# Détection automatique de reformulations - Correspondance de concepts appliquée à la détection du plagiat

Jérémy Ferrero, Alain Simac-Lejeune

# Détection automatique de reformulations - Correspondance de concepts appliquée à la détection du plagiat

Jérémy Ferrero*, Alain Simac-Lejeune**

Compilatio
276, rue du Mont-Blanc
74520 Saint-Félix, France
*jeremyf@compilatio.net
**alain@compilatio.net

**Résumé.** Dans le cadre de la détection du plagiat, la phase de comparaison de deux documents est souvent réduite à une comparaison mot à mot, une recherche de « copier/coller ». Dans cet article, nous proposons une approche naïve de comparaison de deux documents dans le but de détecter automatiquement aussi bien les phrases copiées de l'un des textes dans l'autre que les paraphrases et reformulations, ceci en se focalisant sur l'existence des mots porteurs de sens, ainsi que sur leurs mots de substitution possibles. Nous comparons trois algorithmes utilisant cette approche afin de déterminer la plus efficace pour ensuite l'évaluer face à des méthodes existantes. L'objectif est de permettre la détection des similitudes entre deux textes en utilisant uniquement des mots clefs. L'approche proposée permet de détecter des reformulations non paraphrastiques impossibles à détecter avec des approches conventionnelles faisant appel à une phase d'alignement.

## 1 Introduction

Actuellement, la recherche et la détection de similitudes s'effectuent en deux phases : une première phase de recherche de sources candidates, suivie d'une seconde de comparaison de ces sources possibles avec le document que l'on suspecte d'être un plagiat. La phase de collecte est de plus en plus optimale grâce à l'amélioration de l'efficacité des moteurs de recherche en local et sur le Web. C'est à la seconde phase que cet article s'intéresse. Une fois qu'une source candidate est trouvée, elle doit être comparée avec le document sur lequel pèse les soupçons. À l'heure actuelle, la plupart des logiciels anti-plagiat, une fois une liste de sources candidates constituée, se contentent de comparer mot à mot le document analysé avec chaque source possible. Cette technique permet seulement de détecter les similitudes de types « copier/coller ». Bien que cette approche ait prouvé son efficacité et suffise la plupart du temps, en France près d'un étudiant sur deux a déjà eu recours au « copier/coller » (Gibney, 2006), une énorme faille persiste. En effet, le fait de reformuler ou tout simplement de paraphraser un texte, en utilisant des synonymes par exemple, rend la plupart des techniques actuelles caduques. Certains articles (Callison-Burch et al., 2008; Bannard et Callison-Burch, 2005) se sont tout de même

intéressés à la détection de reformulations paraphrastiques avec des approches d'alignement. Malgré le fait que ces approches soient plus robustes à l'ajout et à la suppression de mots ainsi qu'à l'utilisation de synonymes, elles restent toutefois inefficaces face aux reformulations non paraphrastiques comme le passage de la forme active à la forme passive. L'approche proposée consiste à comparer les deux textes, phrase par phrase, et non plus mot à mot et à rechercher si une phrase de l'un des textes comporte le même sens qu'une phrase dans l'autre texte. *Ceci repose sur l'hypothèse que lorsqu'on paraphrase ou reformule un texte, on garde le sens de celui-ci et ainsi on garde les mots-clés principaux, porteurs du plus de sens de chaque phrase*. Après avoir défini quelques notions et présenté l'état de l'art, nous décrirons d'abord comment segmenter le texte en unités de sens, pour ensuite procéder sur chacune de ces unités à l'extraction des mots porteurs de sens, afin de rechercher des concordances de mots de même concept dans un autre texte. Enfin, nous testerons trois algorithmes utilisant notre approche et nous déterminerons le meilleur seuil pour chacun d'entre eux. Le seuil est le nombre minimum de concepts identiques dans deux phrases permettant d'affirmer que l'une est la reformulation de l'autre. Pour finir, nous présenterons l'évaluation de notre approche en comparant la méthode retenue aux méthodes classiques de détection des paraphrases par alignement.

# 2 La comparaison au-delà du « copier/coller »

## 2.1 La notion de comparaison

La « comparaison de deux documents » est un terme assez vague. Pour comparer correctement deux documents, il faut repérer leurs points communs (leurs similitudes) et leurs différences. Les similitudes étant plus simples à détecter, il est de convention de chercher à repérer celles-ci en premier lieu et d'en déduire ensuite les différences, représentées alors par le reste du document. Cependant, la plupart des comparaisons textuelles se limitent au « copier/coller », or ce ne sont pas les seules similitudes pouvant être recensées dans un texte.

## 2.2 La notion de similitudes

Bien que l'on puisse avoir au sein d'un document des tableaux, images, graphiques ou tout autre type de données, cet article traite seulement des similitudes d'ordre textuel. On distingue plusieurs types de similitudes allant de la ressemblance jusqu'à l'identité même (Simac-Lejeune, 2013b). Les ressemblances sont les types de similitudes les plus difficiles à repérer et sont pour cause le point faible des logiciels anti-plagiat actuels. Dans notre cas, on distingue trois types majeurs de similitudes textuelles, de la plus simple à détecter à la plus complexe :

— la **copie**, 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, qui consiste à reprendre une phrase d'un texte pour la détailler ou l'expliciter. Elle conserve donc l'ordre des éléments évoqués, autorisant simplement le changement de vocabulaire, l'ajout, la suppression et la substitution de mots. Toujours en considérant la phrase de l'exemple

précédent, une paraphrase possible serait :
*« 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 10000. »*
On remarque la conservation des concepts, mais aussi la substitution ou la suppression de certains d'entre eux ;

— la **reformulation**, qui autorise elle toutes modifications textuelles à condition que le sens de la phrase soit conservé. Cela donne souvent lieu à un changement d'ordre des concepts. La reformulation de la phrase exemple serait :
*« 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. »*

## 2.3 La notion de concept

Un concept est une idée, un sens représenté par un mot ou un groupe de mots. Les reformulations et paraphrases exploitent les propriétés paradigmatiques des mots (leur capacité à se substituer mutuellement) et entraînent ainsi des changements de vocabulaire mais elles conservent les concepts et les idées exprimées (Duclaye, 2003). Il est alors, dans le cadre de la détection de similitudes, plus judicieux de représenter un mot par un concept plutôt que par son identité ou sa définition. Par exemple, il est plus judicieux de représenter un mot par un tableau de tous les mots par lesquels il peut être substitué (un tableau de ses synonymes, lui compris) plutôt que seulement par lui-même.

## 2.4 État de l'art

Lorsque les processus anti-plagiat comparent deux documents, ils recherchent les éléments de l'un également présents dans l'autre. Ils tentent de détecter des similitudes, toutes informations communes laissant penser qu'un plagiat a pu avoir lieu. La comparaison mot à mot est certes efficace pour trouver les zones de « copier/coller » mais les plagiaires ne se contentent plus de copier des éléments depuis une source, ils essaient à présent de camoufler leurs emprunts d'idées derrière des modifications syntaxiques. Les recherches de Barron-Cedeño et al. (2013) se concentrant sur la détection de paraphrases appliquée dans le cadre de la détection du plagiat démontrent que le phénomène de paraphrasage nuit aux systèmes anti-plagiat et rend la détection de similitudes plus difficile. Il faut donc tenter de détecter les paraphrases et les reformulations par des moyens différents, car bien que souvent associés ces deux termes représentent des opérations textuelles bien distinctes. Toutefois, les travaux linguistiques ayant portés sur leur définition, s'accordent sur le fait que ce sont des opérations de modifications de texte, certes bien différentes, mais qui conservent toutes deux le sens (Harris, 1957; Martin, 1976; Duclaye, 2003).

Des recherches (Gülich et Kotschi, 1983; Eshkol-Taravella et Grabar, 2014) se sont attardées à chercher des marqueurs de reformulations afin de mieux les repérer par la suite et d'étudier leur fonctionnement et leur construction. D'autres recherches se sont cantonnées à étudier les limites de la détection des paraphrases (Vila et al., 2011) en estimant au contraire qu'il n'existait pas de caractérisation complète sur le plan linguistique et computationnelle de la paraphrase.

Face à ces difficultés, des chercheurs se sont concentrés sur des approches alternatives ne permettant pas de détecter concrètement des reformulations mais de tout de même déterminer qu'un texte en contient :

— les approches stylométriques (Iyer et Singh, 2005) qui suggèrent qu'en analysant des statistiques de fréquences de mots ou bien d'autres 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 (Oberreuter et Velásquez, 2013; van Halteren, 2004; Jardino et al., 2007) ;
— les approches de calcul de distances (Simac-Lejeune, 2013a) qui propose de calculer une distance « sémantique » entre deux textes après avoir extrait les mots clefs de chaque texte, exposant ainsi l'emprunt probable de l'un dans l'autre.

En dehors de ces approches, la majorité des travaux portent sur la détection des reformulations paraphrastiques, comme les recherches de Eshkol-Taravella et Grabar (2014) portant sur leur détection dans des corpus oraux. Les approches les plus répandues sont les méthodes par alignement (Callison-Burch et al., 2008; Bannard et Callison-Burch, 2005). Servant la plupart du temps dans un contexte bi-linguale (alignement d'un texte et de sa traduction), elles consistent à aligner deux textes par leurs mots ou groupes de mots en communs et ainsi de repérer les mots ou groupes de mots différents mais équivalents. Certaines recherches (Shen et al., 2006), visant à produire des paraphrases, se sont également avérées intéressantes. En effet, étudiant la possibilité de générer automatiquement des paraphrases, un processus d'assemblage puis de désassemblage s'est dégagé, remettant ainsi sur le devant de la scène les méthodes d'alignement. Proche de ces méthodes avec alignement, on peut citer le travail de Fenoglio et al. (2007) traitant de la comparaison de versions de documents textuels à la façon des serveurs de versions. Il met en lumière les transformations élémentaires (déplacements, insertions, suppressions et remplacements de blocs de caractères), identifiées depuis longtemps par les spécialistes de la génétique textuelle (de Biasi, 2000; Grésillon, 1994) comme éléments fondateurs d'une paraphrase.

Toutefois, le cadre théorique le plus souvent adopté est la théorie linguistique Sens-Texte (Kahane, 2003) élaborée dans les années 1960 par Mel'čuk, notamment son système de paraphrasage (Žolkovskij et Mel'čuk, 1967; Mel'čuk, 1992; Milićević, 2007) comme dans le travail de Milićević (2010). Ce dernier met également en avant des approches sémantiques qui permettent de s'approcher d'une détection de reformulations. La plupart des règles sémantiques de paraphrasage trouvées jusqu'ici mettent en jeu un découpage du texte en proposition et des liens communicatifs et rhétoriques entre celles-ci (Danlos, 2006), coïncidant ainsi, dans la plupart des cas, à la définition d'une reformulation qui se contente d'être une paraphrase avec changement d'ordre des propositions.

La reformulation non paraphrastique étant bien plus complexe à détecter que sa voisine la paraphrase, les études se concentrant uniquement sur elle se font plus rares. Mais dès lors qu'on sait que la reformulation conserve également le sens du texte (Harris, 1957; Martin, 1976; Duclaye, 2003) et que le mécanisme de paraphrase le plus utilisé est le changement de lexique (Barron-Cedeño et al., 2013), on peut envisager d'appliquer plus ou moins les mêmes approches sémantiques que pour la paraphrase ou bien même, des approches plus naïves de recherche de correspondances de concepts.

# 3 Notre approche

## 3.1 Segmentation

Dans un premier temps, l'idée est de segmenter le document que l'on suspecte être un plagiat. Plusieurs algorithmes de segmentation ont été évalués :

— la segmentation par nombre de blocs : on découpe le document en un certain nombre de blocs de même taille (de même nombre de mots), peu importe la taille finale de chaque bloc ;
— la segmentation par taille de blocs : on découpe le document par blocs d'une certaine taille (un certain nombre de mots), peu importe le nombre de blocs créés ;
— la segmentation par pourcentage que représente un bloc sur l'ensemble du document (e.g. une segmentation comme celle-ci avec en paramètre un pourcentage de 1% pour un bloc reviendrait à une segmentation en 100 blocs, chaque bloc représentant 1%) ;
— la segmentation par granularité (Simac-Lejeune, 2013b) : il s'agit d'une segmentation hiérarchique, on découpe le texte en $nb$ blocs de même taille, puis on redécoupe chaque bloc ainsi obtenu en $nb$ blocs de même taille, et ainsi de suite sur une profondeur limite définie. Ceci permettant d'affiner l'analyse niveau par niveau ;
— la segmentation par paragraphe : chaque segment représentant un paragraphe ;
— la segmentation par phrase : chaque segment représentant une phrase du document ;
— la segmentation par proposition : chaque segment représentant une unité minimale de sens, les délimiteurs étant la ponctuation de fin de phrase mais aussi les virgules, les conjonctions de coordination et divers mots de liaison ou de causalité.

Chaque algorithme a fait l'étude, via de nombreux tests et corrections, à l'optimisation de ses paramètres afin de mettre l'accent sur la rapidité du processus de découpage mais aussi sur la pertinence des métadonnées extraites dans chaque segment. Il est important que chaque segment conserve un sens afin d'être potentiellement sujet à une reformulation. Une segmentation en unité de sens a donc été choisie. Un découpage par phrase ou par proposition est à privilégier car une segmentation à faible granularité, comme celle par paragraphe, donne lieu à des segments trop volumineux pour l'étape d'extraction qui suivra. Au contraire, une segmentation à trop grand niveau de granularité pourrait, en plus d'entraîner un temps d'exécution plus important (plus de segments à traiter), occasionner une perte d'informations dans sa globalité (aucune liaison entre les concepts extraits). C'est pourquoi la segmentation qui a été retenue est une fusion du découpage par phrase et du découpage par taille de blocs : une segmentation par phrase mais d'une taille minimale (en mots). On conserve ainsi une unité de sens (une ou plusieurs phrases) tout en gardant une taille suffisamment importante pour pouvoir obtenir une pertinence raisonnable des métadonnées extraites mais suffisamment petite pour être considérée comme indépendante et donc éventuellement reformulée. Après divers tests, le seuil a été fixé à 15 mots, taille moyenne des phrases dans la langue française. Avec un seuil si faible, c'est l'une des méthodes de segmentation évaluée les plus chronophages mais pour notre étude elle garantit un rapport taille/pertinence optimal.

## 3.2 Extraction de mots clefs (mots porteurs de sens)

La seconde étape du processus est une étape d'extraction des mots porteurs de sens de chaque segment c'est-à-dire des mots représentant les concepts que le plagiaire a été obligé de

réutiliser s'il voulait conserver le sens de la phrase, même s'il a pu les remplacer par des synonymes. Pour déterminer les mots porteurs de sens d'un texte, l'étiqueteur morphosyntaxique TreeTagger (Schmid, 1994) a été utilisé. Il détermine la classe lexicale, le "Part Of Speech" de chaque unité lexicale (token) du texte. De façon plus commune, on peut dire que pour chaque mot ou élément du texte, TreeTagger détermine s'il s'agit d'un nom, d'un verbe, d'un adjectif, d'une ponctuation, etc. L'étiquetage morphosyntaxique permet d'identifier les mots clefs d'un texte par leur classe lexicale. Plutôt que de discriminer les mots vides (stop words) par leur taille, ceci pouvant générer des erreurs (e.g. un mot de moins de trois lettres n'est pas pertinent, un contre exemple est le mot « as » qui peut être important, et le mot « mais » qui est simplement une conjonction), on les discriminera par leur "Part Of Speech".

Dans notre cas, les mots pertinents à conserver sont un peu plus riches sémantiquement que les mots clefs habituels. On ne conserve pas seulement les noms communs et propres, il est important de garder aussi les adjectifs, les verbes et également les adverbes, en réalité tout mot porteur de sens au sein d'une phrase. On néglige donc les méthodes les plus courantes pour extraire des mots clefs, les méthodes fréquentielles (Lee et Baik, 2004) qui consiste pour chaque mot du texte à calculer sa fréquence d'apparition dans le texte. C'est pour cela que le terme de mots clefs est ici un abus de langage et que nous allons préférer le terme de mots porteurs de sens d'une phrase. Tout mot porteur de sens d'une phrase doit être conservé, peu importe son nombre d'occurrences dans le texte.

Un filtre de mots vides a été ajouté à la sortie de TreeTagger afin d'être certain de la pertinence des mots porteurs de sens extraits. Ainsi en couplant les deux techniques, l'efficacité de l'étiquetage est passée d'environ 96% à quasiment 100%.

Considérons la phrase suivante : *« Ce peu de masse disparue crée une grande quantité d'énergie comme le démontre la fameuse formule d'Einstein E=mc2. »*

Ses mots porteurs de sens extraits seraient *« peu, masse, disparue, crée, grande, quantité, énergie, démontre, fameuse, formule, Einstein, E=mc2 »*.

## 3.3 Thésaurus - chargement d'un dictionnaire de synonymes

Parallèlement à cela, un dictionnaire de synonymes est chargé. Pour chaque mot, on a donc accès à un tableau contenant tous les mots de la langue par lesquels il peut être substitué. L'efficacité de notre approche dépendant en grande partie du contenu de cette ressource, il est important de faire la différence entre des synonymes et des mots de substitution possibles. Par exemple, pour le mot « père », « papa » serait un synonyme alors que « parent » serait un mot de substitution envisageable. Autre exemple, le mot « île » a pour synonyme « îlot », « archipel » ou bien encore « atoll » mais aucunement les mots « tâche » ou « pâté » qui eux se trouvent pourtant dans notre tableau et peuvent servir de mot de substitution. En effet, on peut très bien imaginer dans un poème une phrase telle que « cette tâche au milieu de l'océan » faisant référence à un îlot. En règle générale, « îlot » et « tâche » ne sont pas synonymes mais ici, ils représentent le même concept.

Dès lors, un concept est un mot porteur de sens ainsi que tous ses mots de substitution possibles contenus dans son tableau.

Le tableau 1 représente une partie des mots de substitution correspondant aux mots porteurs de sens extraits sur la phrase exemple lors de l'étape précédente. On remarque, comme dans l'exemple de l'îlot cité précédemment, que le terme « énergie » laisse place à « assiduité » qui

n'a strictement rien à voir avec le contexte de notre phrase mais qui dans un autre contexte aurait très bien pu être un synonyme envisageable.

| Mots porteur de sens | Mots de substitution |
|---|---|
| peu | brin, grain, morceau, nuage, larme, miette, ... |
| masse | amas, tas, pile, collection, bloc, monolithe, ... |
| disparue | morte, décédée, passée, trépassée, tuée, ... |
| crée | accouchée, enfantée, fabriquée, composée, ... |
| grande | longue, importante, éternelle, prolixe, ... |
| quantité | avalanche, déluge, multitude, assemblage, amas, tas, ... |
| énergie | constance, persévérance, assiduité, permanence, ... |
| démontre | prouve, montre, justifie, affirme, témoigne, indique, ... |
| fameuse | glorieuse, célèbre, illustre, ... |
| formule | dicton, slogan, équation, proverbe, expression, ... |
| Einstein | - |
| E=mc2 | $E = mc^2$ |

TAB. 1 – *Tableau d'une partie des mots de substitution disponibles pour la phrase étudiée.*

Lors de cette étude, le dictionnaire utilisé fut le thesaurus v.2.3 en date du 20 décembre 2011 de LibreOffice v.3.4. Cette ressource se trouve en accès libre sur internet.

## 3.4 Correspondance

La dernière étape de notre approche consiste à comparer chaque phrase d'une source candidate avec les mots porteurs de sens de chaque segment du texte en cours d'analyse ainsi qu'avec leurs mots de substitution possibles contenus dans le tableau défini précédemment. On appellera *seuil de correspondance* le nombre de concepts communs à partir duquel on peut estimer qu'une phrase est la reformulation d'une autre. S'il y a plus de concepts pertinents communs entre deux phrases que le *seuil de correspondance* défini, c'est sans doute que l'une est une reformulation de l'autre.

Plusieurs algorithmes mettant en œuvre cette méthode ont été développés, certains plus robustes que d'autres face aux changements de genre, de nombre, de casse typographique ou bien d'ordre des mots (e.g. phrase passée de la forme active à la forme passive et vice versa). L'efficacité de la détection dépend de l'algorithme choisi, du *seuil de correspondance* défini, et du nombre et de la pertinence des « synonymes » disponibles dans le dictionnaire chargé.

Nous proposons trois algorithmes, trois implémentations différentes de l'approche décrite précédemment.

— un premier (tableau 2 - A) qui compare la présence des concepts dans l'ordre et tels qu'ils sont présents dans les phrases. Il ne supporte donc ni le changement de casse typographique, ni la dérivation et la flexion ;

— un second (tableau 2 - B) qui compare également la présence des concepts dans l'ordre des phrases mais en comparant leurs lemmes en minuscules, il supporte donc le changement de casse typographique, la dérivation et le changement de genre et de nombre ;

— un troisième (tableau 2 - C), plus naïf, qui reprend le principe du précédent, en comparant cette fois la présence des concepts dans les deux phrases sans prendre l'ordre en compte. Il est ainsi robuste aux reformulations non paraphrastiques de type mise à la forme passive.

Le tableau 2 résume les différentes variations de la langue supportées par chaque algorithme.

| | Algorithme A | Algorithme B | Algorithme C |
|---|---|---|---|
| Casse typographique | NON | OUI | OUI |
| Dérivation | NON | OUI | OUI |
| Genre et nombre | NON | OUI | OUI |
| Ordre des mots | NON | NON | OUI |
| Conjugaison | NON | NON | NON |

TAB. 2 – *Variations de la langue supportées par chaque algorithme.*

Considérons maintenant la phrase :
*« La **célèbre équation** d'**Einstein** $E = mc^2$ **exprime** le phénomène suivant : une **importante quantité** d'**énergie** est apparue et un **peu** de la **masse** a **disparu**. »*
ainsi que sa reformulation :
*« Ce **peu** de **masse disparue** crée une **grande quantité** d'**énergie** comme le **démontre** la **fameuse formule** d'Albert **Einstein E=mc2**. »*
Si on opère la comparaison de type C sur ces deux phrases, on retrouve bien, malgré le changement de vocabulaire et d'ordre des mots, la correspondance de nos concepts, ici en gras.

A noter l'importance du *seuil de correspondance*, il y a dans cette exemple 11 concepts identiques, avec un *seuil de correspondance* de 11 ou moins, la phrase est reconnue comme reformulation, alors qu'avec un *seuil de correspondance* supérieur ce ne sera plus le cas.

# 4 Évaluation et tests

## 4.1 La base de tests et protocole

La base de tests est composée de 150 textes, représentant chacun un passage d'un document, allant de plus de 100 mots pour le plus petit à environ 9000 mots pour le plus grand. Cela représente 400 comparaisons de textes deux à deux. Afin de tester correctement les performances des algorithmes évalués, aussi bien des paraphrases que des reformulations plus complexes sont présentes dans le corpus, ainsi que des textes « pièges » traitant du même sujet et donc employant le même vocabulaire mais n'étant pas pour autant des reformulations d'un autre texte du corpus.

Ci-dessous la répartition des types de textes présents dans le corpus :
— 10 différents chapitres tirés d'un même roman ;
— 20 chapitres de la bible (deux traductions différentes pour 10 chapitres) ;
— 25 textes de Wikipédia (différentes versions à différentes dates de 10 articles) ;
— 35 extraits de travaux d'élèves (avec leurs sources provenant du Web) ;

— 20 textes reformulés générés sur le web ;

Les extraits de travaux d'élèves proviennent pour la plupart de rapports et mémoires scientifiques ou économiques. L'intégralité des textes sont en français.

## 4.2 Résultats

Dans un premier temps, on compare les trois méthodes décrites précédemment, leur efficacité et leur temps moyen d'exécution respectifs étant différents selon le seuil utilisé, on détermine d'abord le seuil optimal pour chacune d'entre elles. Le tableau 3 représente le rapport précision/rappel des trois algorithmes allant du seuil 1 à 7. Un seuillage de 4 semble mieux convenir aux algorithmes A et B, tandis qu'un seuillage de 5 semble idéal pour l'algorithme C. Prenant en compte la F-mesure et privilégiant le rappel plutôt que la précision, l'algorithme C se montre être le plus efficace sur la base de tests.

| | A | | B | | C | |
|---|---|---|---|---|---|---|
| Seuil | P | R | P | R | P | R |
| 1 | 0.407 | 0.648 | 0.425 | 0.867 | 0.114 | 0.852 |
| 2 | 0.502 | 0.608 | 0.551 | 0.852 | 0.226 | 0.851 |
| 3 | 0.511 | 0.597 | 0.615 | 0.827 | 0.425 | 0.843 |
| 4 | **0.586** | **0.524** | **0.708** | **0.788** | 0.615 | 0.818 |
| 5 | 0.650 | 0.447 | 0.722 | 0.761 | **0.745** | **0.807** |
| 6 | 0.692 | 0.311 | 0.784 | 0.654 | 0.811 | 0.720 |
| 7 | 0.706 | 0.218 | 0.833 | 0.601 | 0.847 | 0.589 |

TAB. 3 – *Détermination du seuillage optimal pour chaque méthode.*

Le tableau 4 représente le temps d'exécution de la procédure (segmentation, extraction de mots porteurs de sens, chargement du thésaurus et comparaison) des trois algorithmes en utilisant leur meilleur seuillage en fonction du nombre moyen de mots contenus dans les deux textes à comparer (moyenne du nombre de mots des deux textes). La méthode C s'avère être la plus rapide (200 secondes en moyenne pour un texte d'environ 4000 mots contre 250 pour la méthode A et 212 secondes pour la méthode B) en plus d'avoir un meilleur rapport précision/rappel, respectivement 0.745 et 0.807, car malgré le fait qu'elle soit utilisée avec un seuil plus grand (5 contre 4 pour les deux autres implémentations) et qu'elle fasse donc forcément un plus long parcours, elle ne vérifie pas l'ordre des mots et néglige donc des permutations et suppressions de tableau. On remarque néanmoins une précision générale assez basse due aux faux positifs générés par les propriétés paradigmatiques des mots contenus dans le thésaurus.

Le tableau 5 compare la méthode retenue (l'algorithme C avec un seuil de 5) avec une méthode d'alignement basée sur la méthode de Bannard et Callison-Burch (2005) mais appliquée sur un corpus mono-lingue. Ces deux approches possèdent des performances similaires face à la détection de « copier/coller », environ 84% d'efficacité, en revanche notre méthode montre de biens meilleurs résultats sur la détection des reformulations non paraphrastiques (un rappel de 0.80 contre 0.24 pour une méthode avec alignement).

| Nombre de mots | Algorithmes | | |
|---|---|---|---|
| | A | B | C |
| 280 | 27 | 12 | 13 |
| 375 | 40 | 25 | 26 |
| 555 | 51 | 42 | 40 |
| 615 | 64 | 37 | 33 |
| 745 | 65 | 31 | 32 |
| 1055 | 74 | 44 | 41 |
| 1260 | 81 | 48 | 53 |
| 1425 | 91 | 61 | 81 |
| 3475 | 200 | 159 | 150 |
| 4075 | 250 | 212 | 200 |
| 4525 | 340 | 269 | 247 |
| 7150 | 820 | 724 | 687 |

TAB. 4 – *Temps d'exécution en secondes des trois algorithmes, en utilisant leur meilleur seuil respectif, en fonction du nombre moyen de mots à comparer.*

| | Alignement | | Nous | |
|---|---|---|---|---|
| Documents/Type de similitudes | P | R | P | R |
| Copie | 0.94 | 0.85 | 0.82 | 0.83 |
| Paraphrase | 0.65 | 0.77 | 0.81 | 0.83 |
| Reformulation | 1 | 0.24 | 0.62 | 0.80 |

TAB. 5 – *Evaluation de notre méthode par rapport à une méthode à alignement en fonction des types de similitudes à détecter.*

# 5 Conclusions

La méthode retenue montre des résultats similaires aux méthodes avec alignement sur la détection de copies exactes et de paraphrases et se montre beaucoup plus robuste face aux reformulations. Néanmoins, sa précision est fortement impactée par le thesaurus utilisé, qui peut engendrer des faux positifs pour les raisons évoquées dans la partie *3.2 Extraction de mots clefs*, et la segmentation, qui peut être faussée par du texte enrichi (tableau, liste à puces). Nous conviendrons également que cette technique est plutôt coûteuse en temps et en ressources (chargement du thesaurus en mémoire).

Un seuil adaptatif évoluant en fonction de la taille des phrases pourra également être mis en place dans de futurs travaux. Pour des phrases standards comportant entre 8 et 15 mots, il sera préférable de fixer le seuil à 5, en revanche si la phrase excède la vingtaine de mots, il faudra définir le seuil entre 10 et 12 mots communs.

Au final, cette approche reste naïve et gourmande aussi bien en temps qu'en ressources, néanmoins elle permet de détecter des reformulations jusque là impossibles à détecter avec des méthodes conventionnelles à alignement et constitue donc une alternative intéressante. Elle est à privilégier pour la détection de reformulations non paraphrastiques.

# Références


Bannard, C. et C. Callison-Burch (2005). Paraphrasing with bilingual parallel corpora. In *ACL*, Volume 43, pp. 597–604.

Barron-Cedeño, A., M. Vila, M. A. Martí, et P. Rosso (2013). Plagiarism meets paraphrasing : Insights for the next generation in automatic plagiarism detection. In *Association for Computational Linguistics*, Volume 39, pp. 917–947.

Callison-Burch, C., T. Cohn, et M. Lapata (2008). Parametric : An automatic evaluation metric for paraphrasing. In *COLING*, pp. 97–104.

Danlos, L. (2006). Discourse verbs and discourse periphrastic links. In *Proceedings on the Second International Workshop on Constraints in Discourse*.

de Biasi, P.-M. (2000). *La Génétique des textes*. CNRS éditions 2011.

Duclaye, F. (2003). Apprentissage automatique de relations d'équivalence sémantique à partir du web. In *Thèse de doctorat, Télécom Paris-Tech*.

Eshkol-Taravella, I. et N. Grabar (2014). Repérage et analyse de la reformulation paraphrastique dans les corpus oraux. In *Actes de la 21e conférence sur le Traitement Automatique des Langues Naturelles*, pp. 304–315.

Fenoglio, I., J.-L. Lebrave, et J.-G. Ganascia (2007). EDITE MEDITE : un logiciel de comparaison de versions. In *Item*. [en ligne].

Gibney, E. (2006). I'm No Plagiarist, I Moved a Comma. *The Times Higher Education Supplement : THE*. No. 2104.

Gülich, E. et T. Kotschi (1983). Les marqueurs de la reformulation paraphrastique. In *Cahiers de linguistique française*, Volume 5, pp. 305–351.

Grésillon, A. (1994). *Éléments de critique génétique. Lire les manuscrits modernes* (1 mai 1994 - 1er ed.). Presses Universitaires de France - PUF.

Harris, Z. S. (1957). Co-occurrence and transformation in linguistic structure. In *Language*, Volume 33, pp. 283–340.

Iyer, P. et A. Singh (2005). Document similarity analysis for a plagiarism detection system. In *2nd Indian International Conference and Artificial Intelligence*.

Jardino, M., M. Hurault-Plantet, et G. Illouz (2007). Identification de thème et reconnaissance du style d'un auteur pour une tâche de filtrage de textes. In *DEFT'05*, Volume RNTI-E-10, pp. 107–130.

Kahane, S. (2003). The meaning-text theory. In *Dependency and Valency (Ed.), An International Handbook of Contemporary Research*, Volume 1, pp. 546–570.

Lee, J.-W. et D.-K. Baik (2004). A model for extracting keywords of document using term frequency and distribution. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing*, Volume 2945 of *Lecture Notes in Computer Science*, pp. 437–440. Springer Berlin / Heidelberg.

Martin, R. (1976). Inférence, antonymie et paraphrase. librairie c. In *Paris : Klincksiec*.

Mel'čuk, I. (1992). Paraphrase et lexique : la théorie sens-texte et le dictionnaire explicatif et combinatoire. In *Mel'čuk,I. et al. (1984, 1988, 1992, 2000), Dictionnaire explicatif et com-*

*binatoire du français contemporain. Recherches lexico-sémantiques I-IV. Montréal : Presses de l'Université de Montréal*, pp. 9–59.

Milićević, J. (2007). La paraphrase modelisation de la paraphrase langagière. In *Peter Lang Bern 2007*.

Milićević, J. (2010). Extraction de paraphrases sémantiques et lexico-syntaxiques de corpus parallèles bilingues. In *Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles*.

Oberreuter, G. et J. D. Velásquez (2013). Text mining applied to plagiarism detection : The use of words for detecting deviations in the writing style. In *Expert Systems with Applications*, Volume 40, pp. 3756–3763.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

Shen, S., D. R. Radev, A. Patel, et G. Erkan (2006). Adding syntax to dynamic programming for aligning comparable texts for the generation of paraphrases. In *ACL*, Volume 44, pp. 747–754.

Simac-Lejeune, A. (2013a). Calcul de distance inter-documents par approche mots-clés. In *EGC*.

Simac-Lejeune, A. (2013b). Recherche de documents similaires sur le web par segmentation hiérarchiques et extraction de mots-clés. In *EGC*, pp. 401–406.

van Halteren, H. (2004). Linguistic profiling for author recognition and verification. In *ACL*, Volume 42, pp. 199–206.

Vila, M., M. A. Martí, et H. Rodríguez (2011). Paraphrase concept and typology. a linguistically based and computationally oriented approach. In *Procesamiento del Lenguaje Natural*, Volume 46, pp. 83–90.

Žolkovskij, A. et I. Mel'čuk (1967). O semantičeskom sinteze. In *Problemy kibernetiki*, Volume 19, pp. 177–238. [Traduction française : Sur la synthèse sémantique (1970). TA Informations 2 : 1-85].

## Summary

Comparison of two documents in the plagiarism detection context is often reduced to a word to word comparison, a research of copy and paste. In this article, a naïve approach to compare two documents with the aim of automatically detecting whether a copied sentence from one text to the other, or paraphrases and reformulations, is presented. This is achieved by looking for the existence of meaningful words and their potential substitution words. We compare three algorithms using this approach and retain only the most efficient one to evaluate it with existing methods. The goal is to enable detection of similarities between two texts using only keywords. The proposed approach can detect non paraphrastic reformulations, which are impossible to detect with the conventional alignment approach.