérés (avec ajout automatique de bruit dans le texte afin de rendre le plagiat plus difficilement détectable) et ainsi garantir une diversité dans les alignements proposés dans notre corpus. Pour cela, nous analysons les fichiers XML de chaque document suspect (qui eux sont au format brut) à la recherche d'éléments comportant un attribut *name* qui a pour valeur *plagiarism* et un attribut *type* qui a pour valeur *translation*, c'est-à-dire que nous

8. https://www.uni-weimar.de/en/media/chairs/computer-science-and-media/webis/corpora/corpus-pan-pc-11/ (consulté le 31/05/2017 à 11h)
9. https://www.gutenberg.org/ (consulté le 31/05/2017 à 16h)
10. https://www.mturk.com/ (consulté le 30/05/2017 à 12h)

cherchons les passages plagiés par traduction. Si un tel passage est identifié, nous vérifions que sa langue dans le document suspect (*this_language*) est bien différente de sa langue dans le document source (*source_language*) et que les langues concernées correspondent bien à notre étude (anglais et espagnol –peu importe l'ordre de la "traduction"– le français n'étant pas présent dans ce corpus). Si c'est bien le cas, nous extrayons le passage dans le document suspect grâce aux informations *this_offset* et *this_length*, qui représentent sa position. Nous faisons de même pour le passage correspondant dans le document *source_reference* grâce aux informations *source_offset* et *source_length*. Dans l'exemple précédent, les passages extraits s'étendent du caractère 1 à 1 347 dans le texte suspect et du caractère 33 445 à 34 894 dans le texte source. Divers renseignements supplémentaires (offuscation manuelle ou bruit) sont également conservés. Les deux passages extraits forment alors un alignement (de paires parallèles).

Nous obtenons par cette méthode 2 920 paires anglais-espagnol.

4.1.2.2 Articles TALN et *ACL Anthology*

À notre connaissance, aucun corpus parallèle ou comparable ne contient de documents scientifiques, comme des articles de conférences, potentiellement altérés par des formules, des annotations, des acronymes, *etc.* Toujours dans le but de diversifier le corpus que l'on propose, nous souhaitons y inclure ce type de documents, car ils présentent sans doute, des caractéristiques bien particulières. C'est pourquoi nous décidons de collecter des articles de conférences qui ont été initialement publiés dans une langue et ensuite traduits par leur.s auteur.e.s pour être publiés dans une autre langue. Souhaitant opérer une vérification manuelle après cette collecte, nous nous concentrons uniquement sur le couple de langue français-anglais (langues parlées couramment par mes deux directeurs de thèse et moi-même). Nous collectons donc des articles de conférences francophones qui ont été initialement publiés en français et ensuite traduits par leur.s auteur.e.s pour être publiés dans une conférence internationale en anglais.

Pour mener à bien cette collecte, nous récupérons dans un premier temps des fichiers BibTeX d'actes de conférences francophones sur le traitement automatique du langage. BibTeX est un format de fichier de gestion de références bibliographiques interprété par LaTeX. Un tel fichier contient des informations sur des articles telles que leur titre, leur.s auteur.e.s, leur année de publication, la conférence dans laquelle ils ont été publiés, *etc.* Ci-dessous, un exemple d'entrée BibTeX d'un de nos articles.

```
@InProceedings{ferrero_lrec_2016,
  author = {Jérémy Ferrero and Frédéric Agnès and Laurent Besacier
   and Didier Schwab},
  title = {A Multilingual, Multi-style and Multi-granularity Dataset
   for Cross-language Textual Similarity Detection},
  booktitle = {Proceedings of the Tenth International Conference
   on Language Resources and Evaluation},
  year = {2016},
  pages = {4162-4169},
  address = {Portoroz, Slovenia},
  month = {May},
  publisher = {European Language Resources Association (ELRA)}
}
```

De façon plus détaillée, nous avons collecté :

— les archives de la conférence TALN [11] (*Traitement Automatique des Langues Naturelles*) de 1997 à 2014, rendues disponibles par le travail de Boudin (2013) [12] ;

11. https://www.atala.org/-Conference-TALN-RECITAL- (consulté 16/06/2017 à 12h)
12. http://github.com/boudinfl/taln-archives (consulté 16/06/2017 à 15h)

— les archives des éditions RNTI[13] (*Revue des Nouvelles Technologies de l'Information*) de 2006 à 2011, rendues disponibles lors du challenge de la conférence EGC 2016[14] (*Extraction et Gestion des Connaissances*).

Nous analysons ces fichiers afin de récolter les noms des auteur.e.s de chaque article de tous les actes collectés. Nous effectuons ensuite une recherche dans *Google Scholar* et via l'API *Google Search*[15] en utilisant comme arguments de requêtes les jeux de noms d'auteur.e.s extraits et en spécifiant la langue des résultats attendus (l'anglais). Jusqu'à dix résultats de l'API *Google Search* et jusqu'à quatre résultats de *Google Scholar* peuvent être conservés (en fonction des retours des moteurs de recherche). Nous téléchargeons, au format PDF, les articles des résultats renvoyés par les moteurs de recherche et utilisons l'algorithme de classification de textes de Cavnar *et* Trenkle (1994) pour déterminer la langue de chacun de ces articles. Chaque document candidat téléchargé dont la langue prédite est l'anglais et dont le jeu d'auteur.e.s correspond est alors mis de côté. Nous vérifions ensuite manuellement les articles conservés pour voir si une partie significative de ces articles est bien relative à l'article original français dont il faisait la requête.

Nous avons, par ce moyen, réuni 35 paires d'articles français-anglais.

4.1.3 Plusieurs granularités

Pour permettre une évaluation rigoureuse des méthodes de détection du plagiat translingue, nous souhaitons un corpus disposant de plusieurs granularités d'alignements. Cela permet, en plus d'étudier le comportement de méthodes sur des types de textes différents, de pouvoir également identifier si une méthode est plus performante sur une taille de texte particulière.

Nous proposons donc notre corpus avec trois différentes granularités (tailles de texte) :

— des documents ;
— des phrases ;
— des groupes de mots, aussi appelés syntagmes (*syntagma* ou *chunk* en anglais).

4.1.3.1 Découpage

La première granularité proposée correspond au niveau documentaire. Les alignements restent sous la forme des documents tels que nous les avons collectés (voir section 4.1.1 et section 4.1.2).

La seconde granularité proposée est au niveau phrastique. Il s'agit d'alignements de phrases. Pour obtenir ces alignements, dans un premier temps, nous découpons chaque document du corpus en phrases, suivant les marqueurs de ponctuation de fin de phrase (., ?, !). Un document est donc maintenant sous forme d'une liste de phrases. Ensuite, nous procédons à une phase d'alignement telle que décrite en section 4.1.3.2.

La troisième granularité est constituée de groupes de mots, ou plus exactement de syntagmes nominaux. Nous nous focalisons sur les syntagmes nominaux car nous les considérons comme étant les éléments les plus porteurs de sens au sein d'une phrase et donc étant potentiellement les plus susceptibles d'être ré-utilisés lors d'un plagiat.

Pour extraire ces syntagmes depuis les phrases, nous avons essayé plusieurs solutions d'analyse syntaxique de surface (*shallow parsing* ou *chunking* en anglais). Ces solutions sont répertoriées dans le Tableau 4.1 avec les langues qu'elles supportent nativement et leur format de sortie. Après plusieurs tests, ces solutions ne se sont pas avérées satisfaisantes. Ce ne sont pas exactement des solutions de découpage de phrases en syntagmes mais des solutions d'analyse syntaxique. Un post-traitement est donc nécessaire pour extraire depuis leur sortie, les syntagmes désirés. Étant donné leur format de sortie, ce post-traitement aurait été trop complexe. De plus, pour la plupart d'entre elles, certaines langues ne sont pas supportées. La création d'une grammaire pour une

13. http://editions-rnti.fr/ (consulté le 16/06/2017 à 12h)
14. http://www.egc.asso.fr/Manifestations_dEGC/71-FR-Defi_EGC_2016_Communaute_EGC_quelle_histoire_et_quel_avenir (consulté 16/06/2017 à 12h)
15. https://developers.google.com/custom-search/json-api/v1/using_rest (consulté le 16/06/2017 à 15h)

nouvelle langue ou l'apprentissage de celle-ci depuis un corpus aurait donc été nécessaire. C'est pourquoi notre choix ne s'est pas porté sur l'une de ces solutions mais plutôt sur l'étiqueteur morphosyntaxique *TreeTagger* (à ne pas confondre avec le *chunker TreeTagger* mentionné dans le Tableau 4.1).

	Langue			
	en	fr	es	Format de sortie
StandFord NLP Parser[16] (Klein *et* Manning, 2003, 2002; Green *et al.*, 2011)	✓	✓	✓	Arbre syntaxique
Berkeley Parser[17] (Petrov *et al.*, 2006; Petrov *et* Klein, 2007)	✓	✓		Arbre syntaxique
TreeTagger Chunker[18] (Schmid, 1994)	✓	✓		Arbre syntaxique
YamCha[19] (Kudo *et* Matsumoto, 2001, 2003)	✓			IOB2
MaltParser[20] (Nivre *et al.*, 2006)	✓	✓	✓	CoNLL[21]

Tableau 4.1 – Les différentes solutions d'analyse syntaxique de surface testées avec pour chacune les langues qu'elles supportent et leur format de sortie. Chaque langue est notée par son code à deux lettres sous la norme ISO 639-1 (dit *alpha-2*).

Nous utilisons donc l'étiqueteur morphosyntaxique *TreeTagger* (Schmid, 1994) pour étiqueter la partie du discours (*part-of-speech*) de chaque mot des phrases. Nous normalisons ensuite ces étiquettes avec le jeu d'étiquettes universelles[22] de Petrov *et al.* (2012). Le Tableau 4.2 résume les correspondances entre les parties de discours les plus communes dans les langues à alphabet latin et leur étiquette universelle d'après le travail de Petrov *et al.* (2012).

Étiquette Universelle	Parties de discours concernées
VERB	les verbes conjugués à tous les temps et toutes les personnes
NOUN	les noms propres et communs
PRON	les pronoms
ADJ	les adjectifs
ADV	les adverbes
ADP	les adpositions (prépositions et postpositions)
CONJ	les conjonctions
DET	les déterminants
NUM	les nombres cardinaux
PRT	les particules et autres mots fonctions
X	les mots étrangers, mots inconnus et abréviations
.	les marqueurs de ponctuation

Tableau 4.2 – Correspondances entre les parties de discours les plus communes dans les langues à alphabet latin et les étiquettes auxquelles elles correspondent dans le jeu d'étiquettes universelles de Petrov *et al.* (2012).

Nous concaténons les mots en accord avec leur étiquette pour construire, suivant une grammaire faite maison, des groupes de mots qui peuvent être considérés comme faisant partie d'un même syntagme nominal. Nous construisons ces groupes de mots en concaténant tous les mots ayant pour étiquette NOUN, ADJ, ADP, CONJ, DET ou PRT, dans la limite des 10 mots d'affilée tout en interdisant les prépositions et les articles en fin de syntagme. Un mot n'étant pas alphanumérique (composé uniquement de caractères alphabétiques et numériques) brise la séquence de concaténation et met donc fin au syntagme en cours de construction. Nous avons aussi considéré une taille minimale et maximale pour construire chaque syntagme. Ces tailles sont

16. https://nlp.stanford.edu/software/lex-parser.shtml (consulté 21/06/2017 à 17h)
17. https://github.com/slavpetrov/berkeleyparser (consulté 21/06/2017 à 17h)
18. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (consulté 21/06/2017 à 17h)
19. http://chasen.org/~taku/software/yamcha/ (consulté 21/06/2017 à 17h)
20. http://www.maltparser.org/ (consulté 21/06/2017 à 17h)
21. http://universaldependencies.org/format.html (consulté le 12/07/2017 à 16h)
22. https://github.com/FerreroJeremy/universal-pos-tags (consulté le 21/06/2017 à 15h)

empiriquement fixées à, respectivement, 3 et 10 mots. Pour aligner ces unités, nous procédons de la même façon que pour aligner les phrases (voir section 4.1.3.2).

4.1.3.2 Alignement

Une fois ces deux nouveaux sous-corpus d'éléments constitués (phrases et syntagmes), il reste à les aligner afin d'obtenir des paires parallèles.

Nous avons testé plusieurs outils d'alignement, dont *YouAlign*[23], *Bibtext2tmx*[24], *Felix*[25] et *CafeTran*[26]. Ces outils sont des logiciels avec client lourd graphique. Ils ne disposent pas d'un mode de lancement en ligne de commande par le biais d'un terminal et ne peuvent donc pas être automatisés et monitorés. De plus, ils n'acceptent qu'un fichier à la fois en entrée et certains même avec une limite de taille, ce qui rend leur utilisation inintéressante dans notre cas. C'est pourquoi nous avons plutôt porté notre attention sur *HunAlign*[27] (Varga *et al.*, 2005) et *Champollion*[28] (Ma, 2006). Ces deux solutions sont fondées sur des dictionnaires (de traductions). Comme la plupart des aligneurs, ils ne gèrent pas le changement d'ordre des phrases et ne peuvent donc pas repérer les alignements croisés. *Champollion* ne propose pas de probabilités d'alignements en sortie, ce qui rend un seuillage de post-traitement (afin d'affiner les résultats des alignements) impossible. C'est pourquoi notre choix s'est finalement porté sur *HunAlign* (Varga *et al.*, 2005).

Nous allons ici seulement décrire le procédé d'alignement utilisé pour les phrases, le procédé pour aligner les syntagmes étant scrupuleusement le même.

Nous cherchons donc des phrases à aligner à travers des documents qui étaient à la base des documents parallèles ou comparables (il est donc fort probable que des phrases parallèles s'y trouvent). Nous cherchons à aligner ces phrases par paires ou par triplets, cela dépend du nombre de langues disponibles selon les collections (anglais-français pour TALN-ACL et anglais-français-espagnol pour Europarl, par exemple). *HunAlign* prend en entrée deux textes en deux langues différentes, tout deux segmentés en phrases (deux textes monolingues sous forme de liste de phrases). Ne pouvant donc comparer que deux textes à la fois, nous procédons à trois comparaisons de deux textes deux à deux dans les cas où il y a trois langues disponibles (et donc trois documents à aligner).

HunAlign utilise un dictionnaire d'alignements au niveau mot, combiné avec l'algorithme d'alignement de Gale-Church (Gale *et* Church, 1993) basé sur une notion de longueur de phrases. Nous avons enrichi son dictionnaire avec des traductions tirées de *DBnary*[29] (Sérasset, 2015). Son utilisation a également été couplée avec le modèle de longueur de Pouliquen *et al.* (2003b) présenté dans la section 2.2.1.3. Nous utilisons un seuil empirique que nous ajustons d'un sous-corpus à un autre pour filtrer les sorties de *HunAlign* afin d'assurer le meilleur compromis possible entre le nombre d'alignements retournés et leur qualité.

Il est important de noter qu'étant donné que nous utilisons un outil d'alignement pour constituer les alignements des granularités phrastique et syntagmatique, les alignements conservés à ces granularités tendent à être parallèles, même ceux issus de sous-corpus comparables à la base. En effet, l'outil d'alignement va réussir, de par son fonctionnement, à aligner correctement seulement les phrases et syntagmes qui sont (strictement) parallèles et non les comparables. On peut noter également, pour la même raison, que l'alignement donne de meilleurs résultats (plus de phrases parallèles sont obtenues) sur les corpus parallèles que sur les corpus comparables. Concernant les syntagmes, le seuil du filtre en sortie de *HunAlign* a été défini pour maximiser la qualité des alignements, ce qui peut expliquer le nombre réduit de ces alignements par rapport à ceux des phrases.

23. http://www.youalign.com/ (consulté le 21/06/2017 à 15h)
24. http://bitext2tmx.sourceforge.net/ (consulté le 21/06/2017 à 15h)
25. http://felix-cat.com/tools/align-assist/ (consulté le 21/06/2017 à 15h)
26. https://www.cafetran.com/ (consulté le 21/06/2017 à 15h)
27. http://mokk.bme.hu/resources/hunalign/ (consulté le 14/06/2017 à 10h)
28. http://champollion.sourceforge.net/ (consulté le 26/06/2017 à 14h)
29. http://kaiko.getalp.org/about-dbnary/ (consulté le 30/06/2017 à 15h)

4.1.3.3 Vérification des alignements

Afin de vérifier la qualité des alignements que l'on propose dans notre corpus, un contrôle manuel a été réalisé. Trois annotateurs (mes deux directeurs de thèse et moi-même) ont vérifié manuellement 1 380 paires ou triplets issus des alignements des syntagmes nominaux (la plus petite granularité proposée dans notre corpus). Ces alignements ont été sélectionnés aléatoirement. Chacun des annotateurs s'est chargé d'annoter tous les alignements sélectionnés d'un ou plusieurs corpus. Un alignement n'a donc été annoté que par un seul annotateur.

Cette vérification représente 3,44% du sous-corpus des syntagmes nominaux et montre une confiance de 94,57% dans les alignements considérés comme corrects. Le Tableau 4.3 recense le nombre de paires vérifiées pour chaque sous-corpus ainsi que les taux d'erreurs associés. Nous pouvions augmenter la justesse des alignements en modifiant le seuil utilisé en sortie de *HunAlign* pour filtrer les alignements à conserver. Toutefois, une augmentation de la qualité des alignements retournés donne lieu à une réduction du nombre d'alignements exploitables.

Sous-corpus	Nombre d'alignements total	Nombre d'alignements vérifiés	% d'alignements vérifiés	Nombre d'alignements faux	% d'erreur
JRC-Acquis	10 094	211	2,09	15	7,11
Europarl	25 603	204	0,79	15	7,35
Wikipédia	132	132	100,00	7	5,30
PAN-PC-11	1 360	327	24,04	14	4,28
Webis-CLS-10	2 603	254	9,75	16	6,29
TALN-ACL	272	252	92,64	8	3,17
Total	40 064	1 380	3,44	75	5,43

Tableau 4.3 – Nombre de paires vérifiées pour chaque sous-corpus ainsi que le pourcentage de couverture du corpus et le pourcentage d'erreur qu'elles représentent.

4.1.4 Caractéristiques du corpus constitué

Les différentes caractéristiques de notre corpus sont synthétisées dans le Tableau 4.4, tandis que le Tableau 4.5 présente le nombre d'alignements obtenus pour les trois différentes granularités proposées pour chaque sous-corpus et pour chaque langue. De plus amples détails sur les statistiques de chaque sous-corpus à chaque granularité peuvent être trouvés en Annexe B.

Sous-corpus	Alignement	Auteurs	Traducteurs	Altération	EN
JRC-Acquis	Parallèle	Politiciens	Traducteurs professionnels	Non	3,74%
Europarl	Parallèle	Politiciens	Traducteurs professionnels	Non	7,74%
Wikipédia	Comparable	N'importe qui	N'importe qui	Bruit	8,37%
PAN-PC-11	Parallèle	Auteurs professionnels	N'importe qui	Oui	3,24%
Webis-CLS-10	Parallèle	N'importe qui	Google Translate	Non	6,04%
TALN-ACL	Comparable	Scientifiques en TAL	Scientifiques en TAL	Bruit	9,36%

Tableau 4.4 – Caractéristiques générales par collection de notre corpus au niveau documentaire. Les pourcentages d'entités nommées (EN) présents dans la dernière colonne sont calculés avec le *Stanford Named Entity Recognizer* : `http://nlp.stanford.edu/software/CRF-NER.shtml`. Des chiffres plus détaillés sont donnés dans le Tableau 4.6.

Le Tableau 4.6 donne le nombre et le pourcentage d'entités nommées contenues dans chaque sous-corpus. Ces nombres sont calculés au moyen du *Stanford Named Entity Recognizer*[30].

Nous rendons ce corpus public sur GitHub[31]. Tous les scripts nécessaires à sa construction sont également disponibles à cette adresse.

30. `http://nlp.stanford.edu/software/CRF-NER.shtml` (consulté le 26/06/2017 à 15h)
31. `https://github.com/FerreroJeremy/Cross-Language-Dataset` (consulté le 12/06/2017 à 10h)

Sous-corpus	Langues	Nombre de documents alignés	Nombre de phrases alignées	Nombre de syntagmes alignés
JRC-Acquis	en, fr, es	≃ 10 000	≃ 150 000	≃ 10 000
Europarl	en, fr, es	≃ 10 000	≃ 475 000	≃ 25 600
Wikipédia	en, fr, es	≃ 10 000	≃ 5 000	≃ 150
PAN-PC-11	en, es	≃ 3 000	≃ 90 000	≃ 1 400
Webis-CLS-10	en, fr	≃ 6 000	≃ 23 000	≃ 2 600
TALN-ACL	en, fr	≃ 35	≃ 1 300	≃ 300

Tableau 4.5 – Nombre de documents, de phrases et de syntagmes nominaux alignés dans notre corpus par sous-corpus. Chaque langue est notée par son code à deux lettres sous la norme ISO 639-1 (dit *alpha-2*).

Sous-corpus	Nombre de fichiers analysées	Nombre de mots analysées	Nombre d'entités nommées	% d'entités nommées
JRC-Acquis	10 000	23 595 364	883 347	3,74
Europarl	9 565	54 616 735	4 229 350	7,74
Wikipédia	10 000	18 043 902	1 510 151	8,37
PAN-PC-11	2 920	3 737 788	121 203	3,24
Webis-CLS-10	6 000	563 565	34 020	6,04
TALN-ACL	620	42 923	4 019	9,36

Tableau 4.6 – Pourcentage d'entités nommées par sous-corpus à la granularité documentaire. Les pourcentages d'entités nommées sont calculés avec le *Stanford Named Entity Recognizer* : http://nlp.stanford.edu/software/CRF-NER.shtml.

4.1.5 Perspectives d'évolution

Ce corpus peut évoluer. Nous invitons la communauté à l'améliorer et l'étendre.

Quelques pistes d'améliorations peuvent être :

— ajouter des alignements provenant de nouveaux corpus, comme ceux de SemEval (Cer *et al.*, 2017), BUCC (Zweigenbaum *et al.*, 2016, 2017) ou SNLI (Bowman *et al.*, 2015) ;
— ajouter de nouvelles langues dans les sous-corpus déjà présents, comme JRC-Acquis (Steinberger *et al.*, 2006) ou Europarl (Koehn, 2005) ;
— ajouter des sous-corpus avec des pourcentages plus extrêmes et significatifs d'entités nommées (un sous-corpus avec 0% et un autre avec plus de 10%, par exemple) afin de vérifier le réel impact de cette caractéristique sur l'efficacité des méthodes de détection ;
— ajouter différents types et différents niveaux d'offuscation ;
— ajouter une granularité intermédiaire entre le niveau des syntagmes nominaux et celui des phrases ou bien s'orienter vers l'ajout de syntagmes verbaux ou adverbiaux (aussi porteurs de sens) ;

De plus, nous développons actuellement un outil de génération de corpus qui, à partir d'un corpus translingue comme le nôtre, construit seulement pour l'évaluation des similarités textuelles sémantiques, peut générer un corpus spécifiquement structuré pour l'évaluation de la détection du plagiat comme ceux de la campagne PAN (Potthast *et al.*, 2011b). Nous nous inspirons conceptuellement des travaux qui avaient été présentés lors de la tâche relative à la campagne PAN 2015 (Asghari *et al.*, 2015; Franco-Salvador *et al.*, 2015). Ce générateur est toujours en développement et ne fait pas partie du projet de cette thèse (il n'a donné lieu à aucune publication à ce jour), mais nous décidons de tout de même le rendre disponible sur GitHub [32]. Son interface s'inspire des travaux de Khoshnavataher *et al.* (2015) et la génération des passages offusqués s'inspire, quant à elle, des travaux de Mohtaj *et al.* (2015).

32. https://github.com/FerreroJeremy/Plagiarized-Corpus-Generator (consulté 22/06/2017 à 17h)

4.2 Évaluation de méthodes état de l'art de détection du plagiat translingue à l'aide de notre corpus

Les contributions de cette section ont fait l'objet de deux publications (Ferrero *et al.*, 2016, 2017b).

4.2.1 Protocole d'évaluation

Afin d'évaluer de façon rigoureuse les méthodes de détection du plagiat translingue sur notre corpus, nous avons mis au point un protocole d'évaluation fiable et reproductible.

Soit un corpus parallèle ou comparable noté C, tel que représenté sur la Figure 2.9 et constitué de N paires de documents, tel que pour chaque document $d_i \in D$ un document correspondant $d'_i \in D'$ existe, où i est un entier compris entre 1 et N, nous décidons de comparer la totalité des N documents disponibles dans D à M documents de D' [33]. Cela a pour avantage d'éviter de faire N^2 comparaisons et d'avoir ainsi des temps de calcul beaucoup trop longs dans les cas de corpus trop volumineux. Chaque document d_i est comparé à son document correspondant d'_i et à $M-1$ autres documents sélectionnés aléatoirement avec remise dans D'. Le même document d' peut donc être sélectionné plusieurs fois. M est fixé à 1 000 documents en adéquation avec l'état de l'art (Potthast *et al.*, 2011a).

À des fins de reproductibilité, nous constituons, avant notre évaluation, des masques qui pré-déterminent les M documents d' à comparer pour chaque document d. Ainsi, cela permet de comparer les différentes méthodes sur exactement le même jeu de données (les mêmes comparaisons). De plus, en cas de ré-évaluation de ces travaux, ces masques nous permettent d'obtenir à chaque fois exactement les mêmes comparaisons et donc sensiblement les mêmes résultats [34]. Un masque est donc un sous-corpus commun à plusieurs langues mais propre à un corpus et à une granularité donnés visant à minimiser le nombre de comparaisons.

Nous construisons donc des masques de taille $N \times M$, tels que représentés sur la Figure 4.1. Le premier document d' de chaque ligne (colonne rouge) est le document correspondant au $N^{\text{ième}}$ document d et ensuite viennent les $M-1$ documents tirés aléatoirement.

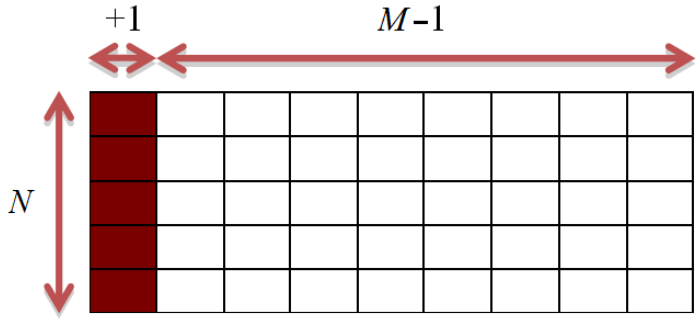


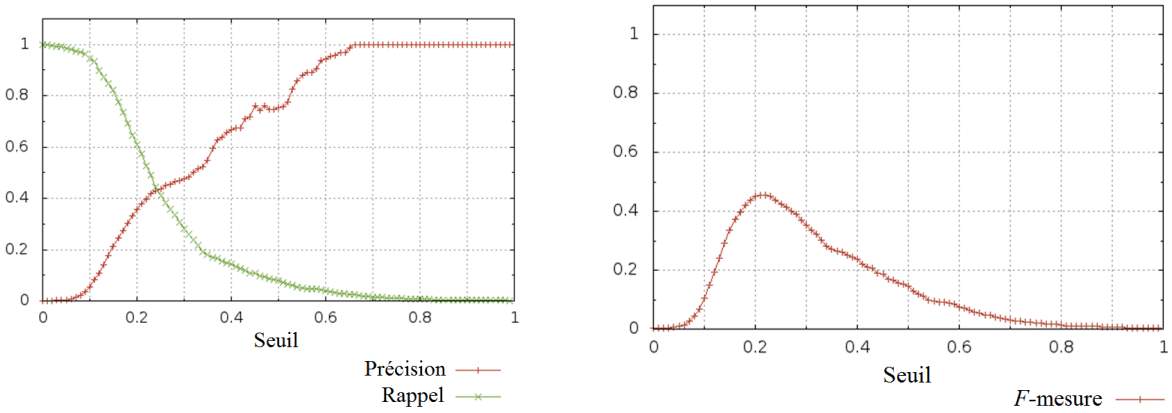
FIGURE 4.1 – Structure d'un masque.

Une fois ces masques constitués, nous lançons la méthode à évaluer sur les comparaisons figurant dans les masques. Un score de similarité est obtenu pour chaque comparaison, ce qui donne une matrice de similarité. Ensuite, nous effectuons un seuillage sur la matrice de similarité en fonction d'un seuil variant de 0 à 1. Si le score obtenu par la méthode est supérieur au seuil, alors on considère la comparaison comme un *Positif*, sinon elle est comptée comme un *Négatif*. Le premier élément de chaque ligne de la matrice (colonne rouge sur la Figure 4.1) est censé

33. En réalité, cela est vrai pour la granularité syntagmatique et phrastique. Pour la granularité documentaire, le nombre de N documents évalués est réduit à 2 000, pour des raisons de temps de traitement.

34. Les résultats risquent tout de même de ne pas être exactement identiques dû au caractère non déterministe de certaines méthodes.

être *Positif*, s'il est prédit comme tel, c'est un *Vrai Positif*, sinon c'est un *Faux Négatif*. Les comparaisons en dehors de cette colonne sont censées être *Négatif*, si elles sont prédites comme telles, alors ce sont des *Vrais Négatifs*, sinon ce sont des *Faux Positifs*. Grâce à cela, nous calculons la précision, le rappel et la F-mesure en fonction du seuil, suivant les formules de la section 3.1.2. Ces valeurs peuvent aussi être représentées sous forme de graphes. La Figure 4.2 représente les courbes de précision et de rappel (a) ainsi que la courbe de la F-mesure (b) des performances du modèle de taille de Pouliquen *et al.* (2003b) sur un sous-échantillon du corpus. Toujours en adéquation avec l'état de l'art (voir chapitre 3), nous considérons la F-mesure comme métrique de performance durant notre évaluation. Nous conservons donc seulement la plus haute F-mesure (atteinte en un certain seuil) pour définir la performance d'une méthode sur un masque. Dans l'exemple de la Figure 4.2, il s'agit de la F-mesure 0,43 atteinte au seuil 0,22.



(a) Précision et rappel en fonction du seuil

(b) F-mesure en fonction du seuil

FIGURE 4.2 – Courbes d'évaluation du modèle de taille de Pouliquen *et al.* (2003b) sur un sous-échantillon du corpus.

Peu importe le couple de langues, pour chaque sous-corpus, pour chacune des granularités, nous construisons dix masques en changeant les $M-1$ documents sélectionnées aléatoirement à chaque fois. Cela donne donc 180 masques (10 masques $\times$ 3 granularités $\times$ 6 sous-corpus). Les masques utilisés pour notre évaluation sont disponibles dans le répertoire GitHub [35] de notre corpus.

Nous appliquons ensuite chaque méthode sur chaque couple de langues disponible pour chaque sous-corpus à chaque granularité. Un masque correspond à un sous-corpus précis à une granularité précise mais peut être utilisé pour toutes les paires de langues disponibles sur ce sous-corpus, nous appliquons donc en fait chaque méthode sur chaque couple de langues sur les dix masques ayant trait à un sous-corpus. La mesure finale utilisée pour évaluer la performance d'une méthode est la moyenne des F-mesures obtenues sur les dix masques testés dans une configuration particulière (un couple de langues ; un sous-corpus ; une granularité). Cette moyenne est reportée dans un tableau bilan et nous profitons ainsi des dix mesures pour calculer un intervalle de confiance autour de cette moyenne.

4.2.2 Méthodes évaluées

Nous nous sommes exclusivement concentrés, lors de notre évaluation, sur les méthodes de chaque approche qui montraient les meilleures performances dans la littérature au commencement de cette thèse. Depuis, la méthode CL-KGA (voir section 2.2.2.4) obtient de meilleurs résultats que les méthode CL-ASA et CL-ESA, notamment. L'approche T+MA (voir section 2.2.5) est en

35. https://github.com/FerreroJeremy/Cross-Language-Dataset/tree/master/masks (consulté le 27/06/2017 à 11h)

perpétuelle évolution, notamment avec l'arrivée des *word embeddings*. Et de façon plus générale, l'ensemble de l'état de l'art des méthodes de détection de similarités textuelles sémantiques translingues ne cesse d'évoluer et de permettre à la détection du plagiat translingue d'évoluer également (voir section 2.2.6).

Les approches que nous décidons d'évaluer ici sont expliquées dans les articles respectifs mais leur code source et les ressources employées pour les faire fonctionner ne sont pas partagés. Cela nous oblige donc à les ré-implémenter pour pouvoir les évaluer. Nous expliquons donc, dans cette section, les spécificités de nos implémentations par rapport aux approches standards présentées en section 2.2.

Vecteurs translingues de n-grammes de caractères (*Cross-Language Character n-Gram, CL-CnG*)

Nous utilisons l'implémentation CL-C3G de Potthast *et al.* (2011a), telle que décrite dans la section 2.2.1.1, mise à part le fait que nous conservons les espaces et que nous remplaçons le comptage brut de la fréquence d'un terme (*raw count*) par sa fréquence normalisée – sa probabilité d'apparition au sein du texte (*standard term frequency*). Nous réduisons dans un premier temps l'alphabet des textes sur un espace $\sum = \{a-z, 0-9, \quad\}$, c'est-à-dire que l'on conserve les caractères alphanumériques, mais aussi les espaces. Les textes sont ensuite transformés sous forme de vecteurs de 3-grammes (séquences de trois caractères contiguës) dont les poids sont définis par le modèle *tf.idf* (*Term Frequency Inverse Document Frequency*) (Salton *et* Buckley, 1988) avant d'être comparés à l'aide d'une mesure de similarité cosinus (Salton, 1989). La formule de la fréquence d'un terme (*tf*) est maintenant définie telle que dans la Formule 4.1 tandis que la formule de la fréquence inverse de document (*idf*) est toujours définie telle que dans la Formule 2.2.

$$tf(t,d) = \frac{f_{t,d}}{\mid d \mid} \tag{4.1}$$

où $f_{t,d}$ est définie comme la fonction qui retourne le nombre d'occurrences du terme t dans le document d.

Nous calculons directement les fréquences inverses de document sur les corpus d'évaluation considérés.

Similarité translingue basée sur des thésaurus (*Cross-Language Conceptual Thesaurus-based Similarity, CL-CTS*)

Nous utilisons l'idée de Pataki (2012) telle que décrite en section 2.2.2.3, qui construit un sac de mots contenant, pour le texte source, toutes les traductions de chaque mot de ce texte, et pour le texte cible, tous les synonymes de chaque mot de ce texte. Nous utilisons *DBnary* (Sérasset, 2015) pour obtenir les traductions et les synonymes des mots. Le sac de mots d'un texte est la fusion des sacs de mots de chacun des mots de ce texte. Pour calculer la similarité entre deux textes, nous utilisons l'indice de Jaccard (Jaccard, 1912) avec une recherche de correspondances approximatives (*fuzzy matching*) (Baeza-Yates *et* Navarro, 1996) entre les deux sacs de mots représentant ces textes. Au cours de la recherche de correspondances, nous utilisons la distance de Damerau–Levenshtein (Damerau, 1964; Levenshtein, 1966) pour déterminer les approximations. Nous considérons un mot, qui possède une distance de Damerau–Levenshtein inférieure ou égale à 0,20 avec un autre mot, comme identique à cet autre mot. C'est-à-dire que le nombre minimal de caractères qu'il faut insérer, supprimer, substituer ou transposer pour transformer l'un des mots en l'autre doit couvrir moins de 20% du mot le plus long. Ce pourcentage a été déterminé empiriquement après des tests sur un corpus échantillon de 2 000 comparaisons.

Similarité translingue basée sur l'alignement (*Cross-Language Alignment-based Similarity Analysis, CL-ASA*)

Nous utilisons l'implémentation de Pinto *et al.* (2009) (voir section 2.2.3.2). Ce modèle a pour but de déterminer combien un texte est potentiellement la traduction d'un autre texte dans une autre langue en se basant sur les probabilités de traductions issues depuis un dictionnaire bilingue créé à partir de corpus parallèles. Notre dictionnaire de probabilité de traductions de mots est créé en appliquant le modèle IBM-1 (Brown *et al.*, 1993) sur la concaténation des corpus parallèles TED [36] (Cettolo *et al.*, 2012) et News [37].

Analyse sémantique explicite translingue (*Cross-Language Explicit Semantic Analysis, CL-ESA*)

Nous utilisons la méthode CL-ESA telle que décrite dans la section 2.2.4.2, qui représente un document par un vecteur, basé sur ses similitudes avec le vocabulaire d'un corpus comparable. Notre implémentation utilise des textes de Wikipédia ne faisant pas partie de notre corpus de test afin de construire les représentations vectorielles.

Modèles à base de traduction suivie d'une analyse monolingue (*Translation + Monolingual Analysis, T+MA*)

Nous utilisons l'approche de Muhr *et al.* (2010) (voir section 2.2.5), qui consiste à remplacer chaque mot d'un texte par ses traductions les plus probables, représentant ainsi le texte sous forme d'un sac de mots. La similarité entre deux textes est ensuite calculée avec une intersection stricte entre les deux sacs de mots représentant ces textes. Nous utilisons *DBnary* (Sérasset, 2015) pour obtenir les traductions et ainsi constituer les sacs de mots.

4.2.3 Résultats et discussions

Dans un premier temps, nous nous concentrons sur les performances des méthodes à travers les diverses paires de langues et tailles de textes. Ensuite, nous effectuons une analyse plus poussée sur une paire de langue particulière (anglais-français) afin de mettre en avant des phénomènes plus précis.

4.2.3.1 À travers les paires de langues

Le Tableau 4.7 rassemble les performances globales des méthodes pour la granularité syntagmatique en fonction des paires de langues. Le Tableau 4.8 fait de même pour la granularité phrastique. Ce que nous appelons la performance globale d'une méthode est en réalité la moyenne pondérée des performances de cette méthode sur les sous-corpus disponibles pour une granularité et pour une paire de langues. Cette moyenne est pondérée en fonction du nombre de comparaisons effectuées par sous-corpus, sachant que la performance obtenue pour un sous-corpus est déjà la moyenne des dix F-mesures obtenues sur ce sous-corpus, c'est -à-dire sur les dix masques évalués relatifs à ce sous-corpus. Au cours de cette étude, chaque langue est notée par son code à deux lettres sous la norme ISO 639-1 (dit *alpha-2*).

On remarque que les méthodes CL-C3G et CL-ESA obtiennent les mêmes résultats pour un couple de langues donné, peu importe la direction de la traduction (même résultat pour CL-C3G et CL-ESA pour la paire en→fr et pour la paire fr→en dans le Tableau 4.8, par exemple). Ces méthodes obtiennent donc les mêmes résultats si l'on inverse la langue source et la langue cible. Ce phénomène est dû à la propriété symétrique de leur mode de fonctionnement (voir section 2.2.1.1 et section 2.2.4.2).

36. https://wit3.fbk.eu/ (consulté le 29/06/2017 à 10h)
37. http://www.statmt.org/wmt13/translation-task.html#download (consulté le 29/06/2017 à 10h)

Méthode	en→fr	fr→en	en→es	es→en	es→fr	fr→es
CL-C3G	0,5071	0,5071	0,4375	0,4375	0,4795	0,4795
CL-CTS	0,4250	0,4116	0,3780	0,3881	0,4203	0,4169
CL-ASA	0,4738	0,4252	0,4083	0,3941	0,3736	0,3540
CL-ESA	0,1499	0,1499	0,1476	0,1476	0,1520	0,1520
T+MA	0,3730	0,3634	0,3177	0,3279	0,3158	0,3140

Tableau 4.7 – Performances moyennes, pondérées par chaque sous-corpus, des méthodes pour la granularité syntagmatique en fonction des paires de langues.

Méthode	en→fr	fr→en	en→es	es→en	es→fr	fr→es
CL-C3G	0,4931	0,4931	0,3819	0,3819	0,4577	0,4577
CL-CTS	0,4734	0,4633	0,3171	0,3204	0,4645	0,4575
CL-ASA	0,3576	0,3523	0,2694	0,2531	0,3098	0,2843
CL-ESA	0,1430	0,1430	0,1337	0,1337	0,1383	0,1383
T+MA	0,3760	0,3692	0,3505	0,3526	0,3673	0,3525

Tableau 4.8 – Performances moyennes, pondérées par chaque sous-corpus, des méthodes pour la granularité phrastique en fonction des paires de langues.

Une autre remarque que nous pouvons faire est que les méthodes se comportent de façon similaire à travers les différentes paires de langues, c'est-à-dire qu'une méthode performante sur une paire de langue, le sera de façon générale tout autant sur les autres paires de langues. Cela est confirmé par le calcul de la corrélation de Pearson (Galton, 1886; Pearson, 1895) entre les performances des méthodes sur les différentes paires de langues. Ces corrélations sont calculées à partir du Tableau 4.7 et reportées dans le Tableau 4.9 pour la granularité syntagmatique et calculées à partir du Tableau 4.8 et reportées dans le Tableau 4.10 pour la granularité phrastique.

en→fr	fr→en	en→es	es→en	es→fr	fr→es	Moyenne	Paire de langues
1,000	0,991	0,998	0,995	0,957	0,940	0,980	en→fr
	1,000	0,990	0,994	0,980	0,971	0,987	fr→en
		1,000	0,996	0,967	0,949	0,983	en→es
			1,000	0,978	0,965	0,988	es→en
				1,000	0,998	0,980	es→fr
					1,000	0,970	fr→es

Tableau 4.9 – Corrélation de Pearson des performances moyennes globales des méthodes à la granularité syntagmatique entre les différentes paires de langues.

en→fr	fr→en	en→es	es→en	es→fr	fr→es	Moyenne	Paire de langues
1,000	1,000	0,929	0,922	0,991	0,982	0,971	en→fr
	1,000	0,931	0,924	0,989	0,981	0,971	fr→en
		1,000	0,997	0,925	0,913	0,949	en→es
			1,000	0,928	0,922	0,949	es→en
				1,000	0,997	0,971	es→fr
					1,000	0,966	fr→es

Tableau 4.10 – Corrélation de Pearson des performances moyennes globales des méthodes à la granularité phrastique entre les différentes paires de langues.

Nous constatons qu'en dépit des variations des langues sources et cibles, les performances restent extrêmement corrélées. À la granularité syntagmatique, par exemple, la corrélation de Pearson minimale observée est de 0,940 pour les paires en→fr *vs.* fr→es et la corrélation maximale observée est de 0,998 pour en→fr *vs.* en→es et es→fr *vs.* fr→es (voir Tableau 4.9). Les corrélations sont du même ordre de grandeur à la granularité phrastique. La corrélation de Pearson minimale est de 0,913 obtenue pour en→es *vs.* fr→es et la maximale est de 0,997 pour en→es *vs.* es→en et es→fr *vs.* fr→es (voir Tableau 4.10). En moyenne, la paire de langues en→fr est corrélée à hauteur de 0,975 avec les autres paires de langues (0,980 à la granularité syntagmatique et 0,971 à la granularité phrastique). Ces résultats sont intéressants car certaines méthodes, bien que dépendantes de la disponibilité de ressources lexicales dont la qualité est hétérogène en fonction des différentes langues, présentent des corrélations laissant supposer qu'elles s'adaptent plutôt bien aux différentes langues malgré cette hétérogénéité. Malgré le fait que les résultats de la plupart des méthodes sont moins bons lorsque que les paires de langues considérées comportent de l'espagnol, comme on peut l'observer sur la méthode CL-CTS qui utilise *DBnary* (qui dispose de moins de traductions espagnoles qu'anglaises ou françaises[38]) par exemple, les corrélations observées restent tout de même très fortes. Ces corrélations suggèrent donc qu'il est possible de paramétrer une méthode sur une certaine langue et de l'appliquer sur d'autres langues si nécessaire, tant qu'il y a suffisamment de ressources lexicales disponibles pour traiter ces langues correctement.

en↔fr	es↔fr
en↔es	
CL-C3G	CL-C3G
CL-ASA	CL-CTS
CL-CTS	CL-ASA

(a) Syntagmes nominaux

en↔fr	en↔es	es→fr
fr→es		
CL-C3G	CL-C3G	CL-CTS
CL-CTS	T+MA	CL-C3G
T+MA	CL-CTS	T+MA

(b) Phrases

Tableau 4.11 – *Top 3* des méthodes en fonction des langues sources et cibles.

Le Tableau 4.11 synthétise le *Top 3* des méthodes pour chaque paire de langues étudiée, en fonction des valeurs du Tableau 4.7 et du Tableau 4.8. Peu importe la langue source ou cible ou encore la granularité, CL-C3G est généralement la méthode la plus performante. Viennent ensuite CL-ASA, CL-CTS et T+MA qui sont toutes trois relativement proches et tout aussi efficaces, même si leur comportement dépend un peu plus de la granularité et des paires de langues considérées. De façon générale, CL-ASA est plus performante sur des textes courts (granularité syntagmatique), suivie alors par CL-CTS et T+MA. Alors qu'au contraire, CL-CTS et T+MA semblent plus efficaces sur la granularité des phrases. Une explication à cela peut être que T+MA dépend de la qualité de l'outil de traduction automatique utilisé, ce qui peut donner des performances pauvres sur des syntagmes isolés, quand un texte court est plus approprié au mode de fonctionnement d'approches comme CL-ASA (multiplication de probabilités de traduction). Toutefois, en dépit des différences de rangs, quand CL-C3G n'est pas la meilleure méthode, l'écart en termes de performances entre cette dernière et la méthode la devançant est très faible. Par exemple, nous pouvons voir que quand CL-CTS est plus performante que CL-C3G, phénomène se produisant dans la colonne es→fr du Tableau 4.8 ainsi que dans le Tableau 4.11 (b), la différence de performance est de seulement 0,0068. C'est aussi vrai dans le cas inverse. Par exemple, lorsque CL-C3G fait mieux que CL-CTS pour la paire de langue fr→es à la granularité phrastique (colonne fr→es dans le Tableau 4.8), l'écart de performance est négligeable avec seulement 0,0002.

Ces résultats sont comparables à ceux que l'on pouvait observer dans la littérature (voir section 2.2.7).

38. http://kaiko.getalp.org/about-dbnary/dataset/ (consulté le 13/07/2017 à 12h)

Le Tableau 4.12 montre les corrélations de Pearson des performances globales des méthodes sur tous les sous-corpus confondus, entre la granularité syntagmatique et phrastique, en fonction des paires de langues (corrélations calculées entre la colonne en→fr du Tableau 4.7 et la colonne en→fr du Tableau 4.8, et ainsi de suite pour toutes les paires de langues). La Figure 4.3 illustre la façon dont ces corrélations sont calculées sur un couple de langues, en considérant les performances de trois méthodes sur trois sous-corpus à deux granularités. Nous pouvons voir, une fois de plus, une forte corrélation des performances entre la granularité des syntagmes et celle des phrases (en moyenne 0,899 avec par exemple 0,907 pour la paire en→fr). Cela montre à nouveau que toutes les méthodes se comportent de façon similaire à la granularité syntagmatique et phrastique en termes de performance sur corpus, indépendamment des langues sur lesquelles elles sont utilisées.

Paire de langues	Corrélation
en→fr	0,907
fr→en	0,946
en→es	0,833
es→en	0,838
es→fr	0,932
fr→es	0,939

Tableau 4.12 – Corrélations de Pearson des résultats des méthodes sur tous les sous-corpus, entre la granularité syntagmatique et phrastique, en fonction des paires de langues. Calculées depuis le Tableau 4.7 et le Tableau 4.8.

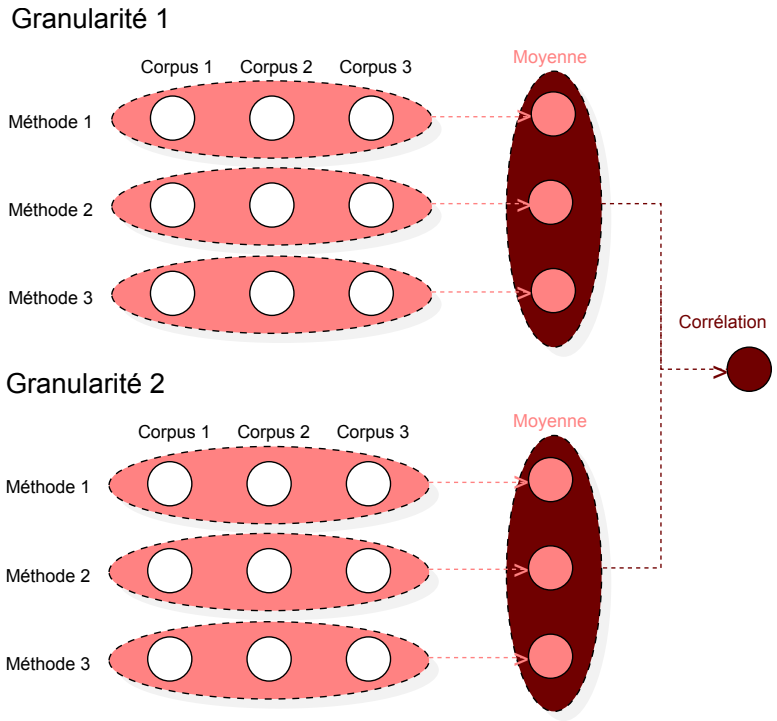


FIGURE 4.3 – Méthode de calcul permettant de mesurer la corrélation des performances moyennes tout corpus confondus entre deux granularités.

D'autre part, nous pouvons voir dans le Tableau 4.13 que si nous calculons les corrélations de Pearson pour les performances de chaque méthode sur tous les corpus et toutes les paires de langues confondus, entre la granularité syntagmatique et phrastique (corrélations calculées entre la colonne CL-C3G du Tableau 4.7 et la colonne CL-C3G du Tableau 4.8, et ainsi de suite pour

toutes les méthodes), certaines méthodes montrent une corrélation plus faible et donc un comportement différent sur les deux granularités. Cela montre cette fois-ci, que, indépendamment des langues sur lesquelles les méthodes sont utilisées, elles peuvent se comporter de façon différente en fonction de la taille des textes traités. Par exemple, c'est le cas pour CL-ASA qui montre une corrélation de (seulement) 0,649 et semble être significativement meilleure sur la granularité syntagmatique (phénomène visible en comparant le Tableau 4.7 et le Tableau 4.8). En revanche, la faible corrélation entre les performances de la méthode CL-ESA aux deux granularités est à nuancer, comme le montre le Tableau 4.14.

Méthode	Corrélation
CL-C3G	0,996
CL-CTS	0,970
CL-ASA	0,649
CL-ESA	0,515
T+MA	0,780

Tableau 4.13 – Corrélations de Pearson des résultats des méthodes sur tous les sous-corpus sur toutes les paires de langues, entre la granularité syntagmatique et phrastique. Calculées depuis le Tableau 4.7 et le Tableau 4.8.

Méthode	en→fr	fr→en	en→es	es→en	es→fr	fr→es	Moyenne
CL-C3G	0,856	0,856	0,947	0,947	1,000	1,000	0,934
CL-CTS	0,757	0,735	0,943	0,901	0,862	0,871	0,845
CL-ASA	0,834	0,683	0,710	0,705	0,587	0,611	0,579
CL-ESA	0,975	0,975	0,991	0,991	1,000	1,000	0,988
T+MA	0,998	0,997	0,956	0,946	0,938	0,927	0,960
Moyenne	0,884	0,849	0,909	0,898	0,877	0,750	

Tableau 4.14 – Corrélations de Pearson des performances sur tous les sous-corpus de chaque méthode prises séparément entre la granularité syntagmatique et phrastique. La dernière ligne et la dernière colonne recensent les moyennes de ces corrélations. La Figure 4.4 illustre la façon dont ces moyennes ont été calculées, de façon opposée à la méthode décrite dans la Figure 4.3.

Le Tableau 4.14 montre les corrélations de Pearson des performances sur chaque sous-corpus de chaque méthode prise séparément entre la granularité syntagmatique et phrastique. Les moyennes de ces corrélations (dernière ligne et dernière colonne du Tableau 4.14) nuancent les valeurs données précédemment dans le Tableau 4.12 et le Tableau 4.13. En effet, nous constatons par exemple ici que, contrairement à ce qui paraissait dans le Tableau 4.13, la méthode CL-ESA se comporte bien, en moyenne, de façon identique sur tous les sous-corpus, peu importe la granularité ou les langues source et cible considérées.

Contrairement aux corrélations de moyennes du Tableau 4.12 et du Tableau 4.13, calculées comme le montre la Figure 4.3, qui caractérisaient les scores moyens sur tous corpus confondus entre les deux granularités, les moyennes de corrélations du Tableau 4.14, calculées comme le montre la Figure 4.4, conservent l'effet corpus et aident donc à vérifier si chaque méthode prise séparément se comporte de la même façon sur chaque sous-corpus sur les deux granularités.

En conclusion, nous pouvons dire que les méthodes évaluées ici, se comportent de façon pratiquement similaire peu importe les langues des textes comparés, tant qu'il y a suffisamment de ressources lexicales disponibles pour traiter ces langues, mais peuvent se comporter différemment en fonction de caractéristiques telles que la taille (granularité) des textes considérés.

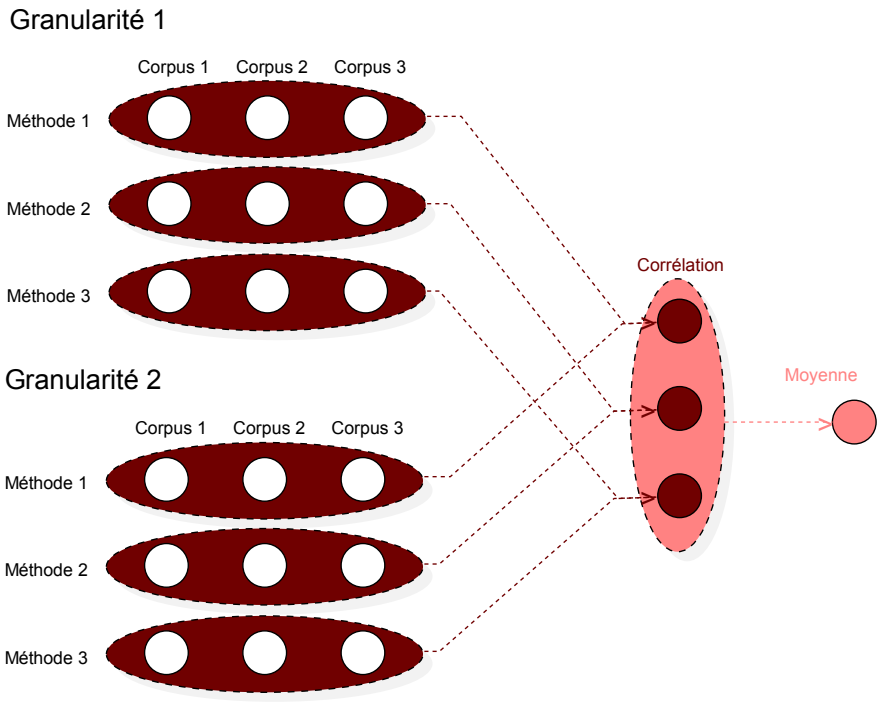


FIGURE 4.4 – Méthode de calcul permettant de mesurer la corrélation des performances prises séparément sur chaque corpus entre deux granularités. L'effet corpus est conservé et la corrélation est calculée entre les deux granularités mais séparément pour chaque méthode.

4.2.3.2 Analyse détaillée pour la paire de langue anglais-français

Nous avons montré que les méthodes se comportent de façon similaire à travers les différentes paires de langues (comportements fortement similaires) et les différentes granularités (comportements moins similaires mais phénomène tout aussi notable). Pour cette raison, nous nous proposons maintenant d'effectuer une analyse détaillée pour les différents corpus, sur seulement la paire de langues en→fr aux trois granularités (syntagmatique, phrastique et documentaire).

Le Tableau 4.15, le Tableau 4.16 et le Tableau 4.17 montrent respectivement les performances des méthodes sur les sous-corpus en→fr à la granularité syntagmatique, phrastique et documentaire. Les résultats pour toutes les autres paires de langues dont dispose notre corpus d'évaluation peuvent être trouvés en Annexe C [39]. Pour rappel, chacune des valeurs de ce tableau est la moyenne, pondérée en fonction du nombre de comparaisons faites selon le sous-corpus, des dix F-mesures obtenues (sur les dix masques) par une méthode, sur un sous-corpus, à une granularité et un couple de langues particulier.

Méthode	Wikipédia (%)	TALN-ACL (%)	JRC-Acquis (%)	Webis-CL-10 (%)	Europarl (%)	Moyenne (%)
CL-C3G	62,91 ±0,815	40,90 ±0,500	36,63 ±0,826	80,30 ±0,703	53,29 ±0,583	50,71 ±0,655
CL-CTS	58,00 ±0,519	33,71 ±0,382	29,87 ±0,815	67,51 ±1,050	44,95 ±1,157	42,50 ±1,053
CL-ASA	23,33 ±0,724	23,39 ±0,432	33,14 ±0,936	26,49 ±1,205	55,50 ±0,681	47,38 ±0,781
CL-ESA	64,89 ±0,664	23,78 ±0,613	14,03 ±0,997	23,14 ±0,777	14,19 ±0,590	14,99 ±0,709
T+MA	58,22 ±0,756	39,13 ±0,551	28,61 ±0,597	73,14 ±0,666	36,95 ±1,502	37,30 ±1,200

Tableau 4.15 – Moyenne et intervalle de confiance des F-mesures des méthodes appliquées sur les sous-corpus en→fr à la granularité syntagme – sur 10 lancers.

39. On y trouve également les performances des méthodes sur les corpus es→en de la tâche STS de SemEval-2016 (Agirre *et al.*, 2016) et SemEval-2017 (Cer *et al.*, 2017) (voir section 3.6).

Méthode	Wikipédia (%)	TALN-ACL (%)	JRC-Acquis (%)	Webis-CL-10 (%)	Europarl (%)	Moyenne (%)
CL-C3G	48,25 ±0,349	48,08 ±0,538	36,68 ±0,693	61,10 ±0,581	52,72 ±0,866	49,31 ±0,798
CL-CTS	46,68 ±0,437	38,67 ±0,552	28,21 ±0,612	50,82 ±1,034	53,21 ±0,601	47,34 ±0,632
CL-ASA	27,63 ±0,330	27,25 ±0,341	35,17 ±0,644	25,53 ±0,795	36,55 ±1,139	35,76 ±0,978
CL-ESA	51,14 ±0,875	14,25 ±0,334	14,44 ±0,341	13,93 ±0,714	13,91 ±0,618	14,30 ±0,551
T+MA	50,57 ±0,888	37,79 ±0,364	32,36 ±0,369	61,94 ±0,756	37,92 ±0,552	37,60 ±0,518

Tableau 4.16 – Moyenne et intervalle de confiance des F-mesures des méthodes appliquées sur les sous-corpus en→fr à la granularité phrase – sur 10 lancers.

Méthode	Wikipédia (%)	TALN-ACL (%)	JRC-Acquis (%)	Webis-CL-10 (%)	Europarl (%)	Moyenne (%)
CL-C3G	51,58 ±1,942	48,67 ±1,662	37,91 ±1,096	57,55 ±1,103	53,86 ±1,330	49,91 ±1,427
CL-CTS	48,45 ±1,867	38,33 ±1,494	27,16 ±0,699	50,60 ±1,771	55,19 ±1,376	43,95 ±1,441
CL-ASA	33,87 ±1,181	26,42 ±1,400	34,08 ±0,944	34,43 ±1,813	36,59 ±1,236	33,08 ±1,315
CL-ESA	53,44 ±1,516	18,03 ±1,261	12,93 ±1,074	13,67 ±0,995	11,73 ±0,963	21,96 ±1,162
T+MA	55,82 ±2,344	34,84 ±1,049	27,27 ±0,771	47,49 ±2,130	32,80 ±1,340	39,64 ±1,527

Tableau 4.17 – Moyenne et intervalle de confiance des F-mesures des méthodes appliquées sur les sous-corpus en→fr à la granularité document – sur 10 lancers.

Comme mentionné plus tôt, CL-C3G est en général la méthode la plus efficace. CL-CTS et T+MA sont assez efficaces et polyvalentes également. CL-ESA semble montrer de meilleurs résultats sur un corpus comparable comme Wikipédia, alors qu'au contraire, CL-ASA obtient de meilleurs résultats sur les corpus parallèles tels que les JRC-acquis ou la collection Europarl. On peut noter que les résultats de ces méthodes sont plutôt bien corrélés entre certains sous-corpus. Par exemple, la corrélation de Pearson des performances des méthodes entre le sous-corpus TALN-ACL et le sous-corpus Webis-CL-10 est de 0,982 à la granularité syntagmatique, 0,937 à la granularité phrastique et 0,960 à la granularité documentaire. Cela signifie qu'une méthode peut être paramétrée sur le corpus Webis-CL-10, qui représente des avis de produits Amazon, et ensuite être appliquée efficacement sur le corpus TALN-ACL, constitué d'articles de conférences sur le traitement automatique du langage.

Il est aussi intéressant de noter que certaines irrégularités sont présentes dans les résultats aux granularités syntagmatique et phrastique, comme le fait que, par exemple, la méthode CL-ASA voit ses performances augmenter sur le sous-corpus Wikipédia en passant de la granularité documentaire à la granularité syntagmatique. Ceci en raison du fait que pour construire les granularités syntagmatique et phrastique de notre corpus, nous avons effectué, comme expliqué en section 4.1.3.2, un ré-alignement des unités textuelles ce qui a très probablement transformé certaines unités comparables en unités parallèles. Il est également important de noter que les performances des méthodes utilisant des connaissances externes comme des ontologies (CL-CTS), des dictionnaires (CL-ASA) ou des corpus (CL-ESA), sont extrêmement dépendantes de la qualité et l'exhaustivité de ces connaissances. On note aussi que les intervalles de confiance sont plus grands sur la granularité documentaire (avec une moyenne de 1,374% contre 0,696% pour la granularité phrastique et 0,880% pour la granularité syntagmatique). Cela sans doute à cause du fait que durant l'évaluation de cette granularité (documentaire), pour des raisons de temps de traitement, le nombre N de documents évalués n'était en fait pas égal à l'intégralité du sous-corpus mais à 2 000 documents (voir section 4.2.1).

Généralement, toutes les méthodes voient leurs performances décroitre au fur et à mesure que la granularité des textes comparés croît. Cependant, nous remarquons que certaines méthodes voient leurs performances stagner entre les phrases et les documents. C'est notamment le cas pour les méthodes CL-CTS ou CL-ESA. Aussi, les résultats tendent à être meilleurs sur les corpus Wikipédia et Webis-CL-10, et légèrement moins bons sur la collection Europarl, ce qui coïncide avec les corpus dont le ratio d'entités nommées est le plus important (voir le Tableau 4.4 et le Tableau 4.6).

Une fois de plus, nous constatons une forte corrélation de Pearson entre les résultats des méthodes sur les trois granularités (en moyenne 0,938), à l'exception des performances de la méthode CL-CTS entre les granularités syntagmatique et phrastique (0,757) et les performances de la méthode CL-ASA entre les granularités phrastique et documentaire (0,493).

4.2.3.3 Étude de la complémentarité des méthodes

Au delà de la capacité des méthodes à correctement prédire si deux textes signifient ou non la même chose (leur justesse de classification), une caractéristique particulièrement intéressante à surveiller est leur capacité à correctement séparer les textes qui signifient ou non la même chose, dans le but de minimiser le doute sur la justesse de classification. Cette capacité caractérise la facilité avec laquelle ils identifient si deux textes signifient ou non la même chose. Pour vérifier cette caractéristique, nous conduisons une seconde expérience avec un nouveau corpus et un nouveau protocole.

Nous constituons donc un nouveau jeu de données en collectant, depuis notre corpus d'origine, 200 paires de chacun des cinq sous-corpus en→fr à la granularité phrastique. Nous comparons 1 000 phrases anglaises (200 phrases × 5 sous-corpus) à leur phrase française relative ainsi qu'à une seule autre phrase française non relative tirée au hasard dans le même sous-corpus que la phrase française relative. Chaque phrase anglaise doit donc strictement donner lieu à une correspondance (un positif) et une non correspondance (un négatif), ce qui donne 2 000 comparaisons avec exactement 1 000 positifs et 1 000 négatifs. Nous répétons cette expérience dix fois pour chaque méthode, en changeant les phrases sélectionnées à chaque fois. Les résultats de cette expérience sont reportés dans le Tableau 4.18, qui montre, pour le meilleur seuil, la moyenne sur les dix lancés de la précision, du rappel et de la F-mesure des méthodes évaluées.

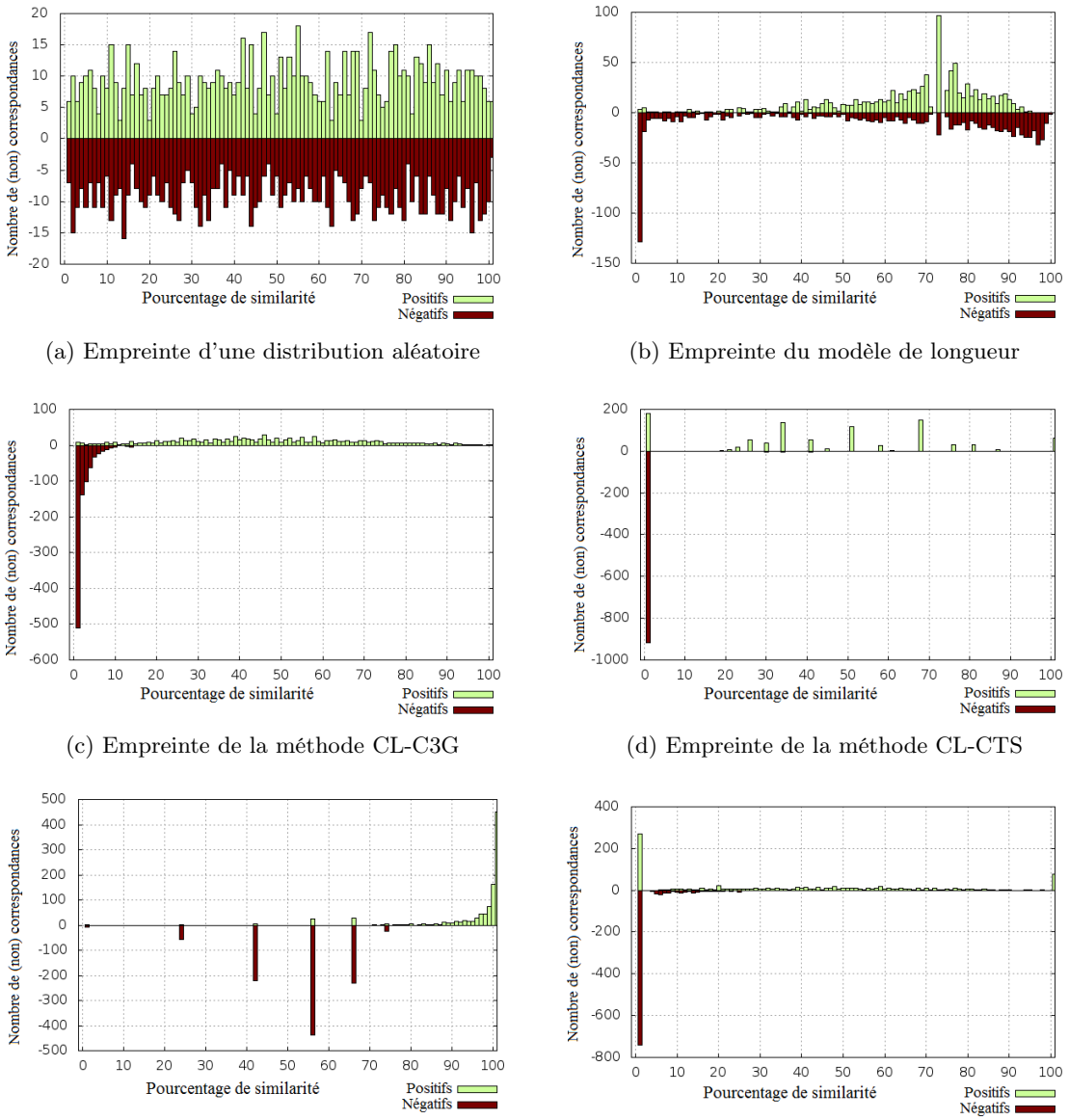
Méthode	Seuil	Précision	Rappel	F-mesure
Distribution aléatoire	0,003	0,501	0,999	0,668
Modèle de longueur	0,203	0,566	0,970	0,714
CL-C3G	0,087	0,972	0,953	0,962
CL-CTS	0,010	0,986	0,808	0,888
CL-ASA	0,762	0,937	0,772	0,847
T+MA	0,157	0,928	0,646	0,762

Tableau 4.18 – La moyenne sur les dix lancés de la précision, du rappel et de la F-mesure des méthodes évaluées, atteint pour le meilleur seuil.

Nous décidons de reporter également les résultats de l'expérience dans la Figure 4.5 qui illustre la distribution des méthodes évaluées pour 1 000 positifs et 1 000 négatifs. Sur cette dernière, l'axe des abscisses représente le score de similarité (en pourcentage) obtenu par les méthodes, tandis que l'axe des ordonnées représente le nombre de comparaisons qui ont été trouvées positives (partie supérieure) ou négatives (partie inférieure) pour un score donné. En vert, dans la partie supérieure des graphiques, les positifs (phrases qu'il fallait trouver comme similaires) et en rouge, dans la partie inférieure des graphiques, les négatifs (phrases qu'il ne fallait pas trouver comme similaires). Plus les positifs se trouvent dans la partie droite des graphes et plus les négatifs se trouvent dans la partie gauche des graphes, plus il est simple d'identifier un seuil optimal, ce qui donnera lieu à une meilleure classification.

Les histogrammes de la Figure 4.5 mettent en lumière le fait que chaque méthode possède sa propre empreinte. Même si deux méthodes semblent équivalentes de prime abord en termes de performances, leur capacité de classification et donc leur empreinte peut être très différente et souligne des modes de fonctionnement bien distincts. Par exemple, nous pouvons voir que la méthode CL-C3G a une distribution concentrée des négatifs et une distribution assez diffuse et étendue des positifs (Figure 4.5 (c)), alors que c'est plutôt l'inverse pour la méthode CL-ASA (Figure 4.5 (e)). Le Tableau 4.18 confirme ce phénomène par le fait que le seuil de décision est

plus élevé pour la méthode CL-ASA (0,762) comparé à celui des autres méthodes (autour de 0,1). Cela signifie que CL-ASA discrimine plus correctement les positifs que les négatifs, quand il semble que ce soit l'inverse pour les autres méthodes. Nous faisons donc l'hypothèse que certaines méthodes sont complémentaires. Ces observations suggèrent qu'une fusion entre ces méthodes, notamment un arbre de décision, devrait conduire à des résultats plutôt prometteurs.



(a) Empreinte d'une distribution aléatoire

(b) Empreinte du modèle de longueur

(c) Empreinte de la méthode CL-C3G

(d) Empreinte de la méthode CL-CTS

(e) Empreinte de la méthode CL-ASA

(f) Empreinte de la méthode T+MA

FIGURE 4.5 – Histogramme des distributions des méthodes de l'état de l'art pour 1 000 positifs et 1 000 négatifs. L'axe des abscisses représente le score de similarité (en pourcentage) obtenu par les méthodes, tandis que l'axe des ordonnées représente le nombre de comparaisons qui ont été trouvées positives (partie supérieure) ou négatives (partie inférieure) pour un score. En vert, dans la partie supérieure des graphiques, les positifs (phrases qu'il fallait trouver comme similaires) et en rouge, dans la partie inférieure des graphiques, les négatifs (phrases qu'il ne fallait pas trouver comme similaires).

De plus, nous pouvons voir que la façon d'opérer de certaines méthodes peut expliquer leur empreinte. Par exemple, CL-CTS semble fonctionner par palier (Figure 4.5 (d)). Cela peut être expliqué par la façon dont sa mesure de similarité est calculée. C'est en fait un indice de Jaccard (Jaccard, 1912), c'est-à-dire un ratio opéré sur des correspondances de concepts, ce qui conduit

principalement à des pourcentages comme 0%, 25%, 33% *etc.* Nous pouvons aussi distinguer quel histogramme correspond à une distribution aléatoire par exemple (Figure 4.5 (a)).

4.3 Conclusion

Pour conclure, nous avons construit un corpus multilingue, multi-genre et multi-granularité pour permettre une évaluation rigoureuse des méthodes de détection du plagiat translingue. Nous avons également mis au point un protocole d'évaluation reproductible. Ce corpus peut s'avérer utile pour de futures tâches d'évaluation de la détection de similarité textuelle sémantique translingue, c'est pour cela que nous le rendons public dans un répertoire GitHub [40]. Tous les scripts nécessaires à sa (re)construction ainsi que les masques pour reproduire l'évaluation sont également disponibles dans ce répertoire.

Nous avons ensuite conduit une étude des performances de certaines méthodes de détection du plagiat translingue sur notre corpus. Nos résultats confirment que les différentes méthodes de l'état de l'art se comportent différemment en fonction des caractéristiques des textes comparés mais aussi que la taille de ces textes impacte leurs performances. Nos résultats montrent un comportement corrélé des méthodes à travers les différentes paires de langues testées. Cela signifie que quand une méthode est plus efficace qu'une autre sur un corpus suffisamment large, elle est en règle générale plus efficace également dans les autres cas. Cela signifie aussi que si une méthode est efficace sur une paire de langues particulière, elle sera tout autant efficace sur une autre paire de langues, tant que des ressources lexicales de qualité sont disponibles pour ces langues.

Finalement, nous avons montré que les méthodes se comportent différemment en matière de classification des correspondances, même si elles semblent similaires en termes de performances. Cela encourage une fusion de méthodes qui pourrait conduire à des résultats bien meilleurs que l'état de l'art. C'est pourquoi dans le prochain chapitre, nous nous sommes intéressés à l'implémentation de nouvelles méthodes et à la fusion de ces dernières avec des méthodes de l'état de l'art déjà éprouvées.

40. https://github.com/FerreroJeremy/Cross-Language-Dataset (consulté le 12/06/2017 à 10h)

5 Introduction de représentations distributionnelles distribuées continues de mots dans la détection de plagiat translingue

« Le désespoir conduit au plagiat bien plus souvent que l'infamie. »

Jeux de mains (traduit par Isabelle Tripault, éd. Calmann-Levy, 1999, p. 435)
— Ruth Rendell (1930-2015)

Nous avons vu en section 2.2.6.1 que les représentations distributionnelles distribuées continues de mots (*word embeddings*) semblent montrer des résultats prometteurs dans plusieurs tâches du TAL. L'idée principale des *word embeddings* est que la représentation d'un mot est obtenue en fonction de son contexte (des mots qui l'entourent dans le texte). Les mots sont projetés dans un espace continu et ceux ayant un contexte similaire se retrouvent proches dans cet espace. Une similarité entre deux vecteurs de mots peut alors être mesurée par une distance vectorielle comme la similarité cosinus. C'est pour cela qu'utiliser des *word embeddings* pour la détection du plagiat nous semble une piste à explorer. Dans ce chapitre, nous présentons dans un premier temps une nouvelle méthode de détection de similarités textuelles sémantiques translingues basée exclusivement sur des *word embeddings*. Ensuite, nous proposons une augmentation de la méthode CL-CTS (voir section 2.2.2.3) en y introduisant également des *word embeddings*. Enfin, nous proposons une notion de pondération morphosyntaxique et fréquentielle de mots qui peut aussi bien être utilisée au sein d'un vecteur qu'au sein d'un sac de mots et nous montrons que son introduction dans les deux nouvelles méthodes présentées augmente leurs performances respectives.

Nous avons également vu dans la section 4.2.3.3 que les méthodes de l'état de l'art de détection du plagiat translingue semblent être complémentaires et dépendantes de certaines caractéristiques des textes sur lesquels elles opèrent. C'est pourquoi, nous essayons ensuite de tirer parti de ces observations pour fusionner les nouvelles méthodes introduites au cours de ce chapitre avec les méthodes de l'état de l'art afin d'améliorer leurs performances globales. Nous étudions les performances de ces méthodes et combinaisons sur notre corpus (présenté dans le chapitre 4) et les comparons aux méthodes de l'état de l'art. Enfin, nous présentons et discutons les résultats de ces méthodes lors de notre soumission à la tâche de détection de similarité textuelle sémantique (STS) translingue espagnol-anglais de la campagne d'évaluation SemEval 2017.

Les contributions de ce chapitre ont fait l'objet de deux publications (Ferrero *et al.*, 2017c,a).

5.1 Nouveaux modèles

5.1.1 Similarité à base de représentations distributionnelles distribuées continues translingues de mots (*Cross-Language Word Embedding-based Similarity, CL-WES*)

La méthode CL-WES est directement basée sur les travaux de Blacoe *et* Lapata (2012). Elle consiste en une similarité cosinus des représentations de deux phrases, la représentation d'une phrase étant la somme des représentations distributionnelles distribuées continues de chacun des mots qu'elle contient.

Soit d un document de longueur n, les n mots du document sont représentés tels que :

$$d = \{w_1, w_2, w_3, ..., w_n\} \tag{5.1}$$

avec w_i le $i^{\text{ème}}$ mot du document d.

La représentation v de ce document est la somme des représentations de chacun de ses mots telle que :

$$v = \sum_{i=1\,,\,w_i \in d}^{n} vector(w_i) \tag{5.2}$$

où w_i est le $i^{\text{ème}}$ mot du document d et *vector* est la fonction qui retourne la représentation d'un mot.

Si d et d' sont deux documents dans deux langues différentes, CL-WES construit leurs vecteurs (représentations vectorielles) v et v', et applique une similarité cosinus (Salton, 1989) entre ces deux vecteurs.

Cette fonction est disponible dans la boîte à outils *MultiVec*[1] (Bérard *et al.*, 2016).

5.1.2 Pondération morphosyntaxique et fréquentielle d'un mot

Les représentations distributionnelles distribuées continues de mots (*word embeddings*) ont fait l'objet de nombreuses recherches ces dernières années. De nombreux travaux se sont concentrés sur comment injecter directement lors de l'apprentissage des *word embeddings* des contraintes ou des données lexico-sémantiques afin d'optimiser leur qualité et leur capacité de discrimination sémantique (Mrkšic *et al.*, 2017; Sugathadasa *et al.*, 2017; Mancini *et al.*, 2017), d'autres travaux encore se sont concentrés sur comment apprendre directement des représentations de phrases optimales (Liu *et* Lapata, 2017; Schwenk *et* Douze, 2017; Wieting *et* Gimpel, 2017; Hill *et al.*, 2016; Wieting *et al.*, 2017; Pagliardini *et al.*, 2017; Lin *et al.*, 2017; Le *et* Mikolov, 2014; Chen, 2017), mais très peu, à notre connaissance, ce sont concentrés sur comment construire des représentations de phrases (ou documents) optimales à partir de représentations de mots déjà existantes. C'est vers cette dernière catégorie de construction de représentations que nous avons orienté nos recherches.

Nous souhaitons, tout en restant dans une conception simpliste, c'est-à-dire la somme de représentations de mots, construire des représentations de phrases les plus pertinentes possibles pour la détection de similarités textuelles sémantiques. Pour cela, nous nous sommes basés sur de nombreuses recherches, comme celles de Ji *et* Eisenstein (2013), qui ont déjà montré que tous les mots n'avaient pas la même importance au sein d'une phrase pour l'identification d'une paraphrase. Les mots les plus importants au sein d'une phrase, ceux qui ont le plus de sens, sont ceux qui sont les plus susceptibles d'être conservés au cours d'un plagiat, d'une reformulation ou d'une traduction et sont donc ceux qui nécessitent un poids plus important au sein de la représentation de cette phrase.

1. https://github.com/FerreroJeremy/multivec (consulté le 19/07/2017 à 16h)

Notre première intuition, s'appuyant sur les travaux de Schwab (2005), où lors de la création du vecteur d'un texte le vecteur d'un mot *gouverneur*[2] (Mel'čuk, 1988) est plus pondéré que les vecteurs des autres mots de ce texte, est que le rôle syntaxique que tient un mot au sein d'une phrase est fortement révélateur de son importance au sein de cette dernière. C'est pourquoi nous nous proposons de pondérer les mots au sein d'une représentation d'un texte en fonction de leur étiquette morphosyntaxique.

En s'inspirant des travaux de Brychcin *et* Svoboda (2016) qui introduisent une pondération fréquentielle (en se servant de la fréquence inverse de document des mots) dans l'indice de Jaccard (Jaccard, 1912) de l'une de leur méthode (voir section 2.2.5), nous décidons aussi d'introduire une pondération fréquentielle en complément de la pondération morphosyntaxique.

D'autres recherches utilisent directement des annotations morphosyntaxiques (Vulić, 2017; Dehouck *et* Denis, 2017) ou bien des annotations fréquentielles de mots (Yin *et* Schütze, 2015) pour construire des modèles de représentations. Il est également important de noter que, contrairement à ce qui est fait dans d'autres recherches comme celles de Chen *et* Manning (2014), nous n'utilisons pas les étiquettes morphosyntaxiques comme entrée de vecteur supplémentaire car nous considérons qu'il est plus utile de les utiliser pour pondérer indépendamment la contribution des mots dans la représentation des phrases.

Nous faisons l'hypothèse que pour un mot donné, nous disposons avant l'étape de comparaison de son étiquette morphosyntaxique ainsi que de sa fréquence inverse de document. Nous étiquetons donc au préalable les documents à comparer avec l'étiqueteur morphosyntaxique *TreeTagger* (Schmid, 1994). Nous normalisons ensuite les étiquettes obtenues en sortie avec le jeu d'étiquettes universelles[3] de Petrov *et al.* (2012). Le jeu d'étiquettes universelles de Petrov *et al.* (2012) dénombre 12 étiquettes. Le Tableau 4.2 résume les correspondances entre ces étiquettes et les parties de discours (*part-of-speech*) les plus communes dans les langues à alphabet latin.

La formule d'agrégation que nous proposons pour déterminer la pondération morphosyntaxique et fréquentielle d'un mot w est la suivante :

$$\Upsilon(w) = weight(pos(w))^{1-\alpha} \times idf(w)^{\alpha} \tag{5.3}$$

où *pos* est la fonction qui retourne l'étiquette morphosyntaxique (*part-of-speech*) universelle d'un mot, *weight* est la fonction qui retourne la pondération attribuée à une certaine étiquette morphosyntaxique, *idf* est la fonction qui retourne la fréquence inverse de document (*inverse document frequency*) d'un mot et le paramètre α est un moyen de contrôler la contribution morphosyntaxique et fréquentielle dans la formule de pondération.

Il s'agit maintenant d'affecter un poids à chacune des 12 étiquettes en fonction de leur importance au sein d'une phrase. Les mots les plus importants au sein d'une phrase, ceux ayant le plus de sens, seront ceux qui sont les plus susceptibles d'être conservés au cours d'une reformulation ou d'une traduction et nécessiteront donc un poids plus important au sein de la représentation de cette phrase. Nous détaillons l'apprentissage des poids optimaux et du paramètre α dans la section 5.3.3.1.

Les travaux de Lioma *et* Blanco (2017) dans la recherche de documents ne s'apparentent pas à ce que nous essayons de faire ici, toutefois ils utilisent également, pour gérer la contribution des mots au sein d'une représentation, une pondération basée sur les parties de discours (*part-of-speech*) qu'ils pondèrent grâce à leur fréquence d'apparition. Ils étiquettent morphosyntaxiquement les documents à comparer, calculent la fréquence (*tf* ou *tf.idf*) de chaque étiquette morphosyntaxique et attribuent un poids à chacune de ces étiquettes en fonction de leur fréquence. De notre côté, étant donné que nous prenons déjà directement en compte la fréquence des mots pour les pondérer et afin donc de ne pas prendre en compte deux fois cette information, nous préférons apprendre un poids idéal pour chaque étiquette morphosyntaxique pour une tâche

2. « Dans chaque syntagme, il y a un constituant qui a un rôle particulier, il a la fonction syntaxique de *gouverneur*. Il s'agit du constituant principal du syntagme. Dans un syntagme nominal, il s'agit du nom, dans un syntagme verbal du verbe, dans une phrase, du syntagme sujet, *etc.* » (Schwab, 2005)

3. https://github.com/FerreroJeremy/universal-pos-tags (consulté le 21/06/2017 à 15h)

spécifique sur un corpus d'entraînement spécifique. L'approche de Lioma *et* Blanco (2017) ne nécessite aucun entraînement en pré-traitement et s'adapte instantanément au corpus considéré, alors que la notre dépend d'une méthode d'optimisation et d'un corpus d'entraînement, ce qui la rend moins flexible mais plus performante.

5.1.3 Similarité morphosyntaxique et fréquentielle à base de représentations distributionnelles distribuées continues translingues de mots (*Cross-Language Word Embedding-based Syntactic and Frequency Similarity, CL-WESFS*)

Cette méthode de comparaison est basée sur la méthode CL-WES mais y introduit la notion de pondération morphosyntaxique et fréquentielle présentée en section 5.1.2. Elle consiste en une similarité cosinus sur les représentations de deux phrases, représentations qui sont elles-mêmes obtenues par la somme pondérée (morphosyntaxiquement et fréquentiellement) des représentations distributionnelles distribuées continues de chacun des mots de ces phrases.

Soit d un document construit suivant le modèle de la Formule 5.1, la représentation vectorielle v du document d s'exprime donc de la façon suivante :

$$v = \sum_{i=1\,,\,w_i \in d}^{n} \Big(vector(w_i) \cdot \Upsilon(w_i) \Big) \tag{5.4}$$

où w_i est le $i^{\text{ème}}$ mot du document d, *vector* est la fonction qui retourne la représentation d'un mot, Υ tel que défini dans Formule 5.3 est un moyen de pondérer morphosyntaxiquement et fréquentiellement la représentation d'un mot et $\cdot$ est le produit scalaire.

Si d et d' sont deux documents dans deux langues différentes, tels que définis dans la Formule 5.1, CL-WESFS construit leurs vecteurs (représentations vectorielles) v et v' suivant la Formule 5.4 (et non plus suivant la Formule 5.2 comme le faisait CL-WES) et applique ensuite une similarité cosinus (Salton, 1989) entre ces deux vecteurs. La Figure 5.1 illustre le fonctionnement de cette méthode sur deux phrases exemples, une en français et l'autre en anglais.

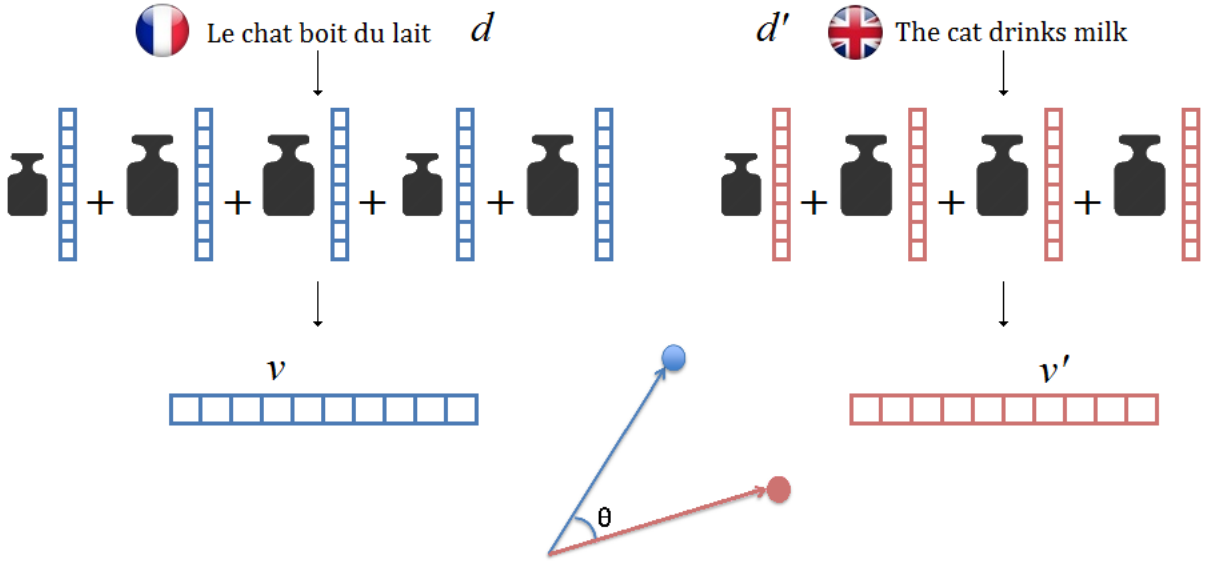


FIGURE 5.1 – Fonctionnement du modèle CL-WESFS. On construit les représentations des deux phrases en faisant la somme pondérée de la représentation de chacun de leurs mots respectifs. Lors de la somme, la pondération d'un mot est déterminée en fonction de son étiquette morphosyntaxique et de sa fréquence inverse de document (*idf*). Les deux représentations construites sont ensuite comparées avec une similarité cosinus.

Cette fonction est également rendue disponible par nos travaux dans la boîte à outils *MultiVec*[4] (Bérard *et al.*, 2016).

5.1.4 Similarité translingue morphosyntaxique et fréquentielle basée sur des thésaurus et des représentations distributionnelles distribuées continues translingues de mots (*Cross-Language Conceptual Thesaurus- and Word Embedding- based Syntactic and Frequency Similarity, CL-CT-WESFS*)

Pour cette méthode, nous nous basons sur l'architecture CL-CTS de Pataki (2012) telle que décrite en section 2.2.2.3, qui consiste à représenter deux documents sous forme de sacs de mots pour pouvoir ensuite les comparer, à laquelle nous ajoutons les améliorations suivantes :

— nous construisons le sac de mots du document source avec toutes les traductions de chaque mot de ce document mais nous construisons également le sac de mots du document cible avec tous les synonymes de chacun des mots de ce document (comme nous l'avions déjà proposé en section 4.2.2). Nous augmentons ainsi les possibilités de correspondances entre documents ;

— pour construire les sacs de mots, nous utilisons conjointement une ressource lexicale liée (comme dans l'architecture originale), ici *DBnary*[5] (Sérasset, 2015), mais également un modèle de représentations distributionnelles distribuées continues de mots (*word embeddings*), toujours dans le but d'augmenter le nombre de correspondances possibles entre les deux documents. D'autres recherches ont précédemment montré l'utilité de croiser des données issues de réseaux lexico-sémantiques, comme *DBnary*, avec des représentations distributionnelles distribuées continues issues de *word2vec* (Servan *et al.*, 2016) ;

— pour mesurer la similarité entre les deux sacs de mots représentant les deux documents à comparer, nous utilisons l'indice de Jaccard (Jaccard, 1912) à recherche de correspondances approximatives (*fuzzy matching*) (Baeza-Yates *et* Navarro, 1996) dans lequel nous introduisons la notion de pondération morphosyntaxique et fréquentielle présentée en section 5.1.2.

Soit d et d' deux documents construits suivant le modèle de la Formule 5.1 et écrits dans deux langues différentes, respectivement L et L'. Le sac de mots S du document d est construit en filtrant les mots vides du document d, et en utilisant une fonction qui retourne pour un mot donné ses 10 mots les plus proches dans le modèle de représentations distributionnelles distribuées continues translingue ainsi que toutes ses traductions vers la langue L' disponibles dans *DBnary*. Le sac de mots S' du document d' est construit en filtrant les mots vides du document d', et en utilisant une fonction qui retourne pour un mot donné ses 10 mots les plus proches dans le modèle de représentations distributionnelles distribuées continues monolingue ainsi que tous les mots issus de ses relations lexicales (hyperonymes, hyponymes, synonymes, *etc.*) dans la langue L' dans *DBnary*. Le sac de mots d'un document est la fusion des sacs de mots de tous les mots contenus dans ce document. Ainsi, les sacs de mots S et S' sont respectivement les représentations conceptuelles (sous la forme de sac de mots) dans la même langue de d et d'.

Comme dit précédemment, pour mesurer la similarité entre deux sacs de mots, nous utilisons une augmentation pondérée morphosyntaxiquement et fréquentiellement de l'indice de Jaccard (Jaccard, 1912) couplée à une recherche de correspondances approximatives (*fuzzy matching*) (Baeza-Yates *et* Navarro, 1996). Au cours de la recherche de correspondances, nous utilisons la distance de Damerau–Levenshtein (Damerau, 1964; Levenshtein, 1966) pour déterminer les approximations. Nous considérons un mot, qui possède une distance de Damerau–Levenshtein inférieure ou égale à 0,20 avec un autre mot, comme identique à cet autre mot. C'est-à-dire que le nombre minimal de caractères qu'il faut insérer, supprimer, substituer ou transposer

4. https://github.com/FerreroJeremy/multivec (consulté le 19/07/2017 à 16h)
5. http://kaiko.getalp.org/about-dbnary/ (consulté le 30/06/2017 à 15h)

pour transformer l'un des mots en l'autre doit couvrir moins de 20% du mot le plus long. Ce pourcentage a été déterminé empiriquement après des tests sur un corpus échantillon de 2 000 comparaisons.

Si d et d' sont deux documents dans deux langues différentes, CL-CT-WESFS construit donc leurs représentations conceptuelles S et S', et estime leur similarité de la façon suivante :

$$Sim(d, d') = J(S, S') = \frac{\Omega(S \Cap S')}{\Omega(d) + \Omega(d')} \tag{5.5}$$

où d et d' sont les deux documents à comparer (représentés sous forme de sacs de mots), S et S' sont leurs représentations conceptuelles également sous forme de sacs de mots, $\Cap$ est l'opérateur d'intersection de deux ensembles avec correspondances approximatives et Ω est la somme des poids des mots d'un ensemble (sac de mots) définie telle que :

$$\Omega(S) = \sum_{i=1\,,\, w_i \in S}^{|S|} \Upsilon(w_i) \tag{5.6}$$

où w_i est le $i^{\text{ème}}$ mot du sac S et Υ est tel que défini dans Formule 5.3, un moyen de pondérer morphosyntaxiquement et fréquentiellement la contribution d'un mot dans l'indice de Jaccard (Jaccard, 1912).

5.2 Fusions et combinaisons de méthodes

Dans la section 4.2.3.3 nous avons vu que certaines méthodes semblent être complémentaires et qu'une fusion de plusieurs de ces méthodes peut donner lieu à des résultats prometteurs. C'est pourquoi, nous proposons deux types de combinaisons de méthodes que nous décrivons dans cette section.

5.2.1 Fusion pondérée

Chaque méthode donnant lieu en sortie à une matrice de similarité (voir section 4.2.1), la fusion pondérée consiste à fusionner ces matrices (par le biais d'une moyenne pondérée) afin d'en constituer une nouvelle. Le but est que la classification de cette nouvelle matrice obtienne de meilleurs résultats que ce que pouvait donner indépendamment chacune des matrices d'entrée. Plus concrètement, durant cette fusion, nous assignons un poids à chaque méthode et nous calculons ensuite la moyenne, pondérée en fonction de ces poids, des scores de similarité de toutes les méthodes pour chaque comparaison (une comparaison donne lieu à un score dans une matrice). La formule de la fusion pondérée peut donc s'exprimer de la façon suivante :

$$\overline{M} = \frac{\sum_{j=1}^{|M|} (P^j \times M^j)}{\sum_{j=1}^{|M|} P^j} \tag{5.7}$$

où M est l'ensemble des méthodes (l'ensemble des matrices), M^j est la matrice de la $j^{\text{ème}}$ méthode et P^j est le poids attribué à la $j^{\text{ème}}$ méthode.

À noter qu'une fusion par moyenne peut facilement être opérée en rendant égaux les poids appliqués à toutes les méthodes, cependant pour que cette fusion soit réellement efficace, il est important que la combinaison des poids attribués aux méthodes soit optimisée. Nous détaillons la procédure d'optimisation de ces poids dans la section 5.3.3.2.

5.2.2 Fusion par arbre de décision

Nous avons vu dans la section 4.2.3.3 que certaines méthodes semblent être complémentaires, or certaines semblent l'être deux à deux, certaines semblent être totalement inutiles ou redondantes avec d'autres et enfin certaines semblent apporter un réel plus en toutes circonstances. Nous allons donc, pour notre deuxième fusion, essayer d'exploiter ces observations en utilisant une combinaison de méthodes fondée sur un arbre de décision, plutôt que seulement sur des combinaisons linéaires, comme décrites en section précédente.

Pour ce faire, nous utilisons la classe *Java J48* de la boîte à outils *Weka 3.8.0* (Hall *et al.*, 2009), qui est une implémentation en code source ouvert (*open source*) de l'algorithme de classification *C4.5* de Quinlan (1993). L'algorithme *C4.5* est un algorithme de classification supervisée publié par Quinlan (1993) et basé sur l'algorithme *ID3* (Quinlan, 1986) auquel plusieurs améliorations ont été apportées, comme la gestion de données d'entraînement avec des valeurs d'attributs manquantes ou bien encore une phase d'élagage (*pruning*) en post-traitement.

Une étude de cette fusion est menée en section 5.3.3.3.

5.3 Évaluation des nouvelles méthodes et des fusions sur notre corpus

5.3.1 Protocole d'évaluation

Nous avons montré en section 4.2.3.1 que les méthodes se comportent de façon similaire à travers les différentes paires de langues (comportements fortement corrélés) et les différentes granularités (comportements moins corrélés mais phénomène tout aussi notable). Pour cette raison, nous nous proposons d'effectuer une évaluation sur les différentes collections de corpus, mais seulement sur la paire de langues en→fr aux granularités syntagmatique et phrastique.

Le protocole utilisé pour cette évaluation est le même que celui présenté en section 4.2.1. Les masques utilisés sont les mêmes que durant l'évaluation de la section 4.2. Cependant, sur les dix masques disponibles pour un sous-corpus à une granularité donnée, cette fois seulement huit sont utilisés pour évaluer les méthodes. La concaténation des deux masques restant de tous les sous-corpus disponibles, donc 10 masques (2 masques pour 5 sous-corpus), forme le corpus de développement utilisé en amont pour l'optimisation des variables nécessaires au bon fonctionnement des méthodes décrites dans ce chapitre (les 12 poids correspondant aux 12 étiquettes morphosyntaxiques et le paramètre α, ainsi que les poids à appliquer à chaque méthode au cours de la fusion pondérée, et l'arbre de décision). En effet, en procédant ainsi, nous utilisons donc un modèle généraliste plutôt qu'un modèle spécifique à un certain type de corpus. Cela correspond au cas industriel, qui concerne Compilatio. Pour les mêmes raisons, nous considérons seulement la granularité phrastique pour apprendre les variables (granularité intermédiaire de notre jeu de données).

5.3.2 Modèle de représentations de mots utilisé

Nous utilisons la boîte à outils *MultiVec*[6] (Bérard *et al.*, 2016) pour générer et manipuler nos représentations distributionnelles distribuées continues de mots (*word embeddings*) monolingues et translingues. Cette boîte à outils inclut les fonctionnalités de *word2vec* (Mikolov *et al.*, 2013a,b,c), *paragraph vector* (Le *et* Mikolov, 2014) et *BiVec* (Luong *et al.*, 2015). En s'appuyant sur l'état de l'art, sur des travaux récents (Ruder, 2017), ainsi que sur des tests empiriques personnels, nous décidons d'utiliser, pour construire notre modèle, une architecture sac de mots continus (*CBOW*) avec une taille de vecteurs de 100, une taille de fenêtres de 5, un paramètre d'échantillonnage négatif (*negative sampling*) de 5 et un alpha de 0,02. Les autres paramètres sont les mêmes que ceux de la configuration par défaut de *MultiVec*. Le corpus utilisé au cours de

6. https://github.com/FerreroJeremy/multivec (consulté le 19/07/2017 à 16h)

cette évaluation pour construire nos représentations est le corpus parallèle *News Commentary*[7] (environ 180 000 phrases pour 77 Mo).

5.3.3 Estimation des paramètres

5.3.3.1 Pondérations morphosyntaxiques

Les mots les plus importants au sein d'une phrase, ceux ayant le plus de sens, seront ceux qui sont les plus susceptibles d'être conservés au cours d'une reformulation ou d'une traduction et nécessiteront donc un poids plus important au sein de la représentation de cette phrase. Intuitivement on peut déjà établir qu'il s'agira d'étiquettes telles que les noms ou verbes plutôt que les déterminants ou prépositions, toutefois nous recherchons la meilleure combinaison de poids possible afin d'obtenir les meilleurs résultats au cours d'une évaluation de détection de similarité sémantique textuelle. Afin d'optimiser au mieux ces variables (les 12 étiquettes et α), nous utilisons l'outil *Condor* (Berghen *et* Bersini, 2005). *Condor* applique un algorithme à régions de confiance basé sur la méthode de Newton (Kelley, 1995, 1999) pour déterminer les valeurs des poids et de α qui optimisent au mieux la F-mesure finale des méthodes sur un corpus de développement donné.

Une optimisation des pondérations affectées à chaque étiquette morphosyntaxique est effectuée sur les corpus de développement à la granularité syntagmatique et phrastique, dans un but de comparaison seulement, car nous rappelons que lors de l'évaluation, nous n'utiliserons que les pondérations trouvées sur la granularité phrastique. Les pondérations résultantes de ces optimisations sont reportées dans le Tableau 5.1, tandis qu'un diagramme en bâtons représentant ces valeurs est reporté en Annexe D. La première remarque que nous pouvons faire est que nous constatons que les noms, verbes, adjectifs, adverbes, numériques et mots étrangers possèdent des poids plus importants que les mots vides, comme les prépositions ou les déterminants. Cela peut s'expliquer par le fait que ces mots peuvent être trouvés en grand nombre et facilement dans toutes phrases, et ne peuvent donc pas être considérés comme importants au sein d'une phrase spécifique. D'ailleurs, en traitement automatique du langage naturel, de nombreuses techniques filtrent ces mots vides en premier lieu avant de passer à une analyse plus profonde.

Étiquettes morphosyntaxiques	corpus syntagmatique	corpus phrastique
VERB - verbes (tous les temps et conjugaisons)	14,17	14,59
NOUN - noms (communs et propres)	19,29	18,81
PRON - pronoms	1,57	1,54
ADJ - adjectifs	16,14	16,31
ADV - adverbes	12,99	12,48
ADP - adpositions (prépositions and postpositions)	1,18	1,15
CONJ - conjonctions	1,18	1,15
DET - déterminants	0,98	1,15
NUM - nombres cardinaux	14,96	14,59
PRT - particules et mots outils	0,98	1,15
X - mots étrangers ou inconnus	15,55	15,36
. - ponctuation	0,98	1,73

Tableau 5.1 – Poids attribués (en %) aux différentes étiquettes morphosyntaxiques après optimisation.

On peut également noter que les pondérations trouvées à l'issue de l'apprentissage sur le sous-corpus syntagmatique sont extrêmement proches et fortement corrélées avec celles trouvées à l'issue de l'apprentissage sur le sous-corpus phrastique. Leur corrélation est de 0,999[8].

7. http://www.statmt.org/wmt14/translation-task.html (consulté le 19/07/2017 à 16h)
8. Nous ne tirons aucune conclusion de cette forte corrélation pour le moment en raison de l'analyse qui suit et qui nuance ce résultat.

Le paramètre α trouvé à l'issue de l'apprentissage est de 0,69 sur le sous-corpus syntagmatique et de 0,78 sur le sous-corpus phrastique. Dans les deux cas, cela montre que pour déterminer l'importance d'un mot au sein d'une phrase, sa rareté dans la langue, induite par sa fréquence inverse de document, semble être une information plus importante que son étiquette morphosyntaxique. Toutefois, la première information ne semble pas écraser complètement la seconde, ce qui prouve que les deux informations sont complémentaires. On peut aisément vérifier cette observation en étudiant les valeurs des pondérations des fréquences inverses de document et des étiquettes morphosyntaxiques de quelques mots sur un corpus de test. Par exemple, au sein du corpus de développement, le mot *secondly* qui se trouve être un adverbe et le mot *schengen* qui est un nom propre ont la même fréquence inverse de document, c'est donc leur étiquette morphosyntaxique qui les départage. Tandis que *solution* et *vector*, qui sont tous les deux des noms communs, ont la même pondération dû à leur étiquette morphosyntaxique identique, mais cette fois c'est donc leur fréquence inverse de document qui les départage. Les pondérations morphosyntaxiques et les fréquences inverses de document sont donc bien deux informations complémentaires, même si naïvement nous aurions pu croire que les fréquences inverses de document couvraient implicitement l'information des étiquettes morphosyntaxiques. Ceci est confirmé par les résultats obtenus dans les travaux de Nagoudi *et al.* (2017b,a) (et par la suite également montré dans le Tableau 5.8).

La différence entre les deux optimisations se situe notamment au niveau des marqueurs de ponctuation, avec un poids de 0,98 pour un apprentissage sur le sous-corpus syntagmatique et un poids de 1,73 pour un apprentissage sur le sous-corpus phrastique. L'augmentation de la pondération pour cette étiquette est donc de plus de 175%, près du double. Cela semble signifier que les marqueurs de ponctuations sont deux fois plus importants au sein d'une phrase qu'au sein de syntagmes. Or, de par leur mode de construction, nous savons que les marqueurs de ponctuation sont très peu présents, si ce n'est inexistants, au sein des syntagmes. Nous décidons donc, afin d'éviter que certaines pondérations soient influencées par la sur-représentation dans les corpus des étiquettes auxquelles elles ont trait, de normaliser ces pondérations par la représentation de leurs étiquettes correspondantes au sein des sous-corpus de développement. Cette normalisation consiste à, pour chaque étiquette morphosyntaxique, diviser sa pondération obtenue après apprentissage (pondération qui se trouve dans le Tableau 5.1) par sa probabilité d'apparition au sein du corpus de développement :

$$rescale(pos) = \frac{weight(pos)}{\left(\frac{f_{pos,D}}{\mid D \mid}\right)} \tag{5.8}$$

où *weight* est la fonction qui retourne la pondération attribuée à une certaine étiquette morphosyntaxique, D est le corpus de développement utilisé pour l'apprentissage des pondérations et $f_{pos,D}$ est la fonction qui retourne le nombre total de mots dans D qui ont pour étiquette morphosyntaxique *pos*.

Le Tableau 5.2 rapporte les probabilités d'apparition des étiquettes dans le corpus de développement. Les fréquences d'apparition détaillées par étiquettes sont présentées en Annexe E. On peut noter que ces fréquences d'apparition concordent avec celles que l'on peut trouver dans la littérature anglaise. Hardie (2007) indique entre autre que, dans un texte, se trouve en moyenne 27∼28% de noms et 8∼9% de pronoms. L'écart entre le français et l'anglais est de l'ordre des 3∼5% pour les étiquettes les plus importantes.

Les probabilités d'apparition des étiquettes au sein du sous-corpus de développement syntagmatique sont corrélées à 0,8725 avec leurs probabilités d'apparition au sein du sous-corpus de développement phrastique. Ceci prouve que, malgré une forte corrélation, il y a bien des disparités entre les distributions des étiquettes morphosyntaxiques sur ces deux sous-corpus et que celles-ci peuvent influencer l'apprentissage. Il y a une plus grande fréquence d'apparition des noms, adjectifs, déterminants et prépositions dans les syntagmes que dans les phrases, et inversement, une plus grande fréquence d'apparition des verbes, adverbes, ponctuations, pronoms et

conjonctions dans les phrases. Par exemple, la fréquence d'apparition des marqueurs de ponctuation dans le sous-corpus phrastique est presque supérieure à 10 fois celle d'apparition dans le sous-corpus syntagmatique. La fréquence d'apparition des noms, quant à elle, est pratiquement divisée par deux entre le sous-corpus syntagmatique et le sous-corpus phrastique.

Étiquettes morphosyntaxiques	corpus syntagmatique	corpus phrastique
VERB - verbes (tous les temps et conjugaisons)	7,09	14,96
NOUN - noms (communs et propres)	42,12	23,82
PRON - pronoms	1,92	7,52
ADJ - adjectifs	10,12	8,30
ADV - adverbes	3,89	4,50
ADP - adpositions (prépositions and postpositions)	20,10	15,21
CONJ - conjonctions	1,18	4,50
DET - déterminants	9,51	8,76
NUM - nombres cardinaux	2,03	2,30
PRT - particules et mots outils	0,00	0,00
X - mots étrangers ou inconnus	0,59	0,65
. - ponctuation	1,47	9,47

Tableau 5.2 – Probabilité d'apparition (en %) des différentes étiquettes morphosyntaxiques au sein des sous-corpus de développement.

Le Tableau 5.3 recense les pondérations attribuées à chaque étiquette une fois normalisées en fonction de la fréquence d'apparition de ces dernières. Un diagramme en bâtons représentant ces valeurs est reporté en Annexe D. On peut constater que cette normalisation a fait évoluer les pondérations, dans le sens où ces dernières ne sont corrélées qu'à seulement 0,43 avec ce qu'elles étaient avant normalisation (0,423 pour les pondérations apprises sur le sous-corpus syntagmatique et 0,437 pour celles apprises sur le sous-corpus phrastique). De plus, on peut voir que les pondérations de nombreuses étiquettes semblent avoir été affaiblies. Cela tend à prouver que l'importance d'une étiquette est bien dépendante de sa fréquence. Toutefois, on distingue encore des étiquettes morphosyntaxiques qui ont plus de poids que d'autres. En effet, en ignorant les numériques et les mots étrangers, qui pour nous sont logiquement des points d'ancrage fort pour n'importe quelle méthode d'alignement, nous observons que, contrairement à l'intuition, les adjectifs et les adverbes sont plus importants que les noms et les verbes. Nous constatons également que les verbes ont toujours un rôle plus important au sein des syntagmes et que les adjectifs tiennent un rôle plus important au sein des phrases.

Étiquettes morphosyntaxiques	corpus syntagmatique	corpus phrastique
VERB - verbes (tous les temps et conjugaisons)	2,00	0,98
NOUN - noms (communs et propres)	0,46	0,79
PRON - pronoms	0,82	0,20
ADJ - adjectifs	1,60	1,97
ADV - adverbes	3,34	2,77
ADP - adpositions (prépositions and postpositions)	0,06	0,08
CONJ - conjonctions	1,00	0,26
DET - déterminants	0,10	0,13
NUM - nombres cardinaux	7,37	6,34
PRT - particules et mots outils	0,00	0,00
X - mots étrangers ou inconnus	26,36	23,62
. - ponctuation	0,67	0,18

Tableau 5.3 – Poids attribués (en %) aux différentes étiquettes morphosyntaxiques après optimisation en normalisant en fonction de la probabilité d'apparition des étiquettes morphosyntaxiques.

Une fois le facteur sur-représentation de certaines classes morphosyntaxiques en fonction des textes considérés pris en compte, des différences de pondération des mêmes classes entre les deux granularités persistent. Il semblerait donc que certaines classes morphosyntaxiques aient réellement plus ou moins d'importance –pour une représentation dans notre tâche en tout cas– en fonction des syntagmes ou des phrases.

5.3.3.2 Poids des méthodes lors de la fusion pondérée

Chaque méthode retourne en sortie une matrice de similarité. Cette fusion consiste à réaliser une moyenne pondérée de ces matrices afin d'en constituer une nouvelle. Le but est qu'en seuillant cette nouvelle matrice, on obtienne sensiblement une meilleure F-mesure que ce que pouvait donner indépendamment le seuillage de chacune des matrices en entrée de la fusion. Il faut donc trouver la combinaison des poids P (voir Formule 5.7) qui donne la matrice de similarité fusionnée finale qui obtient la meilleure F-mesure après seuillage. Nous utilisons, à nouveau, l'outil *Condor* (Berghen *et* Bersini, 2005) pour optimiser la distribution de ces poids. Il utilise un algorithme à régions de confiance basé sur la méthode de Newton (Kelley, 1995, 1999) pour déterminer les poids à appliquer à chaque méthode qui optimisent au mieux la F-mesure de la matrice fusionnée finale sur un corpus de développement donné. Ici, comme décrit en section 5.3.1, ce corpus de développement est la concaténation des deux masques disponibles pour tous les sous-corpus à la granularité phrastique (pour le couple de langue en→fr), c'est donc la concaténation de 10 masques (2 masques pour 5 sous-corpus).

On peut constater dans le Tableau 5.4, qui montre les poids attribués aux méthodes lors de la fusion pondérée, que les méthodes les plus importantes lors d'une telle fusion sont de loin les méthodes CL-C3G, CL-CT-WESFS et CL-WESFS. Ces trois méthodes recouvrent 75% de la pondération attribuée lors de l'apprentissage. Cela signifie que ces trois méthodes, en plus d'être les plus performantes individuellement, sont relativement complémentaires.

Méthode	Poids attribués
CL-C3G	0,34
CL-CTS	0,04
CL-ASA	0,11
CL-ESA	0,00
T+MA	0,02
CL-WESFS	0,17
CL-CT-WESFS	0,24
T+WA-IDF	0,08

Tableau 5.4 – Poids attribués aux méthodes lors de la fusion pondérée.

5.3.3.3 Arbre de décision

Comme dans la section 5.3.3.1 et la section 5.3.3.2, l'apprentissage de l'arbre de décision est effectué sur une concaténation des 10 masques de développement (les 2 masques réservés à l'apprentissage des 5 types de sous-corpus pour le couple de langue en→fr à la granularité phrastique).

La Figure 5.2 représente la sortie de *Weka* lors de l'apprentissage et illustre ainsi l'arbre de décision utilisé pour mener à bien cette fusion.

Durant notre apprentissage, un arbre garantissant au minimum une F-mesure de 0,85 est un arbre qui implique au minimum les méthodes CL-C3G, CL-WESFS et CL-CT-WESFS. En comparant un tel arbre aux pondérations obtenues pour la fusion pondérée (voir section 5.3.3.2), on constate que les méthodes les plus importantes au sein de l'arbre sont les mêmes que celles qui ont le plus de poids au sein de la fusion pondérée. De plus, on peut voir que la méthode CL-ASA se trouve en $4^{\text{ème}}$ profondeur de l'arbre, là où elle obtient la $4^{\text{ème}}$ plus grosse pondération dans la

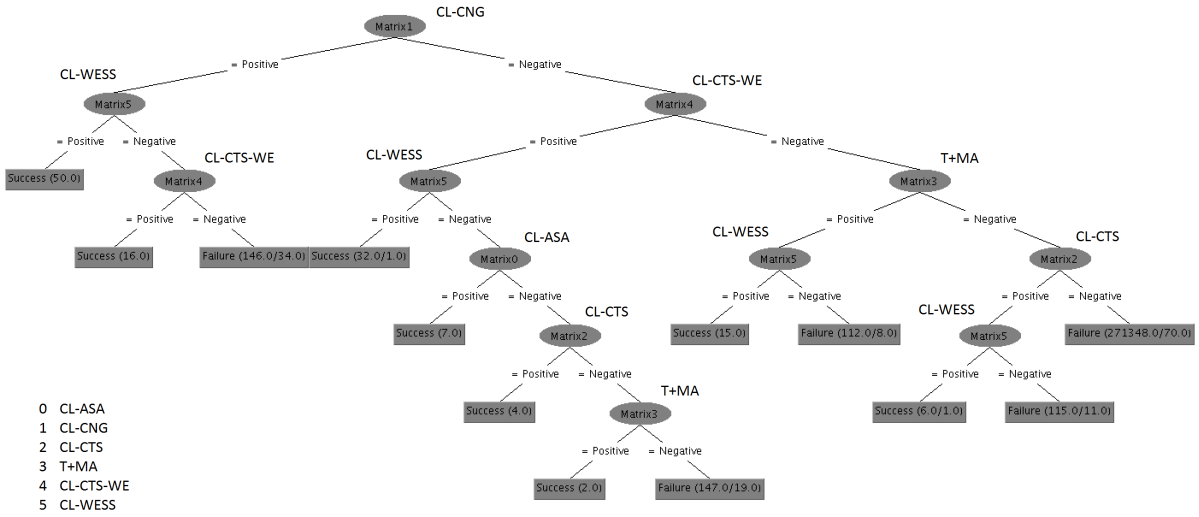


FIGURE 5.2 – Arbre de décision, généré par *Weka 3.8.0* (Hall *et al.*, 2009), utilisé pour la fusion par arbre de décision.

fusion pondérée, ce qui prouve qu'elle est bien complémentaire avec les méthode CL-C3G et les méthodes à base de *word embeddings* et confirme notre intuition révélée dans la section 4.2.3.3. En revanche, dans l'arbre de décision, l'approche T+WA-IDF (définie en section 5.3.4) semble tenir un rôle plus important que dans les pondérations de la fusion pondérée. Cela peut s'expliquer par le fait que la construction de l'arbre cherche à maximiser le taux de classification correcte, cela en faisant intervenir le maximum de méthodes (même celles dont la contribution sera mineure), là où la fusion pondérée est limitée par une combinaison linéaire.

5.3.4 Performances des nouvelles méthodes et des fusions sur notre corpus

Le Tableau 5.5 rapporte les performances, à la granularité syntagmatique, des nouvelles méthodes comparées à celles des méthodes de l'état de l'art (présentées en section 2.2) sur le jeu des huit masques qui servent pour l'évaluation (les deux autres masques ayant servi pour le réglage des paramètres) [9], tandis que le Tableau 5.6 en fait de même pour la granularité phrastique. Pour rappel, chacune des valeurs de ces tableaux est la moyenne, pondérée en fonction du nombre de comparaisons faites selon le sous-corpus, des huit F-mesures obtenues (sur les huit masques) par une méthode sur un sous-corpus à une granularité donnée sur le couple de langues en→fr.

Au cours de cette étude, nous évaluons et reportons également les performances de la méthode introduite par Brychcin *et* Svoboda (2016) au cours de SemEval 2016 et que nous avions présentée en section 2.2.5, qui fait appel à une traduction automatique suivie d'un alignement de mots (*Translation + Word Alignment, T+WA-IDF*) avec pondération par fréquence inverse de document (*inverse document frequency*). Cette méthode s'avère très efficace lors de la tâche de détection de similarité textuelle sémantique de la campagne SemEval (elle a été utilisée par l'équipe classée première en 2016 (Agirre *et al.*, 2016) et a contribué à notre première place en 2017 (voir section 5.4)). En revanche, lors de l'évaluation sur notre corpus, elle montre des performances plus mitigées. En réalité, elle est la meilleure méthode dans seulement deux cas (le sous-corpus Europarl à la granularité syntagmatique et le sous-corpus Webis-CL-10 (*Amazon Product Reviews*) à la granularité phrastique), mais elle se classe tout de même dans le *Top 3* dans la majorité des autres cas.

Nous observons que l'emploi de représentations distributionnelles distribuées continues de mots en complément d'une ressource lexicale liée ainsi que l'introduction en son sein de la pon-

9. Ceci explique pourquoi pour les méthodes état de l'art, les résultats ne sont pas les mêmes que ceux observés dans la section 4.2.3.2, où les dix masques étaient utilisés pour évaluer les méthodes.

Méthode	Wikipédia (%)	TALN-ACL (%)	JRC-Acquis (%)	Webis-CL-10 (%)	Europarl (%)	Moyenne (%)
CL-C3G	63,04 ±0,867	40,80 ±0,542	36,80 ±0,842	80,69 ±0,525	53,26 ±0,639	50,76 ±0,684
CL-CTS	58,05 ±0,563	33,66 ±0,411	30,15 ±0,799	67,88 ±0,959	45,31 ±0,612	42,84 ±0,682
CL-ASA	23,70 ±0,617	23,24 ±0,433	33,06 ±1,007	26,34 ±1,329	55,45 ±0,748	47,32 ±0,852
CL-ESA	64,86 ±0,741	23,73 ±0,675	13,91 ±0,890	23,01 ±0,834	13,98 ±0,583	14,81 ±0,681
T+MA	58,26 ±0,832	38,90 ±0,525	28,81 ±0,565	73,25 ±0,660	36,60 ±1,277	37,12 ±1,043
CL-WES	37,53 ±1,317	21,70 ±1,042	32,96 ±2,351	39,14 ±1,959	46,01 ±1,640	41,95 ±1,842
CL-WESFS	54,48 ±1,424	37,69 ±0,852	46,06 ±1,041	57,60 ±0,952	57,74 ±0,848	54,54 ±0,907
CL-CT-WESFS	63,09 ±0,553	42,87 ±1,405	37,29 ±0,831	77,92 ±1,442	55,14 ±1,355	51,96 ±1,222
T+WA-IDF	60,16 ±1,371	39,28 ±1,255	36,04 ±1,220	73,63 ±1,153	57,74 ±1,227	53,02 ±1,221
Fusion moyenne	81,78 ±1,295	66,94 ±1,412	62,44 ±0,962	92,45 ±0,793	79,89 ±1,144	76,10 ± 1,075
Fusion pondérée	82,74 ±1,992	70,41 ±1,506	67,25 ±1,037	93,17 ±0,945	82,48 ±1,003	79,14 ± 1,015
Arbre de décision	91,76 ±1,727	72,92 ±1,889	71,42 ±1,367	94,68 ±1,132	89,32 ±1,283	84,91 ± 1,301

Tableau 5.5 – Moyenne et intervalle de confiance des F-mesures des méthodes appliquées sur les sous-corpus en→fr à la granularité syntagme – tests sur 8 lancers.

Méthode	Wikipédia (%)	TALN-ACL (%)	JRC-Acquis (%)	Webis-CL-10 (%)	Europarl (%)	Moyenne (%)
CL-C3G	48,24 ± 0,272	48,19 ± 0,520	36,85 ± 0,727	61,30 ± 0,567	52,70 ± 0,928	49,34 ± 0,864
CL-CTS	46,71 ± 0,388	38,93 ± 0,284	28,38 ± 0,464	51,43 ± 0,687	53,35 ± 0,643	47,50 ± 0,601
CL-ASA	27,68 ± 0,336	27,33 ± 0,306	34,78 ± 0,455	25,95 ± 0,604	36,73 ± 1,249	35,81 ± 1,036
CL-ESA	50,89 ± 0,902	14,41 ± 0,233	14,45 ± 0,380	14,18 ± 0,645	14,09 ± 0,583	14,44 ± 0,540
T+MA	50,39 ± 0,898	37,66 ± 0,365	32,31 ± 0,370	61,95 ± 0,706	37,70 ± 0,514	37,42 ± 0,490
CL-WES	28,48 ± 0,865	24,37 ± 0,720	33,99 ± 0,903	39,10 ± 0,863	44,06 ± 1,399	41,43 ± 1,262
CL-WESFS	47,98 ± 0,886	43,43 ± 1,206	49,80 ± 0,949	59,30 ± 1,392	59,29 ± 1,088	56,66 ± 1,073
CL-CT-WESFS	51,18 ± 0,916	48,72 ± 1,410	37,37 ± 0,772	62,33 ± 0,818	60,98 ± 1,145	54,79 ± 1,027
T+WA-IDF	50,70 ± 1,188	41,81 ± 1,236	35,41 ± 1,165	63,20 ± 0,822	56,75 ± 1,231	51,49 ± 1,186
Fusion moyenne	68,84 ±1,049	68,39 ±1,271	63,18 ±0,935	71,32 ±0,942	82,23 ±1,354	77,37 ± 1,241
Fusion pondérée	73,44 ±1,402	75,69 ±1,284	68,02 ±1,056	75,87 ±1,127	88,54 ±1,469	83,27 ± 1,362
Arbre de décision	81,04 ±1,446	81,52 ±1,300	74,35 ±1,115	79,26 ±1,173	94,87 ±1,730	89,50 ± 1,567

Tableau 5.6 – Moyenne et intervalle de confiance des F-mesures des méthodes appliquées sur les sous-corpus en→fr à la granularité phrase – tests sur 8 lancers.

dération morphosyntaxique et fréquentielle des mots augmentent les performances de la méthode état de l'art CL-CTS. En effet, CL-CT-WESFS obtient un gain moyen de performances comparé à CL-CTS de 8,25 sur les syntagmes et de 8,36 sur les phrases.

Les performances moyennes de la méthode CL-WES sont inférieures à trois méthodes de l'état de l'art. En revanche, sa version pondérée morphosyntaxiquement et fréquentiellement (CL-WESFS) semble très prometteuse et augmente les performances globales de CL-WES de 15,25 sur les syntagmes et de 17,96 sur les phrases. Grâce à cette amélioration, CL-WESFS est meilleure que CL-C3G (+0,35 sur les syntagmes et +2,50 sur les phrases) et est la meilleure méthode individuelle (hors combinaison) évaluée jusqu'à présent sur notre corpus. Sur la granularité syntagmatique, elle est équivalente à la seconde méthode (T+WA-IDF) et sur la granularité phrastique, elle fait mieux que la seconde méthode (CL-CT-WESFS) (+1,868).

On remarque que les méthodes à base de représentations distributionnelles distribuées continues de mots produisent des intervalles de confiance nettement supérieurs à ceux des méthodes de l'état de l'art (1,325 de moyenne contre 0,789 pour les méthodes état de l'art sur la granularité syntagmatique et 1,109 de moyenne contre 0,691 pour les méthodes état de l'art sur la granularité phrastique). Cela peut s'expliquer par le fait que le modèle de *word embeddings* utilisé durant l'évaluation est de taille modeste (environ 180 000 phrases pour 77 Mo). Ce modèle est une ressource lexicale comme une autre et sa couverture de vocabulaire peut ainsi s'avérer insuffisante et hétérogène d'un sous-corpus à un autre.

Les résultats des différents systèmes de fusions et combinaisons sont aussi reportés dans le Tableau 5.5 et le Tableau 5.6. Pour chacun de ces systèmes, 8 méthodes servent de base à la fusion (les 5 méthodes de l'état de l'art et les 3 méthodes nouvellement introduites [10]).

Comme mentionné dans la section 5.3.1, les fusions par moyenne pondérée et par arbre de décision sont optimisées sur des corpus de granularité phrastique, ce qui peut expliquer pourquoi elles semblent plus efficaces sur les jeux de données à cette granularité. Il est important de préciser que les résultats observés ici reflètent les performances de nos systèmes de fusions sur des méthodes nouvellement introduites ou des méthodes comportant une variante de notre pondération morphosyntaxique intégrant également une pondération fréquentielle, pondération non présente dans l'article (Ferrero *et al.*, 2017c) qui présentait pour la première fois les résultats de nos fusions. Cela explique pourquoi les résultats observés ici ne sont pas les mêmes que dans cet article.

La fusion par moyenne pondérée obtient de meilleures performances que les méthodes de l'état de l'art, les méthodes à base de représentations distributionnelles distribuées continues de mots et la fusion par moyenne simple, et cela dans tous les cas de figure (tous les sous-corpus sur toutes les granularités). Néanmoins, le système de combinaison par arbre de décision montre des performances encore nettement supérieures avec une différence statistiquement significative (un $p < 0,05$ au *t-test* de Williams (Williams, 1959), évalué par le biais de l'outil *Cocor* [11] (Diedenhofen *et* Musch, 2015)). À la granularité syntagmatique, la fusion par arbre de décision obtient une F-mesure moyenne de 84,91 quand la fusion par moyenne pondérée obtient 79,14 et la meilleure méthode individuelle (CL-WESFS) obtient 54,54. La tendance est la même sur la granularité phrastique, où la fusion par arbre de décision obtient une F-mesure moyenne de 89,50 contre 83,27 pour la fusion par moyenne pondérée et 56,66 pour la meilleure méthode individuelle (CL-WESFS).

Tout comme les méthodes à base de représentations distributionnelles distribuées continues de mots, les fusions montrent des intervalles de confiance beaucoup plus grands que les méthodes de l'état de l'art. Ceci peut s'expliquer par le fait que ce sont des systèmes de combinaisons de différentes méthodes qui montrent toutes des caractéristiques assez différentes, certaines même qui présentent déjà des intervalles de confiance assez importants (les méthodes à base de représentations distributionnelles distribuées continues de mots, par exemple).

Nous pouvons de nouveau étudier les corrélations des méthodes –et plus particulièrement celles des nouvelles méthodes et des systèmes de combinaison– sur le couple de langue en→fr entre les granularités syntagmatique et phrastique. Le Tableau 5.7 présente ces corrélations. Une fois encore, on peut noter de fortes corrélations, cela signifie que les méthodes –y compris les nouvelles– se comportent individuellement de façon similaire sur tous les sous-corpus en→fr entre les granularités syntagmatique et phrastique. Toutefois, après une observation plus minutieuse, on constate que les corrélations des méthodes à base de représentations distributionnelles distribuées continues de mots sont légèrement moins élevées (0,8375 en moyenne contre 0,8904 de moyenne pour les méthodes de l'état de l'art). Ces méthodes se comportent donc de façon un peu plus dépendante de la granularité que les méthodes de l'état de l'art. Cela peut s'expliquer pour la même raison que les écarts plus importants des intervalles de confiance observés précédemment, c'est-à-dire le fait que notre corpus utilisé pour l'apprentissage du modèle des représentations distributionnelles distribuées continues de mots ne doit pas avoir une couverture de vocabulaire suffisante pour couvrir de façon homogène le vocabulaire de tous nos sous-corpus.

10. La méthode CL-WES ne fait pas partie des méthodes individuelles qui servent d'entrée aux différents systèmes de combinaison car par définition (voir section 5.1.1) elle fonctionne de façon similaire à la méthode CL-WESFS, sauf que ses performances sont moindres. Elle sera donc inutilement redondante.

11. http://comparingcorrelations.org/ (consulté le 23/08/2017 à 10h)

Méthode	Corrélation
CL-CNG	0,854
CL-CTS	0,764
CL-ASA	0,862
CL-ESA	0,975
T+MA	0,998
T+WA-IDF	0,960
CL-WES	0,862
CL-WESFS	0,845
CL-CT-WESFS	0,804
Fusion moyenne	0,491
Fusion pondérée	0,407
Arbre de décision	0,397

Tableau 5.7 – Corrélation de Pearson des résultats des méthodes sur tous les sous-corpus entre la granularité syntagmatique et phrastique.

Phénomène plus original, les systèmes de combinaison présentent des corrélations beaucoup plus basses (avec une moyenne de seulement 0,43). En effet, les fusions de méthodes ont tendance à niveler les performances à travers les différents sous-corpus. Il en résulte un « signal de fusion » beaucoup plus « plat ». En d'autres termes, il en résulte de moins grandes différences de performances d'un sous-corpus à un autre. Là où certaines méthodes individuelles pouvaient atteindre jusqu'à plus de 40% de variations de performances d'un sous-corpus à un autre, les performances des fusions oscillent en moyenne à moins de 20% de variations d'un sous-corpus à un autre. Ceci explique que les corrélations entre les deux granularités des performances des fusions sont plus faibles que celles des méthodes individuelles.

Ces résultats confirment que les différentes méthodes proposées sont complémentaires les unes des autres et que l'usage de représentations distributionnelles distribuées continues de mots peut s'avérer utile dans la détection automatique de plagiat translingue.

5.4 Notre participation à SemEval 2017

5.4.1 Présentation de la tâche

La tâche de détection du plagiat translingue extrinsèque n'étant plus proposée dans la campagne d'évaluation PAN, nous avons choisi de participer à la sous-tâche de détection de similarité textuelle sémantique (STS) translingue de la campagne d'évaluation SemEval. Cette tâche consiste, considérant une liste de paires de phrases dans deux langues différentes, à estimer un score pour chacune de ces paires, allant de 0 à 5 en fonction du degré de similarité sémantique entre les deux phrases concernées. Les organisateurs de la tâche définissent ensuite les meilleurs systèmes soumis en mesurant la corrélation entre les scores des systèmes et les scores annotés par des humains. La tâche (corpus, protocole et métrique d'évaluation) a été présentée dans la section 3.6 et est plus amplement décrite dans l'article de présentation écrit par les organisateurs (Cer *et al.*, 2017). Les sous-tâches 4a et 4b sont les seules à proposer une paire de langues qui nous intéresse dans cette thèse (espagnol-anglais), c'est pourquoi nous décidons de participer à ces sous-tâches uniquement.

Cette participation a fait l'objet d'un article (Ferrero *et al.*, 2017a).

À noter que notre méthode CL-WESFS a également été soumise au cours de la sous-tâche arabe-arabe (sous-tâche 1) et que cela a donné lieu à une publication dans laquelle nous sommes co-auteurs (Nagoudi *et al.*, 2017b).

5.4.2 Méthodes soumises

Pour notre soumission, nous décidons majoritairement de nous baser sur des approches de détection du plagiat translingue et plus particulièrement de mettre en avant la notion de pondération morphosyntaxique et fréquentielle introduite au cours de ce chapitre.

Les différentes équipes ont le droit de soumettre jusqu'à trois systèmes différents, nous profitons donc de cela pour soumettre des systèmes supervisés et des systèmes non-supervisés.

Nous adoptons ainsi pour stratégie de soumettre les systèmes suivants :

— la meilleure méthode individuelle (système non-supervisé) ;
— une fusion par moyenne (système non-supervisé) ;
— une fusion complexe par régression linéaire (système supervisé).

5.4.2.1 Méthodes individuelles

Les méthodes individuelles sont la base de notre soumission : non seulement nous soumettons la meilleure d'entre elles mais nous nous servons également de ces méthodes comme entrée pour nos systèmes de combinaison supervisés et non-supervisés. Ces méthodes sont au nombre de quatre :

— la méthode CL-WESFS telle que présentée en section 5.1.3 et paramétrée selon l'étude de la section 5.3.3.1 ;
— la méthode CL-CT-WESFS telle que présentée en section 5.1.4 et paramétrée selon l'étude de la section 5.3.3.1 ;
— la méthode T+WA-IDF telle que présentée en section 2.2.5, qui est basée sur l'aligneur monolingue [12] de Sultan *et al.* (2015) améliorée avec la pondération inverse de document introduite par Brychcin *et* Svoboda (2016) et qui s'était classée première à la tâche STS translingue de l'édition 2016 (Agirre *et al.*, 2016). L'amélioration n'a pas été partagée par ses auteurs initiaux, c'est pourquoi nous partageons notre ré-implémentation sur GitHub [13] ;
— la méthode CL-C3G telle que présentée en section 2.2.1.1, qui se trouvait être la meilleure méthode de l'état de l'art durant notre évaluation en section 4.2. L'implémentation de base de cette méthode est toujours celle de Potthast *et al.* (2011a) où l'on conserve les espaces, telle que décrit en section 4.2.2. Toutefois, pour notre participation à SemEval nous décidons d'améliorer une nouvelle fois cette méthode en remplaçant, pour le calcul de la pondération *tf.idf* des vecteurs, la formule de la fréquence du terme normalisé (*term frequency*) par une double normalisation de 0,5 (*double normalization 0.5*) pour prévenir un biais dans la longueur des documents (Manning *et al.*, 2008). Cette décision a été prise suite aux tests menés sur les corpus dévaluation de SemEval 2016 (voir Tableau 5.8). La formule de la fréquence inverse de document (*idf*) est donc toujours la même (voir Formule 2.2), tandis que la fréquence d'un terme (*tf*) est maintenant calculée grâce à la Formule 5.9 :

$$tf(t,d) = K + (1-K) \cdot \frac{f_{t,d}}{max(f_{(\forall\, t' \in d),d})} \tag{5.9}$$

où $f_{t,d}$ est la fonction qui retourne le nombre d'occurrences du terme t dans le document d, max est la fonction qui retourne la fréquence du terme qui a la fréquence maximale et $K = 0,5$. Nous calculons directement nos fréquences inverses de document (*idf*) sur le corpus d'évaluation considéré.

12. https://github.com/ma-sultan/monolingual-word-aligner (consulté le 10/08/2017 à 16h)
13. https://github.com/FerreroJeremy/monolingual-word-aligner (consulté le 10/08/2017 à 16h)

5.4.2.2 Fusions

Après les résultats obtenus dans la section 5.3.4, nous savons qu'une combinaison de méthodes peut être plus performante que des méthodes individuelles prises séparément. Nous voulons tirer parti de cette observation en soumettant lors de cette campagne deux types de fusions, une fusion par moyenne et donc non-supervisée (voir section 5.2.1), et une fusion plus complexe par régression linéaire qui nécessite un apprentissage sur un corpus. Cette régression est apprise depuis les corpus d'évaluation de la tâche de 2016.

Nous utilisons donc les corpus d'évaluation de l'édition 2016 pour régler nos systèmes fusionnés mais aussi pour décider quelles seront les méthodes retenues à soumettre pour l'évaluation de 2017.

5.4.2.3 Résultats sur les corpus de l'édition de 2016

Le Tableau 5.8 rapporte les résultats des différentes configurations (pondérations utilisées et optimisations apportées) de nos méthodes individuelles et le Tableau 5.9 rapporte les résultats de nos méthodes individuelles optimales et de nos systèmes de combinaison, sur les corpus d'évaluation de la tâche de détection de similarité textuelle sémantique (STS) translingue espagnol-anglais de SemEval 2016 (Agirre *et al.*, 2016). Nous disposons donc d'une comparaison des diverses techniques STS individuelles que nous évaluons, ainsi que des différents systèmes de régression testés afin de pouvoir choisir lesquels soumettre à l'édition 2017.

Les colonnes *Moyenne* contiennent les résultats de la moyenne pondérée des résultats obtenus sur les deux sous-corpus en fonction du nombre de paires disponibles dans ces sous-corpus (respectivement 301 et 294).

Les résultats dans le Tableau 5.8 confirment la pertinence de l'emploi des deux pondérations de mots (morphosyntaxique et fréquentielle) utilisées conjointement. Cela est aussi montré dans les travaux de Nagoudi *et al.* (2017b,a).

Méthode	News	Multi	Moyenne
CL-C3G (fréquence normalisée des termes)	74,47	63,48	69,04
CL-C3G (double normalisation de 0,5)	75,22	65,51	70,42
CL-WES	63,42	39,71	51,71
CL-WES (avec pondération fréquentielle)	68,14	58,58	63,42
CL-WES (avec pondération morphosyntaxique)	67,05	61,33	64,22
CL-WESFS	70,28	63,13	66,75
CL-CTS	84,78	73,56	79,24
CL-CTS (avec pondération fréquentielle)	89,98	79,84	84,97
CL-CTS (avec pondération morphosyntaxique)	88,14	79,93	84,08
CL-CT-WESFS	90,73	82,80	86,81

Tableau 5.8 – Résultats de nos méthodes individuelles, en fonction des pondérations utilisées et des optimisations apportées, sur le corpus d'évaluation de la tâche STS translingue espagnol-anglais de SemEval 2016.

Dans le Tableau 5.9, les résultats en rouge correspondent aux meilleurs scores de corrélation dans chaque catégorie de systèmes (méthode individuelle, fusion non-supervisée et fusion supervisée). Les systèmes marqués d'une • correspondent aux méthodes ayant obtenu le meilleur score moyen dans chaque catégorie de systèmes (méthode individuelle, fusion non-supervisée et fusion supervisée). Ce sont donc ces systèmes que nous décidons de soumettre à l'évaluation 2017.

Les jeux de chiffres caractérisant les fusions par moyenne représentent les méthodes individuelles dont elles font la moyenne. Par exemple, la fusion par moyenne (1-2-4), qui se trouve être

la meilleure fusion par moyenne, utilise les scores des méthodes individuelles CL-C3G (identifiée par le 1), CL-CT-WESFS (identifiée par le 2) et T+WA-IDF (identifiée par le 4).

Les noms des différents systèmes supervisés correspondent aux noms employés dans *Weka 3.8.0* (Hall *et al.*, 2009) pour identifier les différentes approches de régression compatibles avec notre jeu de données (une référence a été reportée lorsqu'elle se trouvait dans l'outil *Weka 3.8.0*). Les scores des méthodes supervisées sont obtenus avec une validation croisée de 10 lancers (*10-fold cross-validation*), étant donné que le corpus d'apprentissage est le même que celui de test. Ces résultats peuvent donc surestimer les performances des méthodes supervisées mais nous nous en sommes tout de même servis pour décider quelle méthode utiliser pour l'évaluation 2017.

Méthode	News	Multi	Moyenne
Systèmes non-supervisés			
CL-C3G (1)	75,22	65,50	70,42
CL-CT-WESFS (2) •	90,72	82,83	86,82
CL-WESFS (3)	70,28	63,12	66,74
T+WA-IDF (4)	90,60	81,44	86,07
T+WA-IDF original (Brychcin *et* Svoboda, 2016)	91,23	80,81	86,08
Fusion par moyenne (1-2-3-4)	85,89	78,24	82,11
Fusion par moyenne (1-2-4) •	90,51	83,47	87,03
Fusion par moyenne (2-3-4)	89,23	82,39	85,85
Fusion par moyenne (2-4)	90,82	82,99	86,95
Systèmes de fusion supervisés			
GaussianProcesses	87,12	78,84	83,03
LinearRegression	90,99	84,14	87,61
MultilayerPerceptron	89,66	79,99	84,88
SimpleLinearRegression	90,48	81,44	86,01
SMOreg (Shevade *et al.*, 2000; Smola *et* Scölkopf, 2004)	90,71	83,75	87,27
Ibk (Aha *et al.*, 1991)	83,96	73,30	78,69
Kstar (Cleary *et* Trigg, 1995)	85,45	81,73	83,61
LWL (Frank *et al.*, 2003)	85,72	75,89	80,86
DecisionTable (Kohavi, 1995)	91,39	80,47	85,99
M5Rules (Holmes *et al.*, 1999; Quinlan, 1992)	91,46	84,06	87,80
DecisionStump	83,29	73,80	78,60
M5P (Quinlan, 1992; Wang *et* Witten, 1997) •	91,54	84,42	88,02
RandomForest (Breiman, 2001)	91,09	84,18	87,68
RandomTree	83,64	72,62	78,19
REPTree	89,72	79,92	84,88
Méthode gagnante de 2016 (Brychcin *et* Svoboda, 2016)	90,62	81,89	86,31

Tableau 5.9 – Résultats de nos systèmes sur le corpus d'évaluation de la tâche STS translingue espagnol-anglais de SemEval 2016.

Les méthodes de régression *LinearRegression*, *SMOreg*, *M5Rules*, *RandomForest* et *M5P* sont les plus performantes sur nos données. Pour cette dernière, c'est environ 1,2 de plus que la meilleure méthode individuelle et 1,71 de plus que la méthode classée première lors de la tâche STS translingue de SemEval 2016. La méthode *M5P* obtient également une meilleure corrélation d'environ 1 point que la meilleure fusion par moyenne testée. La meilleure méthode individuelle est indiscutablement CL-CT-WESFS corrélée à 86,82 avec les annotations de référence (*gold standard*). La deuxième meilleure méthode individuelle, T+WA-IDF (voir section 2.2.5), est notre implémentation de la méthode non-supervisée classée première lors de l'édition 2016. Les résultats de l'implémentation originale des auteurs Brychcin *et* Svoboda (2016) de cette méthode

étaient corrélés en moyenne sur les deux sous-corpus à 86,08 dans les résultats officiels de SemEval 2016 quand la nôtre l'est à 86,07. Cette différence peut paraître minime mais lorsque l'on observe individuellement chacun des sous-corpus on se rend compte que les différences observées oscillent à plus de 0,6. Ces différences sont probablement causées par une évolution de la traduction automatique effectuée par *Google Translate* (son algorithme ou sa base de données) en une année écoulée ou par notre implémentation des fréquences inverses de document (*idf*) dans l'indice de Jaccard (Jaccard, 1912).

Au vu de ces résultats, les trois systèmes soumis lors de l'édition 2017 sont donc :

— la méthode CL-CT-WESFS telle que décrite en section 5.1.3 et paramétrée selon l'étude de la section 5.3.3.1 ;
— la fusion par moyenne (voir section 5.2.1) sur les méthodes CL-C3G, CL-WESFS et T+WA-IDF ;
— le système de régression *M5P* sur toutes les méthodes individuelles (toutes les méthodes décrites en section 5.4.2.1).

Concernant le système *M5P* qui est la méthode de fusion supervisée que nous avons décidé d'employer, elle est en fait l'implémentation dans *Weka 3.8.0* (Hall *et al.*, 2009) du modèle d'arbre M5′ (Wang *et* Witten, 1997). La première implémentation des modèles d'arbre (*model tree*) fut le *M5* proposé par Quinlan (1992) et l'approche fut ensuite redéfinie et améliorée dans un système appelé M5′ par Wang *et* Witten (1997). Les modèles d'arbre possèdent une structure d'arbre de décision conventionnel mais utilisent des fonctions de régression linéaire en leurs feuilles plutôt que de simples étiquettes de classes discrètes.

5.4.3 Résultats de l'évaluation officielle lors de l'édition 2017

5.4.3.1 Résultats

Comme dans la tâche de 2016, le corpus de la sous-tâche translingue espagnol-anglais de 2017 est divisé en deux sous-tâches et donc deux sous-corpus. Les corpus et le système d'évaluation sont présentés dans l'article de description de la tâche (Cer *et al.*, 2017) ainsi que dans la section 3.6.

Le Tableau 5.10 recense les scores de nos méthodes sur le corpus officiel d'évaluation de 2017. Les résultats en rouge correspondent aux meilleurs scores de corrélation dans chaque catégorie de systèmes (méthode individuelle, fusion non-supervisée et fusion supervisée). Les systèmes marqués d'une • correspondent aux méthodes que nous avons soumises, leur classement au sein de la compétition est donné entre parenthèses, et sont donc nos scores officiels présentés dans le papier de description de la tâche des organisateurs (Cer *et al.*, 2017).

Nous constatons que les systèmes que nous avons soumis sont également ceux qui ont les meilleurs résultats moyens. Toutefois, le système *M5P* reste meilleur sur l'ensemble des deux corpus malgré le fait que le système *LinearRegression* fait mieux que ce dernier sur le corpus SNLI.

Nous constatons également que nos systèmes fonctionnent bien sur le sous-corpus SNLI, sur lequel nous nous classons 1^{er} dans la compétition (sur 53 systèmes soumis pour 20 équipes participantes), grâce au système *M5P* (Wang *et* Witten, 1997) qui obtient 83,02 de corrélation de Pearson (Galton, 1886; Pearson, 1895) avec les jugements humains. Inversement, les corrélations sur le sous-corpus WMT, sur lequel nous nous classons seulement $6^{\text{ème}}$ ($11^{\text{ème}}$ système mais $6^{\text{ème}}$ équipe), sont étrangement basses. Toutefois, nous observons que cette différence est notable sur les scores de toutes les équipes participantes (Cer *et al.*, 2017). Le meilleur score pour cette sous-tâche est de 34 quand il avoisine les 85 pour les autres sous-tâches de STS (les sous-tâches 1, 2, 3, 4a, 5 et 6). De même, la méthode *baseline* proposée par les organisateurs obtient un score de 3,20 pour cette sous-tâche quand elle oscille entre 51 et 71 pour les autres sous-tâches. Cela est peut-être dû au fait que les *gold standard* du sous-corpus WMT ont été annotés par seulement un annotateur, quand les corpus des autres tâches l'ont été par plusieurs (5 annotateurs pour la plupart). Nous étudions ce phénomène en section 5.4.3.2, en proposant nos propres annotations et en analysant les corrélations de nos systèmes sur ces dernières.

Méthode	SNLI (4a)	WMT (4b)	Moyenne
Systèmes non-supervisés			
CL-C3G	75,41	8,26	41,84
CL-CT-WESFS •	(9) 76,84	(15) 14,64	45,74
CL-WESFS	47,51	6,57	27,04
T+WA-IDF	76,01	12,45	44,23
Fusion par moyenne •	(3) 79,10	(12) 14,94	47,02
Systèmes de fusion supervisés			
GaussianProcesses	76,43	-12,48	0,3198
LinearRegression	83,23	11,67	47,45
MultilayerPerceptron	75,97	1,73	38,85
SimpleLinearRegression	79,48	0,73	40,10
SMOreg (Shevade *et al.*, 2000; Smola *et* Scölkopf, 2004)	82,91	13,34	48,12
Ibk (Aha *et al.*, 1991)	67,11	10,97	39,04
Kstar (Cleary *et* Trigg, 1995)	75,99	11,05	43,52
LWL (Frank *et al.*, 2003)	71,99	6,88	39,43
DecisionTable (Kohavi, 1995)	76,53	8,92	42,72
M5Rules (Holmes *et al.*, 1999; Quinlan, 1992)	82,74	0,64	41,69
DecisionStump	66,79	-14,86	25,96
M5P (Quinlan, 1992; Wang *et* Witten, 1997) •	(1) 83,02	(11) 15,50	49,26
RandomForest (Breiman, 2001)	78,70	12,04	45,37
RandomTree	64,14	11,62	37,88
REPTree	75,69	-0,43	37,63

Tableau 5.10 – Résultats officiels de nos systèmes sur les corpus de la sous-tâche 4 de SemEval 2017 (53 systèmes soumis par les participants, au total).

5.4.3.2 Discussion

Nous cherchons ici à comprendre pourquoi les résultats des méthodes soumises sont (anormalement) plus faibles sur le sous-corpus (WMT) de la sous-tâche 4b que dans toutes les autres sous-tâches, y compris sur le sous corpus (SNLI) de la sous-tâche 4a. Pour cela, trois personnes (mes deux directeurs de thèse et moi-même) ont annoté manuellement 60 paires (20 chacun) tirées aléatoirement dans chacun des deux sous-corpus de la tâche qui nous intéresse (SNLI et WMT), soit un total de 120 paires (sur 500 disponibles). Nous décidons ensuite d'évaluer les systèmes que nous avons soumis sur ces paires, en comparant les corrélations qu'elles obtiennent sur nos annotations et sur celles des *gold standard* officiels des organisateurs [14].

Nous pouvons voir dans le Tableau 5.11 que, sur le corps SNLI, nos méthodes se comportent de façon similaire sur les deux annotations (une différence de corrélation d'environ 1,3 entre nos annotations et les annotations des organisateurs). Toutefois, la différence de corrélation est importante entre nos annotations et celles de référence des organisateurs sur le corpus WMT, allant de 18,22 à 36,68 (une différence de corrélation de 30 en moyenne).

Nous pouvons également comparer la corrélation de Pearson (Galton, 1886; Pearson, 1895) entre nos annotations et celles des organisateurs. Elle est de 0,8576 sur le corpus SNLI contre 0,2916 sur le corpus WMT. Ces résultats questionnent la validité des annotations du corpus WMT pour la détection de similarité textuelle sémantique translingue (nous avons échangé à ce sujet avec les organisateurs de la tâche).

14. C'est pourquoi les résultats obtenus par nos méthodes ne sont pas les mêmes dans le Tableau 5.10 et le Tableau 5.11. Les résultats observables dans le Tableau 5.10 sont les corrélations obtenues sur 500 paires, tandis que les résultats observables dans le Tableau 5.11 sont les corrélations obtenues sur 120 paires.

Methods	SNLI (4a)	WMT (4b)	Moyenne
Nos annotations			
CL-CT-WESFS	79,81	52,48	66,14
Average	81,05	40,31	60,68
M5P	86,22	53,74	69,98
Annotations de référence de SemEval			
CL-CT-WESFS	81,23	17,39	49,31
Average	82,77	22,09	52,43
M5P	85,36	17,06	51,21

Tableau 5.11 – Corrélations de Pearson entre les résultats de nos systèmes soumis et nos annotations sur 120 paires et sur les *gold standard* des mêmes 120 paires.

5.5 Conclusion

Pour conclure, au cours de ce chapitre, nous avons présenté de nouvelles méthodes de détection de similarités textuelles sémantiques translingues fondées sur des représentations distributionnelles distribuées continues de mots et nous avons également augmenté une méthode de l'état de l'art en y introduisant ces représentations. Nous avons ensuite introduit une notion de pondération morphosyntaxique et fréquentielle de mots qui peut aussi bien être utilisée au sein d'un vecteur que d'un sac de mots et nous montrons que son apport dans les nouvelles méthodes présentées augmente leurs performances respectives. La nouvelle approche la plus prometteuse est une distance cosinus entre les représentations distributionnelles distribuées continues pondérées morphosyntaxiquement et fréquentiellement de deux phrases (CL-WESFS), qui se trouve être significativement plus performante que la meilleure méthode de l'art (CL-C3G) ainsi que toutes les autres méthodes que nous avons pu évaluer jusque là. Par exemple, durant notre évaluation, elle obtient des performances en moyenne meilleures de 4% par rapport à la méthode CL-C3G sur la granularité syntagmatique et des performances en moyenne meilleures de 7% à la granularité phrastique. Finalement, nous avons également démontré que toutes les méthodes précédemment évaluées, aussi bien celles de l'état de l'art que celles nouvellement introduites, sont complémentaires et que leur fusion obtient de meilleures performances que chacune de ces méthodes prises individuellement. Un système de combinaison de méthodes fondé sur un arbre de décision se trouve être la fusion donnant les meilleures performances, avec des F-mesures moyennes de plus de 80 sur les deux granularités.

Afin de mesurer l'efficacité de nos méthodes à un état de l'art plus récent, nous avons décidé de participer à la tâche de détection de similarité textuelle sémantique (STS) translingue de SemEval 2017. Au cours de cette évaluation, nous avons une nouvelle fois prouvé l'utilité de nos méthodes et de leurs fusions en remportant l'une des sous-tâches espagnol-anglais. De plus, notre méthode CL-WESFS s'avère également efficace en monolingue sur des langues telles que l'arabe. En effet, nous sommes également co-auteurs d'une soumission sur la sous-tâche arabe-arabe de la tâche STS de SemEval 2017, soumission qui s'est classée 2^{de} sur 49 systèmes participants en présentant la méthode CL-WESFS sans tenir compte *conjointement* des deux pondérations (morphosyntaxique et fréquentielle).

L'ensemble de ces résultats confirme que les différentes méthodes proposées sont complémentaires et que l'usage de représentations distributionnelles distribuées continues de mots peut s'avérer utile dans la détection automatique de plagiat (translingue).

Conclusion

« Rien de plus original, rien de plus soi que de se nourrir des autres. Mais il faut les digérer. Le lion est fait de mouton assimilé. »

Tel Quel (1941)
— Paul Valéry (1871-1945)

Bilan

Ce manuscrit rapporte le travail de thèse réalisé pour construire un système de mesure de similarité textuelle sémantique translingue dans le but de l'intégrer au sein d'un logiciel de détection de plagiat translingue.

Dès le début de nos travaux, nous avons rencontré des contraintes méthodologiques comme l'absence d'un corpus de données pouvant servir à l'évaluation de méthodes de détection du plagiat translingue et donc également l'absence d'une évaluation rigoureuse des méthodes existantes de détection du plagiat translingue.

La première contribution de la thèse a donc été de constituer un corpus qui répond au problème principal du manque de données déjà existantes pour l'évaluation de la détection du plagiat translingue. En effet, en raison de l'intérêt (trop) récent pour les cas translingues, de la gestion de la vie privée et l'anonymat des données, et des difficultés d'annotations des sources de données, nos recherches de corpus existants suffisamment complets se sont avérées infructueuses. Pris séparément, nous pensons que les corpus déjà existants pour l'évaluation de la détection du plagiat translingue ne sont pas assez diversifiés pour servir à eux seuls à une évaluation rigoureuse de méthodes de détection du plagiat translingue. Ils couvrent, pour la plupart, seulement un domaine (littéraire, législatif, légendes d'images) et sont principalement issus de corpus comparables. Ils sont traduits manuellement, écrits par la même catégorie d'auteurs et disposent de la même granularité de passages plagiés. Il est donc difficile de tirer, des évaluations effectuées sur ces corpus, des conclusions qui peuvent être prises comme cas général et qui peuvent donc être exploitées pour implémenter un outil commercial à destination d'une clientèle exigeante. C'est pourquoi nous avons construit un corpus multilingue, multi-genre et multi-granularité, aussi diversifié que possible, qui permet ainsi une évaluation rigoureuse des méthodes de détection du plagiat translingue. Nous avons également mis au point un protocole d'évaluation reproductible basé sur ce corpus. Le corpus et le protocole peuvent s'avérer utiles pour de futures tâches d'évaluation de la détection de similarité textuelle sémantique translingue. C'est en tout cas dans ce but que nous les partageons avec la communauté [15].

La seconde contribution est une évaluation minutieuse et rigoureuse de certaines méthodes de l'état de l'art à l'aide de notre corpus et de son protocole.

Nous avons conduit une étude des performances de certaines méthodes de l'état de l'art de détection du plagiat translingue sur notre corpus. Peu importe la langue source ou cible ou encore la granularité, l'approche CL-C3G (une méthode à base de similarité de vecteurs de 3-grammes

15. https://github.com/FerreroJeremy/Cross-Language-Dataset (consulté le 31/08/2017 à 17h)

de caractères) est généralement la plus performante. Nos résultats confirment que les différentes méthodes de l'état de l'art se comportent différemment en fonction des caractéristiques des textes comparés mais aussi que la taille de ces textes impacte leurs performances. Notre expérimentation montre également un comportement corrélé des méthodes à travers les différentes paires de langues testées. Cela signifie que si une méthode est efficace sur une paire de langues particulière, elle sera tout autant efficace sur une autre paire de langues, tant que des ressources lexicales de qualité sont disponibles pour ces langues. On remarque également que quand une méthode est plus efficace qu'une autre sur un corpus (suffisamment large), elle est en règle générale plus efficace également sur les autres sous-corpus. Cela signifie qu'une méthode peut être paramétrée sur un corpus et ensuite être appliquée efficacement sur un autre corpus si besoin.

Finalement, nous avons également montré que les méthodes se comportent différemment même si elles semblent similaires en termes de performances. Cela encourage une fusion de méthodes qui conduit à des résultats bien meilleurs que l'état de l'art.

Notre troisième contribution est la proposition de nouvelles méthodes ainsi que des fusions de ces dernières avec des méthodes issues de l'état de l'art.

Nous avons présenté de nouvelles méthodes de détection de similarités textuelles sémantiques translingues fondées sur des représentations distributionnelles distribuées continues de mots (*word embeddings*) et nous avons également amélioré une méthode de l'état de l'art en y introduisant ce type de représentations. Nous avons ensuite introduit une notion de pondération morphosyntaxique et fréquentielle de mots qui peut aussi bien être utilisée au sein d'un vecteur que d'un sac de mots et nous montrons que son apport dans les nouvelles méthodes présentées améliore leurs performances respectives. La nouvelle approche ainsi introduite la plus prometteuse est une distance cosinus entre les représentations continues pondérées morphosyntaxiquement et fréquentiellement de deux phrases (CL-WESFS), qui se trouve être significativement plus performante que toutes les autres méthodes que nous avons pu évaluer jusqu'ici. Par exemple, durant notre évaluation, elle obtient des performances en moyenne meilleures de 5% par rapport à la meilleure méthode de l'état de l'art (CL-C3G). Finalement, nous avons confirmé que toutes les méthodes précédemment évaluées, aussi bien celles de l'état de l'art que celles nouvellement introduites, sont complémentaires et que leurs fusions obtiennent de meilleures performances que chacune d'elles prises individuellement. Nous avons dans un premier temps évalué une fusion linéaire par moyenne pondérée. Cette fusion obtenait déjà de très bonnes performances par rapport à l'état de l'art. Mais notre meilleure contribution est une combinaison de méthodes basée sur un arbre de décision, qui obtient des F-mesures moyennes de plus de 86 sur les sous-corpus évalués.

La dernière contribution de la thèse est donc la confirmation que les différentes méthodes proposées sont complémentaires les unes des autres et que l'usage de représentations continues de mots et des combinaisons de méthodes peuvent s'avérer utiles dans la détection automatique de plagiat translingue.

Pour finir, afin de comparer nos méthodes à un état de l'art plus récent, nous avons décidé de participer aux tâches espagnol-anglais de détection de similarité textuelle sémantique (STS) translingue de SemEval 2017. Au cours de cette évaluation, nous avons une nouvelle fois prouvé l'utilité de nos méthodes et de leurs fusions en remportant l'une des sous-tâches, où nous nous sommes classés 1^{er} sur 53 systèmes soumis par plus de 20 équipes participantes, grâce à un système de régression basé sur un arbre de décision intégrant nos nouvelles méthodes fondées sur des représentations continues de mots. De plus, notre méthode CL-WESFS s'avère également efficace en monolingue sur des langues telles que l'arabe. En effet, nous sommes co-auteurs d'une soumission sur la sous-tâche arabe-arabe, soumission qui s'est classée 2^{de} sur 49 systèmes participants en présentant la méthode CL-WESFS sans tenir compte *conjointement* des deux pondérations (morphosyntaxique et fréquentielle).

Perspectives

Tout d'abord, le corpus que nous avons constitué peut encore évoluer. Dans ce sens, nous invitons la communauté à l'améliorer et à l'étendre. Parmi les piste possibles, on peut citer :

— l'ajout d'alignements provenant de nouveaux corpus, comme ceux de SemEval (Cer *et al.*, 2017), BUCC (Zweigenbaum *et al.*, 2016, 2017) ou SNLI (Bowman *et al.*, 2015) ;
— l'ajout de nouvelles langues dans les sous-corpus déjà présents, comme les langues européennes de JRC-Acquis (Steinberger *et al.*, 2006) ou Europarl (Koehn, 2005) ;
— l'ajout de sous-corpus avec plus d'entités nommées afin de vérifier le réel impact de cette caractéristique sur l'efficacité des méthodes de détection ;
— l'ajout de différents types et différents niveaux d'offuscation ;
— l'ajout de syntagmes verbaux ou adverbiaux (aussi porteurs de sens).

D'autre part, nous avons montré que les représentations continues de mots (*word embeddings*) peuvent être utiles dans le cadre de la détection de similarité textuelle sémantique translingue. Nous pensons qu'il reste des aspects à explorer dans cette voie. Il serait, par exemple, opportun d'optimiser l'apprentissage des représentations en utilisant des corpus plus vastes pour améliorer la couverture de vocabulaire ou au contraire utiliser des corpus plus spécifiques à la tâche. Par ailleurs, nous avons pris comme décision de construire des représentations de phrases à partir de représentations de mots déjà existantes. Nous avons ainsi montré que le fait de pondérer individuellement chaque mot ou groupe de mots peut s'avérer efficace. Cela ouvre la porte à de nouvelles tentatives de pondérations et à un apprentissage de représentations dédiées à la tâche de similarité translingue. Des travaux futurs pourraient s'orienter vers une pondération qui se baserait sur le rôle sémantique des mots (sujet, objet, *etc.*) ou sur leur degré d'ambiguïté (qui dépendrait, par exemple, du nombre de synonymes associés à un mot – en faisant l'hypothèse que moins un mot dispose de synonymes, plus il est représentatif et évocateur de l'idée qu'il exprime). Il serait aussi intéressant de comparer cette approche de pondération du modèle une fois ce dernier construit par rapport à une approche pré-construction (filtrant les mots à injecter dans le modèle en fonction de leur pondération) ou une approche introduisant des informations directement en entrée dans la construction du modèle.

De plus, il devrait également être possible de travailler dans un vrai espace multilingue, c'est-à-dire de disposer de représentations continues de mots pour n langues ($n > 2$) dans un seul et unique espace vectoriel. Pour cela, on pourrait, par exemple, se servir des travaux de Duong *et al.* (2017) qui proposent divers algorithmes pour construire un modèle *word embeddings* multilingue dans un même espace vectoriel unifié. Un tel modèle pourrait, par exemple, être appris à partir de corpus multi-parallèles tels que JRC-Acquis (Steinberger *et al.*, 2006) ou Europarl (Koehn, 2005). Cela rendrait nos mesures robustes à tout changement de langue dans les documents comparés et rendrait également possible des comparaisons à travers des corpus multilingues qui disposent de documents dans plusieurs langues différentes (comme c'est souvent le cas sur Internet).

Nous avons également démontré que certaines méthodes étaient complémentaires et qu'une fusion de ces dernières pouvait obtenir de meilleures performances que l'état de l'art. En effet, certaines méthodes semblent être complémentaires deux à deux, certaines semblent être totalement inutiles ou redondantes avec d'autres et enfin certaines semblent apporter un réel plus en toutes circonstances. Durant notre évaluation, nous avons fait le choix d'un système de régression par arbre de décision (Quinlan, 1993). Notre choix a ensuite été confirmé par de bons résultats lors de notre participation à l'évaluation SemEval 2017 (Ferrero *et al.*, 2017a), où nous avons utilisé un autre système de régression basé sur un arbre de décision (Wang *et* Witten, 1997). Nous suggérons donc de continuer dans cette voie en comparant différents algorithmes de régression par arbre de décision, ces derniers semblant montrer de bons résultats dans cette tâche, ou bien en se tournant vers des algorithmes de fusion par classifieurs faibles (*boosting*) (Breiman, 1996) qui pourraient également répondre à ce type de problématique. Par ailleurs, ces fusions sont des approches par apprentissage automatique supervisé et nécessitent donc des

corpus d'apprentissage de données annotées qui sont particulièrement rares dans ce domaine. Une piste pourrait donc être de mieux alimenter ces systèmes de fusions en leur fournissant de plus amples données d'apprentissage. Ceci nécessiterait de mettre en place une procédure de collecte de données auprès d'utilisateurs réels (autre perspective plus opérationnelle mais néanmois réaliste vu le contexte industriel chez Compilatio).

Enfin, nous rappelons que nous avons, durant cette thèse, traité uniquement de la mesure de similarité textuelle sémantique translingue et que le but final est de détecter du plagiat translingue. Dans un contexte industriel opérationnel de détection de plagiat, il faut en amont effectuer une recherche de sources candidates (sur le Web ou dans une base documentaire interne). Ces documents seront par la suite comparés deux à deux avec le document suspect analysé à l'aide de nos mesures de similarité textuelle sémantique. Dans le cas translingue, cette recherche de sources est un peu plus complexe et de futurs travaux devront donc traiter cette problématique afin qu'une solution de détection du plagiat complète de bout en bout puisse voir le jour.

De plus, dans un contexte applicatif de détection du plagiat, il ne s'agit plus de valider ou non la similarité de deux unités textuelles mais plutôt de comparer des textes complets deux à deux afin d'identifier des possibles passages similaires entre ces deux textes. L'une des façons les plus courantes de procéder au sein de l'état de l'art est de fonctionner par granularité (Potthast *et al.*, 2010b). On découpe les deux textes en fragments et on compare chaque couple de fragments possibles deux à deux. C'est la multiplicité de fragments similaires, plus ou moins proches dans les deux textes à la fois, qui va faire en sorte que l'on étiquettera, comme plagiée ou non, toute la zone formée par ces fragments. En d'autres termes, si deux fragments similaires ou plus apparaissent dans les deux textes, dans le même ordre, et suffisamment proches dans les deux textes à la fois, alors on les regroupe pour former une seule et unique zone plagiée dans les deux textes. Ainsi, un fragment non similaire, qui s'inscrit entre deux fragments similaires, est considéré comme similaire. Ce procédé est illustré dans la Figure 5.3.

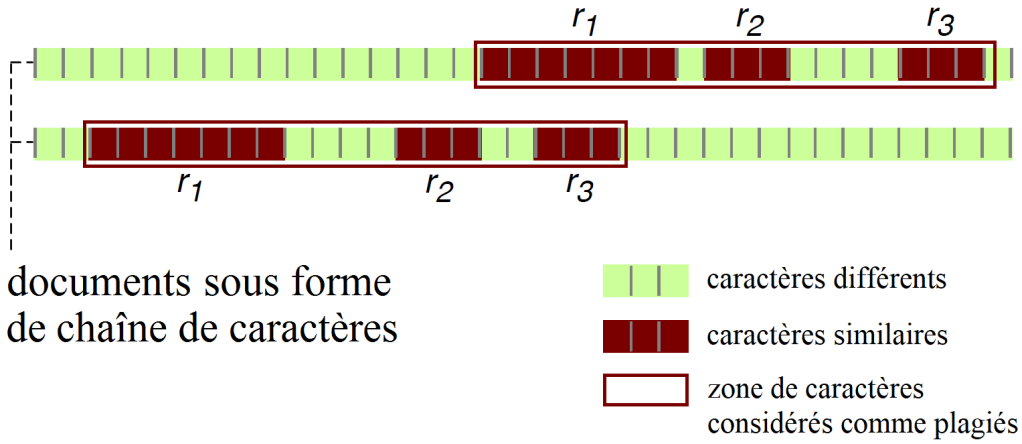


FIGURE 5.3 – Exemple de cas de construction de zone de plagiat à partir de fragments identiques entre deux textes.

L'une des données supplémentaires qui peut donc être prise en compte lors d'une approche de détection de plagiat est cette notion de granularité qui change plus ou moins l'aspect de la tâche. Dans ce cas, si l'on procède ainsi, rien ne nous empêche d'inclure une notion d'incertitude à l'échelle des mesures de similarités textuelles sémantiques. En effet, tout au long de cette thèse, nous avons émis l'hypothèse, à l'aide d'un système de seuil, qu'une unité textuelle était, soit similaire à une autre unité textuelle, soit non similaire. Cette hypothèse est un peu simpliste, une solution serait de rendre nos méthodes capables de considérer l'incertitude sur la similarité de deux fragments, en conservant, par exemple, le pourcentage de similarité mesuré plutôt que de déterminer une étiquette plagiée ou non en fonction de ce pourcentage. Ainsi, une méthode pourra

juger deux fragments comme plus ou moins neutres et les fragments identifiés comme neutres pourront alors être ignorés lors de la phase finale de regroupement des fragments similaires pour former les zones plagiées. Une autre solution pourrait être d'affecter un poids à chaque fragment, qui jouera alors un rôle plus ou moins important dans la prise de décision pour l'étiquetage final de la zone.

Enfin, nos mesures ont montré leur utilité dans l'identification de similarité textuelle sémantique translingue mais elles peuvent aussi s'appliquer à un contexte monolingue, comme démontré dans les travaux de (Nagoudi *et al.*, 2017b). Par exemple, nos mesures peuvent s'appliquer dans la détection de reformulations. Pour cela, il faudra utiliser des dictionnaires de synonymes à la place de dictionnaires de traductions, des représentations continues monolingues plutôt que translingues, ou bien encore adapter certains pré-traitements. De plus, ces mesures peuvent être utiles pour la détection du plagiat, mais peuvent également être utiles pour la désambiguïsation sémantique, la recherche translingue de documents, la classification textuelle ou encore l'évaluation de traduction automatique.

Index

Bibliographie

Yosshi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, *et* Yoav Goldberg. 2017. Fine-grained Analysis of Sentence Embeddings using Auxiliary Prediction Tasks. Dans *P*roceedings of the 5th International Conference on Learning Representations (ICLR 2017). Toulon, France. Avril. 2017, https://openreview.net/pdf?id=BJh6Ztuxl. *citée page 57*

Željko Agić *et* Natalie Schluter. 2017. Baselines and test data for cross-lingual inference. Dans *a*rXiv. Copenhague, Danemark. Septembre. 2017, https://arxiv.org/pdf/1704.05347.pdf. *2 citations pages 68 et 71*

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, *et* Janyce Wiebe. 2016. SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. Dans *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*. San Diego, Californie, États-Unis, pages 497–511. Juin. 2016, http://www.aclweb.org/anthology/S16-1081. *8 citations pages 52, 59, 69, 70, 92, 108, 112, et 113*

David W. Aha, Dennis Kibler, *et* Marc K. Albert. 1991. Instance-based Learning Algorithms. *M*achine Learning 6(1):37–66. https://doi.org/10.1007/BF00153759. *2 citations pages 114 et 116*

Asier Alcázar. 2006. Towards Linguistically Searchable Text. Dans *P*roceedings of the BIDE 2005. Bilbao, Espagne. *citée page 58*

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, *et* Noah A. Smith. 2016. Massively Multilingual Word Embeddings. Dans *a*rXiv. https://arxiv.org/pdf/1602.01925.pdf. *citée page 53*

Gabriella Aragione. 2010. La transmission du savoir entre « tradition » et « plagiat » dans l'Antiquité classique et chrétienne. Dans *Études de lettres*. pages 117–138. https://doi.org/10.4000/edl.388. *citée page 19*

Habibollah Asghari, Khadijeh Khoshnava, Omid Fatemi, *et* Heshaam Faili. 2015. Developing Bilingual Plagiarism Detection Corpus Using Sentence Aligned Parallel Corpus. Dans *W*orking Notes Papers of the CLEF 2015 Evaluation Labs. Toulouse, France, CEUR Workshop Proceedings, pages 8–11. Septembre. 2015, http://ceur-ws.org/Vol-1391/148-CR.pdf. *citée page 83*

Francis R. Bach *et* Michael I. Jordan. 2002. Kernel Independent Component Analysis. *J*ournal of Machine Learning Research 3:1–48. http://www.jmlr.org/papers/volume3/bach02a/bach02a.pdf. *citée page 49*

Ricardo Baeza-Yates *et* Gonzalo Navarro. 1996. A Faster Algorithm for Approximate String Matching. Dans Dan Hirchsberg *et* Gene Myers, éditeurs, *C*ombinatorial Pattern Matching (CPM'96). Irvine, Californie, États-Unis, LNCS 1075, pages 1–23. Juin. 1996, https://doi.org/10.1007/3-540-61258-0_1. *2 citations pages 86 et 101*

Douglas Bagnall. 2016. Authorship clustering using multi-headed recurrent neural networks - Notebook for PAN at CLEF 2016. Dans *N*otebook Papers for PAN at CLEF 2016 LABs and Workshops. http://ceur-ws.org/Vol-1609/16090791.pdf. *citée page 35*

Alberto Barrón-Cedeño. 2012. *On the Mono- and Cross-Language Detection of Text Re-Use and Plagiarism*. Thèse de doctorat, Universitat Politècnica de València, Valence, Espagne. https://doi.org/10.1145/1835449.1835687. *4 citations pages 58, 59, 60, et 65*

Alberto Barrón-Cedeño, Parth Gupta, *et* Paolo Rosso. 2013a. Methods for cross-linguage plagiarism detection. *Knowledge-Based Systems* 50:211–217. https://doi.org/10.1016/j.knosys.2013.06.018. *3 citations pages 58, 59, et 60*

Alberto Barrón-Cedeño, Monica Lestari Paramita, Paul Clough, *et* Paolo Rosso. 2014. A Comparison of Approaches for Measuring Cross-Lingual Similarity of Wikipedia Articles. Dans *36th European Conference on Information Retrieval, ECIR-2014*. Springer Berlin Heidelberg, Amsterdam, Pays-Bas, volume 8416 de *LNCS*, pages 424–429. Avril. 2014, https://doi.org/10.1007/978-3-319-06028-6_36. *3 citations pages 39, 58, et 59*

Alberto Barrón-Cedeño *et* Paolo Rosso. 2009. On Automatic Plagiarism Detection Based on n-Grams Comparison. Dans M Boughanem, éditeur, *Proceedings of the 31st European Conference on Information Retrieval (ECIR'09)*. Springer Berlin, Toulouse, France, LNCS, pages 696–700. Avril. 2009, https://doi.org/10.1007/978-3-642-00958-7_69. *3 citations pages 31, 32, et 52*

Alberto Barrón-Cedeño, Paolo Rosso, Eneko Agirre, *et* Gorka Labaka. 2010. Plagiarism Detection across Distant Language Pairs. Dans *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*. Association for Computational Linguistics, Pékin, Chine, pages 37–45. Août. 2010, http://www.aclweb.org/anthology/C10-1005. *5 citations pages 47, 52, 58, 59, et 60*

Alberto Barrón-Cedeño, Paolo Rosso, Sobha Lalitha Devi, Paul Clough, *et* Mark Stevenson. 2011. PAN@FIRE: Overview of the Cross-Language !ndian Text Re-Use Detection Competition. Dans *Third International Workshop, FIRE 2011*. Springer Berlin Heidelberg, volume 7536 de *Multilingual Information Access in South Asian Languages*, pages 59–70. https://doi.org/10.1007/978-3-642-40087-2_6. *2 citations pages 66 et 71*

Alberto Barrón-Cedeño, Paolo Rosso, David Pinto, *et* Alfons Juan. 2008. On Cross-lingual Plagiarism Analysis using a Statistical Model. Dans Benno Stein and Efstathios Stamatatos and Moshe Koppel, éditeur, *Proceedings of the ECAI'08 PAN Workshop: Uncovering Plagiarism, Authorship and Social Software Misuse*. Patras, Grèce, pages 9–13. http://ceur-ws.org/Vol-377/paper1.pdf. *citée page 46*

Alberto Barrón-Cedeño, Marta Vila, M. Antonia Martì, *et* Paolo Rosso. 2013b. Plagiarism Meets Paraphrasing: Insights for the Next Generation in Automatic Plagiarism Detection. *C*omputational Linguistics 39(4):917–947. https://doi.org/10.1162/COLI_a_00153. *citée page 31*

Thomas Bayes *et* Richard Price. 1763. An Essay towards solving a Problem in the Doctrine of Chance. Dans *P*hilosophical Transactions. Royal Society, 53, pages 370–418. Janvier. 1763, https://doi.org/10.1098/rstl.1763.0053. *citée page 46*

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, *et* Christian Jauvin. 2003. A neural probabilistic language model. *J*ournal of Machine Learning Research (JMLR) 3:1137–1155. http://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf. *2 citations pages 53 et 54*

Alexandre Bérard, Christophe Servan, Olivier Pietquin, *et* Laurent Besacier. 2016. MultiVec: a Multilingual and Multilevel Representation Learning Toolkit for NLP. Dans *P*roceedings of the Tenth International Conference on Language Resources and Evaluation

(LREC'16). European Language Resources Association (ELRA), Portorož, Slovénie, pages 4188–4192. Mai. 2016, http://www.lrec-conf.org/proceedings/lrec2016/pdf/666_Paper.pdf. *5 citations pages 55, 56, 98, 101, et 103*

Frank Vanden Berghen *et* Hugues Bersini. 2005. CONDOR, a new parallel, constrained extension of Powell's UOBYQA algorithm: Experimental results and comparison with the DFO algorithm. *Journal of Computational and Applied Mathematics* 181:157–175. https://doi.org/10.1016/j.cam.2004.11.029. *2 citations pages 104 et 107*

Michael W. Berry *et* Paul G. Young. 1995. Using latent semantic indexing for multilanguage information retrieval. *C*omputers and the Humanities 29(6):413–429. *2 citations pages 47 et 49*

Clive Best, Erik van der Goot, Ken Blackler, Tefilo Garcia, *et* David Horby. 2005. Europe Media Monitor - System Description. Rapport technique, EUR Report 22173-En, Ispra, Italie. *citée page 70*

BIALA. 1910. Actes de la Conférence réunie à Berlin, Berne : Bureau International de l'Association littéraire et artistique. *citée page 20*

William Blacoe *et* Mirella Lapata. 2012. A Comparison of Vector-based Representations for Semantic Composition. Dans *P*roceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics, Jeju-do, Corée du Sud, pages 546–556. Juillet. 2012, http://www.aclweb.org/anthology/D12-1050. *4 citations pages 53, 55, 57, et 98*

John Blitzer, Mark Dredze, *et* Fernando Pereira. 2007. Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. Dans *P*roceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Association of Computational Linguistics, Prague, République tchèque, pages 440–447. Juin. 2007, http://aclweb.org/anthology/P/P07/P07-1056.pdf. *citée page 77*

Piotr Bojanowski, Edouard Grave, Armand Joulin, *et* Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. Dans Hinrich Schutze, éditeur, *T*ransactions of the Association for Computational Linguistics. Association for Computational Linguistics, volume 5, pages 135–146. Juin. 2017, https://transacl.org/ojs/index.php/tacl/article/view/999. *citée page 56*

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, *et* Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. Dans *P*roceedings of the Eighth Workshop on Statistical Machine Translation (WMT13). Association for Computational Linguistics, Sofia, Bulgarie, pages 1–44. http://www.aclweb.org/anthology/W13-2201. *2 citations pages 69 et 71*

Florian Boudin. 2013. TALN Archives: a digital archive of French research articles in Natural Language Processing (TALN Archives : une archive numérique francophone des articles de recherche en Traitement Automatique de la Langue) [in French]). Dans *Proceedings of TALN 2013 (Volume 2: Short Papers)*. Les Sables d'Olonne, France, pages 507–514. Juin. 2013, http://aclweb.org/anthology/F/F13/F13-2001.pdf. *citée page 78*

Samuel R. Bowman, Gabor Angeli, Christopher Potts, *et* Christopher D. Manning. 2015. A Large Annotated Corpus for Learning Natural Language Inference. Dans *P*roceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015). Association for Computational Linguistics, Lisbonne, Portugal. Septembre. 2015, http://aclweb.org/anthology/D/D15/D15-1075.pdf. *4 citations pages 68, 69, 83, et 121*

Leo Breiman. 1996. Bias, Variance, and Arcing Classifiers. Rapport technique. *citée page 121*

Leo Breiman. 2001. Random Forests. *M*achine Learning 45(1):5–32. https://doi.org/10.1023/A:1010933404324. *2 citations pages 114 et 116*

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, *et* Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics* 19(2):263–311. http://www.aclweb.org/anthology/J93-2003. *2 citations pages 46 et 87*

Samuel V. Bruton. 2014. Self-Plagiarism and Textual Recycling: Legitimate Forms of Research Misconduct. *A*ccountability in Research 21(3):176–197. https://doi.org/10.1080/08989621.2014.848071. *citée page 22*

Tomas Brychcin *et* Lukas Svoboda. 2016. UWB at SemEval-2016 Task 1: Semantic textual similarity using lexical, syntactic, and semantic information. Dans *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016)*. San Diego, Californie, États-Unis, pages 588–594. Juin. 2016, https://www.aclweb.org/anthology/S/S16/S16-1089.pdf. *8 citations pages 52, 53, 59, 60, 99, 108, 112, et 114*

Peter Businger *et* Gene H. Golub. 1965. Linear least squares solutions by householder transformations. *N*umerische Mathematik 7(3):269–276. https://doi.org/10.1007/BF01436084. *2 citations pages 47 et 48*

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, *et* Lucia Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. Dans *P*roceeding of WMT 2012. http://aclweb.org/anthology/W/W12/W12-3102.pdf. *citée page 70*

William B. Cavnar *et* John M. Trenkle. 1994. N-Gram-Based Text Categorization. Dans *Proceedings of 3rd Annual Symposium on Document Analysis and Information Retrieval (SDAIR'94)*. pages 161–175. *citée page 79*

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, *et* Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. Dans *P*roceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Association for Computational Linguistics, Vancouver, Canada, pages 1–14. Août. 2017, http://www.aclweb.org/anthology/S17-2001. *9 citations pages 57, 59, 69, 71, 83, 92, 111, 115, et 121*

Mauro Cettolo, Christian Girardi, *et* Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. Dans *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*. pages 261–268. https://wit3.fbk.eu/papers/WIT3-EAMT2012.pdf. *citée page 87*

Danqi Chen *et* Christopher D. Manning. 2014. A Fast and Accurate Dependency Parser using Neural Networks. Dans *P*roceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014). Doha, Qatar, pages 740–750. http://www.aclweb.org/anthology/D14-1082. *citée page 99*

Minmin Chen. 2017. Efficient Vector Representation for Documents through Corruption. Dans *P*roceedings of the 5th International Conference on Learning Representations (ICLR 2017). Toulon, France. Avril. 2017, https://openreview.net/pdf?id=B1Igu2ogg. *citée page 98*

Philipp Cimiano, Antje Schultz, Sergey Sizov, Philipp Sorg, *et* Steffen Staab. 2009. Explicit Versus Latent Concept Models for Cross-Language Information Retrieval. Dans *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI-09)*. Morgan Kaufmann Publishers Inc., Pasadena, Californie, États-Unis, IJCAI'09, pages 1513–1518. Juillet. 2009, https://www.ijcai.org/Proceedings/09/Papers/253.pdf. *3 citations pages 58, 59, et 60*

John G. Cleary *et* Leonard E. Trigg. 1995. K*: An Instance-based Learner Using an Entropic Distance Measure. Dans *P*roceedings of the 12th International Conference on Machine Learning. pages 108–114. *2 citations pages 114 et 116*

Paul Clough. 2003. Old and new challenges in automatic plagiarism detection. Dans *N*ational Plagiarism Advisory Service. pages 391–407. *citée page 61*

Paul Clough *et* Mark Stevenson. 2011. Developing a Corpus of Plagiarised Short Answers. *L*anguage Resources and Evaluation 45(1):5–24. https://doi.org/10.1007/s10579-009-9112-1. *2 citations pages 66 et 70*

Christian Collberg *et* Stephen Kobourov. 2005. Self-Plagiarism in Computer Science. *C*ommunications of the ACM 48(4):88–94. https://doi.org/10.1145/1053291.1053293. *citée page 22*

Collins. 1988. *Cobuild English Dictionary*. Harper Collins Publishers. *citée page 40*

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, *et* Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. Dans *P*roceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017). Copenhague, Danemark, pages 681–691. Septembre. 2017, http://aclweb.org/anthology/D/D17/D17-1071.pdf. *citée page 56*

Nello Cristianini *et* John Shawe-Taylor. 2000. *An introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, New York, États-Unis. *citée page 49*

Fred J. Damerau. 1964. A technique for computer detection and correction of spelling errors. *C*ommunications of the ACM 7(3):171–176. https://doi.org/10.1145/363958.363994. *2 citations pages 86 et 101*

Vera Danilova. 2013. Cross-Language Plagiarism Detection Methods. Dans Galia Angelova, Kalina Bontcheva, *et* Ruslan Mitkov, éditeurs, *Proceedings of the Student Research Workshop associated with RANLP 2013*. Hisarya, Bulgarie, Recent Advances in Natural Language Processing, pages 51–57. Septembre. 2013, http://aclweb.org/anthology/R/R13/R13-2008.pdf. *4 citations pages 36, 37, 45, et 58*

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, *et* Richard Harshman. 1990. Indexing by latent semantic analysis. *J*ournal of the American Society for Information Science 41(6):391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9. *2 citations pages 47 et 49*

Mathieu Dehouck *et* Pascal Denis. 2017. Delexicalized Word Embeddings for Cross-lingual Dependency Parsing. Dans *P*roceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. Association for Computational Linguistics, Valence, Espagne, pages 240–249. Avril. 2017, http://aclweb.org/anthology/E/E17/E17-1023.pdf. *citée page 99*

John DeNero *et* Dan Klein. 2007. Tailoring Word Alignments to Syntactic Machine Translation. Dans *P*roceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Association for Computational Linguistics, Prague, République tchèque, pages 17–24. Juin. 2007, http://www.aclweb.org/anthology/P07-1003. *citée page 52*

Birk Diedenhofen *et* Jochen Musch. 2015. cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. *P*LoS ONE 10(6). https://doi.org/10.1371/journal.pone.0121945. *citée page 110*

Florence Duclaye. 2003. *Apprentissage automatique de relations d'équivalence sémantique à partir du Web*. Thèse de doctorat, Télécom Paris-Tech. https://tel.archives-ouvertes.fr/pastel-00001119/. *citée page 40*

Susan T. Dumais, Todd A. Letsche, Michael L. Littman, *et* Thomas Landauer. 1997. Automatic Cross-language Retrieval Using Latent Semantic Indexing. Dans *AAAI-97 Spring Symposium Series: Cross-Language Text and Speech Retrieval*. pages 18–24. https://www.aaai.org/Papers/Symposia/Spring/1997/SS-97-05/SS97-05-003.pdf. *citée page 47*

Long Duong, Hiroshi Kanayama, Tengfei Ma, Steven Bird, *et* Trevor Cohn. 2017. Multilingual Training of Crosslingual Word Embeddings. Dans *P*roceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. Association for Computational Linguistics, Valence, Espagne, pages 893–903. Avril. 2017, http://aclweb.org/anthology/E/E17/E17-1084.pdf. *2 citations pages 56 et 121*

Philip Edmonds *et* Graeme Hirst. 2002. Near-synonymy and lexical choice. *T*he MIT Press. Computational Linguistics 28(2):105–144. https://doi.org/10.1162/089120102760173625. *citée page 40*

Sven Meyer Zu Eissen *et* Benno Stein. 2006. Intrinsic Plagiarism Detection. Dans Lalmas et al. (Eds.): Advances in Information Retrieval, éditeur, *2*8th European Conference on IR Research, ECIR 2006. Londres, Angleterre. https://doi.org/10.1007/11735106_66. *4 citations pages 29, 31, 32, et 34*

Jessica Enright *et* Grzegorz Kondrak. 2007. A Fast Method for Parallel Document Identification. Dans *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'07)*. Association for Computational Linguistics, Rochester, New York, États-Unis, pages 29–32. Avril. 2007, http://aclweb.org/anthology/N/N07/N07-2008.pdf. *citée page 39*

Cristina España-Bonet, Ádám Csaba Varga, Alberto Barrón-Cedeño, *et* Josef van Genabith. 2017. An Empirical Analysis of NMT-Derived Interlingual Embeddings and their Use in Parallel Sentence Identification. Dans *a*rXiv. https://arxiv.org/pdf/1704.05415.pdf. *citée page 57*

Eurovoc. 1995. Thesaurus Eurovoc. Dans *Volume 2: Subject-Oriented Version*. *citée page 41*

Ferric C. Fang, R. Grant Steen, *et* Arturo Casadevall. 2012. Misconduct accounts for the majority of retracted scientific publications. *N*ational Academy of Sciences of the United States of America (PNAS) 109(42):17028–17033. http://www.pnas.org/content/109/42/17028. *citée page 26*

Jérémy Ferrero, Frédéric Agnès, Laurent Besacier, *et* Didier Schwab. 2016. A Multilingual, Multi-style and Multi-granularity Dataset for Cross-language Textual Similarity Detection. Dans *P*roceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). European Language Resources Association (ELRA), Portorož, Slovénie, pages 4162–4169. ISLRN: 723-785-513-738-2. Mai. 2016, http://www.lrec-conf.org/proceedings/lrec2016/pdf/304_Paper.pdf. *3 citations pages 75, 76, et 84*

Jérémy Ferrero, Laurent Besacier, Didier Schwab, *et* Frédéric Agnès. 2017a. CompiLIG at SemEval-2017 Task 1: Cross-Language Plagiarism Detection Methods for Semantic Textual Similarity. Dans *P*roceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Association for Computational Linguistics, Vancouver, Canada, pages 109–114. Août. 2017, http://aclweb.org/anthology/S17-2012. *3 citations pages 97, 111, et 121*

Jérémy Ferrero, Laurent Besacier, Didier Schwab, *et* Frédéric Agnès. 2017b. Deep Investigation of Cross-Language Plagiarism Detection Methods. Dans *Proceedings of the 10th Workshop on Building and Using Comparable Corpora (BUCC)*. Association for Computational Linguistics, Vancouver, Canada, pages 6–15. Août. 2017, http://aclweb.org/anthology/W17-2502. *2 citations pages 75 et 84*

Jérémy Ferrero, Laurent Besacier, Didier Schwab, *et* Frédéric Agnès. 2017c. Using Word Embedding for Cross-Language Plagiarism Detection. Dans *P*roceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. Association for Computer Linguistics, Valence, Espagne, pages 415–421. Avril. 2017, http://aclweb.org/anthology/E/E17/E17-2066.pdf. *2 citations pages 97 et 110*

Jérémy Ferrero *et* Alain Simac-Lejeune. 2015. Détection et regroupement automatique de style d'écriture dans un texte. Dans *1*5ème conférence internationale sur l'extraction et la gestion des connaissances (EGC 2015). Luxembourg, Luxembourg, pages 23–28. Janvier. 2015, https://hal.archives-ouvertes.fr/hal-01108066/document. *citée page 34*

Marc Franco-Salvador, Imene Bensalem, Enrique Flores, Parth Gupta, *et* Paolo Rosso. 2015. PAN 2015 Shared Task on Plagiarism Detection: Evaluation of Corpora for Text Alignment. Dans *N*otebook Papers for PAN at CLEF 2015 LABs and Workshops. Toulouse, France, CEUR Workshop Proceedings. Septembre. 2015, http://ceur-ws.org/Vol-1391/inv-pap12-CR.pdf. *citée page 83*

Marc Franco-Salvador, Parth Gupta, *et* Paolo Rosso. 2013a. Cross-Language Plagiarism Detection using a Multilingual Semantic Network. Dans *3*5th European Conference on Information Retrieval (ECIR'13). Springer Berlin Heidelberg, LNCS 7814, pages 710–713. https://doi.org/10.1007/978-3-642-36973-5_66. *citée page 45*

Marc Franco-Salvador, Parth Gupta, *et* Paolo Rosso. 2013b. Graph-based similarity analysis: a new approach to cross-language plagiarism detection. *J*ournal of the Spanish Society of Natural Language Processing (Sociedad Espaola de Procesamiento del Languaje Natural) 50. *citée page 45*

Marc Franco-Salvador, Parth Gupta, *et* Paolo Rosso. 2013c. Knowledge Graphs as Context Models: Improving the Detection of Cross-Language Plagiarism with Paraphrasing. Dans *Bridging Between Information Retrieval and Databases. PROMISE Winter School 2013*. Springer Berlin Heidelberg, Bressanone, Italie, 8173, pages 227–236. *3 citations pages 40, 41, et 45*

Marc Franco-Salvador, Paolo Rosso, *et* Roberto Navigli. 2014. A Knowledge-based Representation for Cross-Language Document Retrieval and Categorization. Dans *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Göteborg, Suède, pages 414–423. Avril. 2014, http://www.aclweb.org/anthology/E14-1044. *citée page 45*

Marc Franco-Salvador, Paolo Rossoa, *et* Manuel Montes-y-Gómez. 2016. A systematic study of knowledge graph analysis for cross-language plagiarism detection. Dans *Information Processing and Management*. volume 52, pages 550–570. Juillet. 2016, https://doi.org/10.1016/j.ipm.2015.12.004. *2 citations pages 58 et 59*

Marc Franco-Salvadora, Parth Guptaa, Paolo Rossoa, *et* Rafael E. Banchsb. 2016. Cross-language plagiarism detection over continuous-space- and knowledge graph-based representations of language. *K*nowledge-Based Systems 111:87–99. https://doi.org/10.1016/j.knosys.2016.08.004. *citée page 57*

Gil Francopoulo, Joseph Mariani, *et* Patrick Paroubek. 2016. A Study of Reuse and Plagiarism in LREC papers. Dans *P*roceedings of the Tenth International Conference

on Language Resources and Evaluation (LREC'16). European Language Resources Association (ELRA), Portorož, Slovénie, pages 1890–1897. Mai. 2016, http://www.lrec-conf.org/proceedings/lrec2016/pdf/85_Paper.pdf. *citée page 27*

Eibe Frank, Mark Hall, *et* Bernhard Pfahringer. 2003. Locally Weighted Naive Bayes. Dans *P*roceedings of the 19th Conference in Uncertainty in Artificial Intelligence. pages 249–256. *2 citations pages 114 et 116*

Evgeniy Gabrilovich *et* Shaul Markovitch. 2007. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. Dans *Proceedings of the 20th International Joint Conference on Artifical Intelligence (IJCAI'07)*. Morgan Kaufmann Publishers Inc., Hyderabad, Inde, pages 1606–1611. Janvier. 2007, https://www.ijcai.org/Proceedings/07/Papers/259.pdf. *citée page 50*

William A. Gale *et* Kenneth W. Church. 1993. A Program for Aligning Sentences in Bilingual Corpora. *C*omputational Linguistics 19(1):75–102. http://www.aclweb.org/anthology/J93-1004. *citée page 81*

Francis Galton. 1886. Regression towards mediocrity in hereditary stature. *J*ournal of the Anthropological Institute of Great Britain and Ireland 15:246–263. *4 citations pages 70, 88, 115, et 116*

Spandana Gella, Rico Sennrich, Frank Keller, *et* Mirella Lapata. 2017. Image Pivoting for Learning Multilingual Multimodal Representations. Dans *P*roceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017). Copenhague, Danemark, pages 2829–2835. Septembre. 2017, http://aclweb.org/anthology/D/D17/D17-1302.pdf. *citée page 57*

Ulrich Germann. 2001. Aligned Hansards of the 36^{th} Parliament of Canada. Release 2001-1a. https://www.isi.edu/natural-language/download/hansard/. *citée page 58*

Sahar Ghannay, Benoit Favre, Yannick Estève, *et* Nathalie Camelin. 2016. Word Embedding Evaluation and Combination. Dans *P*roceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). European Language Resources Association (ELRA), Portorož, Slovénie, pages 300–305. Mai. 2016, http://www.lrec-conf.org/proceedings/lrec2016/pdf/392_Paper.pdf. *citée page 53*

Elizabeth Gibney. 2006. I'm No Plagiarist, I Moved a Comma. *The Times Higher Education Supplement: THE* No. 2104. http://www.questia.com/magazine/1P3-3034399841/i-m-no-plagiarist-i-moved-a-comma-news#/. *2 citations pages 26 et 28*

G. H. Golub *et* C. Reinsch. 1970. Singular value decomposition and least squares solutions. *N*umerische Mathematik 14(5):403–420. https://doi.org/10.1007/BF02163027. *2 citations pages 47 et 48*

Stephan Gouws, Yoshua Bengio, *et* Greg Corrado. 2015. BilBOWA: Fast Bilingual Distributed Representations without Word Alignments. Dans *Proceedings of the 32nd International Conference on Machine Learning (ICML'15)*. Lille, France, pages 748–756. Juillet. 2015, http://proceedings.mlr.press/v37/gouws15.pdf. *citée page 56*

Stephan Gouws *et* Anders Søgaard. 2015. Simple task-specific bilingual word embeddings. Dans *H*uman Language Technologies : The 2015 Annual Conference of the North American Chapter of the ACL (HLT-NAACL). Association for Computational Linguistics, Denver, Colorado, États-Unis, pages 1386–1390. Mai. 2015, http://aclanthology.coli.uni-saarland.de/pdf/N/N15/N15-1157.pdf. *citée page 56*

Spence Green, Marie-Catherine de Marneffe, John Bauer, *et* Christopher D. Manning. 2011. Multiword Expression Identification with Tree Substitution Grammars: A Parsing tour de force with French. Dans *P*roceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011). Association for Computational Linguistics, Édimbourg, Écosse, pages 725–735. Juillet. 2011, http://aclweb.org/anthology/D/D11/D11-1067.pdf. *citée page 80*

Ján Grman *et* Rudolf Ravas. 2011. Improved implementation for finding text similarities in large collections of data - Notebook for PAN at CLEF 2011. Dans *N*otebook Papers for PAN at CLEF 2011 LABs and Workshops. Amsterdam, Pays-Bas. Septembre. 2011. *citée page 32*

Jeenu Grover *et* Pabitra Mitra. 2017. Bilingual Word Embeddings with Bucketed CNN for Parallel Sentence Extraction. Dans *P*roceedings of the 55th Annual Meeting of the Association for Computational Linguistics- Student Research Workshop. Association for Computational Linguistics, Vancouver, Canada, pages 11–16. Août. 2017, http://aclweb.org/anthology/P17-3003. *citée page 57*

Pascal Guibert *et* Christophe Michaut. 2011. Le plagiat étudiant. *Education et sociétés* 28:214. *citée page 26*

Parth Gupta, Alberto Barrón-Cedeño, *et* Paolo Rosso. 2012. Cross-language High Similarity Search using a Conceptual Thesaurus. Dans *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics*. Springer Berlin Heidelberg, Rome, Italie, pages 67–75. Septembre. 2012, https://doi.org/10.1007/978-3-642-33247-0_8. *2 citations pages 42 et 77*

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, *et* Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations* 11(1):10–18. https://doi.org/10.1145/1656274.1656278. *4 citations pages 103, 108, 114, et 115*

Andrew Hardie. 2007. Part-of-speech ratios in English corpora. *I*nternational Journal of Corpus Linguistics 12(1):55–81. https://doi.org/10.1075/ijcl.12.1.05har. *citée page 105*

Felix Hill, Kyunghyun Cho, *et* Anna Korhonen. 2016. Learning Distributed Representations of Sentences from Unlabelled Data. Dans *P*roceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016). Association for Computational Linguistics, San Diego, Californie, États-Unis, pages 1367–1377. Juin. 2016, http://www.aclweb.org/anthology/N16-1162. *2 citations pages 56 et 98*

Sepp Hochreiter *et* Jürgen Schmidhuber. 1997. Long Short-Term Memory. *T*he MIT Press. Neural Computation 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735. *citée page 57*

Geoffrey Holmes, Mark Hall, *et* Eibe Frank. 1999. Generating Rule Sets from Model Trees. Dans *P*roceedings of the 12th Australian Joint Conference on Artificial Intelligence. Sydney, Australie, pages 1–12. Décembre. 1999, https://doi.org/10.1007/3-540-46695-9_1. *2 citations pages 114 et 116*

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, *et* Hal Daumé III. 2015. Deep Unordered Composition Rivals Syntactic Methods for Text Classification. Dans *P*roceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, Pékin, Chine, pages 1681–1691. Juillet. 2015, http://www.aclweb.org/anthology/P15-1162. *citée page 57*

Paul Jaccard. 1912. The Distribution of the Flora in the Alpine Zone. *N*ew Phytologist 11(2):37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x. *8 citations pages 52, 59, 86, 95, 99, 101, 102, et 115*

Arun Kumar Jayapal *et* Binayak Goswami. 2013. Vector Space Model and Overlap Metric for Author Identification. Dans *N*otebook Papers for PAN at CLEF 2013 LABs and Workshops. http://ceur-ws.org/Vol-1179/CLEF2013wn-PAN-JayapalEt2013.pdf. *citée page 34*

Yangfeng Ji *et* Jacob Eisenstein. 2013. Discriminative Improvements to Distributional Sentence Similarity. Dans *P*roceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013). Seattle, Washington, États-Unis, pages 891–896. Octobre. 2013, http://aclweb.org/anthology/D/D13/D13-1090.pdf. *citée page 98*

Michael Jones *et* Lynnaire Sheridan. 2015. Back translation: an emerging sophisticated cyber strategy to subvert advances in 'digital age' plagiarism detection and prevention. *A*ssessment & Evaluation in Higher Education 40(5):1–7. https://doi.org/10.1080/02602938.2014.950553. *citée page 24*

Josephson Institute. 2011. WHAT WOULD HONEST ABE LINCOLN SAY? Dans *Installment 2: Honesty and Integrity - The Ethics of American Youth: 2010, study by Josephson Institute of Ethics' Report Card on American Youth's Values and Actions*. *citée page 26*

C. T. Kelley. 1995. *Iterative Methods for Linear and Nonlinear Equations*. Frontiers in Applied Mathematics. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611970944. *2 citations pages 104 et 107*

C. T. Kelley. 1999. *Iterative Methods for Optimization*. Frontiers in Applied Mathematics. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611970920. *2 citations pages 104 et 107*

Chow Kok Kent *et* Naomie Salim. 2009. Web Based Cross Language Plagiarism Detection. *J*ournal of Computing 1(1):39–43. *citée page 52*

Chow Kok Kent *et* Naomie Salim. 2010a. Features Based Text Similarity Detection. *J*ournal of Computing 2(1):53–57. *citée page 33*

Chow Kok Kent *et* Naomie Salim. 2010b. Web Based Cross Language Plagiarism Detection. Dans *Second International Conference on Computational Intelligence, Modelling and Simulation (CIMSiM)*. IEEE, Bali, Indonésie, pages 199–204. Juin. 2010, https://doi.org/10.1109/CIMSiM.2010.10. *citée page 52*

Mike Kestemont, Kim Luyckx, *et* Walter Daelemans. 2011. Intrinsic plagiarism detection using character trigram distance scores - Notebook for PAN at CLEF 2011. Dans *N*otebook Papers for PAN at CLEF 2011 LABs and Workshops. Amsterdam, Pays-Bas. *citée page 35*

Khadijeh Khoshnavataher, Vahid Zarrabi, Salar Mohtaj, *et* Habibollah Asghari. 2015. Developing Monolingual Persian Corpus for Extrinsic Plagiarism Detection Using Artificial Obfuscation. Dans *W*orking Notes Papers of the CLEF 2015 Evaluation Labs. Toulouse, France, CEUR Workshop Proceedings. Septembre. 2015, http://ceur-ws.org/Vol-1391/146-CR.pdf. *citée page 83*

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, *et* Sanja Fidler. 2015. Skip-Thought Vectors. Dans *P*roceedings of the 28th International Conference on Neural Information Processing Systems (NIPS 2015). Montréal, Canada, pages 3294–3302. Décembre. 2015, https://papers.nips.cc/paper/5950-skip-thought-vectors.pdf. *citée page 56*

Dan Klein *et* Christopher D. Manning. 2002. Fast Exact Inference with a Factored Model for Natural Language Parsing. Dans *P*roceedings of the 15th Annual Conference on Advances in Neural Information Processing Systems 15 (NIPS 2002). MIT Press, Cambridge, Massachusetts, États-Unis, pages 3–10. https://papers.nips.cc/paper/2325-fast-exact-inference-with-a-factored-model-for-natural-language-parsing.pdf. *citée page 80*

Dan Klein *et* Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. Dans *P*roceedings of the 41st Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Sapporo , Japon, pages 423–430. Juillet. 2003, https://doi.org/10.3115/1075096.1075150. *citée page 80*

Graham Klyne *et* Jeremy J. Carroll. 2004. Resource Description Framework (RDF): Concepts and Abstract Syntax. https://www.w3.org/TR/2004/REC-rdf-concepts-20040210/. *citée page 41*

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. Dans *Conference Proceedings: the tenth Machine Translation Summit*. Phuket, Thaïlande, pages 79–86. *5 citations pages 52, 67, 76, 83, et 121*

Ron Kohavi. 1995. The Power of Decision Tables. Dans *P*roceedings of the 8th European Conference on Machine Learning. Héraklion, Grèce, pages 174–189. Avril. 1995, https://doi.org/10.1007/3-540-59286-5_57. *2 citations pages 114 et 116*

Taku Kudo *et* Yuji Matsumoto. 2001. Chunking with Support Vector Machines. Dans *P*roceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies (NAACL'01). Association for Computational Linguistics, Stroudsburg, Pennsylvanie, États-Unis, pages 1–8. http://aclweb.org/anthology/N/N01/N01-1025.pdf. *citée page 80*

Taku Kudo *et* Yuji Matsumoto. 2003. Fast Methods for Kernel-based Text Analysis. Dans *P*roceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1 (ACL'03). Association for Computational Linguistics, Stroudsburg, Pennsylvanie, États-Unis, pages 24–31. http://aclweb.org/anthology/P/P03/P03-1004.pdf. *citée page 80*

Mikhail Kuznetsov, Anastasia Motrenko, Rita Kuznetsova, *et* Vadim Strijov. 2016. Methods for intrinsic plagiarism detection and author diarization - Notebook for PAN at CLEF 2016. Dans *N*otebook Papers for PAN at CLEF 2016 LABs and Workshops. http://ceur-ws.org/Vol-1609/16090912.pdf. *citée page 35*

Robert Layton, Paul A. Watters, *et* Richard Dazeley. 2013. Local n-grams for Author Identification Notebook for PAN at CLEF 2013. Dans *N*otebook Papers for PAN at CLEF 2013 LABs and Workshops. Valence, Espagne. Septembre. 2013, http://ceur-ws.org/Vol-1179/CLEF2013wn-PAN-LaytonEt2013.pdf. *2 citations pages 34 et 35*

Quoc V. Le *et* Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. Dans *Proceedings of the 31th International Conference on Machine Learning (ICML'14)*. JMLR Proceedings, Pékin, Chine, volume 32, pages 1188–1196. Juin. 2014, http://proceedings.mlr.press/v32/le14.pdf. *5 citations pages 53, 55, 56, 98, et 103*

Michael Lesk. 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. Dans *Proceedings of the 5th Annual International Conference on Systems Documentation (SIGDOC'86)*. ACM, Toronto, Ontario, Canada, pages 24–26. https://doi.org/10.1145/318723.318728. *citée page 43*

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *S*oviet Physics Doklady 10(8):707–710. *2 citations pages 86 et 101*

Gina-Anne Levow, Douglas W. Oard, *et* Philip Resnik. 2005. Dictionary-based Techniques for Cross-language Information Retrieval. *I*nformation Processing and Management 41(3):523–547. https://doi.org/10.1016/j.ipm.2004.06.012. *citée page 43*

Percy Liang, Ben Taskar, *et* Dan Klein. 2006. Alignment by Agreement. Dans *P*roceedings of HLT-NAACL. Association for Computational Linguistics, New York, États-Unis, pages 104–111. Juin. 2006, http://aclweb.org/anthology/N/N06/N06-1014.pdf. *citée page 52*

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, *et* Yoshua Bengio. 2017. A Structured Self-Attentive Sentence Embedding. Dans *P*roceedings of the 5th International Conference on Learning Representations (ICLR 2017). Toulon, France. Avril. 2017, https://openreview.net/pdf?id=BJC_jUqxe. *citée page 98*

Alexis Linard, Béatrice Daille, *et* Emmanuel Morin. 2015. Attempting to Bypass Alignment from Comparable Corpora via Pivot Language. Dans *Proceedings of the 8th workshop on Building and Using Comparable Corpora (BUCC)*. Association for Computational Linguistics, Pékin, Chine, pages 32–37. Juillet. 2015, https://aclweb.org/anthology/W/W15/W15-3405.pdf. *citée page 51*

Christina Lioma *et* Roi Blanco. 2017. Part of Speech Based Term Weighting for Information Retrieval. Dans *P*roceedings of the 31th European Conference on IR Research on Advances in Information Retrieval (ECIR 2009). Toulouse, France, pages 412–423. Avril. 2017, https://doi.org/10.1007/978-3-642-00958-7_37. *2 citations pages 99 et 100*

Michael Littman, Susan T. Dumais, *et* Thomas K. Landauer. 1998. *Automatic Cross-language Information Retrieval Using Latent Semantic Indexing*, Kluwer Academic Publishers, chapitre 5, pages 51–62. https://doi.org/10.1007/978-1-4615-5661-9_5. *citée page 47*

Yang Liu *et* Mirella Lapata. 2017. Learning Structured Text Representations. Dans *a*rXiv. https://arxiv.org/pdf/1705.09207.pdf. *citée page 98*

Minh-Thang Luong, Hieu Pham, *et* Christopher D. Manning. 2015. Bilingual Word Representations with Monolingual Quality in Mind. Dans *Proceedings of the 1st NAACL Workshop on Vector Space Modeling for Natural Language Processing*. Denver, Colorado, États-Unis, pages 151–159. Mai. 2015, http://www.aclweb.org/anthology/W15-1521. *3 citations pages 53, 56, et 103*

Caroline Lyon, James Malcolm, *et* Bob Dickerson. 2001. Detecting short passages of similar text in large document collections. Dans *P*roceedings of the 2001 Conference on Empirical Methods in Natural Language Processing (EMNLP 2001). pages 118–125. http://aclweb.org/anthology/W01-0515. *citée page 33*

Xiaoyi Ma. 2006. Champollion: A Robust Parallel Text Sentence Aligner. Dans *Proceedings of the fifth International Conference on Language Resources and Evaluation (LREC'06)*. Gênes, Italie. Mai. 2006, http://www.lrec-conf.org/proceedings/lrec2006/pdf/746_pdf.pdf. *citée page 81*

Jonathan Mallinson, Rico Sennrich, *et* Mirella Lapata. 2017. Paraphrasing Revisited with Neural Machine Translation. Dans Association for Computational Linguistics, éditeur, *P*roceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. Association for Computational Linguistics, Valence, Espagne, pages 880–892. Avril. 2017, http://aclweb.org/anthology/E/E17/E17-1083.pdf. *citée page 24*

Massimiliano Mancini, Jose Camacho-Collados, Ignacio Iacobacci, *et* Roberto Navigli. 2017. Embedding Words and Senses Together via Joint Knowledge-Enhanced Training. Dans *P*roceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). Association for Computational Linguistics, Vancouver, Canada, pages 100–111. Août. 2017, http://aclweb.org/anthology/K/K17/K17-1012.pdf. *citée page 98*

Christopher D. Manning, Prabhakar Raghavan, *et* Hinrich Schütze. 2008. *Introduction to Information Retrieval*, Cambridge University Press, New York, États-Unis, chapitre 6 - "Scoring, term weighting, and the vector space model", pages 109–133. ISBN: 9780511809071. https://doi.org/10.1017/CBO9780511809071.007. *citée page 112*

Christopher D. Manning *et* Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts, États-Unis. https://nlp.stanford.edu/fsnlp/. *3 citations pages 32, 62, et 63*

Izet Masic. 2012. Plagiarism in Scientific Publishing. Dans *A*cta Informatica Medica. pages 208–213. Décembre. 2012, https://doi.org/10.5455/aim.2012.20.208-213. *citée page 22*

Donald McCabe. 2010. Students' cheating takes a high-tech turn. *2 citations pages 26 et 29*

John Philip McCrae, Philipp Cimiano, *et* Roman Klinger. 2013. Orthonormal Explicit Topic Analysis for Cross-lingual Document Matching. Dans *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, États-Unis, pages 1732–1740. Octobre. 2013, http://www.aclweb.org/anthology/D13-1179. *citée page 47*

John Philip McCrae, Dennis Spohr, *et* Philipp Cimiano. 2011. *The Semantic Web: Research and Applications : 8th Extended Semantic Web Conference, ESWC 2011*, Springer-Verlag Berlin Heidelberg, Héraklion, Grèce, volume 6643 de *I*nformation Systems and Applications, incl. Internet/Web, and HCI, chapitre Linking Lexical Resources and Ontologies on the Semantic Web with Lemon, pages 245–259. https://doi.org/10.1007/978-3-642-21034-1. *citée page 41*

Paul Mcnamee *et* James Mayfield. 2004. Character N-Gram Tokenization for European Language Text Retrieval. *Information Retrieval Proceedings* 7(1-2):73–97. https://doi.org/10.1023/B:INRT.0000009441.78971.be. *citée page 37*

Igor Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. State University of New York Press. *citée page 99*

T. C. Mendenhall. 1887. The Characteristic Curves of Composition. *S*cience 9:237–246. https://doi.org/10.1126/science.ns-9.214S.237. *citée page 34*

Charles E. Metz. 1978. Basic principles of ROC analysis. *S*eminars in Nuclear Medicine 8(4):283–298. https://doi.org/10.1016/S0001-2998(78)80014-2. *citée page 68*

Tomas Mikolov, Kai Chen, Greg Corrado, *et* Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. Dans *T*he Workshop Proceedings of the International Conference on Learning Representations. Scottsdale, Arizona, États-Unis. https://openreview.net/forum?id=idpCdOWtqXd60. *7 citations pages 33, 35, 42, 53, 54, 59, et 103*

Tomas Mikolov, Martin Karafiát, Lukás Burget, Jan Cernocký, *et* Sanjeev Khudanpur. 2010. Recurrent neural network based language model. Dans *1*1th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010. Chiba, Japan, pages 1045–1048. Septembre. 2010, http://www.isca-speech.org/archive/archive_papers/interspeech_2010/i10_1045.pdf. *citée page 54*

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, *et* Jeffrey Dean. 2013b. Distributed Representations of Words and Phrases and their Compositionality. Dans *P*roceedings of the 27th Annual Conference on Neural Information Processing Systems (NIPS'13). Lac Tahoe, États-Unis, pages 3111–3119. Décembre. 2013, https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf. *4 citations pages 53, 54, 55, et 103*

Tomas Mikolov, Wen tau Yih, *et* Geoffrey Zweig. 2013c. Linguistic Regularities in Continuous Space Word Representations. Dans *P*roceedings of NAACL-HLT 2013. Association for Computational Linguistics, Atlanta, Géorgie, États-Unis, pages 746–751. Juin. 2013, http://www.aclweb.org/anthology/N13-1090. *3 citations pages 53, 54, et 103*

George A. Miller. 1985. Wordnet: A Dictionary Browser. Dans *Proceedings of the First Conference of the UW Centre for the New Oxford Dictionary*. Information in Data. *citée page 40*

George A. Miller, Christiane Fellbaum, Judy Kegl, *et* Katherine J. Miller. 1988. WordNet: an electronic lexical reference system based on theories of lexical memory. *Revue quebecoise de linguistique* 17(2):181–213. https://doi.org/10.7202/602632ar. *citée page 40*

Salar Mohtaj, Habibollah Asghari, *et* Vahid Zarrabi. 2015. Developing Monolingual English Corpus for Plagiarism Detection using Human Annotated Paraphrase Corpus. Dans *W*orking Notes Papers of the CLEF 2015 Evaluation Labs. Toulouse, France, CEUR Workshop Proceedings. Septembre. 2015, http://ceur-ws.org/Vol-1391/144-CR.pdf. *citée page 83*

Manuel Montes-y-Gómez, Alexander Gelbukh, Aurelio Lopez-Lopez, *et* Ricardo Baeza-Yates. 2001. Flexible Comparison of Conceptual Graphs. Dans *L*ecture Notes in Computer Science. pages 102–111. Janvier. 2001, https://doi.org/10.1007/3-540-44759-8_12. *citée page 45*

Emmanuel Morin, Amir Hazem, Elizaveta Loginova-Clouet, *et* Florian Boudin. 2015. LINA: Identifying Comparable Documents from Wikipedia. Dans *Proceedings of the 8th workshop on Building and Using Comparable Corpora (BUCC)*. Association for Computational Linguistics, Pékin, Chine, pages 88–91. Juillet. 2015, https://aclweb.org/anthology/W/W15/W15-3413.pdf. *citée page 39*

Nikola Mrkšic, Ivan Vulic, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašic, Anna Korhonen, *et* Steve Young. 2017. Semantic Specialisation of Distributional Word Vector Spaces using Monolingual and Cross-Lingual Constraints. Dans *T*ransactions of the Association for Computational Linguistics. Association for Computational Linguistics. Septembre. 2017, https://transacl.org/ojs/index.php/tacl/article/view/1171. *citée page 98*

Markus Muhr, Roman Kern, Mario Zechner, *et* Michael Granitzer. 2010. External and Intrinsic Plagiarism Detection Using a Cross-Lingual Retrieval and Segmentation System - Lab Report for PAN at CLEF 2010. Dans Martin Braschler, Donna Harman, *et* Emanuele Pianta, éditeurs, *N*otebook Papers for PAN at CLEF 2010 LABs and Workshops. Padoue, Italie. Septembre. 2010, http://ceur-ws.org/Vol-1176/CLEF2010wn-PAN-MuhrEt2010.pdf. *2 citations pages 52 et 87*

El Moatez Billah Nagoudi, Jérémy Ferrero, *et* Didier Schwab. 2017a. Amélioration de la similarité sémantique vectorielle par méthodes non-supervisées. Dans *24*ème conférence sur le Traitement Automatique des Langues Naturelles (TALN 2017). Orléans, France, pages 110–117. juin. 2017, http://taln2017.cnrs.fr/wp-content/uploads/2017/06/actes_TALN_2017-vol2.pdf. *2 citations pages 105 et 113*

El Moatez Billah Nagoudi, Jérémy Ferrero, *et* Didier Schwab. 2017b. LIM-LIG at SemEval-2017 Task1: Enhancing the Semantic Similarity for Arabic Sentences with Vectors Weighting. Dans *P*roceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Association for Computational Linguistics, Vancouver, Canada, pages 134–138. Août. 2017, http://aclweb.org/anthology/S/S17/S17-2017.pdf. *4 citations pages 105, 111, 113, et 123*

Roberto Navigli. 2012. Babelplagiarism: What can BabelNet do for Cross-language Plagiarism Detection. Dans *CLEF 2012, Evaluation Labs and Workshop, Online Working Notes*. Rome, Italie. Septembre. 2012, http://ceur-ws.org/Vol-1178/CLEF2012wn-PAN-Navigli2012.pdf. *citée page 45*

Roberto Navigli *et* Mirella Lapata. 2010. An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32:678–692. https://doi.org/10.1109/TPAMI.2009.36. *citée page 45*

Roberto Navigli *et* Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Dans *Artificial Intelligence Proceedings*. volume 193, pages 217–250. https://doi.org/10.1016/j.artint.2012.07.001. *4 citations pages 40, 41, 45, et 59*

Joakim Nivre, Johan Hall, *et* Jens Nilsson. 2006. MaltParser: A Data-Driven Parser-Generator for Dependency Parsing. Dans *P*roceedings of the fifth International Conference on Language Resources and Evaluation (LREC'06). European Language Resources Association (ELRA), Gênes, Italie, pages 2216–2219. Mai. 2006, http://www.lrec-conf.org/proceedings/lrec2006/pdf/162_pdf.pdf. *citée page 80*

Gabriel Oberreuter *et* Juan D. Velásquez. 2013. Text Mining Applied to Plagiarism Detection: The Use of Words for Detecting Deviations in the Writing Style. *E*xpert Systems with Applications 40(9):3756–3763. https://doi.org/10.1016/j.eswa.2012.12.082. *2 citations pages 31 et 34*

Matteo Pagliardini, Prakhar Gupta, *et* Martin Jaggi. 2017. Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. Dans *a*rXiv. https://arxiv.org/pdf/1703.02507.pdf. *2 citations pages 56 et 98*

Màté Pataki. 2012. A New Approach for Searching Translated Plagiarism. Dans *Proceedings of the 5th International Plagiarism Conference*. Newcastle, Angleterre, pages 49–64. Juillet. 2012. *4 citations pages 43, 44, 86, et 101*

Karl Pearson. 1895. Notes on regression and inheritance in the case of two parents. *P*roceedings of the Royal Society of London 58:240–242. https://www.jstor.org/stable/115794. *4 citations pages 70, 88, 115, et 116*

Jeffrey Pennington, Richard Socher, *et* Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. Dans *P*roceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014). Association for Computational Linguistics, Doha, Qatar, pages 1532–1543. Octobre. 2014, http://aclweb.org/anthology/D/D14/D14-1162.pdf. *2 citations pages 54 et 56*

Rafael Corezola Pereira. 2010. *Cross-Language Plagiarism Detection*. Thèse de master, Instituto de Informática da Universidade Federal do Rio Grande do Sul, Porto Alegre, Brésil. *citée page 67*

Rafael Corezola Pereira, Viviane P. Moreira, *et* Renata Galante. 2010a. A New Approach for Cross-language Plagiarism Analysis. Dans *P*roceedings of the 2010 International Conference on Multilingual and Multimodal Information Access Evaluation: Cross-language Evaluation Forum. Padoue, Italie, CLEF'10, pages 15–26. https://doi.org/10.1007/978-3-642-15998-5_4. *2 citations pages 67 et 71*

Rafael Corezola Pereira, Viviane P. Moreira, *et* Renata Galante. 2010b. UFRGS@PAN2010: Detecting External Plagiarism - Lab Report for PAN at CLEF 2010. Dans Braschler *et* Harman, éditeurs, *N*otebook Papers for PAN at CLEF 2010 LABs and Workshops. Padoue, Italie. Septembre. 2010, http://ceur-ws.org/Vol-1176/CLEF2010wn-PAN-CorezolaPereiraEt2010.pdf. *citée page 51*

Slav Petrov, Leon Barrett, Romain Thibaux, *et* Dan Klein. 2006. Learning Accurate, Compact, and Interpretable Tree Annotation. Dans *P*roceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Sydney, Australie, pages 433–440. Juillet. 2006, https://doi.org/10.3115/1220175.1220230. *citée page 80*

Slav Petrov, Dipanjan Das, *et* Ryan McDonald. 2012. A Universal Part-of-Speech Tagset. Dans *P*roceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA), Istanbul, Turquie, pages 2089–2096. Mai. 2012, http://www.lrec-conf.org/proceedings/lrec2012/pdf/274_Paper.pdf.
4 citations pages 80, 99, 162, et 163

Slav Petrov *et* Dan Klein. 2007. Improved Inference for Unlexicalized Parsing. Dans *H*uman Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference. Association for Computational Linguistics, Rochester, New York, États-Unis, pages 404–411. http://www.aclweb.org/anthology/N07-1051. *citée page 80*

Hieu Pham, Minh-Thang Luong, *et* Christopher D. Manning. 2015. Learning Distributed Representations for Multilingual Text Sequences. Dans *Proceedings of the 1st NAACL Workshop on Vector Space Modeling for Natural Language Processing*. Denver, Colorado, États-Unis. Mai. 2015, http://www.aclweb.org/anthology/W15-1512. *2 citations pages 55 et 56*

David Pinto, Jorge Civera, Alfons Juan, Paolo Rosso, *et* Alberto Barrón-Cedeño. 2009. A Statistical Approach to Crosslingual Natural Language Tasks. Dans *CEUR Workshop Proceedings*. volume 64 de *Journal of Algorithms*, pages 51–60. Janvier. 2009, https://doi.org/10.1016/j.jalgor.2009.02.005. *2 citations pages 46 et 87*

David Pinto, Alfons Juan, *et* Paolo Rosso. 2007. Using Query-Relevant Documents Pairs for Cross-Lingual Information Retrieval. Dans *Text, Speech and Dialogue*. Springer Berlin Heidelberg, volume 4629 de *Lecture Notes in Computer Science*, pages 630–637. https://doi.org/10.1007/978-3-540-74628-7_81. *citée page 46*

Martin Potthast, Alberto Barrón-Cedeño, Andreas Eiselt, Benno Stein, *et* Paolo Rosso. 2010a. Overview of the 2nd International Competition on Plagiarism Detection. Dans & E. Pianta (Eds.) M. Braschler, D. Harman, éditeur, *N*otebook Papers for PAN at CLEF 2010 LABs and Workshops. volume 1176 de *C*EUR Workshop Proceedings. *citée page 64*

Martin Potthast, Alberto Barrón-Cedeño, Benno Stein, *et* Paolo Rosso. 2011a. Cross-Language Plagiarism Detection. *Language Resources and Evaluation* 45(1):45–62. https://doi.org/10.1007/s10579-009-9114-z.
14 citations pages 29, 36, 37, 39, 40, 47, 58, 59, 60, 61, 76, 84, 86, et 112

Martin Potthast, Andreas Eiselt, Alberto Barrón-Cedeño, Benno Stein, *et* Paolo Rosso. 2011b. Overview of the 3rd international Competition on Plagiarism Detection. Dans & P. D. Clough (Eds.) V. Petras, P. Forner, éditeur, *N*otebook Papers for PAN at CLEF 2011 LABs and Workshops. volume 1177 de *C*EUR Workshop Proceedings.
5 citations pages 58, 64, 71, 77, et 83

Martin Potthast, Matthias Hagen, Anna Beyer, Matthias Busse, Martin Tippmann, Paolo Rosso, *et* Benno Stein. 2014. Overview of the 6th International Competition on Plagiarism Detection. Dans *N*otebook Papers for PAN at CLEF 2014 LABs and Workshops. Sheffield, Angleterre, pages 845–876. Septembre. 2014, http://ceur-ws.org/Vol-1180/CLEF2014wn-Pan-PotthastEt2014.pdf. *citée page 64*

Martin Potthast, Benno Stein, *et* Maik Anderka. 2008. A Wikipedia-Based Multilingual Retrieval Model. Dans *30th European Conference on IR Research (ECIR'08)*. Springer, Glasgow, Écosse, volume 4956 de *LNCS of Lecture Notes in Computer Science*, pages 522–530. Mars. 2008, https://doi.org/10.1007/978-3-540-78646-7_51. *2 citations pages 50 et 51*

Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, *et* Paolo Rosso. 2010b. An Evaluation Framework for Plagiarism Detection. Dans Huang *et* Jurafsky, éditeurs, *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*. Pékin, Chine, pages 997–1005. Août. 2010, http://aclweb.org/anthology/C/C10/C10-2115.pdf. *4 citations pages 58, 64, 66, et 122*

Martin Potthast, Benno Stein, Andreas Eiselt, Alberto Barrón-Cedeño, *et* Paolo Rosso. 2009. Overview of the 1st International Competition on Plagiarism Detection. Dans Benno Stein, Paolo Rosso, Stamatatos, Koppel, *et* Agirre (Eds.), éditeurs, *3*rd PAN workshop. Uncovering Plagiarism, Authorship and Social Software Misuse (PAN'09). pages 1–9. http://ceur-ws.org/Vol-502/paper1.pdf. *2 citations pages 63 et 67*

Bruno Pouliquen, Ralf Steinberger, *et* Camelia Ignat. 2003a. Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus. Dans *W*orkshop 'Ontologies and Information Extraction' at the Summer School 'The Semantic Web and Language Technology - Its Potential and Practicalities' (EUROLAN 2003). Bucarest, Roumanie, pages 9–28. Juillet. 2003. *citée page 42*

Bruno Pouliquen, Ralf Steinberger, *et* Camelia Ignat. 2003b. Automatic Identification of Document Translations in Large Multilingual Document Collections. Dans *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP'03)*. Borovets, Bulgarie, pages 401–408. Septembre. 2003. *6 citations pages 39, 42, 43, 47, 81, et 85*

Peter Prettenhofer *et* Benno Stein. 2010. Cross-language Text Classification Using Structural Correspondence Learning. Dans *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Suède, ACL'10, pages 1118–1127. Juillet. 2010, http://www.aclweb.org/anthology/P10-1114. *citée page 76*

J. R. Quinlan. 1992. Learning with continuous classes. Dans Eds. Adams & Sterling, éditeur, *Proceedings of the Fifth Australian Joint Conference on Artificial Intelligence*. World Scientific, Singapour, pages 343–348. *3 citations pages 114, 115, et 116*

J. Ross Quinlan. 1986. *Machine Learning*, Kluwer Academic Publisher, Boston, Massachusetts, États-Unis, chapitre Induction of decision trees, pages 81–106. *citée page 103*

J. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. The Morgan Kaufmann series in machine learning. Morgan Kaufmann Publishers Inc., San Francisco, Californie, États-Unis. https://doi.org/10.1007/BF00993309. *2 citations pages 103 et 121*

Ann M. Rogerson *et* Grace McCarthy. 2017. Using Internet based paraphrasing tools: Original work, patchwriting or facilitated plagiarism? *I*nternational Journal for Educational Integrity 13(2). https://doi.org/10.1007/s40979-016-0013-y. *citée page 24*

Sebastian Ruder. 2017. A survey of cross-lingual embedding models. *citée page 103*

Motaz Saad, David Langlois, *et* Kamel Smaili. 2014. Cross-Lingual Semantic Similarity Measure for Comparable Articles. Dans *Advances in Natural Language Processing - 9th International Conference on NLP, PolTAL 2014*. Springer Berlin Heidelberg, Varsovie, Pologne, pages 105–115. Septembre. 2014, https://doi.org/10.1007/978-3-319-10888-9_11. *citée page 47*

Gerard Salton. 1989. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Longman Publishing Co., Inc., Boston, Massachusetts, États-Unis. *12 citations pages 38, 42, 43, 48, 49, 51, 52, 53, 57, 86, 98, et 100*

Gerard Salton *et* Christopher Buckley. 1988. Term-weighting Approaches in Automatic Text Retrieval. *I*nformation Processing and Management 24(5):513–523. https://doi.org/10.1016/0306-4573(88)90021-0. *5 citations pages 37, 48, 50, 52, et 86*

Miguel A. Sanchez-Perez, Grigori Sidorov, *et* Alexander Gelbukh. 2014. The Winning Approach to Text Alignment for Text Reuse Detection at PAN 2014 - Notebook for PAN at CLEF 2014. Dans Cappellato et al., éditeur, *N*otebook Papers for PAN at CLEF 2014 LABs and Workshops. Sheffield, Angleterre. Septembre. 2014, http://ceur-ws.org/Vol-1180/CLEF2014wn-Pan-SanchezPerezEt2014.pdf. *citée page 33*

Yunita Sari *et* Mark Stevenson. 2016. Exploring Word Embeddings and Character N-Grams for Author Clustering - Notebook for PAN at CLEF 2016. Dans *N*otebook Papers for PAN at CLEF 2016 LABs and Workshops. http://ceur-ws.org/Vol-1609/16090984.pdf. *citée page 35*

Yunita Sari, Andreas Vlachos, *et* Mark Stevenson. 2017. Continuous N-gram Representations for Authorship Attribution. Dans Association for Computational Linguistics, éditeur, *P*roceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. Association for Computational Linguistics, Valence, Espagne, pages 267–273. Avril. 2017, http://aclweb.org/anthology/E/E17/E17-2043.pdf. *citée page 35*

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. Dans *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, Angleterre, pages 44–49. *4 citations pages 80, 99, 162, et 163*

Didier Schwab. 2005. *Approche hybride-lexicale et thématique-pour la modélisation, la détection et l'exploitation des fonctions lexicales en vue de l'analyse semantique de texte*. Thèse de doctorat, Université Montpellier II. Les citations concernent la section 1.1.4.2 en page 24 et la section 2.2.3.1 en page 71. https://tel.archives-ouvertes.fr/tel-00333334/document. *citée page 99*

Holger Schwenk *et* Matthijs Douze. 2017. Learning Joint Multilingual Sentence Representations with Neural Machine Translation. Dans *P*roceedings of the 2nd Workshop on Representation Learning for NLP (RepL4NLP 2017). Association for Computational Linguistics, Vancouver, Canada, pages 157–167. Août. 2017, http://www.aclweb.org/anthology/W17-2619. *citée page 98*

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, *et* Maria Nadejde. 2017. Nematus: a Toolkit for Neural Machine Translation. Dans *P*roceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Valence, Espagne, pages 65–68. Avril. 2017, http://aclweb.org/anthology/E17-3017. *citée page 57*

Gilles Sérasset. 2015. DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. *Semantic Web Journal (special issue on Multilingual Linked Open Data)* 6(4):355–361. https://doi.org/10.3233/SW-140147. *5 citations pages 41, 81, 86, 87, et 101*

Christophe Servan, Zied Elloumi, Hervé Blanchon, *et* Laurent Besacier. 2016. Word2Vec vs DBnary ou comment (ré)concilier représentations distribuées et réseaux lexico-sémantiques ? Le cas de l'évaluation en traduction automatique. Dans *A*ctes de la conférence conjointe JEP-TALN-RECITAL 2016, volume 2 : TALN. Paris, France. Juillet. 2016, https://jep-taln2016.limsi.fr/actes/Actes%20JTR-2016/Papers/T10.pdf. *2 citations pages 41 et 101*

S. K. Shevade, S. S. Keerthi, C. Bhattacharyya, *et* K. R. K. Murthy. 2000. Improvements to the SMO Algorithm for SVM Regression. *I*EEE Transactions on Neural Networks 11(5):1188–1193. https://doi.org/10.1109/72.870050. *2 citations pages 114 et 116*

Tianze Shi, Zhiyuan Liu, Yang Liu, *et* Maosong Sun. 2015. Learning cross-lingual word embeddings via matrix co-factorization. Dans *P*roceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers). Association for Computational Linguistics, Pékin, Chine, pages 567–572. Juillet. 2015, http://www.aclweb.org/anthology/P15-2093. *citée page 56*

Prasha Shrestha, Sebastian Sierra, Fabio A. González, Paolo Rosso, Manuel Montes-y-Gómez, *et* Thamar Solorio. 2017. Convolutional Neural Networks for Authorship Attribution of Short Texts. Dans Association for Computational Linguistics, éditeur, *P*roceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. Association for Computational Linguistics, Valence, Espagne, pages 669–674. Avril. 2017, http://aclweb.org/anthology/E/E17/E17-2106.pdf. *citée page 35*

Prasha Shrestha *et* Thamar Solorio. 2013. Using a Variety of n-Grams for the Detection of Different Kinds of Plagiarism - Notebook for PAN at CLEF 2013. Dans *N*otebook Papers for PAN at CLEF 2013 LABs and Workshops. http://ceur-ws.org/Vol-1179/CLEF2013wn-PAN-ShresthaEt2013.pdf. *citée page 33*

Michel Simard, George F. Foster, *et* Pierre Isabelle. 1993. Using Cognates to Align Sentences in Bilingual Corpora. Dans *Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research: distributed computing (CASCON'93)*. volume 2, pages 1071–1082. *citée page 39*

Alex J. Smola *et* Bernhard Scölkopf. 2004. A tutorial on support vector regression. Dans *S*tatistics and Computing. Kluwer Academic Publishers. Manufactured in The Netherlands, volume 14, pages 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88. *2 citations pages 114 et 116*

Richard Socher, John Bauer, Christopher D. Manning, *et* Andrew Y. Ng. 2013a. Parsing with Compositional Vector Grammars. Dans *P*roceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Sofia, Bulgarie, pages 455–465. Août. 2013, http://www.aclweb.org/anthology/P13-1045. *citée page 56*

Richard Socher, Eric H. Huang, Jeffrey Pennington, Andrew Y. Ng, *et* Christopher D. Manning. 2011. Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection. Dans *P*roceedings of the 24th International Conference on Neural Information Processing Systems (NIPS'11). Grenade, Espagne, pages 801–809. Décembre. 2011, https://papers.nips.cc/paper/4204-dynamic-pooling-and-unfolding-recursive-autoencoders-for-paraphrase-detection.pdf. *citée page 55*

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, *et* Christopher Potts. 2013b. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. Dans *P*roceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013). Seattle, Washington, États-Unis, pages 1631–1642. Octobre. 2013, http://aclweb.org/anthology/D/D13/D13-1170.pdf. *citée page 56*

Sanja Stajner, Marc Franco-Salvador, Simone Paolo Ponzetto, Paolo Rosso, *et* Heiner Stuckenschmidt. 2017. Sentence Alignment Methods for Improving Text Simplification Systems. Dans *P*roceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers). Association for Computational Linguistics, Vancouver, Canada, pages 97–102. Août. 2017, http://aclweb.org/anthology/P17-2016. *citée page 57*

Efstathios Stamatatos. 2009. Intrinsic Plagiarism Detection Using Character n-gram Profiles. Dans Benno Stein, Paolo Rosso, *et* Efstathios Stamatatos, éditeurs, *Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN'09)*. pages 38–46. *citée page 34*

Efstathios Stamatatos. 2017. Authorship Attribution Using Text Distortion. Dans Association for Computational Linguistics, éditeur, *P*roceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. Association for Computational Linguistics, Valence, Espagne, pages 1137–1148. Avril. 2017, http://aclweb.org/anthology/E/E17/E17-1107.pdf. *citée page 35*

Benno Stein *et* Sven Meyer Zu Eissen. 2006. Near Similarity Search and Plagiarism Analysis. Dans *F*rom Data and Information Analysis to Knowledge Engineering. pages 430–437. https://doi.org/10.1007/3-540-31314-1_52. *citée page 33*

Benno Stein *et* Sven Meyer Zu Eissen. 2007a. Fingerprint-based Similarity Search and its Applications. Dans *F*orschung und wissenschaftliches Rechnen. pages 85–98. *citée page 33*

Benno Stein *et* Sven Meyer Zu Eissen. 2007b. Intrinsic Plagiarism Analysis with Meta Learning. Dans Benno Stein, Moshe Koppel, *et* Efstathios Stamatatos, éditeurs, *P*AN. volume 276 de *C*EUR Workshop Proceedings. *citée page 34*

Benno Stein, Nedim Lipka, *et* Peter Prettenhofer. 2011. Intrinsic Plagiarism Analysis. *L*anguage Resources and Evaluation 45(1):63–82. https://doi.org/10.1007/s10579-010-9115-y. *citée page 34*

Ralf Steinberger. 2001. Cross-lingual Keyword Assignment. Dans *C*onference of the Spanish Society for Natural Language Processing (SEPLN'2001). Jaén, Espagne, numéro 27 dans Procesamiento del Lenguaje Natural, pages 273–280. *citée page 42*

Ralf Steinberger, Bruno Pouliquen, *et* Johan Hagman. 2002. Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC. Dans *CICLing*. Springer Berlin Heidelberg, LNCS 2276, pages 415–424. https://doi.org/10.1007/3-540-45715-1_44. *citée page 42*

Ralf Steinberger, Bruno Pouliquen, *et* Camelia Ignat. 2004. Exploiting Multilingual Nomenclatures and Language-Independent Text Features as an Interlingua for Cross-lingual Text Analysis Applications. Dans *Proceedings of the 4th Slovenian Language Technology Conference*. Information Society. *citée page 42*

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufis, *et* Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. Dans *P*roceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006). Gênes, Italie, pages 2142–2147. Mai. 2006, http://www.lrec-conf.org/proceedings/lrec2006/pdf/340_pdf.pdf. *5 citations pages 46, 61, 76, 83, et 121*

G. W. Stewart. 1993. On the Early History of the Singular Value Decomposition. *S*ociety for Industrial and Applied Mathematics (SIAM Review) 35(4):551–566. http://www.jstor.org/stable/2132388. *2 citations pages 47 et 48*

Keet Sugathadasa, Buddhi Ayesha, Nisansa de Silva, Amal Shehan Perera, Vindula Jayawardana, Dimuthu Lakmal, *et* Madhavi Perera. 2017. Synergistic Union of Word2Vec and Lexicon for Domain Specific Semantic Similarity. Dans *P*roceedings of the Seventh international conference on Innovative Computing Technology (INTECH 2017). Luton, Angleterre. Août. 2017. *citée page 98*

Md Arafat Sultan, Steven Bethard, *et* Tamara Sumner. 2015. DLS@CU: Sentence similarity from word alignment and semantic vector composition. Dans *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Denver, Colorado, États-Unis, pages 148–153. Juin. 2015, http://www.aclweb.org/anthology/S15-2027. *4 citations pages 52, 59, 60, et 112*

Junfeng Tian, Zhiheng Zhou, Man Lan, *et* Yuanbin Wu. 2017. ECNU at SemEval-2017 Task 1: Leverage Kernel-based Traditional NLP features and Neural Networks to Build a Universal Model for Multilingual and Cross-lingual Semantic Textual Similarity. Dans *P*roceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Association for Computational Linguistics, Vancouver, Canada, pages 191–197. Août. 2017, http://www.aclweb.org/anthology/S17-2028. *citée page 57*

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. Dans *P*roceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012). Istanbul, Turquie, pages 2214–2218. Mai. 2012, http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf. *2 citations pages 62 et 76*

Diego Antonio Rodríguez Torrejón *et* José Manuel Martín Ramos. 2011. Crosslingual CoReMo System - Notebook for PAN at CLEF 2011. Dans *N*otebook Papers for PAN at CLEF 2011 LABs and Workshops. Amsterdam, Pays-Bas. Septembre. 2011. *citée page 43*

Joseph Turian, Lev-Arie Ratinov, *et* Yoshua Bengio. 2010. Word Representations: A Simple and General Method for Semi-Supervised Learning. Dans *P*roceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Uppsala, Suède, pages 384–394. Juillet. 2010, http://www.aclweb.org/anthology/P10-1040. *citée page 54*

Amos Tversky. 1977. Features of similarity. *Psychological Review* 84(4):327–352. *citée page 43*

Shyam Upadhyay, Manaal Faruqui, Chris Dyer, *et* Dan Roth. 2016. Cross-lingual Models of Word Embeddings: An Empirical Comparison. Dans *P*roceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL'16). Berlin, Allemagne, pages 1661–1670. Août. 2016, https://www.aclweb.org/anthology/P/P16/P16-1157.pdf. *citée page 53*

Hans Van Halteren. 2004. Linguistic Profiling for Author Recognition and Verification. Dans *P*roceedings of the 42^{nd} Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, Stroudsburg, Pennsylvanie, États-Unis, ACL'04. https://doi.org/10.3115/1218955.1218981. *citée page 34*

Dániel Varga, Péter Hálacsy, Viktor Nagy, László Németh, András Kornai, *et* Viktor Trón. 2005. Parallel corpora for Medium Density Languages. Dans *Recent Advances in Natural Language Processing (RANLP'05)*. Borovets, Bulgarie, pages 590–596. Septembre. 2005, https://doi.org/10.1075/cilt.292.32var. *citée page 81*

Zdenek Češka, Michal Toman, *et* Karel Jezek. 2008. Multilingual Plagiarism Detection. Dans *Artificial Intelligence: Methodology, Systems, and Applications*. Springer Berlin Heidelberg, volume 5253 de *Lecture Notes in Computer Science*, pages 83–92. https://doi.org/10.1007/978-3-540-85776-1_8. *citée page 43*

Alexei Vinokourov *et* Mark Girolami. 2002. A Probabilistic Framework for the Hierarchic Organisation and Classification of Document Collections. *J*ournal of Intelligent Information Systems 18(2/3):153–172. https://doi.org/10.1023/A:1013677411002. *citée page 49*

Alexei Vinokourov, John Shawe-Taylor, *et* Nello Cristianini. 2002. Inferring a Semantic Representation of Text via Cross-Language Correlation Analysis. *P*roceedings of the 15th Annual Conference on Advances in Neural Information Processing Systems 15 (NIPS 2002) pages 1473–1480. https://papers.nips.cc/paper/2324-inferring-a-semantic-representation-of-text-via-cross-language-correlation-analysis.pdf. *4 citations pages 49, 58, 59, et 60*

Ivan Vulić. 2017. Cross-Lingual Syntactically Informed Distributed Word Representations. Dans *P*roceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. Association for Computational Linguistics, Valence, Espagne, pages 408–414. Avril. 2017, http://aclweb.org/anthology/E/E17/E17-2065.pdf. *citée page 99*

Ivan Vulić *et* Anna Korhonen. 2016. On the role of seed lexicons in learning bilingual word embeddings. Dans *P*roceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Berlin, Allemagne, pages 247–257. Août. 2016, http://aclweb.org/anthology/P/P16/P16-1024.pdf. *citée page 56*

Ivan Vulic *et* Marie-Francine Moens. 2014. Probabilistic Models of Cross-Lingual Semantic Similarity in Context Based on Latent Cross-Lingual Concepts Induced from Comparable Data. Dans *P*roceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014). Doha, Qatar, pages 349–362. Octobre. 2014, http://www.aclweb.org/anthology/D14-1040. *citée page 50*

Yong Wang *et* Ian H. Witten. 1997. Induction of model trees for predicting continuous classes. Dans *Proceedings of the poster papers of the European Conference on Machine Learning*. Prague, République tchèque, pages 128–137. Octobre. 1997. *4 citations pages 114, 115, 116, et 121*

John Wieting *et* Kevin Gimpel. 2017. Revisiting Recurrent Networks for Paraphrastic Sentence Embeddings. Dans *P*roceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Vancouver, Canada, pages 2078–2088. Août. 2017, http://aclweb.org/anthology/P17-1190. *citée page 98*

John Wieting, Jonathan Mallinson, *et* Kevin Gimpel. 2017. Learning Paraphrastic Sentence Embeddings from Back-Translated Bitext. Dans *P*roceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017). Copenhague, Danemark, pages 274–285. Septembre. 2017, http://aclweb.org/anthology/D/D17/D17-1026.pdf. *citée page 98*

Adina Williams, Nikita Nangia, *et* Samuel R. Bowman. 2017. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. Dans *a*rXiv. Copenhague, Danemark. Septembre. 2017, https://arxiv.org/pdf/1704.05426.pdf. *citée page 68*

E. J. Williams. 1959. The Comparison of Regression Variables. *J*ournal of the Royal Statistical Society. Series B (Methodological) 21(2):396–399. http://www.jstor.org/stable/2983809. *citée page 110*

Viroj Wiwanitkit. 2011. Plagiarism: word, idea, figure, etc. *C*roatian Medical Journal 52(5):657. https://doi.org/10.3325/cmj.2011.52.657. *citée page 21*

Yiming Yang, Jaime G. Carbonell, Ralf D. Brown, *et* Robert E. Frederking. 1998. Translingual Information Retrieval: Learning from Bilingual Corpora. *A*rtificial Intelligence 103(1-2):323–345. https://doi.org/10.1016/S0004-3702(98)00063-0. *citée page 50*

Wenpeng Yin *et* Hinrich Schütze. 2015. Discriminative Phrase Embedding for Paraphrase Identification. Dans *P*roceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2015). Association for Computational Linguistics, Denver, Colorado, États-Unis, pages 1368–1373. Mai. 2015, https://aclweb.org/anthology/N/N15/N15-1154.pdf. *2 citations pages 33 et 99*

Hamed Zamani, Hossein Nasr, Pariya Babaie, Samira Abnar, Mostafa Dehghani, *et* Azadeh Shakery. 2014. Authorship Identification Using Dynamic Selection of Features from Probabilistic Feature Set. Dans *M*ethods for intrinsic plagiarism detection and author diarization - Notebook for PAN at CLEF 2014. pages 128–140. https://doi.org/10.1007/978-3-319-11382-1_13. *citée page 34*

Will Y. Zou, Richard Socher, Daniel Cer, *et* Christopher D. Manning. 2013. Bilingual Word Embeddings for Phrase-Based Machine Translation. Dans *P*roceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013). Association for Computational Linguistics, Seattle, Washington, États-Unis, pages 1393–1398. Octobre. 2013. *citée page 56*

Pierre Zweigenbaum, Serge Sharoff, *et* Reinhard Rapp. 2016. Towards Preparation of the Second BUCC Shared Task: Detecting Parallel Sentences in Comparable Corpora. Dans *Proceedings of the Ninth Workshop on Building and Using Comparable Corpora*

(BUCC). European Language Resources Association (ELRA), Portorož, Slovénie, pages 38–43. Mai. 2016, http://www.lrec-conf.org/proceedings/lrec2016/workshops/LREC2016Workshop-BUCC2016_Proceedings.pdf. *3 citations pages 61, 83, et 121*

Pierre Zweigenbaum, Serge Sharoff, *et* Reinhard Rapp. 2017. Overview of the Second BUCC Shared Task: Spotting Parallel Sentences in Comparable Corpora. Dans *Proceedings of the 10th Workshop on Building and Using Comparable Corpora (BUCC)*. Association for Computational Linguistics, Vancouver, Canada, pages 60–67. Août. 2017, http://aclweb.org/anthology/W17-2512. *5 citations pages 61, 62, 71, 83, et 121*

Annexes

A Bibliographie personnelle

Jérémy Ferrero, Frédéric Agnès, Laurent Besacier, et Didier Schwab. 2016. A Multilingual, Multi-style and Multi-granularity Dataset for Cross-language Textual Similarity Detection. *Dans Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, pages 4162-4169. European Language Resources Association (ELRA). Portorož, Slovénie. 23-28 mai 2016. ISLRN : 723-785-513-738-2. http://www.lrec-conf.org/proceedings/lrec2016/pdf/304_Paper.pdf

Jérémy Ferrero, Laurent Besacier, Didier Schwab, et Frédéric Agnès. 2017. Using Word Embedding for Cross-Language Plagiarism Detection. *Dans Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 2, Short Papers*, pages 415–421. Association for Computer Linguistics. Valence, Espagne. 3-7 avril 2017. http://aclweb.org/anthology/E/E17/E17-2066.pdf

Jérémy Ferrero, Laurent Besacier, Didier Schwab, et Frédéric Agnès. 2017. Deep Investigation of Cross-Language Plagiarism Detection Methods. *Dans Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 6–15. Association for Computational Linguistics. Vancouver, Canada. 3 août 2017. http://aclweb.org/anthology/W17-2502

Jérémy Ferrero, Laurent Besacier, Didier Schwab, et Frédéric Agnès. 2017. CompiLIG at SemEval-2017 Task 1: Cross-Language Plagiarism Detection Methods for Semantic Textual Similarity. *Dans Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 109–114. Association for Computational Linguistics. Vancouver, Canada. 3-4 août 2017. http://aclweb.org/anthology/S17-2012

El Moatez Billah Nagoudi, Jérémy Ferrero, et Didier Schwab. 2017. Amélioration de la similarité sémantique vectorielle par méthodes non-supervisées. *Dans 24e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2017)*, pages 110-117. Orléans, France. 26-30 juin 2017. http://taln2017.cnrs.fr/wp-content/uploads/2017/06/actes_TALN_2017-vol2.pdf

El Moatez Billah Nagoudi, Jérémy Ferrero, et Didier Schwab. 2017. LIM-LIG at SemEval-2017 Task1: Enhancing the Semantic Similarity for Arabic Sentences with Vectors Weighting. *Dans Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 134–138. Association for Computational Linguistics. Vancouver, Canada. 3-4 août 2017. http://aclweb.org/anthology/S/S17/S17-2017.pdf

El Moatez Billah Nagoudi, Jérémy Ferrero, et Didier Schwab. 2017. Word Embedding-Based Approaches for Measuring Semantic Similarity of Arabic-English Sentences. *À paraître dans Proceedings of the 6th International Conference on Arabic Language Processing (ICALP 2017)*. Fès, Maroc. 11-12 octobre 2017.

B Caractéristiques détaillées du corpus constitué

Sous-corpus	Type	Langue	# de fichiers	# de phrases	# de mots	# de caractères	phrases par fichier	mots par fichier	caractères par fichier	Taille moyenne des phrases (en mots)	Taille moyenne des mots (en caractères)	Taille (Mo)
JRC-Acquis	Parallel	FR	10 000	512 665	25 198 937	166 725 545	51,3	2 519,9	16 672,6	49,2	6,6	165,5
		EN	10 000	525 625	23 595 364	147 422 056	52,6	2 359,5	14 742,2	44,9	6,2	146,2
		ES	10 000	526 508	26 608 043	166 709 223	52,7	2 660,8	16 670,9	50,5	6,3	165,5
EuroParl	Parallel	FR	9 442	2 153 834	57 227 654	404 311 275	228,1	6 061,0	42 820,5	26,6	7,1	373,0
		EN	9 565	2 121 685	54 616 735	358 200 651	221,8	5 710,1	37 449,1	25,7	6,6	326,8
		ES	9 431	2 070 442	56 749 713	384 043 899	219,5	6 017,4	40 721,4	27,4	6,8	353,4
Wikipédia	Comparable	FR	10 000	276 870	11 053 888	76 882 428	27,7	1 105,4	7 688,2	39,9	7,0	76,7
		EN	10 000	559 103	18 043 902	120 951 564	55,9	1 804,4	12 095,2	32.3	6,7	120,8
		ES	10 000	244 051	9 601 103	65 730 068	24,4	960,1	6 573,0	39,3	6,8	65,7
PAN-PC-11	Parallel	EN	2 920	173 315	3 737 788	22 761 847	59,4	1 280,1	7 795,2	21,6	6,1	22,6
		ES	2 920	158 693	4 053 139	25 751 371	54,3	1 388,1	8 819,0	25,5	6,4	25,7
Webis-CLS-10	Parallel	FR	6 000	32 564	615 922	3 766 892	5,4	102,7	627,8	18,9	6,1	3,7
		EN	6 000	33 323	563 565	3 227 877	5,6	93,9	538,0	16,9	5,7	3,3
TALN-ACL	Comparable	FR	620	1 968	45 584	311 350	3,2	73,5	502,2	23,2	6,8	0,315
		EN	620	2 020	42 923	272 107	3,3	69,2	438,9	21,2	6,3	0,274
		Total	107 518	9 392 666	291 754 260	1 947 068 153	71,0	2 147,1	14 276,9	30,9	6,5	1 849,5

Tableau B.1 – Statistiques détaillées du sous-corpus des documents de notre jeu de données.

Sous-corpus	Type	Langue	# de fichiers	# de phrases	# de mots	# de caractères	phrases par fichier	mots par fichier	caractères par fichier	Taille moyenne des phrases (en mots)	Taille moyenne des mots (en caractères)	Taille (Mo)
JRC-Acquis	Parallel	FR	9 625	149 506	7 016 268	45 437 915	15,5	729,0	4 720,8	46,9	6,5	45,4
		EN	9 625	149 506	6 965 885	40 401 178	15,5	723,7	4 197,5	46,6	5,8	40,4
		ES	9 625	149 506	7 520 030	46 356 538	15,5	781,3	4 816,3	50,3	6,2	46,4
EuroParl	Parallel	FR	3 716	475 834	14 202 625	97 125 056	128,1	3 822,0	26 137,0	29,8	6,8	97,1
		EN	3 716	475 834	13 823 237	86 533 406	128,1	3 719,9	23 286,7	29,1	6,3	86,5
		ES	3 716	475 834	14 797 590	95 974 395	128,1	3 982,1	25 827,3	31,1	6,5	96,0
Wikipédia	Comparable	FR	2 472	4 792	233 239	1 882 680	1,9	94,4	761,6	48,7	8,1	1,9
		EN	2 472	4 792	213 210	1 735 826	1,9	86,3	702,2	44,5	8,1	1,7
		ES	2 472	4 792	231 910	1 842 072	1,9	93,8	745,2	48,4	7,9	1,8
PAN-PC-11	Parallel	EN	2 669	88 977	2 301 464	13 509 610	33,3	862,3	5 061,7	25,9	5,9	13,5
		ES	2 669	88 977	2 500 605	15 375 164	33,3	936,9	5 760,6	28,1	6,1	15,4
Webis-CLS-10	Parallel	FR	5 479	23 235	485 541	3 044 283	4,2	88,6	555,6	20,9	6,3	3,0
		EN	5 479	23 235	446 558	2 636 697	4,2	81,5	481,2	19,2	5,9	2,6
TALN-ACL	Comparable	FR	547	1 304	31 353	217 228	2,4	57,3	397,1	24,0	6,9	0,217
		EN	547	1 304	29 336	188 307	2,4	53,6	344,3	22,5	6,4	0,188
		Total	64 829	2 117 428	70 798 851	452 260 355	34,4	1 074,2	6 919,7	34,4	6,6	452,1

Tableau B.2 – Statistiques détaillées du sous-corpus des phrases de notre jeu de données.

Sous-corpus	Type	Langue	# de fichiers	# de phrases	# de mots	# de caractères	phrases par fichier	mots par fichier	caractères par fichier	Taille moyenne des phrases (en mots)	Taille moyenne des mots (en caractères)	Taille (Mo)
JRC-Acquis	Parallel	FR	2 964	10 094	57 954	453 489	3,4	19,6	153,0	5,7	7,8	0,453
		EN	2 964	10 094	54 739	419 425	3,4	18,5	141,5	5,4	7,7	0,419
		ES	2 964	10 094	57 512	432 168	3,4	19,4	145,8	5,7	7,5	0,432
EuroParl	Parallel	FR	2 277	25 603	132 987	1 056 280	11,2	58,4	463,9	5,2	7,9	1,100
		EN	2 277	25 603	126 163	995 639	11,2	55,4	437,3	4,9	7,9	0,996
		ES	2 277	25 603	132 297	1 034 666	11,2	58,1	454,4	5,2	7,8	1,000
Wikipédia	Comparable	FR	102	132	709	4 598	1,3	7,0	45,1	5,4	6,5	0,011
		EN	102	132	664	4 451	1,3	6,5	43,6	5,0	6,7	0,011
		ES	102	132	698	4 579	1,3	6,8	44,9	5,3	6,6	0,011
PAN-PC-11	Parallel	EN	692	1 360	5 819	37 972	2,0	8,4	54,9	4,3	6,5	0,044
		ES	692	1 360	5 799	37 888	2,0	8,4	54,8	4,3	6,5	0,044
Webis-CLS-10	Parallel	FR	1 752	2 603	12 309	78 915	1,5	7,0	45,0	4,7	6,4	0,079
		EN	1 752	2 603	11 934	75 630	1,5	6,8	43,2	4,6	6,3	0,076
TALN-ACL	Comparable	FR	185	272	1 302	9 551	1,5	7,0	51,6	4,8	7,3	0,010
		EN	185	272	1 288	9 062	1,5	7,0	49,0	4,7	7,0	0,009
		Total	21 287	115 957	602 174	4 654 313	3,8	19,6	148,5	5,0	7,1	4,7

Tableau B.3 – Statistiques détaillées du sous-corpus des syntagmes de notre jeu de données.

C Résultats de notre évaluation par couple de langues

Méthode	Wikipédia (%)	PAN-PC-11 (%)	JRC-Acquis (%)	Europarl (%)	Moyenne (%)
CL-C3G	72,73 ±0,662	58,80 ±0,579	35,04 ±0,726	46,23 ±0,458	43,75
CL-CTS	62,08 ±1,113	46,86 ±0,552	25,06 ±0,870	42,21 ±0,663	37,80
CL-ASA	46,43 ±0,745	23,21 ±0,386	35,65 ±0,898	43,78 ±0,746	40,83
CL-ESA	67,18 ±0,492	23,93 ±0,620	14,12 ±0,974	14,26 ±0,434	14,76
T+MA	66,64 ±0,656	54,27 ±0,843	26,37 ±0,857	32,53 ±0,746	31,77

Tableau C.1 – Performances des méthodes de l'état de l'art évaluées sur les sous-corpus en→es à la granularité syntagmatique.

Méthode	Wikipédia (%)	PAN-PC-11 (%)	JRC-Acquis (%)	Europarl (%)	Moyenne (%)
CL-C3G	63,36 ±0,543	45,65 ±0,911	35,69 ±0,897	37,33 ±1,231	38,19
CL-CTS	51,14 ±0,949	41,18 ±0,845	27,31 ±0,517	31,13 ±0,665	31,71
CL-ASA	23,34 ±0,547	12,91 ±0,671	11,29 ±0,388	34,51 ±1,463	26,94
CL-ESA	62,92 ±0,882	15,22 ±0,718	12,86 ±0,758	12,68 ±0,934	13,37
T+MA	61,12 ±0,843	44,78 ±0,823	33,62 ±1,055	33,42 ±0,990	35,05

Tableau C.2 – Performances des méthodes de l'état de l'art évaluées sur les sous-corpus en→es à la granularité phrastique.

Méthode	Wikipédia (%)	PAN-PC-11 (%)	JRC-Acquis (%)	Europarl (%)	Moyenne (%)
CL-C3G	72,73 ±0,662	58,80 ±0,579	35,04 ±0,726	46,23 ±0,458	43,75
CL-CTS	62,23 ±0,656	46,97 ±0,609	24,66 ±0,860	43,83 ±0,576	38,81
CL-ASA	44,82 ±1,416	23,10 ±0,351	34,54 ±0,616	42,17 ±0,852	39,41
CL-ESA	67,18 ±0,492	23,93 ±0,620	14,12 ±0,974	14,26 ±0,434	14,76
T+MA	67,05 ±0,677	54,54 ±0,860	25,89 ±0,806	34,18 ±0,765	32,79

Tableau C.3 – Performances des méthodes de l'état de l'art évaluées sur les sous-corpus es→en à la granularité syntagmatique.

Méthode	Wikipédia (%)	PAN-PC-11 (%)	JRC-Acquis (%)	Europarl (%)	Moyenne (%)
CL-C3G	63,36 ±0,543	45,65 ±0,911	35,69 ±0,897	37,33 ±1,231	38,19
CL-CTS	50,94 ±1,393	41,48 ±0,639	28,81 ±5,210	31,10 ±0,748	32,04
CL-ASA	20,21 ±0,839	11,79 ±0,565	12,42 ±0,638	31,94 ±0,420	25,31
CL-ESA	62,92 ±0,882	15,22 ±0,718	12,86 ±0,758	12,68 ±0,934	13,37
T+MA	62,78 ±0,571	44,22 ±0,730	32,82 ±0,507	34,08 ±0,551	35,26

Tableau C.4 – Performances des méthodes de l'état de l'art évaluées sur les sous-corpus es→en à la granularité phrastique.

Méthode	Wikipédia (%)	JRC-Acquis (%)	Europarl (%)	Moyenne (%)
CL-C3G	69,76 ±0,654	39,31 ±1,218	51,24 ±0,619	47,95
CL-CTS	63,36 ±0,801	31,29 ±0,574	46,16 ±0,468	42,03
CL-ASA	39,98 ±0,682	33,99 ±0,634	38,68 ±0,639	37,36
CL-ESA	65,77 ±0,458	15,06 ±0,708	15,00 ±0,407	15,20
T+MA	65,48 ±0,727	24,90 ±0,708	34,04 ±0,823	31,58

Tableau C.5 – Performances des méthodes de l'état de l'art évaluées sur les sous-corpus es→fr à la granularité syntagmatique.

Méthode	Wikipédia (%)	JRC-Acquis (%)	Europarl (%)	Moyenne (%)
CL-C3G	55,98 ±0,388	41,09 ±0,258	47,14 ±0,735	45,77
CL-CTS	51,76 ±0,430	31,82 ±0,611	50,99 ±1,074	46,45
CL-ASA	20,94 ±0,703	13,85 ±0,557	36,46 ±0,897	30,98
CL-ESA	61,51 ±0,521	13,08 ±0,424	13,59 ±0,613	13,83
T+MA	59,74 ±0,740	39,27 ±0,700	35,70 ±0,708	36,73

Tableau C.6 – Performances des méthodes de l'état de l'art évaluées sur les sous-corpus es→fr à la granularité phrastique.

Méthode	Wikipédia (%)	JRC-Acquis (%)	Europarl (%)	Moyenne (%)
CL-C3G	69,76 ±0,654	39,31 ±1,218	51,24 ±0,619	47,95
CL-CTS	63,99 ±0,803	30,81 ±0,829	45,86 ±0,667	41,69
CL-ASA	42,85 ±0,885	35,11 ±0,770	35,48 ±0,571	35,40
CL-ESA	65,77 ±0,458	15,06 ±0,708	15,00 ±0,407	15,20
T+MA	65,79 ±0,699	25,01 ±0,588	33,74 ±0,702	31,40

Tableau C.7 – Performances des méthodes de l'état de l'art évaluées sur les sous-corpus fr→es à la granularité syntagmatique.

Méthode	Wikipédia (%)	JRC-Acquis (%)	Europarl (%)	Moyenne (%)
CL-C3G	55,98 ±0,388	41,09 ±0,258	47,14 ±0,735	45,77
CL-CTS	51,51 ±0,413	31,76 ±0,730	50,09 ±0,751	45,75
CL-ASA	32,72 ±0,911	19,04 ±0,732	31,34 ±0,618	28,43
CL-ESA	61,51 ±0,521	13,08 ±0,424	13,59 ±0,612	13,83
T+MA	58,87 ±0,677	38,67 ±0,387	33,94 ±0,694	35,25

Tableau C.8 – Performances des méthodes de l'état de l'art évaluées sur les sous-corpus fr→es à la granularité phrastique.

Méthode	Wikipédia (%)	TALN-ACL (%)	JRC-Acquis (%)	Webis-CL-10 (%)	Europarl (%)	Moyenne (%)
CL-C3G	62,91 ±0,815	40,90 ±0,500	36,63 ±0,826	80,30 ±0,703	53,29 ±0,583	50,71
CL-CTS	57,63 ±0,701	36,97 ±0,386	29,12 ±0,843	67,79 ±0,623	43,16 ±0,860	41,16
CL-ASA	41,82 ±1,004	42,41 ±0,827	42,81 ±1,910	25,25 ±2,793	44,16 ±1,620	42,52
CL-ESA	64,89 ±0,664	23,78 ±0,613	14,03 ±0,997	23,14 ±0,777	14,19 ±0,590	14,99
T+MA	58,29 ±0,767	39,41 ±0,566	27,23 ±0,713	73,08 ±0,525	36,05 ±1,092	36,34

Tableau C.9 – Performances des méthodes de l'état de l'art évaluées sur les sous-corpus fr→en à la granularité syntagmatique.

Méthode	Wikipédia (%)	TALN-ACL (%)	JRC-Acquis (%)	Webis-CL-10 (%)	Europarl (%)	Moyenne (%)
CL-C3G	48,25 ±0,349	48,08 ±0,538	36,68 ±0,693	61,10 ±0,581	52,72 ±0,867	49,31
CL-CTS	45,58 ±0,549	38,17 ±0,642	27,07 ±0,715	49,96 ±0,925	52,23 ±0,604	46,33
CL-ASA	32,18 ±1,582	25,67 ±0,338	34,81 ±0,413	24,73 ±0,908	35,93 ±1,023	35,23
CL-ESA	51,14 ±0,874	14,25 ±0,334	14,44 ±0,341	13,93 ±0,714	13,91 ±0,618	14,30
T+MA	51,79 ±0,866	37,33 ±0,374	31,78 ±0,388	60,69 ±0,846	37,22 ±0,440	36,92

Tableau C.10 – Performances des méthodes de l'état de l'art évaluées sur les sous-corpus fr→en à la granularité phrastique.

	Corpus de SemEval-2016		Corpus de SemEval-2017		
Méthode	News	Multi-Src	SNLI	WMT	Moyenne
CL-C3G	0,7447	0,6348	0,6562	0,0087	0,5111
CL-CTS	0,7500	0,4768	0,4766	0,0731	0,4441
CL-ASA	0,4383	0,4953	0,5205	-0,0312	0,3557
CL-ESA	0,4126	0,4958	0,4855	-0,0384	0,3389
T+MA	0,4993	0,5249	0,5381	0,0435	0,4014

Tableau C.11 – Corrélations entre les résultats des méthodes et les annotations manuelles des corpus en→es des campagnes SemEval 2016 et 2017.

D Poids attribués aux étiquettes morphosyntaxiques

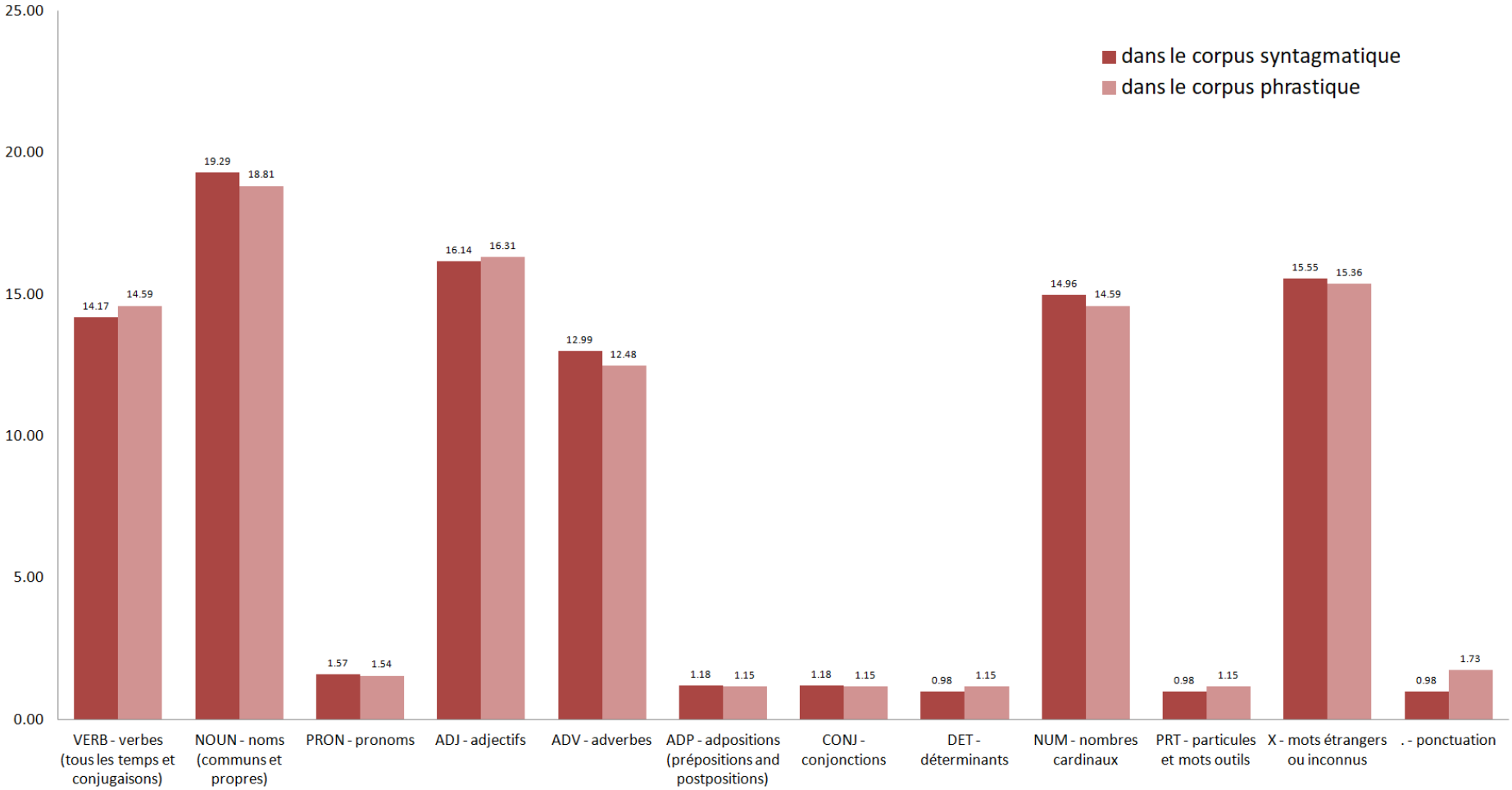


FIGURE D.1 – Poids attribués (en %) aux différentes étiquettes morphosyntaxiques après optimisation.

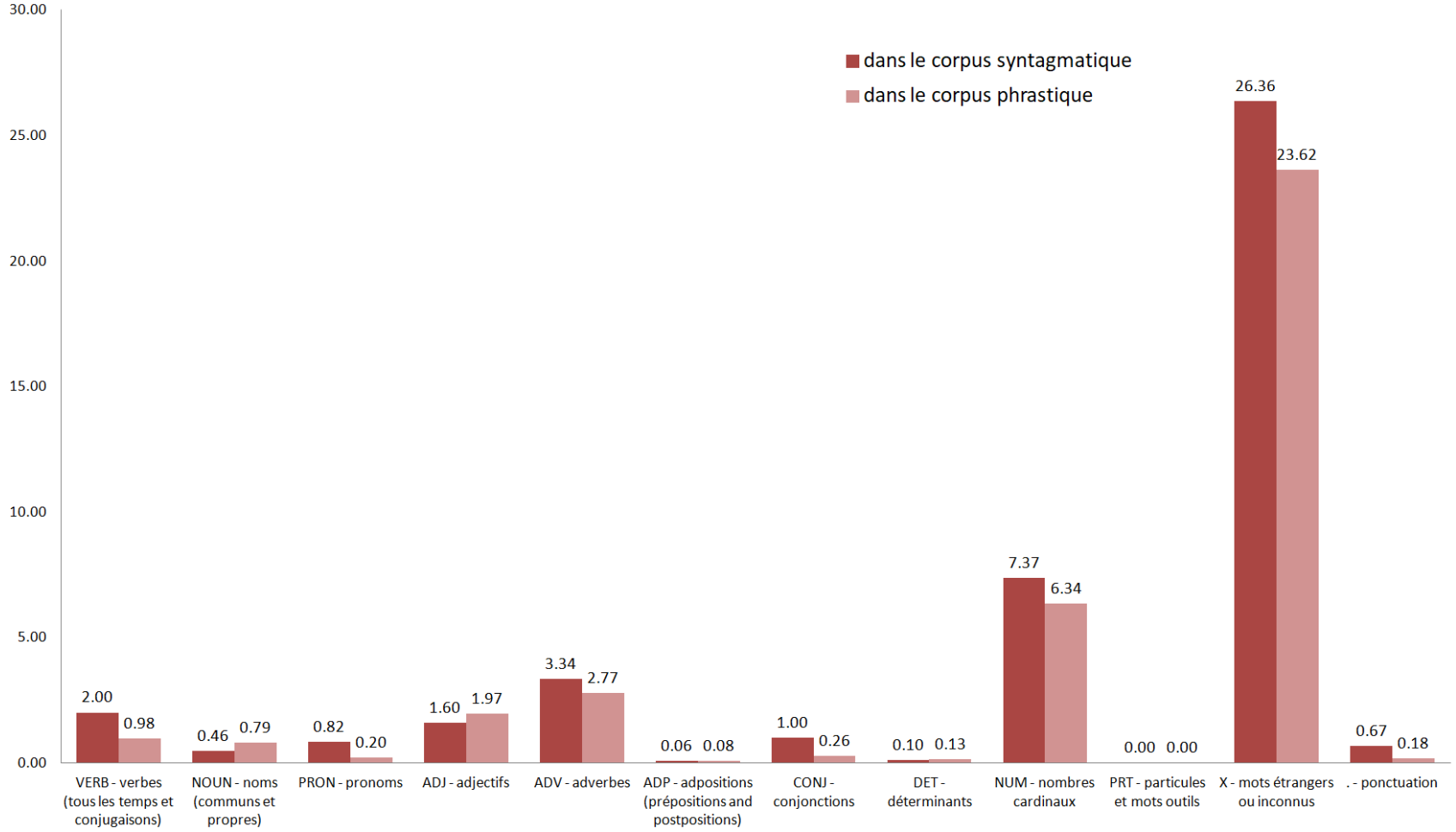


FIGURE D.2 – Poids attribués (en %) aux différentes étiquettes morphosyntaxiques après optimisation en les normalisant en fonction de la probabilité d'apparition des étiquettes morphosyntaxiques.

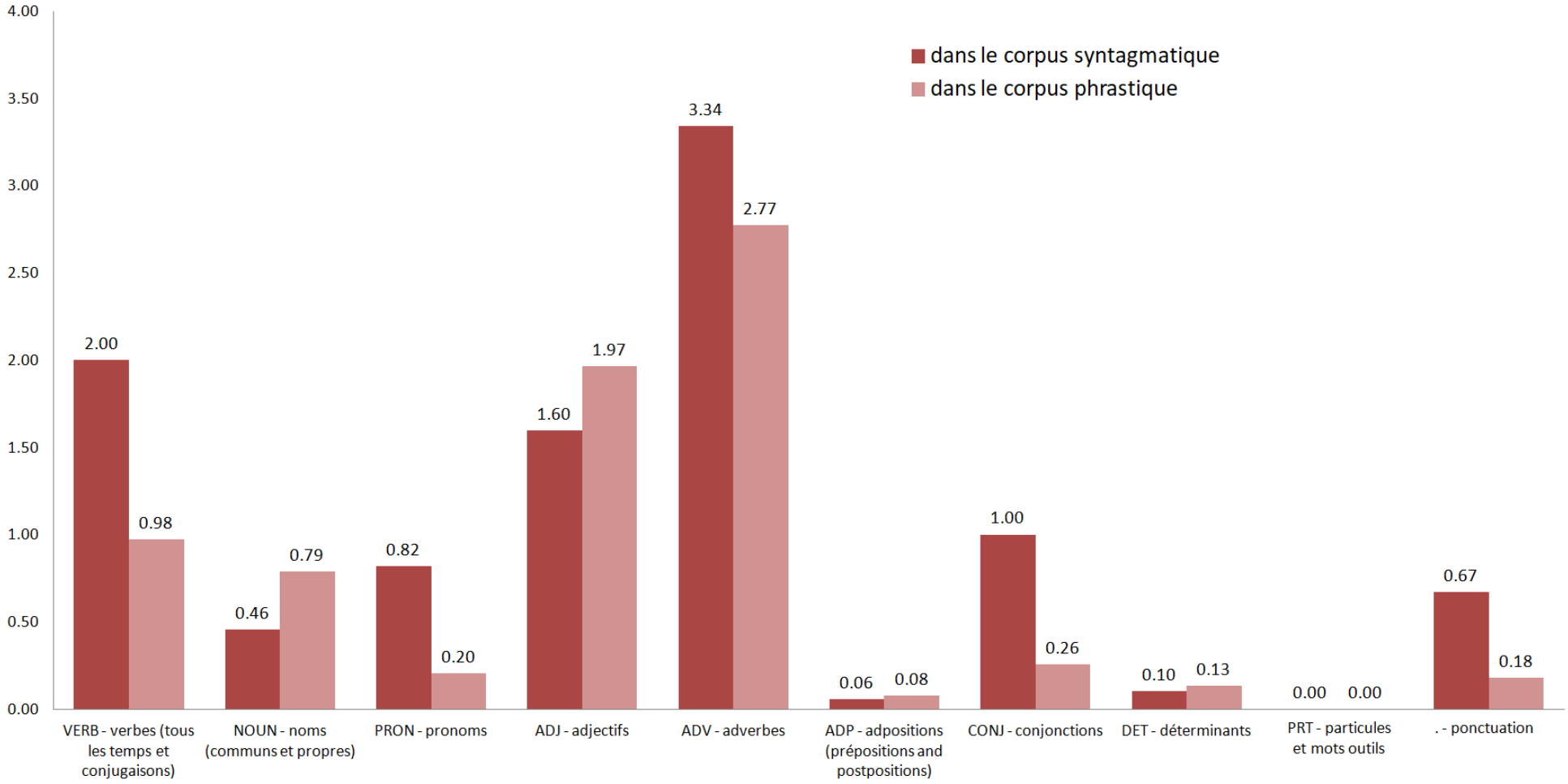


FIGURE D.3 – Poids attribués (en %) aux différentes étiquettes morphosyntaxiques après optimisation en les normalisant en fonction de la probabilité d'apparition des étiquettes morphosyntaxiques, sans les étiquettes les plus fréquentes.

E Probabilités d'apparition des étiquettes morphosyntaxiques au sein du corpus de développement

Étiquettes morphosyntaxiques à la sortie de *TreeTagger* (Schmid, 1994)	Étiquettes morphosyntaxiques universelles de Petrov *et al.* (2012)	Probabilités d'apparition dans le corpus (%)
PUN	.	1,47
ADJ	ADJ	10,12
PRP:det	ADP	3,01
PRP	ADP	17,09
ADV	ADV	3,89
KON	CONJ	1,18
DET:ART	DET	9,51
NOM	NOUN	42,02
NAM	NOUN	0,1
NUM	NUM	2,03
DET:POS	PRON	1,19
PRO:DEM	PRON	0,73
VER:subp	VERB	0,62
VER:infi	VERB	0,86
VER:simp	VERB	0,89
VER:subi	VERB	0,72
VER:pper	VERB	1,44
VER:pres	VERB	0,94
VER:ppre	VERB	0,68
VER:futu	VERB	0,76
PRO:IND	VERB	0,14
PRO:PER	VERB	0,02
PRO:REL	VERB	0,02
INT	X	0,59
	Total	100,02

Tableau E.1 – Probabilités d'apparition des étiquettes morphosyntaxiques dans le corpus de développement à la granularité syntagmatique. Les étiquettes qui n'apparaissent pas dans le corpus (avec 0% comme probabilité d'apparition) ne sont pas reportées ici.

Étiquettes morphosyntaxiques à la sortie de *TreeTagger* (Schmid, 1994)	Étiquettes morphosyntaxiques universelles de Petrov *et al.* (2012)	Probabilités d'apparition dans le corpus (%)
SENT	.	2,88
PUN	.	6,4
PUN:cit	.	0,19
ADJ	ADJ	8,3
PRP	ADP	11,54
PRP:det	ADP	3,67
ADV	ADV	4,5
KON	CONJ	4,5
DET:ART	DET	8,76
NOM	NOUN	21,17
NAM	NOUN	2,65
NUM	NUM	2,3
PRO:PER	PRON	3,18
PRO:DEM	PRON	1,57
DET:POS	PRON	0,88
PRO:REL	PRON	1,26
PRO:IND	PRON	0,61
PRO:POS	PRON	0,01
PRO	PRON	0,01
VER:pres	VERB	5,08
VER:ppe	VERB	3,44
VER:simp	VERB	0,52
VER:futu	VERB	1,3
VER:subp	VERB	0,22
VER:subi	VERB	0,58
VER:infi	VERB	2,76
VER:impf	VERB	0,23
VER:cond	VERB	0,4
VER:ppre	VERB	0,43
ABR	X	0,57
INT	X	0,01
SYM	X	0,07
	Total	99,99

Tableau E.2 – Probabilités d'apparition des étiquettes morphosyntaxiques dans le corpus de développement à la granularité phrastique. Les étiquettes qui n'apparaissent pas dans le corpus (avec 0% comme probabilité d'apparition) ne sont pas reportées ici.

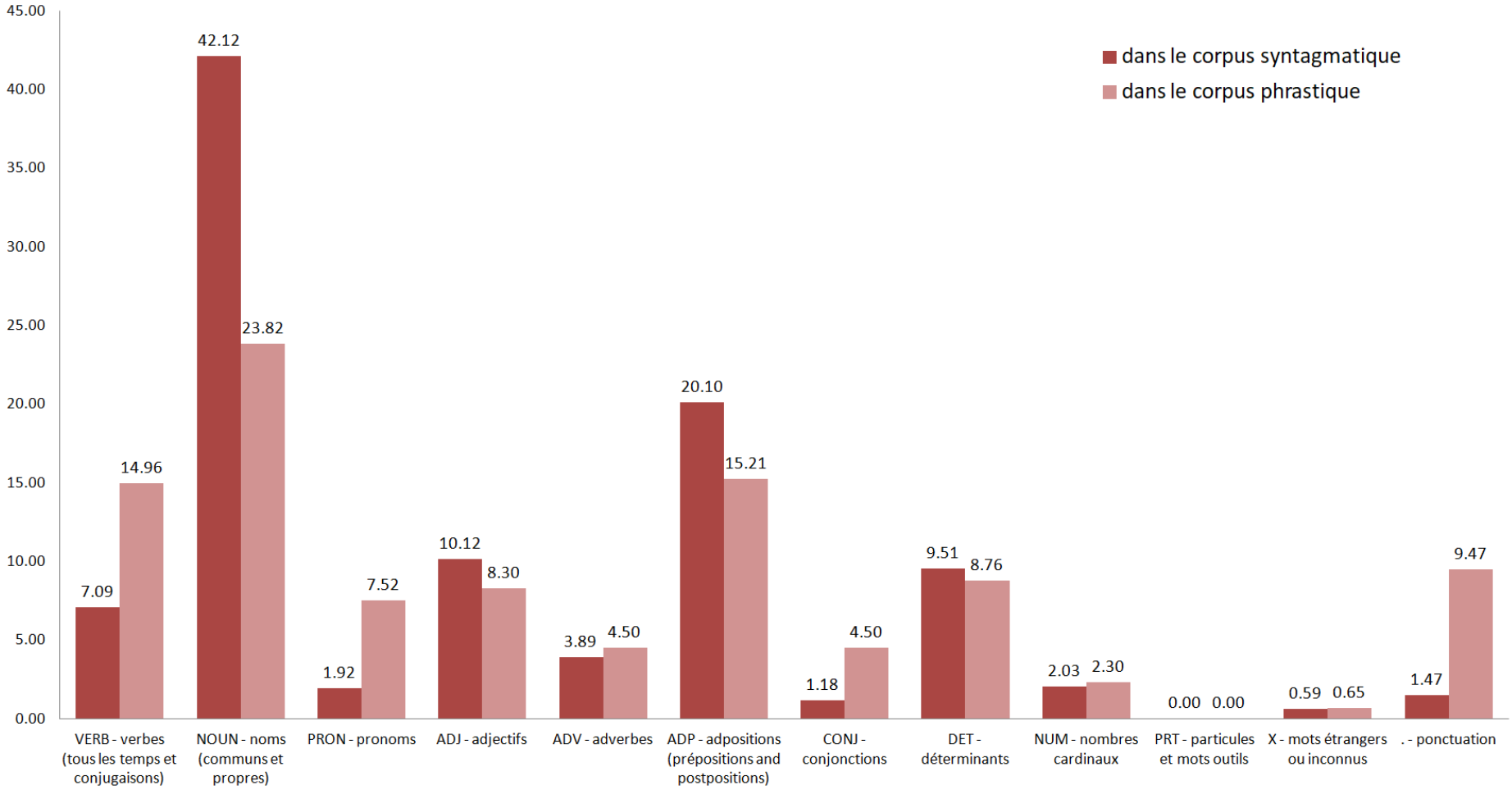


FIGURE E.1 – Probabilités d'apparition (en %) des différentes étiquettes morphosyntaxiques au sein du corpus de développement.