Simplification-induced transformations: typology and some characteristics

Anaïs Koptient[1] **Rémi Cardon**[2] **Natalia Grabar**[2]

1. {firstname.lastname}.etu@univ-lille.fr

2. {firstname.lastname}@univ-lille.fr

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

Abstract

The purpose of 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

The purpose of automatic text simplification is to provide a simplified version for a given text. Simplification can be done at lexical, syntactic, semantic but also pragmatic and stylistic levels. Simplification can be useful in two main contexts: as help provided to human readers, which guarantees better access and understanding of the content of documents (Son et al., 2008; Paetzold and Specia, 2016; Chen et al., 2016; Arya et al., 2011; Leroy et al., 2013), and as a pre-processing step for other NLP tasks and applications, which makes easier the work of other NLP modules and may improve the overall results (Chandrasekar and Srinivas, 1997; Vickrey and Koller, 2008; Blake et al., 2007; Stymne et al., 2013; Wei et al., 2014; Beigman Klebanov et al., 2004). We can see that potentially this task may play an important role.

Three main types of methods are currently exploited in text simplification:

- *Methods based on knowledge and rules*. For instance, the use of WordNet (Miller et al., 1993) may provide equivalent expressions for difficult words (Carroll et al., 1998; Bautista et al., 2009), or help with the choice of synonyms using their frequency (Devlin and Tait, 1998; De Belder and Moens, 2010; Drndarevic et al., 2012) or their length (Bautista et al., 2009). One limitation of such methods is their weak recall performance (De Belder and Moens, 2010) and confusion between difficult and simple words (Shardlow, 2014);

- *Methods based on distribution probabilities*, like word embeddings (Mikolov et al., 2013; Pennington et al., 2014), are used to acquire a lexicon and substitution rules for simplification. When trained on relevant data (Wikipedia, Simple Wikipedia, PubMed Central...), word embeddings can contain simpler equivalents, that can be exploited to perform the simplification (Glavas and Stajner, 2015; Kim et al., 2016). Nonetheless, such methods require consequent filtering to keep only the best candidates. Those methods generally provide good coverage and, when the filtering is efficient, good precision;

- *Methods issued from machine translation* tackle the problem as translation from technical to simple text. A growing number of works propose to exploit this type of method to English texts (Zhao et al., 2010; Zhu et al., 2010; Wubben et al., 2012; Sennrich et al., 2016; Xu et al., 2016; Wang et al., 2016a,b; Zhang and Lapata, 2017; Nisioi et al., 2017). They exploit corpora made of parallel and aligned sentences, that mainly derive from the Simple English Wikipedia - English Wikipedia corpus (SEW-EW). Globally, those methods seem to maintain a balance between the quality of the simplification, good coverage and precision.

Almost all the existing works address text simplification in English, while other languages are poorly described. Yet, whatever the method and language it is necessary to have available suitable resources for making the transformations required by the task. This work is intended as a basis to design a method and to use it for preparing linguistic data for the creation of transformation rules.

2 Linguistic Data

We exploit an existing corpus with comparable documents[1] differentiated by their technicity: technical documents and the corresponding simplified documents. The corpus is composed of documents from three sources: information on drugs, encyclopedia articles and abstracts from systematic reviews. We use *simple* and *simplified* interchangeably in our work. Yet, a *simplified* document is the result of the simplification process of a technical document, like the simplified abstracts from systematic reviews; while a *simple* document is issued from an independently written simple document, like drug information and encyclopedia articles. In the used corpus, the technical part contains over 2.8M occurrences, and the simplified part contains over 1.5M occurrences. A subset of this corpus has been manually aligned at the level of sentences, which provides 663 pairs of parallel sentences exploited in our work. These pairs of sentences show two types of relations:

- *Semantic equivalence:* two sentences of a pair have the same or very close meaning:
  - les sondes gastriques sont couramment utilisées pour administrer des médicaments ou une alimentation entérale aux personnes ne pouvant plus avaler (feeding tubes are often used to administer medicine or enteral nutrition to people who cannot swallow)
  - 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 (feeding tubes are often used to administer medicine and food directly into the gastrointestinal tract (a tube that allows to digest food))

  With the semantic equivalence, simplification is mainly performed at lexical level, typically using lexical substitutions. Simplification can also be done by adding information and, in this case, complex notions are followed by their explanations, like *le tractus gastro-intestinal (un tube permettant de digérer les aliments)*. Often, those two processes (substitution and addition of information) are applied jointly;

- *Semantic inclusion:* the meaning of one sentence is included in the meaning of the other sentence. The inclusion is oriented: the technical sentence as well as the simplified sentence can be inclusive or included. In this example, the technical sentence is inclusive and informs in addition on the number of participants and the evaluation metric:
  - 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 (only a few data (43 participants) were available concerning the detection of a bad placement (specificity), due to the weak incidence of bad placements)
  - cependant, peu de données étaient disponibles concernant les sondes placées incorrectement et les complications possibles d'une sonde mal placée (however, only a few data were available concerning the badly placed probes and the potential complications of a bad probe placement)
  With inclusion, simplification is also peformed at the syntactic level, as the example above illustrates. Typically, subordinate and inserted clauses, information between brackets, some adjectives of adverbs are deleted during the simplification, like the information between brackets (*43 participants* and *la spécificité*). Semantic inclusion also involves enumerations: technical sentences with coordination can be segmented into lists with separate items in the simplified versions. Yet, enumerations with comma-separated items can be found in either technical and simplified documents. We should also point out that syntactic and lexical transformations often occur together.

[1] http://natalia.grabar.free.fr/resources.php

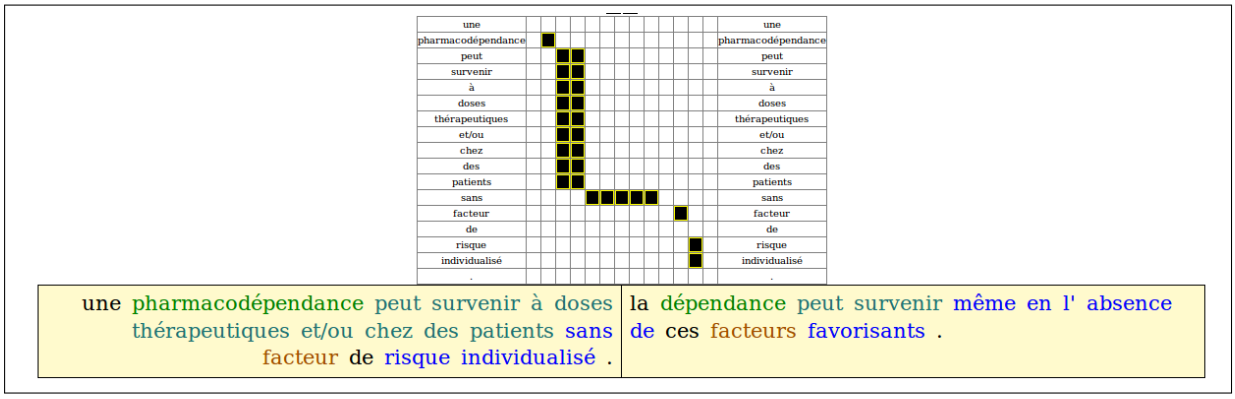

Figure 1: Matrix-based alignment of words within YAWAT

3 Methods

The methods for annotating and preparing the linguistic data for the description of simplification-induced transformations rely on three main dimensions: (1) control of the semantic inclusion relations, when sentences are split or merged during the simplification (Section 3.1); (2) semantic annotation of pairs of sentences to describe more precisely the transformations (Section 3.2); and (3) syntactic tagging and analysis for joining the semantic and syntactic information (Section 3.3).

3.1 Merging and Splitting of Sentences

One typical strategy applied during text simplification consists in merging or splitting the technical sentences when creating simple sentences (Brouwers et al., 2014). When merged, the technical sentences become shorted, which allows their merging into one sentence which yet remains readable in the simplified version. On contrary, when a given technical sentence contains more than one clause, like one main and one secondary, it can be split into two sentences by transforming the secondary clause into the main clause of another sentence. Sometimes, the splitting should be blocked because it can make the understanding of the main clause more difficult (Brunato et al., 2014).

In our corpus, merged and split sentences are detected using their proximity in the corpus and multiple alignments, like in these examples:

T_1 *elle impose l' arrêt du traitement et contre-indique toute nouvelle administration de clindamycine.* *(It forces the stopping of the cure and it is contra-indicated to administer once again the clindamycin.)*

S_1 *prévenez votre médecin immédiatement car cela impose l' arrêt du traitement.* *(tell it to your doctor immediately for it forces the stopping of the cure.)*
cette réaction va contre-indiquer toute nouvelle administration de clindamycine. *(this response contra-indicates any new administration of the clindamycin.)*

T_2 *abcès.* *(abscess)*
douleurs. *(pain)*

S_2 *douleurs ou accumulation de pus au niveau du site d' injection* *(pain or accumulation of pus at the injection site)*

Note that in the case of merging, the complex sentences when they are merged get also other simplifications, such as synonymy for instance.

3.2 Semantic Annotation

The simplification-induced transformations are annotated semantically using YAWAT (Yet Another Word Alignment Tool) (Germann, 2008). YAWAT permits to visualize and manipulate parallel texts. The tool was designed for working with parallel bilingual texts related to mutual translations (Yu et al., 2012). We propose to exploit it with monolingual parallel texts related to simplification. YAWAT displays the two parallel and aligned sentences side by side. The annotator can then align the words using the matrix (Figure 1), and to assign the type of transformation to each pair of text segments considered. The number of squares displayed vertically correspond to the number of words that are counted in the sentence on the left (that is, the technical sentence). The number of squares displayed horizontally correspond to the number of words that are counted in the sentence on the right (that is, the simple

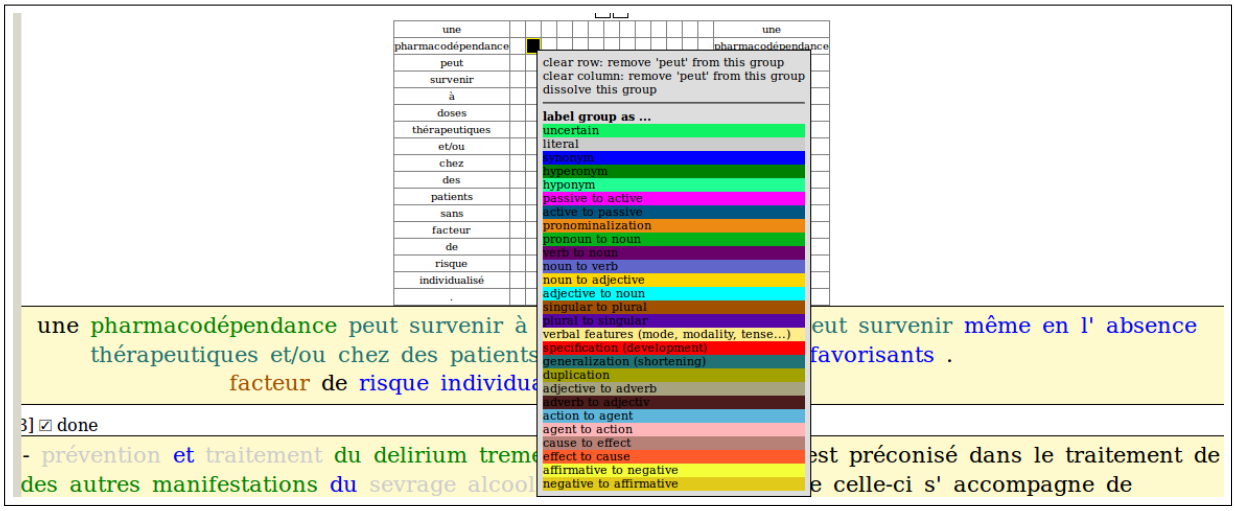

Figure 2: Annotation scheme within YAWAT

sentence). Then, in order to match word/group of words that correspond to a simplification phenomenon, the user has to click on the square that corresponds in both sentences. The fact that the text in the two columns are the same permits the user to click on the right square easily. The transformation types permits to describe more precisely their semantic nature. We defined a set of transformation types using previous similar work and observations on our corpus (Brunato et al., 2014). The proposed typology contains up to 25 transformations (Figure 2):

- *literal* is the default value which is kept when the words are identical in both sentences,
- *synonym*: substitution of technical term by its synonym {*effets négligeables* ; *effets délétères*} ({*insignificant effects* ; *deleterious effects*}),
- *hyperonym*: technical term is replaced by its hyperonym {*clindamicine* ; *médicament*} ({*clindamycin* ; *drug*}),
- *hyponym*: technical term is replaced by its hyponym {*benzodiazépines* ; *bromazepam*} ({*benzodiazepine* ; *bromazepam*})
- *p2a* (and *a2p*): verb in passive voice in technical sentence is replaced by its active voice {*ne doit jamais être utilisé* ; *ne prenez jamais*} ({*should never be used* ; *never use*}), and the contrary {*n'a aucun* ; *n'est pas attendu*} ({*does not have* ; *is not expected*})
- *pronominalization*: substitution by pronouns {*l'antibioprophylaxie* ; *elle*} ({*the antibiotic prophylaxis* ; *it*}),
- *p2n*: substitution of pronoun by its reference {*elles* ; *ce médicament*} ({*they* ; *this drug*}),
- *v2n* (and *n2v*): substitution of verbs by nouns {*conduire* ; *conduite*} ({*to drive* ; *driving*}), and the contrary {*l'arrêt du traitement* ; *arrêter brutalement*} ({*cessation of treatment* ; *stop at once*})
- *n2a* (and *a2n*): substitution of nouns by their adjectives {*allergies* ; *allergiques*} ({*allergies* ; *allergic*}), and the contrary {*cardiaque* ; *du coeur*} ({*cardiac* ; *of the heart*}),
- *s2p* (and *p2s*): substitution of singular forms by plural forms {*de tout antibiotique* ; *d'antibiotiques*} ({*of any antibiotic* ; *of antibiotics*}), and the contrary {*les enfants* ; *l'enfant*} ({*the children* ; *the child*}),
- *specification*: adding explanation to technical term {*bêta-lactamines* ; *bêta-lactamines (pénicilline, céphalosporine)*} ({*beta-lactam* ; *beta-lactam (penicillin, cephalosporins)*}). The difference with *synonymy* is that, instead of substitution, the technical term remains and its explanation (definition, examples) is added,
- *generalization*: removal of some information {*arrêt du traitement et contre indique toute nouvelle administration du clindamycine* ; *arrêt du traitement*} ({*cessation of treatment*

and contra-indicate any new administration of clindamycine ; *cessation of treatment*}),

- *duplication*: two or more occurrences of a given term in simple sentence,
- *adj2adv* (and *adv2adj*): substitution of adjectives by adverbs {*récente* ; *récemment*} ({*recent* ; *recently*}), and the contrary {*tardif* ; *tard*} ({*late* ; *late*}),
- *agt2act* (and *act2agt*): substitution of agent by the action {*conducteurs* ; *conduite*} ({*drivers* ; *driving*}), and the contrary {*conduite* ; *conducteurs*} ({*driving* ; *drivers*}),
- *cau2eff* (and *eff2cau*): substitution of cause by its effect {*prescrits* ; *utilisés*} ({*prescribed* ; *used*}), and the contrary {*dans le traitement* ; *chez les patients atteints*} ({*during the treatment* ; *in affected patients*}),
- *aff2neg* (and *neg2aff*): substitution of affirmative form of information by negative form with the same meaning {*présentant une absence complète* ; *n'avez aucune*} ({*indicating total absence* ; *you have no*}), and the contrary {*ne pas* ; *éviter*} ({*should not* ; *avoid*}).

Since it is common that some sequences can be tagged with several concurrent tags, we defined the priority rules, such as *a2n* > *synonym*, like in {*cardiaque* ; *du coeur*} ({*cardiac* ; *of the heart*}), because it describes the transformation more precisely. Since it is common that some sequences can be tagged with several concurrent tags, we prioritized part-of-speech related tags over synonymy because it is more precize, like in {*cardiaque* ; *du coeur*} ({*cardiac* ; *of the heart*}). Similarly, pronominalization is prioritized over verbal features, and also all the lexical transformations over syntactic transformations.

3.3 Syntactic Analysis

Syntactic analysis permits to linguistically annotate the parallel sentences and to mark within them the syntactic groups. Syntactic processing is done with Cordial (Laurent et al., 2009), which performs tokenization, POS-tagging, lemmatisation and syntactic analysis into constituents. In Table 1, we provide an example of Cordial tagging and analysis for the sentence *dalacine n'a aucun effet ou qu'un effet négligeable sur l'aptitude à conduire des véhicules et à utiliser des machines.*

nb.	*form*	*POS tag*	*synt. group*
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	NCFS	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	-

Table 1: Example of syntactic annotation by Cordial (word position in the sentence, word form, POS tag, and syntactic group)

(dalacine has no effect or the effect is insignificant on the capacity to drive vehicles and to use machines.) We can see that the sequence *un effet négligeble (insignificant effect)* belongs to the same syntactic group, stated in column *synt. group*. Besides, the syntactic head is *effet (effect)*, which has the same number as the syntactic groupe (9) and, being common noun (*NC*), it characterizes this group as nominal phrase.

4 Results and Discussion

4.1 Merging and Splitting of Sentences

We counted 51 cases in which two of more technical sentences are merged into one simple sentence, and 16 cases in which technical sentences are split into two or more simple sentences. In a previous work, it was noticed that the merging of sentences during the simplification is rare (Brouwers et al., 2014). Yet, in our corpus, we observe the contrary: much more technical sentences are merged than split. We can see several explanations:

- The cited work (Brouwers et al., 2014) is

done on articles from Wikipedia and Vikidia. Vikidia is designed for 8-13 year old children and relies on strong guidelines when creating the articles. One of the rules is to use short and clear sentences. In our work, Wikipedia and Vikidia correspond to the encycopedia part of the corpus. The two other subcorpora (drug leaflets and scientific abstracts) do not respect same writing principles.

- Drug leaflets frequently use coordinations with disorders, known adverse effects, functions, etc. Often, they are presented as itemized lists in technical documents, while in simplified documents then occur within coordinated sentences.

- In abstracts of systematic reviews, technical sentences are often shortened during their simplification in order to keep the main information. Then, possibly as consequence of it, the sentences may be merged. Notice also that there is no clear guidelines when writing plain-language abstracts and that each editor may apply its own principles.

4.2 Semantic Annotation

In Figure 3, we present the typology of the simplification-induced transformations. The Figure also contains information on prevalence of each transformation in terms of its frequence and percentage. We distinguish several high-level transformations, which may also be present in the existing typologies (Brunato et al., 2014; Brouwers et al., 2014): lexical substitution, lexical addition, lexical deletion, syntactic substitution, pronominalization and use of affirmative and negated forms. The biggest set of transformations (965 occurrences, 69%) is related to lexical substitutions, within which we differentiate substitutions with semantic shift (hyponymy and hyperonymy) and without semantic shift (synonymy and morphological transformation). We subsequently have lexical additions or specifications (199 occurrences, 14%), when explanations are added to technical terms in simplifed sentences, and lexical deletions or generalizations (132 occurrences, 9%), when some information is shortened and removed during the simplification. Then we consider that the only pure syntactic substitutions correpospond to active and passive voices of verbs. Hence, singular/plural and other verbal features belong to lexical substitutions without semantic shift. Pronominalization, and use of positive and negative equivalent expressions correspond to distinct small types of transformations.

By comparison with the typology from (Brouwers et al., 2014), we separated synonymy from hyperonymy because they have fundamental differences (semantic equivalence or subsumption) and require specific methods and resources. We differentiate several syntactic and morphological transformations, while in the citetd work, only the passive/active transformation is considered. Another difference is that we do not differentiate betweel lexical and semantic transformations: semantics becomes a feature of lexical substitutions.

By comparison with the typology from (Brunato et al., 2014), the authors differentiate several types of word insertion and deletion, according to the syntactic nature of these words (verb, noun...). We do not make this differentiation because, in most cases, insertions and deletions apply to syntactic clauses. Besides, we considered the shift of grammatical categories as lexical substitution, which we describe with detail according to the POS categories. Unlike in the cited work, we consider separately hyperonymy, hyponymy and synonymy, because they have fundamental differences and require specific methods and resources.

Finally, by comparison with the typology from (Vila et al., 2011), which is dedicated to the general description of paraphrases and does not specifically aim transformations due to the simplification, we notice several similarities. The main difference is that the authors separated lexical substitutions and morphological derivations, which we keep together because they all apply at the word level. Yet, we can differentiate them through the use of syntactic infomartion.

On the whole, we count 1,394 transformations, which gives 2.1 transformations per pair of sentences on average. In Table 2, we indicate the frequency of the most frequent types of transformations according to whether they occur in split or merged sentences, or generally in the corpus (the *total* column). As in Figure 3, the most frequent transformations are related to the use of synonyms, and the specification and generalization of contents. These types are frequent in the whole corpus and, by consequence, in merged and split sentences. There is no real association between

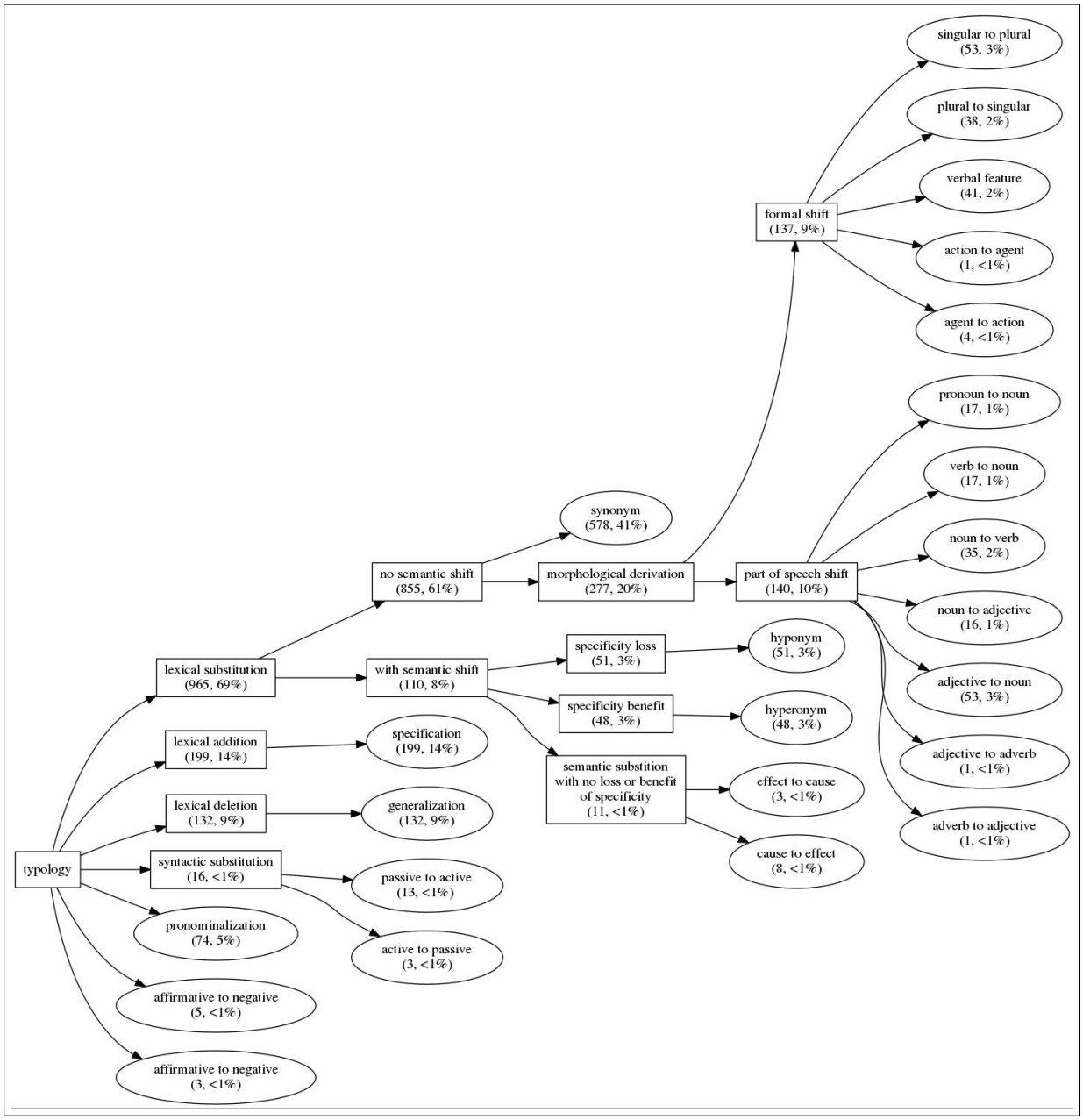

Figure 3: Typology of simplification-induced transformations

sentence splitting or merging and transformations. At the more fine-grained level, we observe that:

- *a2n* (adjective → noun) transformations (53 occ.) may be necessary to replace adjectives, often coined on suppletive bases (*cardiac*), by the corresponding nouns, often coined on alternative native bases (*heart*),
- *hyperonymy* transformations (48 occ.) permit to use words with broader meaning, which may make the understanding easier,
- *hyponymy* transformations (51 occ.) permit to use instanciations and terms with narrower meaning, which may also make the understanding easier,
- *n2v* (noun → verb) transformations (35 occ.) make the sentence less abstract by replacing concept by the action, and hence easier to understand.

It may seem counter-intuitive that there are more cases of hyponymy than hyperonymy in simplification, however, this can be explained. Indeed, in the simple side of the drug corpus, the exact name of the drug is given, when on the technical side of the drug corpus, the name given is the

tag	*split*	*merge*	*total*
syno	24	112	578
hypero	1	10	48
hypo	0	13	51
pronoun	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

Table 2: Frequency of the most frequent transformations in split and merged sentences, and in all the parallel sentences

therapeutical class of the drug. For instance, there is a case where we have *IEC (ACE inhibitor)* on the technical side and *Moex* (the name of a drug) on the simple side. Since *Moex* is a kind of *IEC*, then *Moex* is a hyponym for *IEC*.

4.3 Syntactic Analysis

Syntactic analysis permitted to associate semantic and syntactic information. One issue is that, with the substitutions, the POS-tags or syntactic groups remain identical in 221 cases. In several other cases, the original syntactic group is completed with other groups (*GN* $\rightarrow$ *GP GN*, *GN* $\rightarrow$ *GN GAdj*). Besides, up to 531 transformations start from nominal groups, up to 190 from prepositional groups and up to 174 from verbal groups. Overall, this means that: (1) the syntactic analysis may provide important clues for the detection of frontiers of the sequences to transform; (2) words and expressions of various syntactic nature are involved in transformations (nouns, verbs, adjectives...); (3) nouns and noun groups, often corresponding to concepts, occupy important place among the transformations.

5 Conclusion and Future Work

We proposed to work with parallel sentences differentiated by their technicity: technical and simplified contents are put in parallel. The main purpose is to describe the transformations involved during the simplification. Hence, the sentences are characterized on three dimensions: splitting and merging of sentences, semantic annotation of the transformations, and their syntactic annotation. We also propose a typology of transformations and quantify them. For instance, our work indicates that among the most frequent transformations we can find: synonymy, specification (insertion of additional information), generalization (removal of information), pronominalization, substitution of adjectives by the corresponding nouns, and swich between singular and plural forms. The material prepared will be used for the creation of transformation rules joining syntactic, lexical and semantic information. These rules will be later used for the simplification of biomedical texts.

Acknowledgments

This work was funded by the French National Agency for Research (ANR) as part of the *CLEAR* project (*Communication*, *Literacy*, *Education*, *Accessibility*, *Readability*), ANR-17-CE19-0016-01.

The authors would like to thank the reviewers for their questions and comments that helped in improving the article.

References

Diana J. Arya, Elfrieda H. Hiebert, and P. David Pearson. 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.

Susana Bautista, Pablo Gervás, and R. Ignacio Madrid. 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.

B Beigman Klebanov, K Knight, and D Marcu. 2004. Text simplification for information-seeking applications. In R Meersman and Z Tari, editors, *On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE*. Springer, LNCS vol 3290, Berlin, Heidelberg.

Catherine Blake, Julia Kampov, Andreas Orphanides, David West, and Cory Lown. 2007. Query expansion, lexical simplification, and sentence selection strategies for multi-document summarization. In *DUC*.

Laetitia Brouwers, Delphine Bernhard, Anne-Laure Ligozat, and Thomas François. 2014. Syntactic sentence simplification for French. In *PITR workshop*, pages 47–56.

Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2014. Defining

an annotation scheme with a view to automatic text simplification. In *CLICIT*, pages 87–92.

J Carroll, G Minnen, Y Canning, S Devlin, and J Tait. 1998. Practical simplification of English newspaper text to assist aphasic readers. In *AAAI-98 Workshop on Integrating Artifficial Intelligence and Assistive Technology*, pages 7–10.

R Chandrasekar and B Srinivas. 1997. Automatic induction of rules for text simplification. *Knowledge Based Systems*, 10(3):183–190.

Ping Chen, John Rochford, David N. Kennedy, Soussan Djamasbi, Peter Fay, and Will Scott. 2016. Automatic text simplification for people with intellectual disabilities. In *AIST*, pages 1–9.

Jan De Belder and Marie-Francine Moens. 2010. Text simplification for children. In *Workshop on Accessible Search Systems of SIGIR*, pages 1–8.

Siobhan Devlin and John Tait. 1998. The use of psycholinguistic database in the simplification of text for aphasic readers. In *Linguistic Database*, pages 161–173.

B Drndarevic, S Stajner, and H Saggion. 2012. Reporting simply: A lexical simplification strategy for enhancing text accessibility. In *Easy to read on the web*, pages 1–6.

Ulrich Germann. 2008. Yawat: Yet another word alignment tool. In *ACL-08: HLT Demo Session*, pages 20–23, Columbus, USA.

Goran Glavas and Sanja Stajner. 2015. Simplifying lexical simplification: Do we need simplified corpora? In *ACL-COLING*, pages 63–68.

Yea-Seul Kim, Jessica Hullman, Matthew Burgess, and Eytan Adar. 2016. SimpleScience: Lexical simplification of scientific terminology. In *EMNLP*, pages 1–6.

D Laurent, S Nègre, and P Séguéla. 2009. L'analyseur syntaxique Cordial dans Passage. In *Traitement Automatique des Langues Naturelles (TALN)*.

Gondy Leroy, David Kauchak, and Obay Mouradi. 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.

T Mikolov, K Chen, G Corrado, and J Dean. 2013. Efficient estimation of word representations in vector space. In *Workshop at ICLR*.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1993. Introduction to wordnet: An on-line lexical database. Technical report, WordNet.

Sergiu Nisioi, Sanja Stajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Ann Meeting of the Assoc for Comp Linguistics*, pages 85–91.

Gustavo H. Paetzold and Lucia Specia. 2016. Benchmarking lexical simplification systems. In *LREC*, pages 3074–3080.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP 2014*, pages 1532–1543.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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.

Matthew Shardlow. 2014. A survey of automated text simplification. *Int J Advanced Computer Science and Applications*, 1:1–13.

Ji Y. Son, Linda B. Smith, and Robert L. Goldstone. 2008. Simplicity and generalization: Short-cutting abstraction in children's object categorizations. *Cognition*, 108:626–638.

S Stymne, J Tiedemann, C Hardmeier, and J Nivre. 2013. Statistical machine translation with readability constraints. In *NODALIDA*, pages 1–12.

D Vickrey and D Koller. 2008. Sentence simplification for semantic role labeling. In *Annual Meeting of the Association for Computational Linguistics-HLT*, pages 344–352.

Marta Vila, M Antònia Mart, and Horacio Rodríguez. 2011. Paraphrase concept and typology. A linguistically based and computationally oriented approach. *Procesamiento del Lenguaje Natural*, 46:83–90.

Tong Wang, Ping Chen, Kevin Amaral, and Jipeng Qiang. 2016a. An experimental study of LSTM encoder-decoder model for text simplification. In *IJCAI*, pages 1–7.

Tong Wang, Ping Chen, John Rochford, and Jipeng Qiang. 2016b. Text simplification using neural machine translation. In *Proc of the AAAI Conference on Artificial Intelligence (AAAI-16)*, pages 4270–4271.

Chih-Hsuan Wei, Robert Leaman, and Zhiyong Lu. 2014. SimConcept: A hybrid approach for simplifying composite named entities in biomedicine. In *BCB '14*, pages 138–146.

S Wubben, A van den Bosch, and E Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Annual Meeting of the Association for Computational Linguistics*, pages 1015–1024.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Qian Yu, Aurélien Max, and François Yvon. 2012. Revisiting sentence alignment algorithms for alignment visualization and evaluation. In *BUCC workshop*, pages 1–7.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proc of the Conf on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark.

S Zhao, H Wang, and T Liu. 2010. Leveraging multiple MT engines for paraphrase generation. In *COLING*, pages 1326–1334.

Z Zhu, D Bernhard, and I Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *COLING 2010*, pages 1353–1361.