Typologie de transformations dans la simplification de textes

Anaïs Koptient, Natalia Grabar

Typologie de transformations dans la simplification de textes

Anaïs Koptient, et *Natalia* Grabar

CNRS, Univ. Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France

Résumé. L'objectif de la simplification automatique de textes est de transformer un texte technique ou difficile à comprendre en un document plus compréhensible. Le sens doit être préservée lors de cette transformation. La simplification automatique peut être effectuée à plusieurs niveaux (lexical, syntaxique, sémantique, ou encore stylistique) et repose sur des connaissances et ressources correspondantes (lexique, règles, …). Notre objectif consiste à proposer des méthodes et le matériel pour la création de règles de transformation acquis à partir d'un échantillon de paires de phrases parallèles différenciées par leur technicité. Nous proposons également une typologie de transformations et les quantifions. Nous travaillons avec des données en langue française liées au domaine médical, même si nous estimons que notre méthode peut s'appliquer à n'importe quelle langue et n'importe quel domaine de spécialité.

Abstract. Typology of transformations in the text simplification. The purpose of the automatic text simplification is to transform technical or difficult to understand texts into a more friendly version. The semantics must be preserved during this transformation. Automatic text simplification can be done at different levels (lexical, syntactic, semantic, stylistic...) and relies on the corresponding knowledge and resources (lexicon, rules...). Our objective is to propose methods and material for the creation of transformation rules from a small set of parallel sentences differentiated by their technicity. We also propose a typology of transformations and quantify them. We work with French-language data related to the medical domain, although we assume that the method can be exploited on texts in any language and from any domain.

1. Introduction

La simplification automatique de textes a pour objectif de créer une version simplifiée d'un texte donné. La simplification peut être effectuée aux niveaux lexicaux, syntaxiques ou sémantiques, mais aussi aux niveaux pragmatiques ou stylistiques. La simplification peut être utilisée dans deux contextes principaux : d'abord comme une aide proposée à des utilisateurs humains, ce qui garantit une meilleure accessibilité et une meilleure compréhension du contenu des documents [25, 21, 9, 1, 17], mais aussi en pré-traitement pour d'autres tâches et programmes de traitement automatique des langues (TAL), ce qui facilite le travail d'autres modules de TAL et permet de donner de meilleurs résultats [8, 27, 4, 26, 31, 3]. Nous pouvons donc remarquer que cette tâche peut jouer un rôle important. La simplification est différente de la vulgarisation dans le sens où la simplification a pour objectif de simplifier un texte, tandis que lors de la vulgarisation, la simplification de concepts et de notions techniques est visée. La simplification a donc une ampleur plus importante. Cependant, dans un domaine de spécialité, comme la médecine, la simplification porte souvent sur les termes techniques.

Trois types de méthodes sont exploités en simplification automatique de textes :

- *Les méthodes basées sur les connaissances et les règles.* Par exemple, l'utilisation de WordNet [19] peut donner des expressions équivalentes à des mots difficiles [2, 7] ou aider à choisir un synonyme en utilisant la fréquence [10, 11, 12] ou la longueur [2] des candidats. Une des limitations de ces méthodes est leur faible rappel (nombre de phrases correctement simplifiées par rapport au nombre de phrases qui auraient dû être simplifiées) [10] et la confusion entre les mots difficiles et les mots simples [24] ;
- *Les méthodes basées sur les probabilités distributionnelles.* Par exemples, les *word embeddings* [18, 22], ou plongements lexicaux, peuvent être utilisés pour obtenir un lexique de substitution. Les plongements lexicaux permettent de représenter chaque mot d'un dictionnaire ou du corpus par un vecteur, ce qui permet d'associer les mots ayant des contextes proches (et partageant des vecteurs similaires) ensemble. Lorsque les plongements lexicaux sont entraînés sur des données pertinentes (Wikipedia, Simple Wikipedia, PubMed Central…), ils peuvent contenir des équivalents plus simples qui peuvent être exploités lors de la simplification [14, 15]. Néanmoins, ce genre de méthodes demande un travail conséquent de filtrage pour ne garder que les meilleurs candidats. Ces méthodes présentent généralement une bonne couverture et, lorsque le filtrage est efficace, une bonne précision (nombre de phrases correctement simplifiées sur le nombre de phrases totalement simplifiées) ;
- *Les méthodes basées sur la traduction automatique.* Ces méthodes abordent la problématique comme s'il s'agissait de la traduction monolingue : le passage de la langue technique à la langue générale. Plusieurs travaux proposent ainsi d'exploiter ce type de méthodes sur les textes en anglais [36, 37, 32, 23, 33, 29, 30, 35, 20]. Ils utilisent des corpus de phrases parallèles alignées qui proviennent principalement des *Simple English Wikipedia* et *English Wikipedia* (SEW-EW). Globalement, ces méthodes semblent

maintenir un bon équilibre entre la qualité de la simplification, la couverture et la précision. Elles requièrent cependant de très gros volumes de données.

Presque tous les travaux existants en simplification automatique sont dédiés à la langue anglaise et très peu aux autres langues. Pourtant, quelles que soient les méthodes ou les langues, il est nécessaire de disposer de ressources pertinentes pour effectuer les transformations de simplification. Dans notre travail, nous proposons de telles ressources qui permettent de créer des méthodes dédiées et des données linguistiques appropriées pour les règles de transformation.

2. Données linguistiques

Nous exploitons le corpus CLEAR existant composé de textes comparables (http://natalia.grabar.free.fr/resources.php) qui se différencient par leur technicité : des textes techniques et les textes simplifiés correspondants. Le corpus est composé de documents issus de trois sources différentes : des notices de médicaments, des articles d'encyclopédies et des résumés de revues systématiques. Chacun de ces sous-corpus comporte deux versions du même texte : une version « technique » et une version « simple ». Ainsi, le sous-corpus des notices de médicaments est composé de documents techniques (notices de médicaments destinées aux praticiens) et simples (notices de médicaments destinées aux patients – celles que l'on trouve dans les boîtes de médicaments). Pour les articles d'encyclopédies, les documents techniques proviennent d'articles du domaine de la médecine de l'encyclopédie libre Wikipédia et les documents simples proviennent d'articles issus de l'encyclopédie libre Vikidia destinée aux enfants de 8-13 ans). Enfin, pour les résumés de revues systématiques, les documents techniques correspondent aux versions techniques des résumés et les documents simples aux versions simplifiée de ces résumés.Dans notre travail, nous utilisons les termes *simple* et *simplifié* de manière interchangeable. Néanmoins, un texte *simplifié* est le résultat d'un processus de simplification d'un texte technique, comme le sont les résumés simplifiés de revues systématiques, alors qu'un texte *simple* provient d'un texte qui a été rédigé indépendamment, comme dans le cas de notices de médicaments ou d'articles d'encyclopédies. Dans le corpus utilisé, la partie technique est composée de 2,8 millions d'occurrences et la partie simplifiée de 1,5 millions d'occurrences. Un échantillon de ce corpus a été manuellement traité et aligné, ce qui a donné 663 paires de phrases parallèles. Ces paires de phrases nous montrent deux principaux types de relations :

- *L'équivalence sémantique* : les deux phrases de la paire ont une signification presque identique :

1) *Les sondes gastriques sont couramment utilisées pour administrer des médicaments ou une alimentation entérale aux personnes ne pouvant plus avaler.*
2) *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.*

Avec l'équivalence sémantique, la simplification est principalement effectuée au niveau lexical, en utilisant typiquement des substitutions lexicales. La simplification peut également être effectuée en ajoutant des informations. Dans ce cas, des notions complexes à comprendre sont suivies par leurs explications, comme dans *le tractus gastro-intestinal (un tube permettant de digérer les aliments).* Ces deux procédés (substitution et ajout d'information) sont souvent utilisés conjointement ;

- *L'inclusion sémantique* : la signification d'une phrase est incluse dans la signification de l'autre phrase. L'inclusion est orientée : la phrase technique, tout comme la phrase simple, peut être incluante ou inclue. Dans l'exemple suivant, la phrase technique est inclue et informe en plus du nombre de participants et de la métrique d'évaluation :

3) *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.*
4) *Cependant, peu de données étaient disponibles concernant les sondes placées incorrectement et les complications possibles d'une sonde mal placée.*

Avec l'inclusion sémantique, la simplification est également effectuée au niveau syntaxique, comme le montre l'exemple ci-dessus. Généralement, ce sont les subordonnées, les insertions, certains adverbes et adjectifs et les informations qui se trouvent entre parenthèses qui sont supprimées, comme *(43 participants)* de l'exemple 3). L'inclusion sémantique concerne aussi les énumérations : les phrases techniques avec des conjonctions de coordination peuvent être segmentées en listes dans la version simplifiée. Néanmoins, il est toujours possible de trouver des énumérations séparées par des virgules dans les textes techniques ou simplifiés. Notons aussi que les transformations syntaxiques et lexicales sont souvent utilisées en même temps.

3. Méthodologie

La méthode utilisée pour annoter et préparer les données linguistiques pour la description des transformations observables lors de la simplification repose sur trois dimensions principales : (1) le contrôle de relations d'inclusion sémantique, lorsque les phrases sont fusionnées ou segmentées lors de la simplification (section 3.1.) ; (2) l'annotation sémantique des paires de phrases pour décrire plus précisément la sémantique des transformations (section 3.2.) ; (3) l'annotation et l'analyse syntaxique pour regrouper les informations sémantiques et syntaxiques (section 3.3.). Ces dimensions réunissent donc les niveaux lexical et sémantique (section 3.2) et syntaxique (sections 3.1 et 3.3.).

3.1. Fusion et segmentation de phrases

L'une des stratégies souvent appliquées lors de la simplification de textes consiste en la fusion ou la segmentation de phrases techniques [5]. Lors de la fusion de deux phrases techniques chacune d'elles est raccourcie, ce qui permet d'obtenir leur version simplifiée

compréhensible. Au contraire, lorsqu'une phrase technique contient plus d'une proposition (typiquement une proposition principale et une secondaire), elle peut être segmentée en deux phrases en transformant la proposition secondaire en proposition principale indépendante. Néanmoins, il ne faut pas toujours diviser une phrase en deux car, dans certains cas, la proposition principale peut ne plus être comprise sans sa proposition secondaire [6].

Dans notre corpus, les phrases qui ont été segmentées et fusionnées ont été retrouvées sur la base de leur proximité dans le corpus et de leurs alignements, comme dans les exemples suivants :

5) **Phrase technique :** *elle impose l'arrêt du traitement et contre-indique toute nouvelle administration de clindamycine.*

6) **Phrase simple 1 :** *Prévenez votre médecin immédiatement car cela impose l'arrêt du traitement.*
Phrase simple 2 : *Cette réaction va contre-indiquer toute nouvelle administration de clindamycine.*

7) **Phrase technique 1 :** *abcès.*
Phrase technique 2 : *douleurs.*

8) **Phrase simple :** *Douleurs ou accumulation de pus au niveau du site d'injection.*

Comme nous le voyons, en cas de fusion ou de segmentation, les phrases techniques subissent également d'autres transformations et en particulier des substitutions lexicales.

3.2. Annotation sémantique

Les transformations sont annotées sémantiquement avec YAWAT (Yet Another Word Alignement Tool) [13]. YAWAT permet de visualiser et de manipuler des textes parallèles. Cet outil a été développé pour travailler avec des textes parallèles bilingues issus de la traduction [34]. Nous avons choisi d'utiliser cet outil avec des textes parallèles monolingues issus de la simplification car nous supposons qu'il s'agit du même type de transformations. YAWAT présente les deux phrases parallèles et alignées côte à côte. L'annotateur peut ensuite aligner les mots/groupes de mots qui correspondent en utilisant la matrice des deux phrases (voir figure 1) et leur assigner le type de transformation auquel ces mots/groupes de mots correspondent.

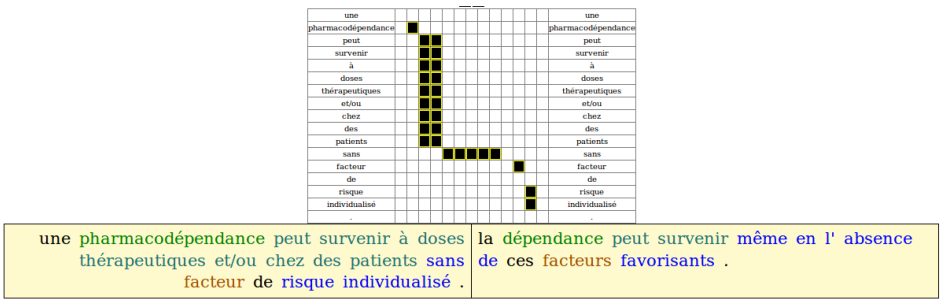

Fig. 1. La matrice de YAWAT

Le nombre de carrés qui se présentent verticalement correspond au nombre de mots dans la phrase de gauche (c'est-à-dire, la phrase technique), et le nombre de carrés qui se présentent horizontalement correspond au nombre de mots dans la phrase de droite (c'est-à-dire la phrase simple). Ensuite, pour apparier les mots/groupes de mots transformés lors de la simplification, l'annotateur clique sur le ou les carrés correspondant. Les types de transformation apparaissent et permettent de décrire plus précisément leur nature sémantique. Nous avons établi un ensemble de transformations sur la base de travaux existants [6] et d'observations sur notre corpus. Nous utilisons un ensemble de 25 transformations, comme présenté dans la figure 2.

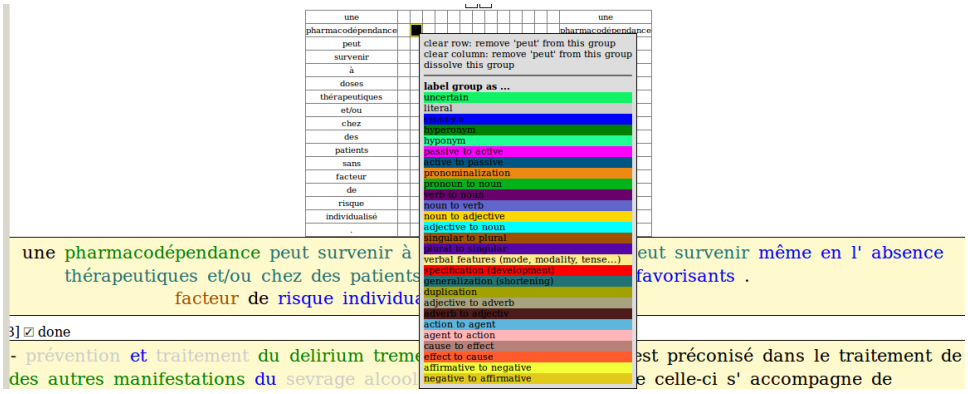

Fig. 2. Ensemble des types de transformations

Nous décrivons ces transformations ci-dessous :

- *literal* est la valeur par défaut ; elle est utilisée lorsque les mots des deux phrases sont identiques ;
- *synonym* : substitution d'un mot technique par son synonyme (*céphalées* > *mots de tête) ;*

- *hyperonym* : le terme technique est remplacé par son hyperonyme (*clindamicine* > *médicament)* ;
- *hyponym* : le terme technique est remplacé par un hyponyme (*benzodiazépines* > *bromazepam)* ;
- *p2a* (et *a2p*) : le verbe à la voix passive dans la phrase technique est remplacé par un verbe à la voix active (*ne dois jamais être utilisé* > *ne prenez jamais)* et inversement (*n'a aucun* > *n'est pas attendu*) ;
- *pronominalization* : le terme dans la phrase technique est remplacé par un pronom (*l'antibioprophylaxie* > *elle*) ;
- *p2n* : le pronom dans la phrase technique est remplacé par son référent (*elles* > *ce médicament*) ;
- *v2n* (et *n2v*) : le verbe dans la phrase technique est remplacé par un nom (*conduire* > *conduite*) et inversement (*l'arrêt du traitement* > *arrêter brutalement*) ;
- *n2a* (et *a2n*) : le nom dans la phrase technique est remplacé par un adjectif (*allergies* > *allergiques*) et inversement (*cardiaque* > *du cœur*) ;
- *s2p* (et *p2s*) : une forme au singulier dans le texte technique est remplacée par une forme au pluriel (*de tout antibiotique* > *d'antibiotiques*) et inversement (*les enfants* > *l'enfant*) ;
- *specification* : on ajoute des informations spécifiques au terme technique dans la version simple (*bêta-lactamines* > *bêta-lactamines (pénincilline, céphalosporine)*) ;
- *generalization* : suppression d'informations dans la version simple (*arrêt du traitement et contre-indique toute nouvelle administration du clindamycine* > *arrêt du traitement*) ;
- *duplication* : le terme technique est répété plusieurs fois dans la version simple ;
- *adj2adv* (et *adv2adj*) : un adjectif dans la phrase technique est remplacé par un adverbe (*récente* > *récemment*) et inversement (*tardif* > *tard*) ;
- *agt2act* (et *act2agt*) : l'agent dans la phrase technique est remplacé par l'action (*conducteurs* > *conduite*) et inversement (*conduite* > *conducteurs*) ;
- *cau2eff* (et *eff2cau*) : la cause dans la phrase technique est remplacée par son effet (*prescrits* > *utilisés*) et inversement (*dans le traitement* > *chez les patients atteints*) ;
- aff2neg (et neg2aff) : une forme affirmative est remplacée par une forme négative (*présentant une absence complète* > *n'avez aucune*) et inversement (*ne pas* > *éviter*).

Comme il est possible que certains segments puissent être annotés avec plusieurs étiquettes, nous avons établi la priorité d'étiquettes. Ainsi, les étiquettes qui décrivent un changement de nature du mot (comme *a2n*), et sont donc plus précises, sont prioritaires sur les étiquettes qui décrivent les synonymes (*synonym*).

Cet aspect de l'annotation est ainsi focalisé sur les aspects lexical et sémantique de la simplification. Yawat est en effet conçu comme un outil d'annotation lexicale : il ne nous est donc pas possible de quantifier les aspects syntaxiques (analyse syntaxique et division/fusion de phrases) en même temps.

3.3. Analyse syntaxique

L'analyse syntaxique a pour objectif d'annoter linguistiquement les phrases parallèles et de marquer les groupes syntaxiques qu'elles comportent. L'annotation est effectuée à l'aide de Cordial [16] qui fournit l'étiquetage morpho-syntaxique, la lemmatisation et l'analyse syntaxique en constituants et en dépendances. Le tableau 1 montre un exemple d'annotation et d'analyse de Cordial de la phrase *dalacine n'a aucun effet ou qu'un effet négligeable sur l'aptitude à conduire et à utiliser des machines*.

Tableau 1. Exemple d'annotation syntaxique de Cordial (position du mot dans la phrase, forme du mot, étiquette morpho-syntaxique et numéro du groupe syntaxique).

Numéro	Forme	Nature	Groupe Syntaxique
1	dalacine	NCI	1
2	n'	ADV	3
3	a	VINDP3S	3
4	aucun	ADJIND	5
5	effet	NCMS	5
6	ou	COO	-
7	qu'	ADV	3
8	un	DETIMS	9
9	effet	NCMS	9
10	négligeable	ADJSIG	9
11	sur	PREP	13
12	l'	DETDFS	13
13	aptitude	NCDS	13
14	à	PREP	15
15	conduire	VINF	15
16	des	DETDPIG	17
17	véhicules	NCMP	17

18	et	COO	-
19	à	PREP	20
20	utiliser	VINF	20
21	des	DETDPIG	22
22	machines	NCFP	22
23	.	PCTFORTE	-

Nous pouvons voir que la séquence *un effet négligeable* appartient au même groupe syntaxique (numéro 5), comme l'indique la colonne *groupe syntaxique*. De plus, la tête du groupe est *effet* car son numéro dans la phrase est le même que le numéro de ce groupe syntaxique (5). Comme c'est un nom (*NC*), il s'agit d'un groupe nominal.

4. Résultats et Discussion

4.1. Fusions et segmentations de phrase

Nous avons trouvé 51 cas dans lesquels au moins deux phrases techniques ont été fusionnées en une phrase simple et 16 cas où les phrases techniques ont été segmentées en plusieurs phrases simples. Dans un travail antérieur, il a été remarqué que la fusion de phrases est rare en simplification [5], alors que, dans notre corpus, nous avons observé le contraire : il y a beaucoup plus de cas de fusion de phrases techniques que de leur segmentation. Nous pouvons y voir plusieurs explications :

- Le travail existant [5] a été effectué à partir d'articles issus de Wikipédia et Vikidia. Vikidia est une encyclopédie en ligne destinée aux enfants de 8-13 ans. Elle a des règles de rédaction très strictes concernant la création d'articles. Dans notre travail, Wikipédia et Vikidia font également partie du corpus (la partie encyclopédie), mais les deux autres sous-corpus (notices de médicaments et résumés de revues systématiques) n'ont pas les mêmes principes de rédaction ;

- Dans la partie technique des notices de médicaments, les listes de maladies, effets secondaires ou encore fonctions sont présentées sous forme de listes à puces, tandis que dans la partie simple, ces listes sont coordonnées dans la même phrase :

9) **Segment technique :**

- hypomagnésémie.
- hypocalcémie.
- hyperglycémie.
- hyponatrémie.

10) **Segment simple :**

Diminution du sodium, du magnésium ou du calcium dans le sang, augmentation du sucre dans le sang.

- Dans les résumés de revues systématiques, les phrases techniques sont souvent raccourcies afin de garder l'information principale. Il est ainsi possible de fusionner ces phrases raccourcies. Il est également à noter qu'il n'y a pas de lignes directrices pour rédiger ces résumés et que chaque éditeur choisi ses propres principes.

4.2. Annotations sémantiques

La figure 3 présente la typologie des transformations de simplifications que nous proposons. Elle indique également des informations sur la fréquence et le pourcentage de chaque type de transformations.

Nous différencions plusieurs niveaux de transformations, dont certaines peuvent être présentes dans les typologies existantes [5, 6] (substitution lexicale, ajout lexical, suppression lexicale, substitution syntaxique, pronominalisation et utilisation de formes affirmatives et négatives). Le type de transformations le plus important (965 occurrences, 69 %) concerne les substitutions lexicales au sein desquelles nous avons différencié les substitutions avec changement sémantique (hyponymie et hyperonymie) et les substitutions sans changement de sens (synonymie et transformations morphologiques). Nous observons également que les ajouts lexicaux (spécifications) représentent 199 occurrences, soit 14 % des transformations, et que les suppressions lexicales (généralisations) représentent 132 occurrences, soit 9 % des transformations. Nous avons considéré que seuls les changement de voix active/passive correspondent aux substitutions syntaxiques. Ainsi, les changements singulier/pluriel et les changement de temps verbaux correspondent à des substitutions lexicales sans changement sémantique. Enfin, la pronominalisation et les changements affirmatifs/négatifs correspondent à d'autres petits types de transformations.

Si nous comparons notre typologie avec celle proposée dans un travail existant [5], nous pouvons noter plusieurs différences :

- nous avons différencié les synonymes des hyperonymes car ces deux types de transformations montrent des différences fondamentales (équivalence sémantique versus subsomption) et ont besoin de méthodes et ressources spécifiques,

- nous avons d'autres transformations syntaxiques et morphologiques, tandis que dans [5], seule la transformation de la voix passive à la voix active et inversement sont prises en compte,

- nous ne faisons pas la différence entre les transformations lexicales et sémantiques : dans notre travail, les transformations sémantiques font partie des transformations lexicales.

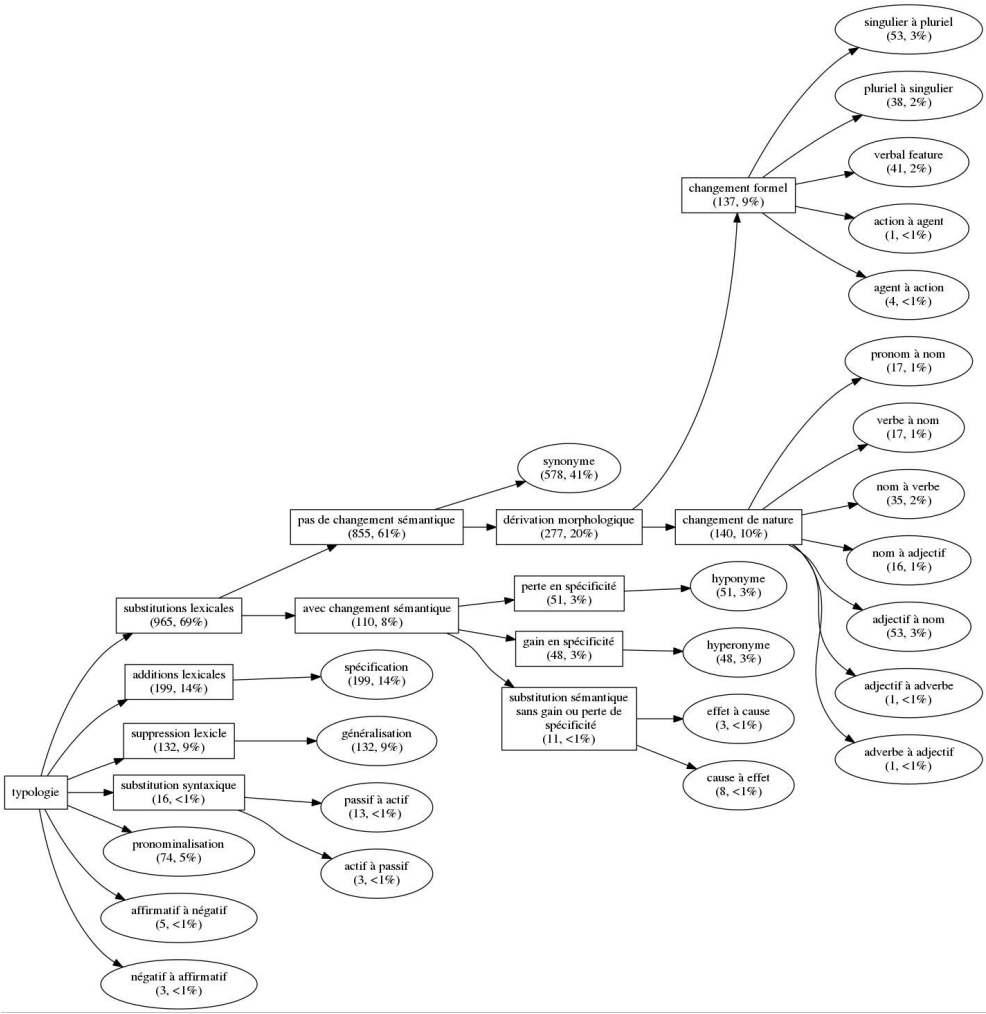

Fig. 3. Typologie des transformations

Si nous comparons notre typologie avec celle proposée dans [6], nous pouvons remarquer qu'ici les auteurs ont considéré qu'il existait plusieurs types d'insertions et de suppressions, selon la nature de mots en question (nom, verbe, etc.). Nous n'avons pas trouvé pertinent de faire cette différence dans notre travail car, dans la plupart des cas, les ajouts et suppressions correspondent à des groupes syntaxiques que nous annotons avec Cordial. De plus, nous considérons le changement de catégories grammaticales comme des substitutions lexicales dont nous décrivons précisément la nature syntaxiquement. Contrairement à [6], nous avons différencié les synonymes des hyperonymes et des hyponymes car, comme dit précédemment, nous considérons que ces trois types de transformations ont des différences fondamentales et ont besoin de méthodes et de ressources spécifiques.

Enfin, si nous comparons notre typologie avec celle proposée dans [28], nous pouvons remarquer que cette typologie est dédiée à la description de paraphrases en général et n'est pas spécifiquement destinée à la description de la simplification. La principale différence est que les auteurs ont séparé les substitutions lexicales des dérivations morphologiques. Nous avons pour notre part préféré les laisser ensemble car elles sont effectuées au niveau du mot et nous pouvons tout de même les différencier grâce à l'utilisation d'informations syntaxiques de Cordial et d'étiquettes spécifiques de la typologie.

Nous avons annoté en tout 1 394 occurrences de transformations, ce qui nous donne en moyenne 2,1 transformations par paire de phrases. Le tableau 2 indique les fréquences des types de transformations selon qu'ils sont présents dans des phrases qui ont été segmentées ou fusionnées, ou globalement dans le corpus (colonne *Tout le corpus*).

Tableau 2. Fréquences des transformations les plus fréquentes dans les phrases fusionnées et segmentées, ainsi que dans tout le corpus.

Type de transformation	**Phrases segmentées**	**Phrases fusionnées**	**Tout le corpus**
syno	24	112	578
hypero	1	10	48
hypo	0	13	51
pronom	9	2	74
v2n	1	1	17
n2v	2	5	35
n2a	2	0	16
a2n	2	17	53
s2p	0	6	53
p2s	5	3	38
vfea	0	4	41
specif	12	34	199
gener	14	10	132

Comme nous l'avons vu dans la figure 3, les types de transformations les plus fréquents sont les synonymes, la spécification et la généralisation. Ces types de transformations sont les plus fréquents dans le corpus en général et, par conséquent, dans les cas de fusion et segmentation de phrases. Il n'existe cependant pas de lien entre les types de transformations et la fusion ou segmentation de phrases.

À un niveau plus fin, nous pouvons observer que :

- les transformations adjectif>nom (53 occurrences) peuvent être nécessaires pour remplacer les adjectifs, souvent créés sur des bases d'origine savante (comme *card-* dans *cardiaque*) par le nom qui leur correspond, souvent construit sur une base alternative (comme *coeur*),
- les transformations terme>hyperonyme (48 occurrences) permettent d'utiliser un mot avec un sens plus large (*médicament*) à la place du terme technique (*clindamicyne*) et donc garantir une meilleure compréhension,
- les transformations terme>hyponyme (51 occurrences) permettent d'utiliser des exemples et des termes avec un sens plus fin (*frissons et tremblements*) à la place du terme au sens plus large (*syndrome pseudo-grippal*), ce qui peut également rendre la compréhension plus facile et contextuelle,
- les transformations nom>verbe (35 occurrences) permettent de rendre la phrase moins abstraite en remplaçant le concept par l'action et ainsi de rendre la phrase plus compréhensible.

Le fait qu'il y ait davantage de transformations par hyponymie que de transformations par hyperonymie peut sembler contre productif, mais cela peut être expliqué par les particularités du corpus. En effet, la partie simple du sous corpus de notices de médicaments contient souvent le nom exact du médicament, alors que la partie technique contient la classe thérapeutique de médicaments. Par exemple, il y a un cas dans le corpus où du côté technique il y a le mot *IEC* (*inhibiteur de l'enzyme de conversion)*, qui est un type de médicament, tandis que du côté simple, il y a *Moex* (le nom d'un médicament). Comme le Moex est une sorte d'IEC, alors *Moex* est un hyponyme de *IEC*.

4.3. Analyse syntaxique

L'analyse syntaxique nous a permis d'associer les informations sémantiques et les informations syntaxiques. Ainsi, dans beaucoup de cas (221), la nature des groupes syntaxiques reste la même. Dans plusieurs autres cas, le groupe syntaxique de départ est complété par d'autres groupes dans la phrase simple (GN —> GP GN ; GN —> GN GAdj). 531 transformations visent à modifier des groupes nominaux (*pustulose exanthématique aiguë généralisée* > *éruption sur la peau pouvant être accompagnée de fièvre*), 190 des groupes prépositionnels (*aux premier et deuxième trimestre de la grossesse* > *en début de grossesse*) et 174 des groupes verbaux (*peut entraîner l'apparition* > *en cas*). Cela montre que : (1) l'analyse syntaxique permet de donner d'importantes indications pour la détection des frontières des séquences à transformer ; (2) des mots et des expressions de différentes natures syntaxiques peuvent être transformés (noms, verbes, adjectifs…) ; (3) les noms et les groupes nominaux, qui correspondent à des concepts, occupent une place importante dans les transformations et sont souvent transformés lors de la simplification.

5. Conclusion et travaux futurs

Dans ce travail, nous avons proposé de travailler avec des phrases parallèles différenciées selon leur technicité : des phrases techniques et des phrases simples mises en parallèle. L'objectif principal du travail est de mettre en évidence les transformations impliquées dans la simplification. Ainsi, les phrases sont caractérisées sur trois dimensions : la fusion et la segmentation de phrases, les annotations sémantiques des transformations, et leurs annotations syntaxiques. Nous proposons également une typologie de transformations et nous quantifions ces transformations. Par exemple, notre travail montre que, parmi les transformations les plus fréquentes, nous pouvons trouver la synonymie, des spécifications (insertion d'information), la généralisation (suppression d'informations), la pronominalisation, la substitution d'adjectifs par les noms correspondants, et des substitutions entre les singuliers et les pluriels (et inversement). Cette typologie permettra de créer des règles de transformation qui lient les informations syntaxiques, lexicales et sémantiques. Ces règles seront ensuite utilisées pour simplifier des textes biomédicaux.

Références

1. Arya, D. J., Hiebert, E. H., and Pearson, P. D. (2011). The effects of syntactic and lexical complexity on the comprehension of elementary science texts. *Int Electronic Journal of Elementary Education*, 4(1):107–125.
2. Bautista, S., Gerva ́s, P., and Madrid, R. I. (2009). Feasibility analysis for semi-automatic conversion of text to improve readability. In *Int Conf on Inform and Comm Technology and Accessibility (ICTA)*, pages 33–40.
3. Beigman Klebanov, B., Knight, K., and Marcu, D. (2004). Text simplification for information-seeking applications. In Meersman, R. and Tari, Z., editors, *On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE*. Springer, LNCS vol 3290, Berlin, Heidelberg.
4. Blake, C., Kampov, J., Orphanides, A., West, D., and Lown, C. (2007). Query expansion, lexical simplification, and sentence selection strategies for multi-document summarization. In *DUC*.
5. Brouwers, L., Bernhard, D., Ligozat, A.-L., and Franc ̧ois, T. (2014). Syntactic sentence simplification for French. In *PITR workshop*, pages 47–56.
6. Brunato, D., Dell'Orletta, F., Venturi, G., and Montemagni, S. (2014). Defining an annotation scheme with a view to automatic text simplification. In *CLICIT*, pages 87–92.
7. Carroll, J., Minnen, G., Canning, Y., Devlin, S., and Tait, J. (1998). Practical simplification of English newspaper text to assist aphasic readers. In *AAAI98 Workshop on Integrating Artifficial Intelligence and Assistive Technology*, pages 7–10.
8. Chandrasekar, R. and Srinivas, B. (1997). Automatic induction of rules for text simplification. *Knowledge Based Systems*, 10(3):183–190.
9. Chen, P., Rochford, J., Kennedy, D. N., Djamasbi, S., Fay, P., and Scott, W. (2016). Automatic text simplification for people with intellectual disabilities. In *AIST*, pages 1–9.
10. De Belder, J. and Moens, M.-F. (2010). Text simplification for children. In *Workshop on Accessible Search Systems of SIGIR*, pages 1–8.
11. Devlin, S. and Tait, J. (1998). The use of psycholinguistic database in the simplification of text for aphasic readers. In *Linguistic Database*, pages 161– 173.
12. Drndarevic, B., Stajner, S., and Saggion, H. (2012). Reporting simply: A lexical simplification strategy for enhancing text accessibility. In *Easy to read on the web*, pages 1–6.
13. Germann, U. (2008). Yawat: Yet another word alignment tool. In ACL, editor, *ACL-08: HLT Demo Session*, pages 20–23, Columbus, USA.
14. Glavas, G. and Stajner, S. (2015). Simplifying lexical simplification: Do we need simplified corpora? In *ACL-COLING*, pages 63–68.

15. Kim, Y.-S., Hullman, J., Burgess, M., and Adar, E. (2016). SimpleScience: Lexical simplification of scientific terminology. In *EMNLP*, pages 1–6.
16. Laurent, D., Negre, S., and Ségue´la, P. (2009). L'analyseur syntaxique Cordial dans Passage. In *Traitement Automatique des Langues Naturelles (TALN)*.
17. Leroy, G., Kauchak, D., and Mouradi, O. (2013). A user-study measuring the effects of lexical simplification and coherence enhancement on perceived and actual text difficulty. *Int J Med Inform*, 82(8):717– 730.
18. Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *Workshop at ICLR*.
19. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. (1993). Introduction to wordnet: An on-line lexical database. Technical report, WordNet.
20. Nisioi, S., Stajner, S., Ponzetto, S. P., and Dinu, L. P. (2017). Exploring neural text simplification models. In *Ann Meeting of the Assoc for Comp Linguistics*, pages 85–91.
21. Paetzold, G. H. and Specia, L. (2016). Benchmarking lexical simplification systems. In *LREC*, pages 3074–3080.
22. Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *EMNLP 2014*, pages 1532–1543.
23. Sennrich, R., Haddow, B., and Birch, A. (2016). Improving neural machine translation models with monolingual data. In *Proc of the Ann Meeting of the Assoc for Comp Linguistics*, pages 86–96, Berlin, Germany.
24. Shardlow, M. (2014). A survey of automated text simplification. *Int J Advanced Computer Science and Applications*, 1:1–13.
25. Son, J. Y., Smith, L. B., and Goldstone, R. L. (2008). Simplicity and generalization: Short-cutting abstraction in children's object categorizations. *Cognition*, 108:626–638.
26. Stymne, S., Tiedemann, J., Hardmeier, C., and Nivre, J. (2013). Statistical machine translation with readability constraints. In *NODALIDA*, pages 1–12.
27. Vickrey, D. and Koller, D. (2008). Sentence simplification for semantic role labeling. In *Annual Meeting of the Association for Computational Linguistics-HLT*, pages 344–352.
28. Vila, M., Anto`nia Mart, M., and Rodr´ıguez, H. (2011). Paraphrase concept and typology. A linguistically based and computationally oriented approach. *Procesamiento del Lenguaje Natural*, 46:83–90.
29. Wang, T., Chen, P., Amaral, K., and Qiang, J. (2016a). An experimental study of LSTM encoder-decoder model for text simplification. In *IJCAI*, pages 1–7.
30. Wang, T., Chen, P., Rochford, J., and Qiang, J. (2016b). Text simplification using neural machine translation. In *Proc of the AAAI Conference on Artificial Intelligence (AAAI-16)*, pages 4270–4271.
31. Wei, C.-H., Leaman, R., and Lu, Z. (2014). SimConcept: A hybrid approach for simplifying composite named entities in biomedicine. In *BCB '14*, pages 138–146.
32. Wubben, S., van den Bosch, A., and Krahmer, E. (2012). Sentence simplification by monolingual machine translation. In *Annual Meeting of the Association for Computational Linguistics*, pages 1015– 1024.
33. Xu, W., Napoles, C., Pavlick, E., Chen, Q., and Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.
34. Yu, Q., Max, A., and Yvon, F. (2012). Revisiting sentence alignment algorithms for alignment visualization and evaluation. In *BUCC workshop*, pages 1–7.
35. Zhang, X. and Lapata, M. (2017). Sentence simplification with deep reinforcement learning. In ACL, editor, *Proc of the Conf on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark.
36. Zhao, S., Wang, H., and Liu, T. (2010). Leveraging multiple MT engines for paraphrase generation. In *COLING*, pages 1326–1334.
37. Zhu, Z., Bernhard, D., and Gurevych, I. (2010). A monolingual tree-based translation model for sentence simplification. In *COLING 2010*, pages 1353– 1361.