# Construction d'un corpus parallèle à partir de corpus comparables pour la simplification de textes médicaux en français

**Rémi Cardon*** — **Natalia Grabar***

* *UMR 8163 STL – CNRS / Université de Lille, F-59000 LILLE*
*{remi.cardon, natalia.grabar}@univ-lille.fr*

RÉSUMÉ. *La simplification automatique a pour objectif de produire une version de textes plus facile à comprendre à destination d'un public identifié. Nous nous intéressons à la simplification de textes médicaux. Le plus souvent, le lexique et les règles de simplification sont acquis à partir de corpus parallèles. Comme de tels corpus n'existent pas en français, nous proposons des méthodes pour les construire à partir de corpus comparables. Notre méthode repose sur une étape de filtrage, destinée à ne garder que les meilleures phrases candidates à l'alignement, et une étape d'alignement considérée comme un problème de catégorisation. Il s'agit de décider si une paire de phrases est alignable ou non. Nous exploitons différents types de descripteurs (essentiellement basés sur le lexique et les corpus) et obtenons jusqu'à 0,97 de F-mesure avec les données équilibrées.*

ABSTRACT. *The purpose of automatic simplification is to create version of texts which is easier to understand for a given targeted population. We aim at simplifying medical texts. Usually, lexicon and rules required for the simplification are acquired from parallel corpora. Since such corpora are not available for French, we propose methods for their creation from comparable corpora. Our method relies on filtering step, which purpose is to keep the best sentence candidates for alignment, and alignment step considered as categorization problem. The aim is to decide whether a pair of sentences is alignable or not. We exploit different types of features (mainly issued from lexicon and corpora) and get up to 0.97 F-measure with balanced data.*

MOTS-CLÉS : *simplification automatique, textes médicaux, corpus de phrases parallèles, constitution de ressources.*

KEYWORDS: *automatic simplification, medical texts, corpus with parallel sentences, resource building.*

## 1. Introduction

La simplification automatique vise à fournir une version simplifiée des textes à destination d'une population donnée. La simplification peut concerner le lexique, la syntaxe, la sémantique mais aussi la pragmatique et l'organisation des textes. La simplification peut être vue comme une aide fournie aux lecteurs ou comme un prétraitement dans les applications de TAL. Dans le cas d'aide aux lecteurs, la simplification vise différents types d'utilisateurs : les enfants (De Belder et Moens, 2010), les personnes non ou mal alphabétisées, les lecteurs étrangers (Paetzold et Specia, 2016), les personnes handicapées ou ayant des pathologies neurodégénératives (Chen *et al.*, 2016), ou les personnes non spécialistes face à des documents spécialisés (Leroy *et al.*, 2013). Dans le cas d'applications de TAL, la simplification produit une version de textes plus facile à traiter par d'autres modules de TAL, comme l'analyse syntaxique (Chandrasekar et Srinivas, 1997 ; Jonnalagadda *et al.*, 2009), l'annotation sémantique (Vickrey et Koller, 2008), le résumé automatique (Blake *et al.*, 2007), la traduction automatique (Stymne *et al.*, 2013 ; Štajner et Popović, 2016), l'indexation (Wei *et al.*, 2014) et la recherche et extraction d'information (Beigman Klebanov *et al.*, 2004).

Au moins deux types de connaissances sont nécessaires pour effectuer la simplification : un lexique pour la simplification au niveau lexical, et un ensemble de règles de transformation pour la simplification syntaxique. De telles ressources peuvent être construites manuellement, provenant de l'expertise d'un spécialiste, ce qui était propre aux premiers travaux de simplification, ou bien être acquises à partir de données réelles, ce qui correspond aux approches actuelles et nécessite de gros corpus parallèles (Nisioi *et al.*, 2017). Deux corpus avec des données en anglais sont fréquemment utilisés : *WikiLarge* (Zhang et Lapata, 2017), un corpus libre d'utilisation qui contient environ 300 000 paires de phrases, et *Newsela* (Xu *et al.*, 2015 ; Hwang *et al.*, 2015). Ce type de corpus n'existe pas en français, alors qu'il pourrait fournir des informations précieuses pour la simplification automatique. Ainsi, les paires de phrases, qui se différencient par leur degré de technicité, peuvent fournir des informations utiles, comme dans les exemples (1), avec le texte technique, et (2), avec le texte simplifié.

(1) *L'hématome aigu de l'oreille est une affection qui se caractérise par la formation d'une collection sanguine sous le périchondre du pavillon. Il est souvent provoqué par un traumatisme contondant. En l'absence de traitement, il finit par entraîner une difformité couramment appelée oreille en chou-fleur ou oreille du boxeur.*

(2) *L'hématome aigu de l'oreille est une affection qui se caractérise par la formation d'une collection sanguine dans le pavillon (oreille externe), souvent à la suite d'un traumatisme contondant. S'il n'est pas traité, il entraîne une difformité appelée oreille en chou-fleur ou oreille du boxeur.*

L'objectif principal de notre travail consiste à détecter automatiquement des phrases parallèles, avec le contenu similaire ou identique, au sein de documents monolingues

comparables en français et distingués par leur technicité. À notre connaissance, le seul travail de ce type en français a été effectué avec un alignement de phrases manuel (Brouwers *et al.*, 2014) mais les phrases alignées et les règles de transformation syntaxiques ne sont pas disponibles. Par rapport à nos travaux précédents (Cardon et Grabar, 2019 ; Cardon et Grabar, 2020), nous proposons une méthode plus complète (avec une étape de filtrage et une étape d'alignement), exploitons un ensemble plus riche de descripteurs (basés sur le lexique, les indices formels, la similarité et les corpus), effectuons plus d'expériences en alignement de phrases parallèles, et présentons globalement de meilleurs résultats.

Dans la suite de ce travail, nous présentons d'abord les travaux existants (section 2). Nous présentons ensuite notre approche (sections 3 et 4) et les résultats obtenus (section 5). Nous terminons avec une conclusion et des perspectives (section 6).

## 2. Travaux existants

Dans les corpus parallèles, l'alignement de phrases parallèles peut se baser sur des indices de surface comme la longueur relative des phrases (Gale et Church, 1993) ou les informations lexicales (Chen, 1993), tandis que dans les corpus comparables, les phrases ont une sémantique relativement proche et, de plus, elles ne sont pas forcément ordonnées de la même manière. D'autres difficultés proviennent du fait que le degré du parallélisme peut varier en allant des corpus presque parallèles, avec beaucoup de phrases parallèles, aux corpus assez éloignés (*very-non-parallel corpora*) (Fung et Cheung, 2004) et que de tels corpus peuvent contenir des données parallèles à différents niveaux de granularité : documents, phrases, segments sous-phrastiques (Hewavitharana et Vogel, 2011). Plusieurs travaux sont positionnés en traduction automatique : les corpus comparables bilingues sont exploités pour créer des corpus parallèles et alignés. Ces travaux requièrent l'utilisation de lexiques bilingues ou de systèmes de traduction automatique et reposent en général sur trois étapes :

1) détection de documents comparables au sein d'un corpus grâce aux métriques de similarité par exemple (Utiyama et Isahara, 2003 ; Fung et Cheung, 2004), ce qui permet de réduire l'espace de recherche de phrases parallèles ;

2) détection de phrases ou de segments candidats à l'alignement en exploitant des systèmes de recherche d'information cross-langue (Utiyama et Isahara, 2003), des arbres d'alignement de séquences (Munteanu et Marcu, 2002) ou des traductions automatiques mutuelles (Yang et Li, 2003 ; Abdul-Rauf et Schwenk, 2009) ;

3) sélection de bonnes propositions en exploitant des classifieurs binaires (Ștefănescu *et al.*, 2012), des mesures de similarité (Fung et Cheung, 2004), le taux d'erreurs (Abdul-Rauf et Schwenk, 2009), des modèles génératifs (Zhao et Vogel, 2002) ou des règles spécifiques (Yang et Li, 2003).

Plus récemment, cette tâche est également explorée dans le contexte monolingue : la similarité sémantique textuelle (*semantic text similarity - STS*) est calculée au niveau de phrases ou de segments sous-phrastiques. Cette tâche a attiré l'attention des cher-

cheurs car ce type d'information fournit des indications précieuses pour la détection du plagiat, les questions-réponses ou la pondération des réponses, par exemple. Ainsi, la compétition *SemEval* propose une tâche dédiée à la similarité sémantique textuelle (Agirre *et al.*, 2013) et poursuit l'objectif suivant : étant donné une paire de phrases, les systèmes doivent prédire si ces phrases sont similaires sémantiquement et leur attribuer un score de similarité allant de 0 (sémantique indépendante) à 5 (sémantique identique). Plusieurs types de méthodes sont exploités par les participants :

– *les méthodes basées sur le lexique*, qui exploitent les chaînes de caractères et de mots ou la traduction automatique (Clough *et al.*, 2002 ; Zhang et Patrick, 2005 ; Nelken et Shieber, 2006 ; Qiu *et al.*, 2006 ; Zhu *et al.*, 2010 ; Zhao *et al.*, 2014) ;

– *les méthodes basées sur les connaissances*, qui exploitent des sources lexicales externes, comme WordNet ou la ressource PPDB avec les paraphrases (Mihalcea *et al.*, 2006 ; Fernando et Stevenson, 2008 ; Lai et Hockenmaier, 2014) ;

– *les méthodes basées sur la syntaxe*, qui exploitent la modélisation syntaxique des phrases (Wan *et al.*, 2006 ; Severyn *et al.*, 2013 ; Tai *et al.*, 2015 ; Tsubaki *et al.*, 2016) ;

– *les méthodes basées sur les corpus*, qui exploitent les modèles distributionnels, LSA, etc. (Barzilay et Elhadad, 2003 ; Guo et Diab, 2012 ; Zhao *et al.*, 2014 ; He *et al.*, 2015 ; Mueller et Thyagarajan, 2016).

Nous nous intéressons à cette tâche car elle permet de construire des ressources (un lexique, des règles de transformation, etc.) utilisables en simplification automatique.

## 3. Données linguistiques

Nous exploitons un corpus monolingue comparable disponible (section 3.1), les données de référence issues de l'alignement manuel au niveau des phrases (section 3.2) et une liste de mots vides (décrite dans la section 4.1).

### 3.1. *Corpus monolingue comparable*

Le corpus monolingue comparable[1] contient des textes provenant de trois sources de données : (1) les articles de deux encyclopédies collaboratives disponibles en ligne Wikipédia[2] et Vikidia[3], (2) les informations sur les médicaments de la base publique de médicaments [4] gérée par le ministère de la Santé, (3) les résumés de revues systématiques de la fondation Cochrane[5]. Trois genres sont donc couverts dans ce corpus : les articles d'encyclopédie, les informations sur les médicaments et la littérature scientifique. Dans chaque source et pour un sujet donné, les textes techniques et simplifiés

1. http://natalia.grabar.free.fr/resources.php
2. https://fr.wikipedia.org
3. https://fr.vikidia.org
4. http://base-donnees-publique.medicaments.gouv.fr/
5. http://www.cochranelibrary.com/

sont disponibles. Ce corpus contient plus de 55 M de mots dans la partie technique et plus de 35 M de mots dans la partie simplifiée.

### 3.2. *Données de référence*

Pour créer les données de référence, nous sélectionnons aléatoirement 14 articles encyclopédiques, 12 médicaments et 13 revues Cochrane. Les documents sont segmentés en phrases. L'alignement est effectué manuellement et indépendamment par deux annotateurs (les auteurs) au niveau de la paire de documents. L'alignement des 39 paires de documents a pris environ 20 heures par annotateur, à quoi s'ajoutent environ 5 heures de consensus où toutes les annotations ont été passées en revue. L'annotation a été menée sans guide. L'accord inter-annotateur (Cohen, 1960) est de 0,76.

Un accord est compté lorsqu'un alignement est proposé par les deux annotateurs, et un désaccord lorsqu'un alignement est proposé par un seul annotateur. Les phrases non alignées ne sont pas considérées. Dans le tableau 1, nous indiquons la taille des données de référence avant (colonne « *brut* ») et après l'alignement (colonne « *aligné* »), ainsi que le taux d'alignement, c'est-à-dire le nombre de phrases d'un corpus qui trouvent un alignement par rapport au nombre total de phrases. Il s'agit des résultats d'alignement consensuel. Les séances de consensus ont été l'occasion d'identifier les types d'alignement conservés et de trouver un accord sur les données utiles pour la simplification. Nous avons observé que les désaccords venaient de la prise en compte ou non de certains types d'alignement, comme les phrases identiques ou l'intersection sémantique. Dans les exemples, la phrase technique est suivie par la phrase simplifiée :

1) les deux phrases, technique et simplifiée, doivent contenir un verbe.

2) les phrases ne sont pas identiques et diffèrent par le lexique ou la morphologie des mots, mais pas uniquement par la ponctuation ou les mots vides. Ces phrases ont une sémantique équivalente : *{Les sondes gastriques sont couramment utilisées pour administrer des médicaments ou une alimentation entérale aux personnes ne pouvant plus avaler}{Les sondes gastriques sont couramment utilisées pour administrer des médicaments et de la nourriture directement dans le tractus gastro-intestinal (un tube permettant de digérer les aliments) pour les personnes ne pouvant pas avaler}*

3) le sens d'une phrase est intégralement inclus dans le sens de l'autre phrase. Il s'agit de l'inclusion sémantique. Cela permet de repérer les cas de simplification syntaxique (fusion ou découpage de phrases) ainsi que les ajouts et suppressions. Dans cet exemple, la phrase technique indique le nombre de participants et la mesure d'évaluation en plus : *{Peu de données (43 participants) étaient disponibles concernant la détection d'un mauvais placement (la spécificité) en raison de la faible incidence des mauvais placements}{Cependant, peu de données étaient disponibles concernant les sondes placées incorrectement et les complications possibles d'une sonde mal placée}*

4) les cas d'intersection sémantique, où chaque phrase apporte des informations spécifiques propres, sont rejetés. L'intersection sémantique est en effet plus difficile à généraliser pour en dégager des règles de transformation : *{Des études à plus grande échelle sont nécessaires pour déterminer la possibilité d'événements indésirables lorsque les ultrasons sont utilisés pour confirmer le positionnement des sondes}{Des études à plus grande échelle sont nécessaires pour déterminer si les ultrasons pourraient remplacer les rayons x pour confirmer la mise en place d'une sonde gastrique, et pour évaluer si les ultrasons pourraient permettre de réduire les complications graves, telles que la pneumonie résultant d'un tube mal placé}*

| | | *Technique* | | | | *Simplifié* | | | | *Alignement* | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *Brut* | | *Aligné* | | *Brut* | | *Aligné* | | *(%)* | |
| *Corpus* | *Doc.* | *Ph.* | *Occ.* | *Ph.* | *Occ.* | *Ph.* | *Occ.* | *Ph.* | *Occ.* | *Tech.* | *Simp.* |
| *Médicaments* | 12 × 2 | 4 391 | 44 684 | 143 | 4 227 | 2 710 | 27 804 | 143 | 8 481 | 3,25 | 5,27 |
| *Cochrane* | 13 × 2 | 426 | 8 852 | 84 | 2 278 | 227 | 4 688 | 84 | 2 466 | 19,71 | 36,56 |
| *Encyclopédie* | 14 × 2 | 2 416 | 36 703 | 39 | 873 | 235 | 2 659 | 39 | 710 | 1,61 | 16,6 |

**Tableau 1.** *Taille des données de référence avec l'alignement consensuel*

Selon le tableau 1, nous pouvons voir que les phrases alignées sont relativement plus rares dans les corpus *Médicaments* et *Encylopédie*, alors que le corpus *Cochrane* en offre plus par rapport à sa taille : le taux d'alignement est entre 1,61 et 36,56. Ceci peut être expliqué par les spécificités des corpus : (1) la ligne directrice de rédaction des versions simplifiées des résumés *Cochrane* affiche explicitement une volonté de simplifier le contenu de ses résumés d'origine pour le grand public. Les rédacteurs prennent donc comme point de départ les résumés techniques et les simplifient ; (2) l'objectif de Vikidia est de traiter des sujets présents dans Wikipédia mais pour un public d'enfants. La création d'articles de Vikidia est rarement basée sur les articles de Wikipédia et, le plus souvent, il s'agit d'une écriture indépendante ; (3) quant au corpus *Médicaments*, en respect avec la législation, les informations sur les médicaments sont créées à destination des professionnels de santé et des patients. Certaines de ces informations sont propres à la version technique (composition plus détaillée, action sur l'organisme, molécules, détail sur les effets indésirables...), alors que d'autres sont propres à la version simplifiée (précautions d'emploi, mises en garde...).

Suite à l'alignement manuel, nous gardons deux types d'alignement :

1) *équivalence sémantique.* Les deux phrases, technique et simplifiée, ont le même sens ou des sens proches, comme dans ces phrases du corpus *Cochrane* : *{Les sondes gastriques sont couramment utilisées pour administrer des médicaments ou une alimentation entérale aux personnes ne pouvant plus avaler}{Les sondes gastriques sont couramment utilisées pour administrer des médicaments et de la nourriture directement dans le tractus gastro-intestinal (un tube permettant de digérer les aliments) pour les personnes ne pouvant pas avaler}*. Dans le cas d'équivalence sémantique, la simplification est essentiellement effectuée au niveau lexical. Elle repose alors sur la substitution de termes, comme c'est observable à travers les paires *{technique}{simplifié}* :

*{alimentation}{nourriture}*, *{entérale}{directement dans le tractus gastro-intestinal}*. La simplification peut également être effectuée grâce à l'ajout d'informations et, dans ce cas, les notions complexes sont suivies par leurs équivalents, souvent entre parenthèses, comme *le tractus gastro-intestinal (un tube permettant de digérer les aliments)*. Dans plusieurs cas, ces deux procédés (substitution et ajout d'informations) sont employés conjointement parce qu'ils apportent des informations différentes et complémentaires ;

2) *inclusion sémantique.* Le sens d'une phrase se trouve inclus dans le sens de l'autre phrase de la paire. L'inclusion est orientée : la phrase technique ou la phrase simplifiée peuvent être incluantes. Nous traitons les deux sens de l'inclusion comme un seul type d'alignement. Leur distinction n'a pas montré de différences significatives lors des expériences. Dans l'exemple qui suit, la phrase technique est incluante et indique en plus le nombre de participants et la mesure d'évaluation : *{Peu de données (43 participants) étaient disponibles concernant la détection d'un mauvais placement (la spécificité) en raison de la faible incidence des mauvais placements}{cependant, peu de données étaient disponibles concernant les sondes placées incorrectement et les complications possibles d'une sonde mal placée}*. Dans le cas d'inclusion, la simplification est effectuée également au niveau syntaxique. Typiquement, les subordonnées, les incises, les informations entre parenthèses, certains adjectifs ou adverbes peuvent être supprimés. Dans l'exemple cité, les informations entre parenthèses (*43 participants* et *la spécificité*) sont omises. Dans d'autres cas, les phrases complexes syntaxiquement sont segmentées. L'inclusion sémantique concerne également les énumérations. Ainsi, une phrase technique coordonnée peut être segmentée en une liste d'items séparés dans la version simplifiée. Notons aussi que la simplification syntaxique est souvent accompagnée par des transformations lexicales, comme *{incidence}{complications possibles}*, *{mauvais placement}{placé incorrectement}* ou *{mauvais placement}{mal placé}* dans le dernier exemple.

## 4. Méthodologies pour l'alignement de phrases parallèles

Dans notre corpus, les documents comparables sont déjà associés entre eux. En revanche, comme les textes techniques et simplifiés sont souvent rédigés de manière indépendante, l'ordre des phrases dans les documents n'est pas significatif. L'accent principal de la méthode est donc mis sur la recherche de phrases parallèles. Notre méthode se compose de plusieurs étapes : le prétraitement dont le filtrage de phrases, l'alignement de phrases et l'évaluation des alignements. Nous décrivons également les différentes expériences effectuées.

### 4.1. *Prétraitement*

Tous les documents sont étiquetés avec TreeTagger (Schmid, 1994), ce qui permet d'en obtenir leurs versions lemmatisées. Les documents sont ensuite segmentés en phrases en exploitant la ponctuation forte (*. ? ! ; :*). D'autres prétraitements sont dédiés

au filtrage des phrases pour ne retenir que les meilleurs candidats à l'alignement. Nous exploitons trois méthodes pour le filtrage basées sur la forme et la syntaxe :

1) méthode basée sur le nombre de mots dans les phrases : chaque phrase candidate doit contenir au moins cinq mots, ce qui correspond à la longueur de la phrase la plus courte dans les données de référence ;

2) suppression de paires avec les phrases identiques ;

3) exploitation d'informations syntaxiques : nous nous inspirons d'un travail existant qui mesure la similarité entre les phrases dans un corpus monolingue grâce aux constituants syntaxiques (Duran *et al.*, 2014). Le score de similarité est alors calculé sur la base des nœuds syntaxiques similaires qui contiennent des mots similaires. Il est difficile d'adapter cette méthode, essentiellement parce qu'elle se base sur une table de similarité entre les constituants, alors que cette table est créée pour l'anglais et que de plus les auteurs ne donnent pas d'indications sur les principes de sa création. Nous supposons cependant que l'adoption d'une approche similaire permettra d'éliminer les paires de phrases indésirables pour l'alignement. Ainsi, au lieu de calculer le score de similarité, nous effectuons un filtrage binaire : garder ou non une paire de phrases candidates. Pour une paire donnée, nous calculons l'arbre syntaxique de chacune des phrases. Ensuite, nous comparons les feuilles (c'est-à-dire les mots) des arbres, sauf celles qui contiennent les mots vides. La liste de mots vides contient 83 entrées (mots grammaticaux comme les déterminants ou prépositions). Lorsque nous trouvons deux mots identiques, nous vérifions leurs nœuds pères : s'ils sont identiques, nous gardons la phrase comme candidate à l'alignement. Le processus est illustré par l'algorithme 1. Nous exploitons également une variante de la méthode : au lieu de nous arrêter lorsque les nœuds pères ne sont pas identiques, nous continuons de remonter l'arbre jusqu'au troisième nœud tant que les nœuds précédents n'ont pas donné de résultats positifs. La comparaison s'arrête lorsque les nœuds sont identiques et la phrase est retenue pour l'alignement ou lorsque la profondeur est supérieure à 3. Cette approche est illustrée par l'algorithme 2. La considération de nœuds parents de profondeur 3 permet également d'observer comment la profondeur de l'arbre influence le filtrage. L'analyse syntaxique des phrases est obtenue avec le Berkeley Neural Parser et le modèle de langue pour le français de la librairie python `benepar` (Kitaev et Klein, 2018). Nous utilisons la librairie NLTK `Tree` pour la manipulation d'arbres syntaxiques (Bird *et al.*, 2009).

### 4.2. *Alignement de phrases*

Nous abordons la recherche de phrases parallèles comme une problématique de catégorisation : pour une paire de phrases présélectionnées lors de l'étape précédente, il faut décider s'il faut les mettre dans la catégorie *aligné* ou non.

Nous utilisons plusieurs classifieurs linéaires de `scikit-learn` (Pedregosa *et al.*, 2011) avec leurs paramètres par défaut, s'il n'est pas indiqué autrement : `Perceptron` (Rosenblatt, 1958), `Perceptron multicouche (MLP)` (Rosenblatt, 1961), `Random`

**Data:** Deux arbres syntaxiques ($T_1$ et $T_2$), une liste de *mots vides* (*SW*)
**Result:** Booléen
Booléen ← False;
**if** *au moins un verbe est dans chaque arbre* **then**
  **foreach** *feuille de $T_1$ ($L_1$) absente de SW* **do**
    **foreach** *feuille de $T_2$ ($L_2$) absente de SW* **do**
      **if** *$L_1$ est identique à $L_2$* **then**
        **if** *l'étiquette du père de $L_1$ est identique à l'étiquette du père de $L_2$* **then**
          Booléen ← True ;
        **else**
          rien ;
        **end**
      **else**
        rien ;
      **end**
    **end**
  **end**
**else**
  rien ;
**end**
**return** Booléen ;
**Algorithm 1:** Filtrage par la comparaison des pères immédiats des feuilles

`Forest (RF)` (Ho, 1995) `Linear discriminant analysis (LDA)` (Fisher, 1936) avec le solveur `LSQR`, `Quadratic discriminant analysis (QDA)` (Cover, 1965), `Logistic regression` (Berkson, 1944), modèle log-linéaire appris avec `Stochastic gradient descent (SGD)` (Ferguson, 1982), et `SVM linéaire` (Vapnik et Lerner, 1963).

Pour avoir une méthode assez générique et pouvoir l'évaluer sur d'autres jeux de données, nous utilisons des descripteurs qui seraient facilement calculables. Nous exploitons cinq types de descripteurs. Par rapport à la typologie présentée dans la section 2, ces descripteurs sont essentiellement liés au lexique et au corpus. Les descripteurs sont calculés sur les formes et les lemmes :

1) *BL : descripteurs de base (baseline)* :

– *nombre de mots communs, hors mots grammaticaux*, ce qui permet de calculer l'intersection lexicale de base entre les phrases (Barzilay et Elhadad, 2003) ;

– *ratio longueur de la phrase la plus courte sur la longueur de la phrase la plus longue*. Ce descripteur suppose que la simplification peut impliquer une association stable avec la longueur des phrases ;

– *différence de la longueur moyenne des mots entre les deux phrases* pour estimer l'utilisation de mots longs, jugés spécifiques au langage technique ;

2) *L : descripteurs issus de la distance de chaînes d'édition* (Levenshtein, 1966) :

```
Data: Deux arbres syntaxiques (T1 et T2), une liste de mots vides (SW)
Result: Booléen
Booléen ← False;
if au moins un verbe est dans chaque arbre then
    foreach feuille de T1 (L1) absente de SW do
        foreach feuille de T2 (L2) absente de SW do
            if L1 est identique à L2 then
                if l'étiquette du père de L1 (P1) est identique à l'étiquette du
                  père de L2 (P2) then
                    Booléen ← True ;
                else
                    if l'étiquette du père de P1 (PP1) est identique à l'étiquette
                      du père de P2 (PP2) then
                        Booléen ← True;
                    else
                        if l'étiquette du père de PP1 est identique à l'étiquette
                          du père de PP2 then
                            Booléen ← True;
                        else
                            rien;
                        end
                    end
                end
            else
                rien ;
            end
        end
    end
else
    rien ;
end
return Booléen ;
```

**Algorithm 2:** Filtrage par la comparaison d'ancêtres des feuilles (profondeur 3)

– *distance d'édition calculée au niveau des caractères*. Il s'agit de l'acception classique de la mesure. Elle prend en compte les opérations d'édition de base (insertion, suppression et substitution). Le coût de chaque opération est de 1 ;

– *distance d'édition calculée au niveau des mots*. Ce descripteur est calculé avec des mots comme unité. Il prend en compte les mêmes opérations d'édition avec le coût de 1. Le descripteur permet de calculer le coût de la transformation lexicale ;

3) *S : descripteurs basés sur les similarités lexicales* avec la *similarité au niveau des mots calculée selon trois scores (cosinus, Dice et Jaccard)*. Ce descripteur fournit une indication plus sophistiquée sur l'intersection lexicale entre les deux phrases. Le

poids de chaque mot est de 1 ;

4) *N : descripteurs basés sur les n-grammes (bigrammes et trigrammes) de caractères en commun*, ce qui permet de prendre en compte la présence de séquences de caractères communs ;

5) *PL : descripteurs basés sur les plongements lexicaux.* Deux descripteurs sont utilisés : *WAVG* (Stajner *et al.*, 2018), où la moyenne des vecteurs de mots de chacune des deux phrases est calculée et ces vecteurs sont comparés pour attribuer un score de similarité ; et *CWASA* (Franco-Salvador *et al.*, 2016) pour (*continuous word alignment-based similarity analysis*). Ces descripteurs sont exploités avec des plongements entraînés sur le corpus CLEAR à l'aide de Word2Vec[6] (Mikolov *et al.*, 2013), alors que les scores sont calculés avec l'outil CATS (Stajner *et al.*, 2018). Nous avons mené les mêmes expériences avec des vecteurs préentraînés avec Fast Text[7] (Grave *et al.*, 2018) et n'avons pas noté de différences significatives.

### 4.3. *Évaluation*

L'évaluation est effectuée par rapport aux données de référence. L'entraînement du système est effectué sur 70 % de paires de phrases et le test est effectué sur le reste des données. Plusieurs classifieurs et plusieurs combinaisons de descripteurs sont testés. Les mesures d'évaluation classiques sont calculées : précision, rappel, F-mesure, EQM (erreur quadratique moyenne) et vrais positifs. Avec les données déséquilibrées, l'évaluation est effectuée sur 50 tirages différents afin de mieux évaluer les performances de l'alignement des phrases. Nous rapportons uniquement les scores pour la catégorie de phrases alignées car, d'une part, c'est le principal résultat visé et, d'autre part, avec les données déséquilibrées et une très grande quantité de phrases non alignables, les résultats globaux sont toujours très élevés.

### 4.4. *Expériences*

Les données de référence fournissent 266 paires de phrases parallèles comme exemples positifs et nous choisissons aléatoirement des exemples négatifs à partir des mêmes documents : 266 paires de phrases non parallèles pour les expériences avec des données équilibrées et d'autres paires de phrases pour des expériences avec des données non équilibrées. Les exemples négatifs sont obtenus en appariant aléatoirement des phrases d'un document technique et de son pendant simple, en vérifiant que ces paires ne font pas partie de la classe positive (équivalence ou inclusion). Il n'y a pas d'intersection entre les phrases alignées et non alignées. Plusieurs expériences sont effectuées, où nous étudions les effets des descripteurs et du déséquilibre.

6. Hyperparamètres de Word2Vec : `-size 300 -window 7 -sample 1e-5 -hs 1 -negative 50 -mincount 20 -alpha 0.025 -cbow 0`
7. `https://fasttext.cc/docs/en/crawl-vectors.html`

#### 4.4.1. *Baseline*

Notre *baseline* correspond à la combinaison de descripteurs lexicaux traditionnellement utilisés pour l'alignement de phrases : la longueur des phrases et l'intersection lexicale entre les phrases. Cette expérience est effectuée avec les données équilibrées.

#### 4.4.2. *Détection de phrases parallèles avec une distribution équilibrée*

Le nombre d'exemples positifs et négatifs est comparable, ce qui correspond à une distribution équilibrée des paires de phrases entre les deux catégories. Cette expérience permet de tester différents descripteurs et leurs combinaisons.

#### 4.4.3. *Détection de phrases parallèles selon la sémantique des paires*

Les paires de phrases de référence sont divisées en deux sous-ensembles, selon le lien sémantique qui existe au sein de la paire :

– *E* : 130 paires avec équivalence sémantique ;

– *I* : 136 paires où le contenu de la phrase technique est compris dans la phrase simplifiée ou l'inverse. Ceci représente les cas de découpage ou de fusion de phrases, ainsi que la suppression ou l'ajout d'informations, lors de la simplification.

#### 4.4.4. *Détection de phrases parallèles avec une distribution déséquilibrée*

Comme le montre le tableau 1, les phrases parallèles sont plutôt rares et il existe beaucoup plus de phrases non alignables. La distribution de phrases parallèles n'est donc pas élevée ni constante : le taux d'alignement varie selon les corpus, les paires de documents et le sens d'alignement. Ainsi, l'objectif de cette expérience est de voir quelles sont les performances du système lorsque les données traitées s'approchent de la distribution naturelle de phrases alignables. Pour chaque sous-ensemble (*E, I*), nous prenons d'abord autant de paires équilibrées que d'exemples négatifs sélectionnés aléatoirement. Ensuite, nous augmentons progressivement le nombre de paires non alignables jusqu'à un ratio de $200 : 1$, proche de celui des données réelles après le filtrage. Ceci correspond à l'ensemble déséquilibré $D$ avec les 136 (*E*) ou 130 *I* paires alignées et le ratio croissant de paires non alignées. Le ratio et les données changent donc à chaque itération. Nous utilisons aussi l'ensemble réel $R$, qui comporte toute la combinatoire possible de paires de phrases après filtrage (21 428), alignées et non alignées. L'ensemble $R$ est toujours le même. Nous procédons ainsi en raison du faible nombre d'exemples positifs. Il est donc à noter que le score de rappel en sera artificiellement augmenté. Cela dit, le score de précision évalue la robustesse du modèle à ne pas produire de faux positifs, ce qui nous semble important en raison du grand déséquilibre en faveur d'exemples négatifs. À chaque point de déséquilibre de l'ensemble $D$, nous faisons deux séries d'expériences :

1) *DD* : entraînement et test au sein de l'ensemble déséquilibré $D$ ;

2) *DR* : entraînement sur l'ensemble $D$ et test sur les données réelles $R$ (environ 21 428 paires de phrases après filtrage).

## 5. Résultats

Nous présentons les résultats de différentes expériences : (1) l'effet du filtrage sur le corpus, (2) la méthode de *baseline* pour l'alignement de phrases avec l'utilisation de descripteurs basiques, (3) la détection de phrases parallèles avec une distribution équilibrée de paires de phrases parallèles et non parallèles, (4) la détection de phrases parallèles selon la sémantique des paires en distinguant l'équivalence sémantique et l'inclusion, (5) la détection de phrases parallèles avec une distribution déséquilibrée s'approchant de la distribution réelle moyenne de phrases alignables. Nous faisons également une analyse des erreurs et présentons quelques limitations actuelles.

### 5.1. *Filtrage*

| Paires restantes | *Originales* | *IF* | *Syntaxe 1* | *Syntaxe 3* |
|---|---|---|---|---|
| Total | 1 164 407 | 409 530 | 16 879 | 21 428 |
| Équivalence | 136 | 136 | 94 | 94 |
| Inclusion | 130 | 130 | 94 | 100 |

**Tableau 2.** *Effet du filtrage sur le corpus*

La première colonne du tableau 2 indique le nombre de paires de phrases originales, la seconde le nombre de paires qui restent après l'utilisation des indices formels liés à la présence du verbe et l'élimination des paires avec des phrases identiques (IF), et les deux dernières indiquent le nombre de paires qui restent après l'utilisation du filtre syntaxique, en remontant respectivement au premier et au troisième père. Les indices formels sont appliqués avant les filtres syntaxiques. Les filtres syntaxiques sont appliqués indépendamment l'un de l'autre.

Nous observons que les indices formels réduisent le nombre total de paires de 65 % : on passe de 1 164 407 à 409 530. Nous voyons qu'avec ces indices nous ne perdons aucun exemple positif. À partir des 409 530 paires obtenues après le premier filtre, nous observons une autre grande réduction du volume de paires avec chacun des filtres syntaxiques. Le filtre de profondeur 1 laisse 16 879 paires ($\sim$ 96 % de réduction) et celui de profondeur 3 laisse 21 428 paires ($\sim$ 95 % de réduction). Le défaut de ce type de filtre est qu'un nombre non négligeable d'exemples positifs est perdu : 42 sur 136 ($\sim$ 30 %) pour les couples équivalents avec les deux filtres syntaxiques, 36 sur 130 ($\sim$ 27 %) pour la profondeur 1, et 32 sur 130 ($\sim$ 24 %) pour la profondeur 3 pour les couples avec l'inclusion. Nous présentons deux exemples, où le premier est conservé alors que le deuxième est rejeté à tort après filtrage IF et syntaxe 3 :

– *{L'apparition de signes cliniques tels qu'un mal de gorge, une fièvre, une pâleur, un purpura ou un ictère pendant le traitement par la sulfasalazine peut faire suspecter une myélosuppression, une hémolyse ou une hépatotoxicité.}{L'apparition de signes cliniques tels qu'un mal de gorge, une fièvre, une pâleur, de petites taches rouges sur*

*la peau ou une jaunisse pendant le traitement par la sulfasalazine peut faire suspecter une diminution du nombre de cellules du sang, une destruction des globules rouges ou une toxicité du foie.}*

*– {L'allaitement doit être interrompu en cas de traitement par capécitabine.}{Vous ne devez pas allaiter si vous êtes traitée par capecitabine eg.}*

Nous voyons que les phrases avec des substitutions lexicales, comme *{hémolyse}{destruction des globules rouges}*, sont conservées. Ceci est important car il a été observé que ce procédé représente environ 70 % des transformations dans les textes médicaux (Koptient *et al.*, 2019). En revanche, les transformations syntaxiques, comme la modification de parties du discours *{allaitement}{allaiter}* ou de la voix du verbe *{passive}{active}*, sont plus difficiles à conserver. Pour ces cas, un travail plus poussé sur les structures syntaxiques comparables et la morphologie sera nécessaire.

Tous les traitements décrits ci-après sont effectués sur les données filtrées avec les indices formels et le filtre syntaxique de profondeur 3.

### 5.2. *Alignement de phrases parallèles*

| *Classifieur* | *P* | *R* | *F1* | *EQM* | *VP* |
|---|---|---|---|---|---|
| `Perceptron` | 0,90 | 0,93 | 0,92 | 0,08 | 28 |
| `MLP` | 0,93 | 0,93 | 0,93 | 0,06 | 28 |
| `RF` | **1,00** | **0,97** | **0,98** | **0,02** | **29** |
| `LDA` | 0,93 | 0,87 | 0,90 | 0,09 | 26 |
| `QDA` | 0,96 | 0,90 | 0,93 | 0,06 | 27 |
| `LogReg` | 0,97 | **0,97** | 0,97 | 0,03 | **29** |
| `SGD` | 0,90 | 0,93 | 0,92 | 0,08 | 28 |
| `LinSVM` | 0,97 | 0,93 | 0,95 | 0,04 | 28 |

**Tableau 3.** *Résultats d'alignement : différents classifieurs, ensemble des descripteurs, ensemble de test, texte non lemmatisé, ratio des classes* 1 : 1. *en-têtes de colonnes : précision (P), rappel (R), erreur quadratique moyenne (EQM), vrais positifs (VP)*

Les résultats globaux se trouvent dans le tableau 3. Il s'agit de l'exploitation de l'ensemble des descripteurs sur l'ensemble de test avec le texte non lemmatisé. Les résultats sont présentés en termes de rappel $R$, précision $P$, F-mesure $F$, erreur quadratique moyenne $EQM$ et vrais positifs $VP$ (sur un total de 30 paires de phrases alignées dans l'ensemble de test). Nous pouvons voir que tous les classifieurs testés sont compétitifs avec une F-mesure entre 0,92 et 0,98. Pour tous les classifieurs, nous indiquons les scores moyens de 20 itérations. La précision et le rappel sont équilibrés. `Random Forest` semble être le meilleur classifieur : F-mesure de 0,98 (précision 1 et rappel 0,97), le plus grand nombre de vrais positifs (56) et l'erreur quadratique moyenne la plus faible (0,02). `Régression Logistique` est presque aussi performant

avec 0,97 de précision, rappel et F-mesure. Les expériences qui suivent sont effectuées avec `Random Forest`.

5.2.1. *Baseline*

Pour la *baseline*, nous exploitons les descripteurs le plus souvent utilisés : longueur des phrases et intersection lexicale entre les phrases. Les résultats sont présentés dans la première ligne du tableau 4 : nous obtenons une F-mesure de 0,95, ce qui indique que les descripteurs traditionnels sont en effet assez efficaces pour cette tâche.

5.2.2. *Détection de phrases parallèles avec une distribution équilibrée*

| *Descripteurs* | *R* | *P* | *F1* | *EQM* | *VP* | *Descripteurs* | *R* | *P* | *F1* | *EQM* | *VP* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *BL* | 0,97 | 0,93 | 0,95 | 0,05 | 28 | *BL + L + S* | 1,00 | 0,97 | 0,98 | 0,02 | 29 |
| *S* | 0,97 | 0,97 | 0,97 | 0,03 | 29 | *BL + L + N* | 1,00 | 0,97 | 0,98 | 0,02 | 29 |
| *L* | 0,90 | 0,93 | 0,92 | 0,09 | 28 | *BL + L + PL* | 1,00 | 0,97 | 0,98 | 0,02 | 29 |
| *N* | 0,97 | 0,93 | 0,95 | 0,05 | 28 | *BL + S + N* | 1,00 | 0,97 | 0,98 | 0,02 | 29 |
| *PL* | 0,97 | 0,97 | 0,97 | 0,03 | 29 | *BL + S + PL* | 1,00 | 0,97 | 0,98 | 0,02 | 29 |
| *L + S* | 1,00 | 0,93 | 0,97 | 0,03 | 28 | *BL + N + PL* | 1,00 | 0,97 | 0,98 | 0,02 | 29 |
| *L + N* | 1,00 | 0,97 | 0,98 | 0,02 | 29 | *L + S + N* | 1,00 | 0,97 | 0,98 | 0,02 | 29 |
| *L + PL* | 0,97 | 0,97 | 0,97 | 0,03 | 29 | *L + S + PL* | 1,00 | 0,97 | 0,98 | 0,02 | 29 |
| *S + N* | 1,00 | 0,97 | 0,98 | 0,02 | 29 | *L + N + PL* | 1,00 | 0,97 | 0,98 | 0,02 | 29 |
| *S + PL* | 1,00 | 0,97 | 0,98 | 0,02 | 29 | *BL + L + S + N* | 1,00 | 0,97 | 0,98 | 0,02 | 29 |
| *BL + L* | 1,00 | 0,97 | 0,98 | 0,02 | 29 | *BL + L + S + PL* | 1,00 | 0,97 | 0,98 | 0,02 | 29 |
| *BL + S* | 1,00 | 0,97 | 0,98 | 0,02 | 29 | *BL + L + N + PL* | 1,00 | 0,97 | 0,98 | 0,02 | 29 |
| *BL + N* | 1,00 | 0,97 | 0,98 | 0,02 | 29 | *BL + S + N + PL* | 1,00 | 0,97 | 0,98 | 0,02 | 29 |
| *BL + PL* | 1,00 | 0,97 | 0,98 | 0,02 | 29 | *L + S + N + PL* | 1,00 | 0,97 | 0,98 | 0,02 | 29 |
| *N + PL* | 1,00 | 0,97 | 0,98 | 0,02 | 29 | *BL + L + S + N + PL* | 1,00 | 0,97 | 0,98 | 0,02 | 29 |

**Tableau 4.** *Résultats d'alignement : différents ensembles de descripteurs,* `Random Forest`*, texte non lemmatisé, ratio des classes* 1 : 1*. en-têtes de colonnes : précision (P), rappel (R), erreur quadratique moyenne (EQM), vrais positifs (VP). Rangées : baseline (BL), similarité (S), Levenshtein (L), plongements lexicaux (PL).*

Le tableau 4 présente les résultats de la détection de phrases parallèles avec une distribution équilibrée des données. Nous testons différents types de descripteurs et leurs combinaisons. Les meilleurs résultats sont obtenus par l'ensemble *S* (mesures de similarité) avec une F-mesure de 0,97.

Les moins bons résultats sont obtenus avec l'ensemble *L* (distance de Levenshtein) avec une F-mesure de 0,92. Les différentes combinaisons de descripteurs permettent d'améliorer ces résultats, ce qui indique que chaque type de descripteurs apporte des informations complémentaires. La plupart des combinaisons atteignent les résultats les plus élevés. Les expériences qui suivent sont effectuées avec tous les descripteurs.

5.2.3. *Détection de phrases parallèles selon la sémantique des paires avec des données équilibrées*

| *Ensemble* | *P* | *R* | *F1* | *EQM* | *VP* |
|---|---|---|---|---|---|
| *Équivalence E* | 1,00 | 0,97 | 0,98 | 0,02 | 29 |
| *Inclusion I* | 1,00 | 0,94 | 0,97 | 0,03 | 29 |

**Tableau 5.** *Résultats d'alignement : les deux ensembles de données équilibrées (l'équivalence sémantique et les inclusions), ensemble de test, tous les descripteurs,* `Random Forest`*, texte non lemmatisé, ratio des classes* $1:1$*. en-têtes de colonnes : précision (P), rappel (R), erreur quadratique moyenne (EQM), vrais positifs (VP).*

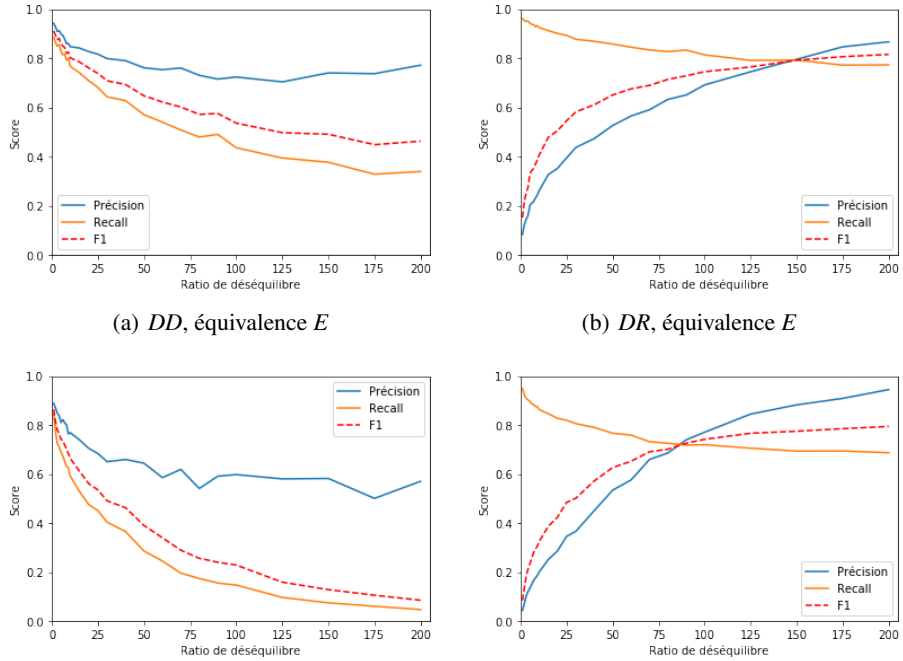
Dans cette série d'expériences sur les données équilibrées, nous voulons voir s'il existe une différence selon le type de relation sémantique au sein des paires de phrases. Dans l'ensemble de test, nous comptons 30 couples en équivalence et 31 en inclusion. Selon le tableau 5, il existe une légère différence : il est un peu plus facile de détecter les phrases en relation d'équivalence que les phrases en relation d'inclusion. Nous supposons que les paires d'inclusion couvrent une plus grande variété de situations, ce qui est plus difficile à modéliser avec le volume de données dont nous disposons.

5.2.4. *Détection de phrases parallèles avec une distribution déséquilibrée*

Comme les documents comparables peuvent contenir un taux variable de phrases parallèles, nous faisons des tests avec des données déséquilibrées. Nous testons différents taux de déséquilibre. Les résultats sont présentés à la figure 1 : l'axe $x$ représente l'augmentation du déséquilibre (seule la première position 1 correspond aux données équilibrées), alors que l'axe $y$ représente les scores de précision, rappel et F-mesure. Les résultats pour les deux ensembles sont présentés : équivalence (figures 1(a) et 1(b)) et inclusion (figures 1(c) et 1(d)). La colonne de gauche présente les résultats $DD$, lorsque l'entraînement et le test sont effectués sur des données avec le même rapport de déséquilibre $D$. La colonne de droite présente les résultats $DR$ obtenus par les mêmes modèles $D$ mais testés sur l'ensemble des données $R$ (toutes les paires de phrases possibles). Les résultats présentés sont les moyennes de 50 itérations.

Nous pouvons faire plusieurs observations. Comme indiqué dans la section précédente, les paires équivalentes (figures 1(a) et 1(b)) sont plus faciles à catégoriser que les inclusions. Les scores de précision et de rappel sont alors plus élevés à différents points de déséquilibre.

Ce résultat est positif car les phrases équivalentes fournissent les informations les plus utiles et complètes sur les transformations requises lors de la simplification. Sans surprise, l'augmentation du déséquilibre mène vers des performances réduites durant l'entraînement. Cela signifie que le déséquilibre crée de la confusion entre les paires alignables et non. Cependant, pour atteindre notre objectif, qui consiste à identifier le peu d'exemples positifs présents dans une masse de paires non alignables, il vaut mieux utiliser un modèle plus robuste face au déséquilibre des données, même s'il




(a) *DD*, équivalence *E*

(b) *DR*, équivalence *E*

(c) *DD*, inclusion *I*

(d) *DR*, inclusion *I*

**Figure 1.** *Précision, rappel et F-mesure obtenus pour les deux séries d'expériences (DD et DR) sur les données déséquilibrées*

fonctionne moins bien à l'entraînement. Il s'agit ici de modèles entraînés avec un déséquilibre important. Finalement, notons que la même tâche a été effectuée avec un réseau neuronal à propagation avant basique[8], avec différentes expériences faisant varier le nombre de couches cachées et leur taille ainsi que le nombre d'*epochs*. Les résultats étaient plus élevés lors de l'entraînement sur les données équilibrées : jusqu'à 0,98 de précision, rappel et F-mesure. Cependant, à partir du déséquilibre $5:1$, toutes les phrases étaient systématiquement classées dans la catégorie « *non aligné* », ce qui montre les limites de l'architecture neuronale utilisée. Nous avons également fait des essais avec un système d'alignement existant qui utilise un réseau de neurones récurrent bidirectionnel (Grégoire et Langlais, 2018), mais les résultats obtenus n'étaient pas exploitables. Finalement, nous avons ajouté les descripteurs à base de plongements lexicaux comme proposé dans un travail existant (Kajiwara et Komachi, 2016) mais cette expérience n'a pas montré d'évolution dans les résultats. Nous

8. Un réseau à propagation avant avec fonction d'activation `ReLU`, optimiseur `ADAM`, fonction d'erreur `BCEWithLogitsLoss` (sigmoïde + `BCELoss`).

pensons que le faible volume d'exemples à notre disposition (136 exemples d'équivalence et 130 exemples d'inclusion) représente un frein à l'utilisation des méthodes neuronales. Un travail plus avancé sur ce type de méthodes et l'accroissement du volume des données de référence font partie de nos perspectives.

### 5.3. *Analyse des erreurs*

| | Équivalence | Inclusion | Intersection | Faux positifs |
|---|---|---|---|---|
| Nb de paires | 75 | 15 | 2 | 8 |

**Tableau 6.** *Analyse des 100 premiers alignements par le modèle entraîné sur les paires équivalentes à un ratio de* $125:1$*, appliqué sur un ensemble aléatoire de paires non vues pendant l'entraînement*

Le tableau 6 présente les différents types de paires de phrases considérées comme alignables par le modèle entraîné sur un ratio de $125:1$ de paires équivalentes, qui montre le meilleur équilibre entre rappel et précision sur les données réelles. Les 100 premiers alignements, absents des données d'entraînement, sont analysés. Nous pouvons observer que 75 % de paires correspondent en effet à l'équivalence, ce qui est en adéquation avec la figure 1(b). Nous remarquons également que 15 % des paires alignées relèvent de l'inclusion, qui est une catégorie plus difficile à identifier de manière ciblée. Nous trouvons enfin que 2 % des alignements correspondent à d'intersection. Nous ne recherchons pas spécifiquement ce type de couples de phrases pour la simplification car nous le considérons plus difficile à traiter. Cependant, de tels alignements peuvent être utiles à l'identification de paraphrases, par exemple. Nous avons ainsi 90 à 92 % d'alignements exploitables et 8 à 10 % de bruit. Certaines paires des corpus *Cochrane* et *Médicament* sont plus difficiles à aligner car ces corpus combinent les transformations lexicales spécifiques du domaine, l'utilisation de l'opposition et du contraire, et des transformations syntaxiques :

– *{Les médicaments inhibant le péristaltisme sont contre-indiqués dans cette situation.}{Dans ce cas, ne prenez pas de médicaments destinés à bloquer ou ralentir le transit intestinal.}*

– *{Aucune preuve n'indique que les agents gonflants sont efficaces dans le traitement du SCI}{Nous avons observé que les agents gonflants n'étaient pas efficaces dans le traitement du SCI}*

Pour aider ce type d'alignement, il serait nécessaire de capter la similarité lexicale et sémantique avec des ressources et connaissances complémentaires. Elles peuvent venir de ressources externes ou bien être acquises sur le corpus.

Finalement, dans les faux positifs, nous trouvons de bons candidats à l'alignement qui ne sont pas dans les données de référence, comme :

*– {Trois études ont rapporté des résultats mitigés concernant l'association entre le début des cours plus tardif et la vigilance des étudiants.}{Différentes études rapportaient des résultats mitigés concernant l'association entre le début plus tardif des cours et une augmentation de la fréquentation et de la vigilance des étudiants.}*

### 5.4. *Limitations de l'étude actuelle*

La limitation principale des expériences présentées est liée aux descripteurs exploités. Parmi les quatre types de descripteurs distingués dans les travaux existants, nous exploitons les descripteurs principalement basés sur le lexique et le corpus. L'exploitation de descripteurs faciles à calculer et ne nécessitant pas de ressources externes était un des objectifs. Cependant, cet aspect doit évoluer car actuellement les similarités lexicales sont assez faibles dans les phrases différenciées par leur degré de technicité. Ce point a été relevé lors de l'analyse des erreurs d'alignement : les faux négatifs contiennent souvent des phrases sémantiquement similaires mais contenant un lexique et des structures syntaxiques différents. Une meilleure intégration de plongements lexicaux et d'une architecture neuronale fait partie des perspectives.

Une autre limitation est liée à la catégorisation binaire des paires de phrases selon qu'elles sont alignables ou non. Cette catégorisation est motivée par la tâche poursuivie, où nous avons besoin de paires de phrases parallèles pour induire des règles de transformation nécessaires pour la simplification. Cependant, comme dans les données STS, nous pouvons aussi viser de caractériser les paires de phrases sur une échelle de similarité et disposer ainsi de données de référence plus fines. Notons que nous avons effectué un travail de ce type sur des données issues du corpus CLEAR et des articles de Wikipédia et Vikidia en langue générale. Ces données ont été exploitées lors de la compétition DEFT 2020.

## 6. Conclusion

Nous avons proposé une série d'expériences en alignement de phrases parallèles à partir de corpus monolingues comparables en français. La dimension comparable est due à la technicité des documents et contraste les versions techniques et simplifiées des documents et des phrases. Nous exploitons un corpus comparable existant lié au domaine biomédical et contenant des documents de trois genres (encyclopédique, scientifique et notices de médicaments). Les données de référence sont construites manuellement. La recherche de phrases parallèles est abordée comme une problématique de catégorisation : nous devons décider si une paire de phrases peut être alignée ou non. Plusieurs classifieurs sont exploités. Nos résultats atteignent une F-mesure de 0,97 sur les données équilibrées en français, avec un bon équilibre entre la précision et le rappel. Les meilleurs résultats sont obtenus avec `Random Forest`. Deux autres expériences s'intéressent aux types de relations au sein des paires de phrases (les paires de phrases avec la relation d'équivalence sont plus faciles à aligner que les phrases

en relation d'inclusion) et sur l'équilibre entre les paires alignables et non alignables dans les ensembles d'entraînement et de test.

Comme nous l'avons vu, dans les données de référence, la distance lexicale entre les phrases techniques et simplifiées est assez élevée. En conséquence, d'autres descripteurs doivent être utilisés pour mieux cerner les phrases alignables. Par exemple, nous comptons utiliser des connaissances externes, comme les terminologies médicales (Côté *et al.*, 1993 ; Lindberg *et al.*, 1993) et le lexique ReSyf (Billami *et al.*, 2018), ou des ressources constituées à partir de corpus. Nous comptons également tester les représentations de phrases avec FlauBERT (Le *et al.*, 2020) et CamemBERT (Martin *et al.*, 2020). Une étude plus poussée des descripteurs pourrait également être intéressante : les performances selon les types d'alignement (équivalence et inclusion), l'impact des descripteurs individuels et non pas par sous-ensembles. Comme le déséquilibre est une caractéristique naturelle des données que nous traitons, le travail à venir pourra également enrichir les filtres pour éliminer un maximum de phrases non alignables, *a priori* et/ou *a posteriori* de l'alignement.

Les meilleurs modèles générés sont actuellement exploités pour enrichir l'ensemble de phrases parallèles. En plus des corpus comparables liés au domaine médical, nous exploitons également un corpus similaire de la langue générale qui regroupe l'ensemble d'articles comparables de Wikipédia et de Vikidia. Nous avons également l'intention d'essayer la méthode sur des corpus spécialisés d'autres domaines que celui de la santé. La ressource constituée dans ce travail, un corpus avec plusieurs milliers de phrases parallèles, sera mise à disposition des chercheurs. En dehors de la simplification automatique, les phrases parallèles peuvent aussi être intéressantes pour d'autres applications de TAL, comme l'étude de la similarité textuelle, les systèmes de questions-réponses, la recherche d'information ou l'implication textuelle.

Remerciements

Ce travail s'inscrit dans le cadre du projet ANR-17-CE19-0016-01 CLEAR (*Communication*, *Literacy*, *Education*, *Accessibility*, *Readability*) financé par l'Agence nationale de la recherche. Nous remercions les relecteurs pour leurs commentaires et remarques détaillés qui ont permis d'améliorer la qualité du présent article.

## 7. Bibliographie

Abdul-Rauf S., Schwenk H., « On the Use of Comparable Corpora to Improve SMT performance », *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, Association for Computational Linguistics, Athens, Greece, p. 16-23, March, 2009.

Agirre E., Cer D., Diab M., Gonzalez-Agirre A., Guo W., « *SEM 2013 shared task : Semantic Textual Similarity », *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1 : Proceedings of the Main Conference and the Shared Task : Seman-*

*tic Textual Similarity*, Association for Computational Linguistics, Atlanta, Georgia, USA, p. 32-43, June, 2013.

Barzilay R., Elhadad N., « Sentence Alignment for Monolingual Comparable Corpora », *in* ACL (ed.), *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, Sapporo, Japan, p. 25-32, 2003.

Beigman Klebanov B., Knight K., Marcu D., « Text Simplification for Information-Seeking Applications », *in* R. Meersman, Z. Tari (eds), *On the Move to Meaningful Internet Systems 2004 : CoopIS, DOA, and ODBASE*, Springer, LNCS vol 3290, Berlin, Heidelberg, 2004.

Berkson J., « Application of the Logistic Function to Bio-Assay », *Journal of the American Statistical Association*, vol. 39, n° 227, p. 357-365, 1944.

Billami M. B., François T., Gala N., « ReSyf : a French lexicon with ranked synonyms », *in* ACL (ed.), *27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, p. 2570-2581, 2018.

Bird S., Klein E., Loper E., *Natural Language Processing with Python : Analyzing Text with the Natural Language Toolkit*, O'Reilly, Beijing, China, 2009.

Blake C., Kampov J., Orphanides A. K., West D., Lown C., « Unc-ch at duc 2007 : Query expansion, lexical simplification and sentence selection strategies for multi-document summarization », *Proceedings of Document Understanding Conference (DUC) Workshop*, Rochester, New York, USA, 2007.

Brouwers L., Bernhard D., Ligozat A.-L., François T., « Syntactic Sentence Simplification for French », *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, Association for Computational Linguistics, Gothenburg, Sweden, p. 47-56, April, 2014.

Cardon R., Grabar N., « Parallel Sentence Retrieval From Comparable Corpora for Biomedical Text Simplification », *Proceedings of Recent Advances in Natural Language Processing*, Varna, Bulgaria, p. 168-177, september, 2019.

Cardon R., Grabar N., « Reducing the Search Space for Parallel Sentences in Comparable Corpora », *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*, European Language Resources Association, Marseille, France, p. 44-48, May, 2020.

Chandrasekar R., Srinivas B., « Automatic induction of rules for text simplification », *Knowledge Based Systems*, vol. 10, n° 3, p. 183-190, 1997.

Chen P., Rochford J., Kennedy D. N., Djamasbi S., Fay P., Scott W., « Automatic text simplification for people with intellectual disabilities », *Artificial Intelligence Science and Technology*, 2016.

Chen S. F., « Aligning Sentences in Bilingual Corpora Using Lexical Information », *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*, ACL '93, Association for Computational Linguistics, USA, p. 9–16, 1993.

Clough P., Gaizauskas R., Piao S. S., Wilks Y., « Measuring Text Reuse », *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, p. 152-159, July, 2002.

Cohen J., « A Coefficient of Agreement for Nominal Scales », *Educational and Psychological Measurement*, vol. 20, n° 1, p. 37, 1960.

Côté R. A., Rothwell D. J., Palotay J. L., Beckett R. S., Brochu L., *The Systematised Nomenclature of Human and Veterinary Medicine : SNOMED International*, College of American Pathologists, Northfield, 1993.

Cover T., « Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition », *IEEE Transactions on Electronic Computers*, vol. 14, nº 3, p. 326-334, 1965.

De Belder J., Moens M.-F., « Text Simplification for Children », *Workshop on Accessible Search Systems of SIGIR*, Geneva, Switzerland, p. 1-8, 2010.

Duran K., Rodriguez J., Bravo M., « Similarity of sentences through comparison of syntactic trees with pairs of similar words », *11th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE)*, Campeche, p. 1-6, 09, 2014.

Ferguson T., « An inconsistent maximum likelihood estimate », *Journal of the American Statistical Association*, vol. 77, nº 380, p. 831-834, 1982.

Fernando S., Stevenson M., « A semantic similarity approach to paraphrase detection », *Comp Ling UK*, p. 1-7, 2008.

Fisher R., « The Use of Multiple Measurements in Taxonomic Problems », *Annals of Eugenics*, vol. 7, nº 2, p. 179-188, 1936.

Franco-Salvador M., Gupta P., Rosso P., Banchs R. E., « Cross-language plagiarism detection over continuous-space- and knowledge graph-based representations of language », *Knowledge-Based Systems*, vol. 111, p. 87 - 99, 2016.

Fung P., Cheung P., « Mining Very-Non-Parallel Corpora : Parallel Sentence and Lexicon Extraction via Bootstrapping and EM », *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Barcelona, Spain, p. 57-63, July, 2004.

Gale W. A., Church K. W., « A Program for Aligning Sentences in Bilingual Corpora », *Comp Linguistics*, vol. 19, nº 1, p. 75-102, 1993.

Grave E., Bojanowski P., Gupta P., Joulin A., Mikolov T., « Learning Word Vectors for 157 Languages », *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association (ELRA), Miyazaki, Japan, May, 2018.

Grégoire F., Langlais P., « Extracting Parallel Sentences with Bidirectional Recurrent Neural Networks to Improve Machine Translation », *Proceedings of the 27th International Conference on Computational Linguistics*, Association for Computational Linguistics, Santa Fe, New Mexico, USA, p. 1442-1453, August, 2018.

Guo W., Diab M., « Modeling Sentences in the Latent Space », *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Association for Computational Linguistics, Jeju Island, Korea, p. 864-872, July, 2012.

He H., Gimpel K., Lin J., « Multi-Perspective Sentence Similarity Modeling with Convolutional Neural Networks », *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Lisbon, Portugal, p. 1576-1586, September, 2015.

Hewavitharana S., Vogel S., « Extracting Parallel Phrases from Comparable Data », *Proceedings of the 4th Workshop on Building and Using Comparable Corpora : Comparable Corpora and the Web*, Association for Computational Linguistics, Portland, Oregon, p. 61-68, June, 2011.

Ho T. K., « Random Decision Forests », *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1*, ICDAR '95, IEEE Computer Society, USA, p. 278-282, 1995.

Hwang W., Hajishirzi H., Ostendorf M., Wu W., « Aligning Sentences from Standard Wikipedia to Simple Wikipedia », *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, Association for Computational Linguistics, Denver, Colorado, p. 211-217, May–June, 2015.

Jonnalagadda S., Tari L., Hakenberg J., Baral C., Gonzalez G., « Towards Effective Sentence Simplification for Automatic Processing of Biomedical Text », *Proceedings of Human Language Technologies : The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume : Short Papers*, Association for Computational Linguistics, Boulder, Colorado, p. 177-180, June, 2009.

Kajiwara T., Komachi M., « Building a Monolingual Parallel Corpus for Text Simplification Using Sentence Similarity Based on Alignment between Word Embeddings », *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, The COLING 2016 Organizing Committee, Osaka, Japan, p. 1147-1158, December, 2016.

Kitaev N., Klein D., « Constituency Parsing with a Self-Attentive Encoder », *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Association for Computational Linguistics, Melbourne, Australia, July, 2018.

Koptient A., Cardon R., Grabar N., « Simplification-induced transformations : typology and some characteristics », *Proceedings of the 18th BioNLP Workshop and Shared Task*, Association for Computational Linguistics, Florence, Italy, p. 309-318, August, 2019.

Lai A., Hockenmaier J., « Illinois-LH : A Denotational and Distributional Approach to Semantics », *Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland, p. 239-334, 2014.

Le H., Vial L., Frej J., Segonne V., Coavoux M., Lecouteux B., Allauzen A., Crabbé B., Besacier L., Schwab D., « FlauBERT : Unsupervised Language Model Pre-training for French », *Proceedings of The 12th Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, p. 2479-2490, May, 2020.

Leroy G., Kauchak D., Mouradi O., « A user-study measuring the effects of lexical simplification and coherence enhancement on perceived and actual text difficulty », *Int J Med Inform*, vol. 82, n° 8, p. 717-730, 2013.

Levenshtein V. I., « Binary Codes Capable of Correcting Deletions, Insertions and Reversals », *Soviet Physics Doklady*, vol. 10, p. 707, February, 1966.

Lindberg D., Humphreys B., McCray A., « The Unified Medical Language System », *Methods Inf Med*, vol. 32, n° 4, p. 281-291, 1993.

Martin L., Muller B., Ortiz Suárez P. J., Dupont Y., Romary L., de la Clergerie É., Seddah D., Sagot B., « CamemBERT : a Tasty French Language Model », *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, p. 7203-7219, July, 2020.

Mihalcea R., Corley C., Strapparava C., « Corpus-Based and Knowledge-Based Measures of Text Semantic Similarity », *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, AAAI'06, AAAI Press, Boston, Massachusetts, p. 775–780, 2006.

Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J., « Distributed Representations of Words and Phrases and their Compositionality », *Advances in Neural Information Processing Systems 26 : 27th Annual Conference on Neural Information Processing Systems*, Lake Tahoe, Nevada, USA, p. 3111-3119, 2013.

Mueller J., Thyagarajan A., « Siamese Recurrent Architectures for Learning Sentence Similarity », *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, AAAI Press, Phoenix, Arizona, p. 2786–2792, 2016.

Munteanu D. S., Marcu D., « Processing Comparable Corpora With Bilingual Suffix Trees », *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, Association for Computational Linguistics, Philadelphia, PA, USA, p. 289-295, July, 2002.

Nelken R., Shieber S. M., « Towards Robust Context-Sensitive Sentence Alignment for Monolingual Corpora », *11th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, Trento, Italy, April, 2006.

Nisioi S., Štajner S., Ponzetto S. P., Dinu L. P., « Exploring Neural Text Simplification Models », *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, Association for Computational Linguistics, Vancouver, Canada, p. 85-91, July, 2017.

Paetzold G., Specia L., « Benchmarking Lexical Simplification Systems », *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, European Language Resources Association (ELRA), Portorož, Slovenia, p. 3074-3080, May, 2016.

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E., « Scikit-learn : Machine Learning in Python », *Journal of Machine Learning Research*, vol. 12, p. 2825-2830, 2011.

Qiu L., Kan M.-Y., Chua T.-S., « Paraphrase recognition via dissimilarity significance classification », *Empirical Methods in Natural Language Processing*, Sydney, Australia, p. 18-26, 2006.

Rosenblatt F., « The Perceptron : a probabilistic model for information storage and organization in the brain », *Psychological Review*, vol. 65, n° 6, p. 386-408, 1958.

Rosenblatt F., *Principles of Neurodynamics : Perceptrons and the Theory of Brain Mechanisms*, Spartan Books, Washington DC, 1961.

Schmid H., « Probabilistic Part-of-Speech Tagging Using Decision Trees », *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK, 1994.

Severyn A., Nicosia M., Moschitti A., « Learning Semantic Textual Similarity with Structural Representations », *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, Association for Computational Linguistics, Sofia, Bulgaria, p. 714-718, August, 2013.

Stajner S., Franco-Salvador M., Ponzetto S. P., Rosso P., « CATS : A Tool for Customized Alignment of Text Simplification Corpora », *Proceedings of the 11th Language Resources and Evaluation Conference, LREC*, Miyazaki, Japan, 2018.

Ștefănescu D., Ion R., Hunsicker S., « Hybrid Parallel Sentence Mining from Comparable Corpora », *16th Conference of the European Association for Machine Translation EAMT*, Trento, Italy, p. 137-144, 2012.

Stymne S., Tiedemann J., Hardmeier C., Nivre J., « Statistical Machine Translation with Readability Constraints », *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA)*, Linköping University Electronic Press, Sweden, Oslo, Norway, p. 375-386, May, 2013.

Tai K. S., Socher R., Manning C. D., « Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks », *Annual Meeting of the Association for Computational Linguistics*, Beijing, China, p. 1556-1566, 2015.

Tsubaki M., Duh K., Shimbo M., Matsumoto Y., « Non-Linear Similarity Learning for Compositionality », *AAAI Conference on Artificial Intelligence*, p. 2828-2834, 2016.

Utiyama M., Isahara H., « Reliable Measures for Aligning Japanese-English News Articles and Sentences », *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Sapporo, Japan, p. 72-79, July, 2003.

Vapnik V., Lerner A., « Pattern Recognition using Generalized Portrait Method », *Automation and Remote Control*, vol. 24, p. 709-715, 1963.

Vickrey D., Koller D., « Sentence Simplification for Semantic Role Labeling », *Proceedings of ACL-08 : HLT*, Association for Computational Linguistics, Columbus, Ohio, p. 344-352, June, 2008.

Štajner S., Popović M., « Can Text Simplification Help Machine Translation ? », *Baltic J. Modern Computing*, vol. 4, n° 2, p. 230-242, 2016.

Wan S., Dras M., Dale R., Paris C., « Using Dependency-based Features to Take the "Para-farce" out of Paraphrase », *Australasian Language Technology Workshop*, p. 131-138, 2006.

Wei C.-H., Leaman R., Lu Z., « SimConcept : A Hybrid Approach for Simplifying Composite Named Entities in Biomedicine », *BCB '14*, p. 138-146, 2014.

Xu W., Callison-Burch C., Napoles C., « Problems in Current Text Simplification Research : New Data Can Help », *Transactions of the Association for Computational Linguistics*, vol. 3, p. 283-297, 2015.

Yang C. C., Li K. W., « Automatic construction of English/Chinese parallel corpora », *J. Am. Soc. Inf. Sci. Technol.*, vol. 54, n° 8, p. 730-742, 2003.

Zhang X., Lapata M., « Sentence Simplification with Deep Reinforcement Learning », *in* ACL (ed.), *Proc of the Conf on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, p. 584-594, 2017.

Zhang Y., Patrick J., « Paraphrase identification by text canonicalization », *Australasian Language Technology Workshop*, p. 160-166, 2005.

Zhao B., Vogel S., « Adaptive parallel sentences mining from web bilingual news collection », *IEEE International Conference on Data Mining*, p. 745-748, 2002.

Zhao J., Zhu T., Lan M., « ECNU : One Stone Two Birds : Ensemble of Heterogenous Measures for Semantic Relatedness and Textual Entailment », *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Association for Computational Linguistics, Dublin, Ireland, p. 271-277, August, 2014.

Zhu Z., Bernhard D., Gurevych I., « A Monolingual Tree-based Translation Model for Sentence Simplification », *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, Coling 2010 Organizing Committee, Beijing, China, p. 1353-1361, August, 2010.