Transformations syntaxiques entre niveaux de simplification dans le corpus Newsela

Rita Hijazi

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Transformations syntaxiques entre niveaux de simplification dans le corpus Newsela

Rita Hijazi[1, 2]
(1) LPL, 13090 Aix-en-Provence, France
(2) LIS, 13397 Marseille, France
rita.hijazi@etu.univ-amu.fr

RÉSUMÉ

La simplification de textes est une tâche complexe du traitement automatique des langues. Depuis quelques années, des corpus parallèles de textes originaux et simplifiés sont proposés, permettant d'apprendre différents types d'opérations de simplification à partir de corpus. Dans le but de pouvoir développer et évaluer des systèmes de simplification automatique de textes, cet article s'intéresse au corpus Newsela, un corpus parallèle de textes en langue anglaise avec quatre niveaux de simplification. Nous présentons en détail ce corpus et étudions les différentes transformations caractérisant le passage d'un niveau de simplification à l'autre sur un sous-ensemble de textes, en nous intéressant plus particulièrement aux transformations syntaxiques.

ABSTRACT

Syntactic transformations between simplification levels in the Newsela corpus.

Text simplification is a complex task of Natural Language Processing. Research into this topic has many potential practical applications. For several years now, parallel corpora of complex–simple paired sentences have been developed to provide examples of structural transformations and particularly suitable for text simplification. To develop and evaluate ATS systems, this article focuses on the Newsela corpus, a parallel corpus of English texts with four levels of simplification. We present this corpus in detail, and we study the different transformations characterizing the passage from one level of simplification to another on a subset of texts. We specifically focus on syntactic transformations.

MOTS-CLÉS : Corpus parallèle, simplification de textes, analyse du corpus.
KEYWORDS: Parallel corpora, text simplification, corpus analysis.

1 Introduction

La simplification de textes (SAT) est un domaine du traitement automatique des langues (TAL) qui suscite l'intérêt dans la communauté depuis quelques années. L'objectif est de rendre des textes plus abordables tout en garantissant l'intégrité sémantique de leur contenu. Saggion (2017) définit la SAT comme étant le processus de transformation d'un texte en un autre texte qui véhicule le même contenu sémantique, afin de le rendre plus facile à lire et à comprendre par un public cible.

Volume 3 : Rencontre des Étudiants Chercheurs en Informatique pour le TAL, pages 137–150.
HAL : hal-02786194.

La SAT peut être adressée à des lecteurs humains faisant face à différents types de difficultés de lecture, par exemple, les apprenants de langues (Petersen et Ostendorf, 2007 ; Burstein, 2009), les personnes souffrant d'aphasie (Devlin and Tait, 1998 ; Carroll et al., 1998), de dyslexie (Rello et al., 2013) ou d'autisme (Evans et al., 2014). La SAT peut aussi être utilisée comme étape de prétraitement pour d'autres tâches de TAL, telles que l'analyse syntaxique (Chandrasekar et al., 1996), le résumé automatique (Vanderwende et al., 2007 ; Silveira et Branco, 2012) et la traduction automatique (Hasler et al., 2017).

Un des problèmes majeurs pour la construction de systèmes de SAT efficaces est l'absence de corpus annotés en niveaux de difficulté ou, tout au moins, des versions parallèles originales et simplifiées. En effet, une étape préalable et importante vers la construction des systèmes de SAT est l'analyse et la comparaison de versions parallèles de textes simplifiés originaux, afin d'examiner quels types de changements doivent être appliqués et pour quel public et quelles ressources sont nécessaires pour les mettre en place automatiquement. Nous nous intéressons ici au corpus Newsela (Newsela, 2016) dédié à cette tâche. Il s'agit d'un corpus parallèle de textes en anglais avec quatre niveaux de simplification. Nous présentons en détail ce corpus et étudions les différentes transformations caractérisant le passage d'un niveau de simplification à l'autre sur un sous-ensemble de textes, en nous intéressant plus particulièrement aux transformations syntaxiques.

Cet article est structuré de la façon suivante. Dans la section 2, nous définissons le contexte du travail et nous présentons en détail le corpus Newsela, un corpus parallèle dédié à la simplification de textes en anglais avec quatre niveaux de simplification. Dans la section 3 nous présentons la méthodologie que nous avons retenue pour l'étude des différents niveaux de simplification de ce corpus. Dans la section 4 nous proposons une analyse des opérations de simplification mises en œuvre pour passer d'un niveau de simplification à l'autre dans le corpus Newsela. Nous concluons en présentant plusieurs perspectives à ce travail.

2 Contexte

2.1 Des corpus pour la simplification des textes

Force est de constater que les corpus parallèles version originale *vs* version simplifiée ne sont pas nombreux, bien que pour l'anglais, des initiatives différentes ont vu le jour dans différents domaines (pas forcément en TAL). L'initiative Plain English[1], par exemple, est née dans les années 1970 pour faciliter la compréhension des textes officiels administratifs. Il s'agit de directives qui peuvent en principe s'appliquer à toutes les langues, par exemple : garder le sujet, le verbe et l'objet ensemble ; expliquer une seule idée par phrase ; utiliser des phrases courtes ; utiliser la voix active ; etc.
En TAL, des ressources existent pour le français, comme Vikidia[2] (Wikipédia Junior) qui a été utilisée par Brouwers et ses collaborateurs (2012) afin d'établir une typologie des règles de simplification. Vikidia est un corpus destiné aux jeunes de huit à treize ans et rassemble des articles plus accessibles, tant au niveau de la langue que du contenu (Brouwers et al., 2012). Ce corpus comprend à ce jour plus de 29 900 articles.
Une autre ressource existante est le corpus Alector[3] (Gala et al., 2020a). Il s'agit d'une collection de 79 textes littéraires (contes, histoires) et scientifiques (documentaires) originaux ainsi que leurs

1 https://www.plainlanguage.gov/
2 http://fr.vikidia.org/
3 https://corpusalector.huma-num.fr/

équivalents simplifiés. Les textes ont été choisis parmi une variété de supports disponibles pour les élèves des écoles primaires françaises, particulièrement les apprenants du cours élémentaire 1 et 2 (CE1 et CE2), et cours moyen 1 (CM1). Les 79 textes originaux ont tous subi des simplifications aux niveaux lexical, morphologique, syntaxique et discursif. Les simplifications ont été faites manuellement par une équipe d'experts, les corpus ont été testés dans des écoles (plus de mille élèves) dans le but d'obtenir des résultats sur l'impact des simplifications sur la lecture et la compréhension[4].

Au niveau de la SAT, on distingue plusieurs étapes. L'une des sous-tâches est également la simplification lexicale (Specia et al., 2012), qui consiste à remplacer les mots par des synonymes plus simples. La simplification est parfois une forme de paraphrase dans laquelle une phrase est reformulée en une phrase linguistiquement plus simple tout en conservant le sens de la phrase d'origine. Les paraphrases pour la simplification sont généralement extraites à partir de corpus parallèles. En anglais, Creutz (2018) a proposé Opusparcus[5] (*OpenSUbtitleSPARaphraseCorpUS*), un corpus de paraphrases pour six langues européennes : l'allemand, l'anglais, le finnois, le français, le russe et le suédois. Les ensembles de données ont été extraits d'OpenSubtitles2016[6] (Lison et Tiedemann, 2016), qui est une collection de sous-titres traduits de films et d'émissions de télévision. Pour chaque langue, les données sont divisées en trois types d'ensembles de données apprentissage, développement et évaluation. Les données d'apprentissage sont composées de millions de paires de phrases, et ont été compilés automatiquement. Les ensembles de développement et d'évaluation sont constitués de paires de phrases qui ont été vérifiées manuellement ; chaque ensemble contient environ 1000 paires de phrases (Creutz, 2018). Ce corpus, du fait de son contenu (paraphrases) peut être utilisé à des fins de développement ou de test d'un système de SAT.

La paraphrase a déjà été prise en compte dans des systèmes de SAT à des fins éducatives. Ces systèmes reposent souvent sur des règles de transformation définies par des experts. Inui et ses collaborateurs (2003) répondent aux besoins des apprenants sourds de l'anglais et du japonais écrits en paraphrasant des textes en supprimant les structures syntaxiques difficiles pour ce groupe d'apprenants (Inui et al., 2003). Le but du projet *Practical Simplification of English Text* (PSET) est de paraphraser les textes des journaux pour les personnes aphasiques (Canning et al., 2000). Max et ses collaborateurs (2006) ciblent les rédacteurs de textes pour les lecteurs souffrant de troubles du langage avec un système de simplification de texte interactif intégré dans un traitement de texte suggérant des simplifications tout en permettant à l'auteur du texte de garder le contrôle sur son contenu (Max, 2006). Le service de test pédagogique (*Educational Test Service* ETS) a développé l'outil d'adaptation automatique de texte (*Automatic Text Adaptation* ATA ; Burstein et al., 2007). Ce système ne simplifie pas directement le texte d'origine mais fournit plutôt une aide à la lecture via des adaptations de texte en anglais et/ou en espagnol qui sont affichées avec le texte original. Ces adaptations incluent la prise en charge du vocabulaire, les notes marginales et la synthèse vocale.

Une deuxième sous-tâche de SAT est la simplification syntaxique ayant pour but d'identifier et de transformer de longues phrases contenant des phénomènes syntaxiques qui peuvent nuire à la lisibilité pour certaines personnes en paraphrases plus simples qui ne contiennent pas ces phénomènes. La majorité des méthodes de simplification syntaxique proposées reposent sur un ensemble de règles de transformation définies manuellement pour être appliquées aux phrases. Le

---

4 Plus d'informations sur le projet Alector : https://alectorsite.wordpress.com/

5 https://korp.csc.fi/download/opusparcus/

6 OpenSubtitles2016 est un sous-ensemble de la collection OPUS ("... *the open parallel corpus*") http://opus.nlpl.eu/, et fournit un grand nombre de corpus parallèles alignés sur les phrases dans 65 langues.

processus de simplification peut ainsi être considéré comme un processus de reconnaissance de paraphrase dirigée ou d'implication textuelle, avec une contrainte de lisibilité sur le texte simplifié. La relation de simplification est donc asymétrique, contrairement à la paraphrase, elle se rapproche ainsi de l'implication textuelle. Les opérations de simplification visent à préserver la plupart des informations contenues dans le texte. Par conséquent, les informations périphériques sont supprimées, d'où le fait que les techniques de résumé jouent un rôle important dans la SAT.

Enfin, au cours des dernières années, la disponibilité des corpus parallèles de textes originaux et simplifiés a rendu possible un ensemble d'approches permettant d'apprendre différents types d'opérations de simplification à partir de corpus. Cependant, la SAT peut être appréhendée avec des méthodes de traduction automatique et d'apprentissage automatique dont les modèles statistiques sont construits à partir de corpus parallèles de textes originaux et simplifiés (Zhu et al., 2010; Specia, 2010; Woodsend et Lapata, 2011). Notamment, la disponibilité du corpus PWKP (*Parallel Wikipedia Simplification Corpus*) constitué par Zhu et ses collaborateurs (2010) a eu un impact considérable. Sa taille et sa disponibilité en ont fait le jeu de données de référence des travaux de simplification pour l'anglais. Il se compose d'un texte aligné de Wikipédia anglais[7] et de son équivalent dans Simple English Wikipedia[8]. L'ensemble de données contient 108 016 paires de phrases, avec 25,01 mots en moyenne par phrase 'complexe' et 20,87 mots par phrase simple.
Cependant, Xu et ses collaborateurs (2015) ont présenté une prise de position, dans laquelle ils décrivent plusieurs lacunes de cette ressource et contient de bruits. Ils ont approfondi cette étude par une annotation manuelle de 200 alignements de phrases choisies aléatoirement. Ils ont montré que seuls 50% des paires de phrases correspondent à une simplification. Afin de répondre au problème du bruit présent dans les alignements du corpus Wikipédia anglais, les auteurs ont introduit une nouvelle ressource : le corpus Newsela[9] (Xu et al., 2015). Dans ce travail, nous étudions les différentes transformations caractérisant le passage d'un niveau de simplification à l'autre sur un sous-ensemble de textes, en nous intéressant plus particulièrement aux transformations syntaxiques.

2.2 Le corpus Newsela

Newsela est un jeu de données pour la simplification de texte qui a comme avantage le fait de proposer différentes versions de simplification à partir d'un texte original (la plupart des corpus parallèles existants, mentionnés plus haut, sont des corpus parallèles des textes originaux et simplifiés). Il s'agit d'un corpus d'articles de presse réécrits par des éditeurs professionnels. Le public cible considéré était les enfants de différents niveaux scolaires. Le corpus contient 1.130 articles de presse, chacun d'eux réécrit 4 fois (tableau 1) pour les enfants de différents niveaux, obtenant quatre versions de simplification (simp-1 est la moins simple et simp-4 la plus simple).

Texte original	**The athletic shoe and apparel maker** said Thursday it will provide free design resources to schools looking to shelve Native American mascots, nicknames, imagery or symbolism. **The German company** also pledged to provide financial support to ensure the cost of changing is not prohibitive.
Simp_1	**The athletic shoe and apparel maker** said Thursday it will provide free design resources to schools looking to give up their Native American mascots, nicknames, imagery or symbolism. **It** also pledged to provide financial help to ensure the cost of changing is not excessive.
Simp_2	**Adidas** officials said Thursday it will help the schools create new mascots, nicknames, images or symbols. **It** also promised to help cover the cost to make sure

7 https://www.wikipedia.org/
8 https://simple.wikipedia.org/
9 https://newsela.com/data/

	that the change is not too expensive.
Simp_3	On Thursday, **Adidas** offered to help high schools change their mascots. A mascot is usually a person or animal. It represents a group or school, and many people think it brings good luck. Most sports teams have a mascot. **The shoe and clothing company** will also help to pay for the change. New uniforms, mascots and signs can be expensive for schools.
Simp_4	**Adidas** said it will help schools make new mascots. A mascot can be a person or an animal. Most sports teams have a mascot. **Adidas** will help schools design new uniforms. It will also help them to design new logos. Logos are the pictures on uniforms or signs. It costs a great deal of money to change logos and mascots. Adidas will help schools pay for it.

TABLE 1. Exemple de phrases écrites à plusieurs niveaux de complexité de texte à partir de l'ensemble de données Newsela.

3 Méthodologie

Comme nous venons de le voir, le corpus Newsela est un corpus parallèle dédié à la simplification de textes, proposant pour un texte donné, quatre niveaux de simplification. Nous nous intéressons dans cet article aux différentes transformations caractérisant le passage d'un niveau de simplification à l'autre. Xu et ses collaborateurs (2015) ont effectué une analyse systématique de l'ensemble du corpus en se focalisant sur l'aspect lexical. Notre but est d'analyser ce corpus en nous focalisant sur les aspects syntaxiques, c'est-à-dire, étudier les changements syntaxiques qui ont été faits lors du passage d'un niveau de difficulté à un autre plus simple. Une analyse qualitative a été réalisée afin de cibler différents types d'opérations de simplification.

Le travail que nous décrivons dans cette proposition est basé sur un sous-ensemble de textes composés de 107 phrases tirées de 6 textes choisis d'une façon aléatoire du corpus Newsela, en nous intéressant plus particulièrement aux transformations syntaxiques. Nous avons aligné ces originaux avec les 4 niveaux de difficulté et nous avons repéré les opérations effectuées au niveau syntaxique.

3.1 Analyse quantitative du corpus Newsela

Le Tableau 1 récapitule le nombre de tokens et de phrases dans l'ensemble des 6 textes étudiés. Ce tableau montre le nombre total de phrases et de mots et la longueur moyenne de la phrase (en mots) des textes simplifiés originaux et les 4 niveaux de simplification : de l'original au Simp_1, du Simp_1 au Simp_2, du Simp_2 au Simp_3 et du Simp_3 au Simp_4.

Il y a une réduction importante au niveau de la longueur du texte en passant du niveau 2 au niveau 3 et une autre plus grande (27 %) en passant du niveau 3 au niveau 4, ce qui était attendu pour ce type de public (les apprenants d'une langue seconde). Dans ce type de corpus, des ajouts sont considérés comme utiles pour améliorer la lecture et la compréhension de textes mais privilégiant la suppression des informations supplémentaires et redondantes en gardant toujours des phrases courtes.

	Tokens		Phrases		Tokens/phrase	
	Total	Réduction	Total	Augmentation	Moyenne	Réduction
Textes originaux	4.866		107		45,5	
Simp_1	4.577	6 %	118	10 %	38,8	15 %
Simp_2	4.358	5 %	165	40 %	26,4	32 %
Simp_3	3.768	14 %	201	22 %	18,7	29 %
Simp_4	2.740	27 %	192	-4 %	14,3	24 %

TABLE 1 : Nombre de tokens et de phrases dans 6 textes du corpus Newsela : original → Simp_1, Simp_1 → Simp_2, Simp_2 → Simp_3 et Simp_3 → Simp_4

Le pourcentage d'augmentation du nombre de phrases est clairement important en passant du niveau 1 au niveau 2 (40 %). En comparant le nombre de phrases entre le texte original et sa version la plus simplifiée (Simp_4), le nombre augmente de 79 %. Ces différences s'expliquent essentiellement par les opérations de découpage des phrases longues appliquées. Le nombre de mots par phrase diminue en passant d'un niveau à un autre, ce qui revient à des phrases plus courtes qui maintiennent la structure SVO et qui présentent une seule idée par phrase.

Nous avons utilisé l'outil CollateX[10] (Dekker et Middell, 2011) pour repérer les transformations (découpage de phrases, suppression et substitution morpho-syntaxiques, insertion d'informations, réorganisation et fusion) effectuées dans les versions simplifiées par rapport au texte original (Original Simp_1, du Simp_1 au Simp_2, du Simp_2 au Simp_3 et du Simp_3 au Simp_4). CollateX est un outil utilisé dans les humanités numériques qui implémente des algorithmes d'alignement et fournit une visualisation statique pour les graphiques de variantes de texte. Dans cet outil, quatre étapes de base sont définies et appliquées dans l'ordre et/ou de manière itérative. La première est la tokenisation des textes numériques à comparer. La deuxième étape est l'alignement des tokens de différents textes et implication des opérations d'édition. La troisième étape est l'analyse de l'alignement calculé, les opérations d'édition étant désormais qualifiées (par exemple, suppression, ajout ou déplacement). La quatrième et dernière étape est la sortie ou visualisation des résultats. La figure 1 montre la fréquence des différentes opérations de transformation dans les 6 textes du corpus.

De ces résultats d'analyse, nous pouvons déduire les spécificités de chaque niveau de simplification :

- Le ***niveau 1*** privilégie la réorganisation de la phrase et substitution morpho-syntaxique : des clauses peuvent être échangées afin que la présentation de l'information soit plus lisible. De plus, lorsque des structures complexes ne sont pas supprimées, elles sont généralement déplacées pour faciliter la compréhension. Les opérations de remplacements et de réorganisations représentent respectivement 35 % et 32 % des transformations syntaxiques dans ce niveau de difficulté.

[10] https://collatex.net/demo/

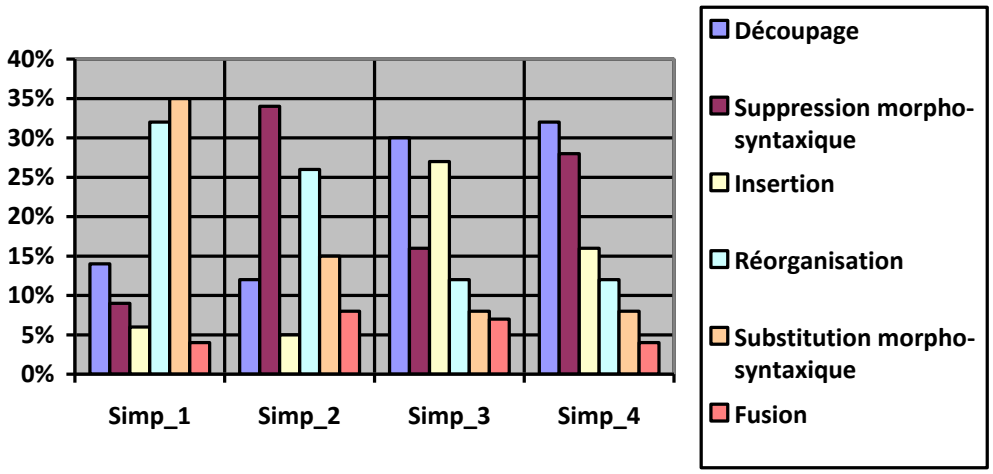

FIGURE 1 : Fréquence des opérations de transformation syntaxique dans les 6 textes du corpus Newsela

- Le ***niveau 2*** privilégie les suppressions morpho-syntaxiques (34 %) : les informations d'importance secondaire sont supprimées des phrases.

- Le ***niveau 3*** privilégie le découpage de phrases et les insertions : les auteurs choisissent de diviser des phrases trop longues (quand il y a coordination ou signe de ponctuation). Le découpage et les insertions représentent respectivement 30 % et 27% des transformations syntaxiques dans ce niveau de complexité.

- Le ***niveau 4*** privilégie les découpage et suppressions morpho-syntaxiques : les découpages des phrases complexes représentent 32 % des transformations et les suppressions 28%.

4 Analyse des opérations de simplification

Nous avons observé manuellement 107 phrases alignées du corpus dans ses 4 niveaux de difficultés afin d'en dégager les différents phénomènes intervenant dans la simplification de texte.

Les opérations de simplification présentées dans le corpus Newsela s'appliquent à plusieurs niveaux : lexical, morpho-syntaxique et discursif, ce qui reste classique dans le cadre de la simplification de textes (Gala et al., 2018). Dans cet article, nous présentons uniquement une caractérisation du niveau syntaxique. Sur la base de différentes classifications des opérations de simplification proposée dans la littérature (Brunato et al., 2015 ; Bott et Saggion, 2014 ; Coster et Kauchak, 2011 ; Caseli et al., 2009, Medero et Ostendorf, 2011 et Zhu et al., 2010, Gala et al. 2020b), nous avons identifié six classes principales d'opérations observées dans le corpus Newsela, à savoir : (4.1) des découpages, (4.2) des fusions, (4.3) des réorganisations, (4.4) des insertions (rajouts), (4.5) des suppressions et (4.6) des substitutions.

4.1 Découpage de phrases

C'est l'opération la plus fréquente en SAT, pour les applications à la fois humaines et automatiques. En règle générale, le découpage supprime les conjonctions, les deux points, les points-virgules, les énumérations et les appositives afin d'obtenir deux phrases indépendantes. Dans (1) et (2), nous donnons deux exemples de découpage. Nous pouvons observer qu'elles s'appliquent à partir du premier niveau de simplification.

(1) *Originale*. The advocacy group says about a dozen schools have dropped Native American mascots over the past two years **and** an additional 20 are considering a change.
Simpl_1. The advocacy group says about a dozen schools have dropped Native American mascots over the past two years. An additional 20 are considering a change.
Simpl_2. The advocacy group said about a dozen schools have dropped these mascots over the past two years. An additional 20 are considering a change.
Simpl_3. About a dozen of the schools decided to pick new mascots in the last two years. Another 20 are thinking about it.
Simpl_4. In the last two years, about 12 schools stopped using them. Another 20 are thinking about it.

(2) *Originale.* **The company, which** has its North American headquarters in Portland, Oregon, also said it will be a founding member of a coalition that addresses Native American mascots in sports.
Simpl_1. Adidas also said it will be a founding member of a coalition that addresses the problem of Native American mascots in sports**. The German company** has its North American headquarters in Portland, Oregon.
Simpl_2. Adidas also said it will be a founding member of a group that deals with the problem of Native American mascots in sports. The German company makes shoes and clothing. **Its** North American headquarters is in Portland, Oregon.
Simpl_3. Adidas said it will help start a new group. It will deal with the problem of Native American mascots in sports.
Simpl_4. Phrases Supprimées.

Les appositions ne sont pas supprimées, elles sont transformées en phrases indépendantes :

(3) *Originale***.** Eric Liedtke**, Adidas head of global brands,** traveled to the conference. He said sports must be inclusive.
Simpl_1. Eric Liedtke, **Adidas head of global brands**, said sports must be inclusive.
Simpl_2. Eric Liedtke **is** the Adidas head of global brands. He said that sports must include everyone.
Simpl_3. Eric Liedtke, **who works for Adidas**, attended the Tribal Nations conference. Sports must include everyone, he said.
Simpl_4. Eric Liedtke **works for Adidas**. Sports must include everyone, he said.

Les exemples 2 et 3 montrent que les opérations ne sont pas indépendantes les uns des autres, puisque la substitution d'une phrase entraîne un changement syntaxique.

4.2 Fusion de phrases

Cette opération est conçue comme l'inverse de la division, c'est l'opération par laquelle deux (ou plusieurs) phrases originales sont fusionnées en une phrase simplifiée unique. Une telle

transformation est moins fréquente que la division des phrases (< 8% dans tous les niveaux de simplification).

(4) *Originale*. Eric Liedtke, Adidas head of global brands, traveled to the conference. **He said sports must be inclusive.**
Simpl_1. Eric Liedtke, Adidas head of global brands, **said sports must be inclusive.**

4.3 Réorganisation

Lorsque certaines structures complexes ne sont pas supprimées, elles sont souvent déplacées ou modifiées dans le texte dans le but de maintenir une structure SVO (Gala et al., 2020b). Cette opération marque le changement de position de mots entre la phrase d'origine et son équivalent simplifié (ex. 5).

(5) *Originale*. **According to** the group Change the Mascot, there are about 2,000 schools nationwide that have Native American mascots.
Simpl_1. **According to** the group Change the Mascot, there are about 2,000 schools nationwide that have Native American mascots.
Simpl_2. About 2,000 schools nationwide have Native American mascots, **according to** the group Change the Mascot.
Simpl_3. Change the Mascot is a group that wants schools to stop using Native American mascots. About 2,000 American schools have them, **the group said**.
Simpl_4. Change the Mascot is a Native American group. **It wants** schools to drop Native American mascots. About 2,000 American schools have these mascots.

4.4 Insertion

Le processus de simplification peut entraîner une phrase plus longue, en raison de rajout de mots ou d'expressions qui fournissent des informations de clarification (ex. 6). Nous avons observé un seul type d'informations pour marquer les insertions : le rajout d'explications et de définitions. Ce procédé est approprié pour les corpus destinés à des apprenants normo-lecteurs ou à des adultes illettrés ; il ne l'est pas pour d'autres types de public cible, par exemple les enfants faibles-lecteurs ou dyslexiques (Rello, 2014).

(6) *Originale*. The NFL's Washington **Redskins** have resisted appeals by Native American and civil rights groups to change their name and mascot.
Simpl_1. The NFL's Washington **Redskins** have resisted appeals by Native American and civil rights groups to change the name and mascot.
Simpl_2. Native American and civil rights groups have asked NFL's Washington **Redskins** to change their name and mascot. The football team has refused.
Simpl_3. They have repeatedly asked the Washington Redskins football team to change its name and mascot. **Redskins is an old word for Native Americans, who feel that it is unkind**. The football team has refused to change its name.
Simpl_4. Americans have asked the Washington Redskins to change the team's name. **The Redskins is a football team. Redskins is a very old word for Native Americans. It is not a nice word.** The football team has refused again and again.

4.5 Suppression morpho-syntaxique

Les informations secondaires ou redondantes, généralement considérées comme supplémentaires au niveau syntaxique, ne sont pas incluses dans les textes simplifiés (ex. 7 à 10). Un texte devrait être simplifié en éliminant les informations redondantes. Les phrases simplifiées contiennent moins d'adverbes ou d'adjectifs que les phrases originales. Certains adverbes et adjectifs, entre autres, sont omis. Nous proposons six types d'informations qui peuvent être supprimées : les informations entre parenthèses, les exemples, les constructions appositives, certains modifieurs, quelques relatives ainsi que les expressions temporelles et locatives dans certains cas.

(7) *Originale*. Some colleges kept their nicknames by obtaining permission from tribes, **including the Florida State Seminoles and the University of Utah Utes.**
Simpl_1. Several colleges kept their nicknames by obtaining permission from tribes, **such as the Florida State Seminoles and the University of Utah Utes.**
Simpl_2. Some colleges kept their nicknames by getting permission from tribes. Two teams that received permission were the **Florida State Seminoles and the University of Utah Utes**.
Simpl_3. Some colleges were able to keep their names, though. They received permission from tribes.
Simpl_4. Some colleges kept their names, though. They asked for permission from Native American groups.

(8) *Originale*. […] Ray Halbritter applauded Adidas' move in a **joint statement**.
Simpl_1. […] Ray Halbritter applauded Adidas' announcement in a **joint statement**.
Simpl_2. […] Ray Halbritter applauded Adidas' announcement in a **statement**.

(9) *Originale*. **In 2005**, the NCAA warned schools that they would face sanctions if they didn't change Native American logos or nicknames.
Simpl_1. **In 2005**, the NCAA warned schools that they would face sanctions if they did not change Native American logos or nicknames.
Simpl_2. **In 2005**, the National Collegiate Athletic Association (NCAA) warned colleges that they would face penalties if they did not change Native American logos or nicknames.
Simpl_3. **In 2005**, the National Collegiate Athletic Association (NCAA) told colleges to stop using Native American mascots.
Simpl_4. It told the colleges to get new mascots. If not, the colleges could be punished.

(10) *Originale*. Adidas announced the initiative in conjunction with the White House Tribal Nations Conference on Thursday **in Washington**.
Simpl_1. Adidas announced the initiative in conjunction with the White House Tribal
Simpl_2. Nations Conference on Thursday **in Washington**, **D.C**.
Adidas announced the project as the White House Tribal Nations Conference met **in Washington, D.C.**, on Thursday.
Simpl_3. The White House Tribal Nations Conference took place on Thursday.

4.6 Substitution morpho-syntaxique

Nous avons observé des substitutions de nature morpho-syntaxique : transformer les phrases passives en phrases actives, privilégier les propositions positives à la place des propositions négatives, ainsi que les formes personnelles à la place des formes impersonnelles.

(11) *Originale*. The NFL's Washington Redskins have resisted appeals **by** Native American and civil rights groups to change their name and mascot.
Simpl_4. Native Americans have asked the Washington Redskins to change the team's name.

(12) *Originale*. "Today's announcement is a great way for us to offer up our resources to schools that want to do what's right — to administrators, teachers, students and athletes who want to make a difference in their lives and in their world," Liedtke said in a statement to The Associated Press. "Our intention is to help break down any barriers to change — change that can lead to a more respectful and inclusive environment for all American athletes."
Simpl_3. Liedtke said that many schools want to do what is right. Teachers and students want to make a difference in their lives and in the world. Adidas wants to make it easier for them, he said. He said the new school mascots will respect all American athletes.

En plus de ces variations linguistiques, les versions simplifiées de Newsela présentent des variations typographiques[11] notamment avec des variations des nombres. Ce procédé est assez courant pour alléger la charge cognitive pendant la lecture (spécialement pour les lecteurs en difficulté de lecture, voir Rello, 2014 ; Gala et al. 2020a) :

(13) *Originale***.** The advocacy group says about a **dozen** schools have dropped Native American mascots over the past two years and an additional 20 are considering a change.
Simpl_4. In the last two years, about **12** schools stopped using them. Another 20 are thinking about it.

5 Conclusion

La disponibilité de corpus parallèles monolingues est fondamentale pour à la recherche en simplification automatique du texte (SAT). Ces corpus constituent, malgré leur rareté et la difficulté de leur construction, les corpus les plus appropriés et adaptés pour l'étude de la simplification de SAT. La majorité des méthodes de simplification syntaxique reposent sur un ensemble de règles de transformation définies manuellement pour être appliquées aux phrases. Dans Newsela, l'existence de quatre versions correspondant à quatre niveaux de difficulté rend le corpus intéressant à étudier dans le but de comprendre les transformations et pouvoir les implémenter plus tard dans un système de SAT.

Dans cet article, nous avons présenté une étude de transformations syntaxiques qui nous permettent de décrire 6 textes originaux en anglais et leurs simplifications. Notre étude du corpus Newsela est à la fois qualitative et quantitative, sur les transformations appliquées lors du passage d'un niveau de difficulté à un autre. Nous nous sommes basées sur un sous-ensemble de textes restreint de Newsela. Dans le but de généraliser nos résultats, il sera intéressant de mener des expériences sur la totalité du corpus.

Par la suite, notre objectif sera d'appréhender la simplification automatique de textes selon une approche à base de représentations sémantiques, en utilisant le formalisme sémantique UCCA (*Universal Cognitive Conceptual Annotation* ; Abend et Rappoport, 2013 ; Sulem et al., 2018). Les

[11] Bouamor (2012) a défini les variations typographiques dans la catégorisation des paraphrases sous-phrastiques qui peut être aussi une classe définie pour la catégorisation de la simplification syntaxique.

informations sémantiques sont fondamentales d'où notre intérêt à déterminer automatiquement, au moyen d'un formalisme, quelles informations sont secondaires. Ce faisant elles pourront être supprimées et il sera alors possible de rendre les informations primordiales plus visibles aux lecteurs en difficulté (travaux en cours).

Remerciement

Je tiens à remercier Núria GALA (LPL) et Bernard ESPINASSE (LIS) pour leur aide précieuse et pour leurs contributions à la réalisation de ce travail.

Références bibliographiques

ABEND, O., & RAPPOPORT, A. (2013, August). Universal conceptual cognitive annotation (ucca). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 228-238).

ALUÍSIO, S. M., SPECIA, L., PARDO, T. A., MAZIERO, E. G., CASELI, H. M., & FORTES, R. P. (2008, September). A corpus analysis of simple account texts and the proposal of simplification strategies: first steps towards text simplification systems. In *Proceedings of the 26th annual ACM international conference on Design of communication* (pp. 15-22).

ALVA-MANCHEGO, F., BINGEL, J., PAETZOLD, G., SCARTON, C., & SPECIA, L. (2017, November). Learning how to simplify from explicit labeling of complex-simplified text pairs. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 295-305).

BOTT, S., & SAGGION, H. (2014). Text simplification resources for Spanish. *Language Resources and Evaluation*, *48*(1), 93-120.

BOUAMOR, H. (2012). *Étude de la paraphrase sous-phrastique en traitement automatique des langues* (Doctoral dissertation).

BROUWERS, L., BERNHARD, D., LIGOZAT, A. L., & FRANÇOIS, T. (2012, June). Simplification syntaxique de phrases pour le français (Syntactic Simplification for French Sentences) [in French]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN* (pp. 211-224).

BROUWERS, L., BERNHARD, D., LIGOZAT, A. L., & FRANÇOIS, T. (2014, April). Syntactic sentence simplification for French. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)* (pp. 47-56).

BRUNATO, D., DELL'ORLETTA, F., VENTURI, G., & MONTEMAGNI, S. (2015, June). Design and annotation of the first Italian corpus for text simplification. In *Proceedings of The 9th Linguistic Annotation Workshop* (pp. 31-41).

BRUNATO, D., CIMINO, A., DELL'ORLETTA, F., & VENTURI, G. (2016, November). Paccss-it: A parallel corpus of complex-simple sentences for automatic text simplification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 351-361).

BURSTEIN, J., SHORE, J., SABATINI, J., LEE, Y. W., & VENTURA, M. (2007, April). The automated text adaptation tool. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)* (pp. 3-4).

BURSTEIN, J. (2009, March). Opportunities for natural language processing research in education. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 6-27). Springer, Berlin, Heidelberg.

CANNING, Y., TAIT, J., ARCHIBALD, J., & CRAWLEY, R. (2000, September). Cohesive generation of syntactically simplified newspaper text. In *International Workshop on Text, Speech and Dialogue* (pp. 145-150). Springer, Berlin, Heidelberg.
CARROLL, J., MINNEN, G., CANNING, Y., DEVLIN, S., & TAIT, J. (1998, July). Practical simplification of English newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology* (pp. 7-10).
CASELI, H. M., PEREIRA, T. F., SPECIA, L., PARDO, T. A., GASPERIN, C., & ALUÍSIO, S. M. (2009). Building a Brazilian Portuguese parallel corpus of original and simplified texts. *Advances in Computational Linguistics, Research in Computer Science*, *41*, 59-70.
CHANDRASEKAR, R., DORAN, C., & SRINIVAS, B. (1996, August). Motivations and methods for text simplification. In *Proceedings of the 16th conference on Computational linguistics-Volume 2* (pp. 1041-1044). Association for Computational Linguistics.
COSTER, W., & KAUCHAK, D. (2011, June). Learning to simplify sentences using wikipedia. In *Proceedings of the workshop on monolingual text-to-text generation* (pp. 1-9). Association for Computational Linguistics.
CREUTZ, M. (2018). Open Subtitles Paraphrase Corpus for Six Languages. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA). *arXiv preprint arXiv:1809.06142.*
CROSSLEY, S. A., LOUWERSE, M. M., MCCARTHY, P. M., & MCNAMARA, D. S. (2007). A linguistic analysis of simplified and authentic texts. *The Modern Language Journal*, *91*(1), 15-30.
DEKKER, R. H. ET MIDDELL, G. (2011). Computer-supported collation with CollateX : Managing textual variance in an environment with varying requirements. *Supporting Digital Humanities*, (pp.17–18).
DEVLIN, SIOBHAN AND JOHN TAIT. (1998). The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic Databases* (pp. 161–173).
EVANS, R., ORASAN, C., & DORNESCU, I. (2014). An evaluation of syntactic simplification rules for people with autism. Association for Computational Linguistics.
FENG, L. (2008). Text simplification: A survey. *The City University of New York, Tech. Rep*.
GALA, N., FRANÇOIS, T., JAVOUREY-DREVET, L., & ZIEGLER, J. C. (2018). La simplification de textes, une aide à l'apprentissage de la lecture. *Langue française*, (3), 123-131.
GALA, N., TACK, A., JAVOUREY-DREVET, L., FRANÇOIS, T., & ZIEGLER, J. C. (2020a). Alector: A Parallel Corpus of Simplified French Texts with Alignments of Misreadings by Poor and Dyslexic Readers. In *Language Resources and Evaluation for Language Technologies (LREC)*.
GALA N., TODIRASCU, A., BERNHARD, D., WILKENS, R. ET MEYER, J.-P. (2020b) Transformations syntaxiques pour une aide à l'apprentissage de la lecture : typologie, adéquation et corpus adaptés. Actes du *Congrès Mondial de Linguistique Française* (CMLF 2020). Montpellier, France.
GASPERIN, C., SPECIA, L., PEREIRA, T., & ALUÍSIO, S. (2009). Learning when to simplify sentences for natural text simplification. In *Proceedings of ENIA*, 809-818.
HASLER, E., DE GISPERT, A., STAHLBERG, F., WAITE, A., & BYRNE, B. (2017). Source sentence simplification for statistical machine translation.
INUI, K., FUJITA, A., TAKAHASHI, T., IIDA, R., & IWAKURA, T. (2003, July). Text simplification for reading assistance: a project note. In *Proceedings of the second international workshop on Paraphrasing-Volume 16* (pp. 9-16). Association for Computational Linguistics.
KOPTIENT, A., CARDON, R. ET GRABAR, N. (2019) Simplification-induced transformations: typology and some characteristics. In *Proceedings of the 18th BioNLP Workshop and Shared Task, Association for Computational Linguistics*, Florence, Italy (pp. 309-318).
LISON, P. AND TIEDEMANN, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016).*

MAX, A. (2005). Simplification interactive pour la production de textes adaptés aux personnes souffrant de troubles de la compréhension. In *Proceedings of Traitement Automatique des Langues Naturelles (TALN)*.
MAX, A. (2006). Writing for language-impaired readers. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 567-570). Springer, Berlin, Heidelberg.
MEDERO, J., & OSTENDORF, M. (2011). Identifying targets for syntactic simplification. In *Speech and Language Technology in Education*.
NEWSELA. (2016). Newsela article corpus. https://newsela.com/data. Version : 2016-01-29.
PETERSEN, S. E., & OSTENDORF, M. (2007). Text simplification for language learners: a corpus analysis. In *Workshop on Speech and Language Technology in Education*.
RELLO, L., BAYARRI, C., GÒRRIZ, A., BAEZA-YATES, R., GUPTA, S., KANVINDE, G., ... & TOPAC, V. (2013, May). DysWebxia 2.0! More accessible text for people with dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility* (pp. 1-2).
RELLO, L. (2014). *DysWebxia: a text accessibility model for people with dyslexia* (Doctoral dissertation, Universitat Pompeu Fabra).
SAGGION, H. (2017). Automatic Text Simplification: Synthesis Lectures on Human Language Technologies, vol. 10 (1). California, Morgan & Claypool Publishers.
SIDDHARTHAN, A. (2006). Syntactic simplification and text cohesion. *Research on Language and Computation*, *4*(1), 77-109.
SILVEIRA, S. B., & BRANCO, A. (2012). Enhancing multi-document summaries with sentence simplification. In *Proceedings on the International Conference on Artificial Intelligence (ICAI)* (p. 1). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).
SPECIA, L. (2010, April). Translating from complex to simplified sentences. In *International Conference on Computational Processing of the Portuguese Language* (pp. 30-39). Springer, Berlin, Heidelberg.
SPECIA, L., JAUHAR, S. K., & MIHALCEA, R. (2012, June). Semeval-2012 task 1: English lexical simplification. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation* (pp. 347-355). Association for Computational Linguistics.
SULEM, E., ABEND, O., & RAPPOPORT, A. (2018). Simple and effective text simplification using semantic and neural methods. *arXiv preprint arXiv:1810.05104.*
VAJJALA, S., & MEURERS, D. (2014). Readability assessment for text simplification: From analysing documents to identifying sentential simplifications. *ITL-International Journal of Applied Linguistics*, *165*(2), 194-222.
VANDERWENDE, L., SUZUKI, H., BROCKETT, C., & NENKOVA, A. (2007). Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing & Management*, *43*(6), 1606-1618.
WOODSEND, K., & Lapata, M. (2011, July). Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 409-420). Association for Computational Linguistics.
XU, W., CALLISON-BURCH, C., & NAPOLES, C. (2015). Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, *3*, 283-297.
ZHU, Z., BERNHARD, D., & GUREVYCH, I. (2010, August). A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd international conference on computational linguistics* (pp. 1353-1361). Association for Computational Linguistics.