CLEAR – Simple Corpus for Medical French

Natalia Grabar, Rémi Cardon
CNRS, UMR 8163, F-59000 Lille, France;
Univ. Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France
natalia.grabar@univ-lille.fr, remi.cardon@univ-lille.fr

Abstract

Availability of corpora with technical and simplified contents is crucial for the development and test of methods for text simplification. We describe this kind of corpus for the French medical language. The corpus contains texts from three sources: encyclopedia, drug leaflets and scientific summaries. Each source proposes comparable information in specialized and plain languages. A subset of this corpus has been processed manually in order to find and align parallel sentences. This subset currently contains 663 pairs with parallel sentences. Alignment has been done by two annotators and shows 0.76 inter-annotator agreement. The corpus with comparable data is available for research (http://natalia.grabar.free.fr/resources.php).

1 Introduction

Research performed in text simplification provides tools and resources for the creation of simplified versions of texts. Simplification can be positioned at different levels (*ie.* lexical, syntactic, semantic, pragmatic and structural). It can be useful for different kinds of human users: children (Son et al., 2008; De Belder and Moens, 2010; Vu et al., 2014), foreigners or poor-readers (Paetzold and Specia, 2016), people with neurodegenerative disorders (Chen et al., 2016), lay people reading specialized documents (Arya et al., 2011; Leroy et al., 2013). In these cases, simplification may guarantee a better access to the contents of documents. Simplification may also be exploited as a pre-processing step of documents undergoing other NLP treatments: syntactic analysis (Chandrasekar and Srinivas, 1997; Jonnalagadda et al., 2009), semantic annotation (Vickrey and Koller, 2008), summarization (Blake et al., 2007), machine translation (Stymne et al., 2013; Štajner and Popović, 2016), indexing (Wei et al., 2014), information retrieval and extraction (Beigman Klebanov et al., 2004). The purpose is then to provide more easily processable versions of text and to improve the overall results of NLP tools.

Often, the feasibility and success of such works depend on the existence of the required corpora. Yet, in some languages and specialized fields such corpora may be missing.

The purpose of our work is to introduce and describe the CLEAR corpus, which gathers complex and simplified versions of documents related to medical topics in French. In what follows, we first present some existing work in corpora building for simplification (Section 2), we then describe our contribution to this area (Sections 3 and 4), and conclude (Section 5).

2 Corpora for Simplification

If the first works in development of simplification tools have mainly relied on manually crafted simplification rules following the linguistic intuition of researchers (Chandrasekar et al., 1996; Siddharthan, 2006; Max, 2008), recent works are mostly guided by linguistic data and rely on dedicated corpora. Most often, parallel corpora are exploited in this task. They provide original texts together with their simplified versions. Sometimes, aligned corpora are also available, in which the correspondence is done at the level of sentences. This kind of corpora provide direct correspondence between complex and simple (or simplified) sentences. Notice that comparable corpora, containing complex and simple documents addressing the same topics, are more easily available but require specific methods or pre-processings before

they can be exploited for simplification work.

Several parallel corpora for several languages have been created, mainly thanks to the manual simplification of their contents: Spanish (Bott et al., 2014), Italian (Brunato et al., 2014), Brazilian Portuguese (Caseli et al., 2009), Danish (Klerke and Sgaard, 2012), and of course English (Chandrasekar and Srinivas, 1997; Daelemans et al., 2004; Petersen and Ostendorf, 2007; Specia et al., 2012). Yet, these parallel corpora are seldom freely available. Some of these corpora also explicitly indicate what has been simplified and how (removal, segmentation...). Hence, a multi-axial annotation schema has been proposed for this purpose with several simplification classes: split, merge, reorder, insert (verbs, subjects and other components), delete (verbs, subjects and other components), transform (lexical substitution, replacement of anaphora, noun-verb, verb-noun, passive-active, verbal features...) (Brunato et al., 2014). This annotation schema covers lexical and syntactic simplification.

Comparable corpora of this kind are also available, among which the most frequently used is the pair built with English Wikipedia[1] and English Simple Wikipedia[2]. This corpus is widely used by researchers (Zhu et al., 2010; Biran et al., 2011; Coster and Kauchak, 2011). A similar comparable corpus also exists in French and can be built fromq French Wikipedia[3] and Vikidia[4], which has been created for children. This source in French has been used for the detection of rules for syntactic transformations (Brouwers et al., 2012). Besides, researchers working on English also exploit history of revisions of articles from Simple Wikipedia (Yatskar et al., 2010), simplified versions of scientific articles[5] (Elhadad and Sutaria, 2007), simplified versions of novels [6] (Vajjala and Meurers, 2015), as well as simplified versions of educational and news articles[7].

[1] https://en.wikipedia.org/wiki/Main_Page

[2] https://simple.wikipedia.org/wiki/Main_Page

[3] https://fr.wikipedia.org

[4] https://fr.vikidia.org

[5] http://www.reutershealth.com

[6] www.onestopenglish.com

[7] https://newsela.com/

3 Comparable Medical Simplified French Corpus

For the building of the corpus, we propose to exploit three types of French sources related to the medical field: articles from online encyclopedia (Section 3.1), drug leaflets with drug description and their optimal use (Section 3.2), and summaries from systematic reviews as provided by the Cochrane collaboration (Section 3.3). These sources provide documents from different textual genres: encyclopedia articles, scientific articles and drug description close to clinical texts. These three sources are available under free license (license not allowing modifications of the data in the case of the Cochrane reviews), and can be used for research purposes. Finally, these sources provide comparable corpora, distinguished by their technicality, on different topics: medical topics in encyclopedia, various drugs in drug leaflets, and questions related to treatment and diagnosis of disorders in Cochrane summaries. A part of these data have been aligned manually at the level of sentences (Section 4).

3.1 Encyclopedia Articles

This source provides articles from two collaborative encyclopedia in French available online: Wikipedia and Vikidia. French Wikipedia is intended for French-speaking people, while Vikidia has been created for providing similar information for 8 to 13 year old children. These two encyclopedia provide articles on a great variety of topics: politics, economics, medecine, culture, geography, etc. Wikipedia shows a better coverage than Vikidia: it is older and more popular. Creation of articles in these encyclopedia has to respect precise guidelines: they must be clear and understandable, be formal, with no use of jargon from specialized areas. Yet, as Vikidia is intended for children, the articles must contain as well: simple definitions and introduction, clear development, examples, sources and external links, and, if possible, pictures, schema, audio and video. It is also suitable to make children participate in the creation of the articles[8]. Even if articles from these two sources may be related to common topics, they are created independently from each other.

Articles from encyclopedia have been collected from the corresponding dumps in September 2017

[8] https://fr.vikidia.org/wiki/Aide:Comment_cr%C3%A9er_un_article

for Wikipedia and in August 2017 for Vikidia. Overall, Wikipedia contains 1,906,251 articles, and Vikidia contains 46,721 articles. Among the Wikipedia articles, we keep only 20,972 articles related to medicine and the medical portal. Among these, 575 articles exist in Wikipedia and Vikidia with identical titles. These 575 topics and pairs of articles are collected for building the corpus. Wikipedia articles contain 2,293,078 word occurrences, and Vikidia articles contain 197,672 word occurrences.

3.2 Drug Leaflets

Each drug marketed in France is provided together with a leaflet describing for instance its composition, prescription indications, known adverse effects, and precautions. This information is created in two versions. One version is intended for health professionals, and contains technical and comprehensive information on a given drug. Besides, this version presents a specific structure and makes use of a very rich medical terminology. Another version is intended for patients, and contains essential and simplified information on drugs. The style is personal. It addresses the patient directly and commonly using expressions like *votre santé (your health)*, *votre médecin (your physician)*, or *vous pouvez (you can)*. Information is structured as questions and answers: *Qu'est-ce que c'est ? (What is this?)*, *Quels sont les effets indésirables éventuels ? (What are the possible adverse effects?)*. These simplified versions are created systematically for each marketed drug, and later inserted into the drug boxes.

This corpus is built from documents available in the *public drug base*[9] managed by the Ministry of Health in France. These documents have been downloaded in June 2017. The corpus contains 11,800 drugs with technical and simplified leaflets. The technical part contains 52,313,126 word occurrences, and the simplified part contains 33,682,889 word occurrences.

3.3 Cochrane Summaries

The purpose of the Cochrane foundation is to provide high evidence medical information (Sackett et al., 1996). For several years, researchers of the domain have been working on creation of systematic reviews on various medical questions often in relation with diagnostics and treatment of disorders. Existing work on a given question are collected and read by experts. A synthesis is created, which methodological and scientific validity is higher than the one of each individual work. This also provides information with a higher evidence for medical professionals. For each extensive review, a short summary is also created. In addition to technical summaries for the experts, simplified summaries (*Plain language summary*) are created for lay people.

This corpus is built with documents available on the online library of Cochrane[10]. The documents have been downloaded in November 2017. The corpus contains 8,789 systematic reviews. Among these, 3,815 reviews provide technical and simplified versions of summaries. The technical part of the corpus contains 2,840,003 word occurrences and the simplified part contains 1,515,051 word occurrences.

4 Parallel Medical Simplified French Corpus

A subset of the whole comparable corpus has been aligned at the level of sentences. We randomly selected 14 encyclopedia articles, 12 drug leaflets, and 13 Cochrane summaries. The alignment has been performed manually by two annotators with the NLP training and used to the medical area texts. We have determined several criteria for alignment or non-alignment of two sentences, technical and simplified. They are illustrated with examples from the *Cochrane* corpus:

1. Identical sentences and sentences varying only by punctuation or stopwords are not aligned. Even if such pairs of sentences provide very close or identical semantic contents, we consider indeed that such pairs are not helpful for the creation of transformation rules useful for the simplification of contents;

2. Sentences within an aligned pair must have the same or very close meaning (semantic equivalence), and they must show lexical and/or syntactic adaptations, at least:

   - *Preterm infants are at risk of periventricular haemorrhage (PVH).*
   - *Babies born very early (before 34 weeks) are at risk of bleeding in the brain (periventricular haemorrhage).*

[9] http://base-donnees-publique.medicaments.gouv.fr/

[10] http://www.cochranelibrary.com/

		Technical				*Simplified*			
		source		*aligned*		*source*		*aligned*	
corpus	*doc.*	*sent.*	*occ.*	*sent.*	*occ.*	*sent.*	*occ.*	*sent.*	*occ.*
Drug	12*2	4,416	44,709	502	5,751	2,736	27,820	502	10,398
Cochrane	13*2	553	8,854	112	3,166	263	4,688	112	3,306
Encyclopedia	14*2	2,494	36,002	49	1,100	238	2,659	49	853

Table 1: Size of the reference data and their consensual alignment at the level of sentences.

3. The meaning of one sentence can be fully included in another sentence. This is the case of semantic inclusion. In the following example, the content of the simplified sentence is included in the technical sentence:

   - *We found no studies that reported the effect of whole grain diets on total cardiovascular mortality or cardiovascular events (total myocardial infarction, unstable angina, coronary artery bypass graft surgery, percutaneous transluminal coronary angioplasty, total stroke).*
   - *We found no studies reporting on the effect of whole grains on deaths from cardiovascular disease or cardiovascular events.*

4. Semantic intersection, where each sentence of the pair brings its own additional information, is not accepted:

   - *However, over the past two decades endovascular aneurysm repair (EVAR) has gained popularity as a treatment option.*
   - *However, over the past 20 years, a newer, 'key hole' technique has been used, in which the AAA is repaired without the need for open surgery - a thin tube is passed via the blood vessels in the groin to the site of the AAA.*

The alignment has been done independently by two annotators. Agreement occurs when the annotators propose the same alignment of sentences, and disagreement occurs when a given pair is only aligned by one of the annotators. As a second step, the disagreements are discussed in order to reach the consensus when possible. As a result, a given pair of sentences can be approved for the alignment or rejected.

Table 1 indicates the size of the source and aligned sets with consensual alignments. We obtain a total of 663 pairs of aligned sentences. This is a small set of parallel data, but it is intended to grow up thanks to the design and use of suitable models for the automatic alignment of sentences. The 663 already aligned pairs of sentences provide the necessary reference data.

Semantic annotation is one of the hardest annotation tasks and usually shows low annotation agreement (Artstein and Poesio, 2008), which has been particularly highlighted for word sense tagging (Véronis, 1998; Mihalcea et al., 2004; Palmer et al., 2007). Hence, the annotation of semantic closeness between two sentences is also complicated. In our experiment, the inter-rater agreement is 0.76 (Cohen, 1960). It is computed within the set of the aligned sentences from the two annotators. Such inter-annotator agreement is qualified as substantial according to the usual interpretation scale (Landis and Koch, 1977) and may indicate a good reliability of the obtained data.

Another interesting point is related to the parallelism between the technical and simple versions of documents. It has been indeed observed that the degree of parallelism in comparable corpora may vary from almost parallel corpora, with many parallel sentences, to very-non-parallel corpora (Fung and Cheung, 2004). In the CLEAR corpus, we can observe that aligned sentences are rarer in the *Drugs* and *Encylopedia* corpora than in the *Cochrane* corpus. Indeed, these three sources have different principles involved during the creation of their contents:

- Summaries of systematic reviews from Cochrane are intentionally simplified by researchers starting from the original technical summaries;

- Vikidia articles are written independently from Wikidia articles, even if they address the same topics: there is no adaptation of one content into another. Besides, as Vikidia articles are created for children, their content is adapted for them;

- In the *Drugs* corpus, the same drugs are described for health professionals and for patients, which provides good common ground. Yet, several kinds of information are specific either to the professional version (precise composition, action on the organism, molecules, detailed information on adverse effects...) or to the patient's (precautions of use, warnings...).

It would be interesting to formalize the notion of parallelism between two corpora, which should be indicative of the rate of parallel sentences they may provide.

The first observations of parallel sentences indicate that they provide mainly syntactic and lexical transformations, and that the simplification principles differ according to the document sources. For instance, sentence splitting is applied in drug leaflets and encyclopedia articles, while the sentences are usually merged during the simplification process in Cochrane summaries. These and other simplification features are being analyzed. They will allow to propose adaptation rules that apply at lexical and syntactic levels. As for the semantic and especially structural levels of adaptation, we assume that information available from parallel sentence pairs is not sufficient and that more global observations and datasets should be exploited.

5 Conclusion and Future Work

In this paper, we introduced the CLEAR corpus with technical and simplified contents in French from the medical field. This kind of corpora is indeed very useful for preparing work on automatic text simplification. The corpus contains texts from three sources: encyclopedia, drug leaflets and summaries of systematic reviews. The source texts are comparable: they propose information on the same topics. The corpus totalizes 16,190 pairs of documents, which corresponds to over 57M word occurrences in the technical part and over 35M word occurrences in the simplified part. A subset of this corpus has been aligned at the sentence level by two annotators with 0.76 inter-annotator agreement. This subset provides 663 pairs of sentences.

In the future, the parallel dataset will be extended automatically further to the design and use of suitable language models. Hence, comparable and parallel datasets will be exploited for designing and testing methods for simplification of medical documents in French. This is an important issue because health-related documents typically contain specialized terminology and notions, which are difficult to be understood by lay people (AMA, 1999; McCray, 2005; Jucks and Bromme, 2007; Kickbusch et al., 2013). In addition to this lexical level, transformations at syntactic level may also be helpful.

The CLEAR corpus with comparable data is available for research and can be found online[11].

6 Acknowledgements

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. We would like to thank the anonymous reviewers for their comments.

References

AMA. 1999. Health literacy: report of the council on scientific affairs. Ad hoc committee on health literacy for the council on scientific affairs, American Medical Association. *JAMA*, 281(6):552–7.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

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.

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.

Or Biran, Samuel Brody, and Noémie Elhadad. 2011. Putting it simply: a context-aware approach to lexical simplification. In *Annual Meeting of the Association for Computational Linguistics*.

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*.

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

Stefan Bott, Horacio Saggion, and Simon Mille. 2014. Text simplification tools for Spanish. In *LREC 2014*, pages 1–7.

Laetitia Brouwers, Delphine Bernhard, Anne-Laure Ligozat, and Thomas Franois. 2012. Simplification syntaxique de phrases pour le français. In *Traitement Automatique des Langues Naturelles (TALN)*, pages 211–224.

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.

Helena M. Caseli, Tiago F. Pereira, Lucia Specia, Thiago A. S. Pardo, Caroline Gasperin, and Sandra M. Aluisio. 2009. Building a Brazilian Portuguese parallel corpus of original and simplified texts. In *CICLING*, pages 1–12.

R. Chandrasekar, C. Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *COLING*, pages 1041–1044.

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.

J Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

William Coster and David Kauchak. 2011. Simple English wikipedia: A new text simplification task. In *Annual Meeting of the Association for Computational Linguistics*, pages 665–669.

Walter Daelemans, Anja Höthker, and Erik Tjong Kim Sang. 2004. Automatic sentence simplification for subtitling in Dutch and English. In *LREC*, pages 1045–1048.

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

N Elhadad and K Sutaria. 2007. Mining a lexicon of technical terms and lay equivalents. In *BioNLP*, pages 49–56.

Pascale Fung and Percy Cheung. 2004. Mining very non-parallel corpora: Parallel sentence and lexicon extraction vie bootstrapping and em. In *Conference on Empirical Methods in Natural Language Processing*, pages 57–63.

Siddhartha Jonnalagadda, Luis Tari, Jörg Hakenberg, Chitta Baral, and Graciela Gonzalez. 2009. Towards effective sentence simplification for automatic processing of biomedical text. In *NAACL HLT 2009*, pages 177–180.

R Jucks and R Bromme. 2007. Choice of words in doctor-patient communication: an analysis of health-related internet sites. *Health Commun*, 21(3):267–77.

I Kickbusch, JM Pelikan, F Apfel, and AD Tsouros. 2013. Health literacy. the solid facts. Technical report, WHO.

Sigrid Klerke and Anders Sgaard. 2012. DSim, a Danish parallel corpus for text simplification. In *LREC*, pages 4015–4018.

JR Landis and GG Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

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.

A Max. 2008. Local rephrasing suggestions for supporting the work of writers. In *GOTAL*, pages 324–335.

A McCray. 2005. Promoting health literacy. *J of Am Med Infor Ass*, 12:152–163.

Rada Mihalcea, Timothy Chklovski, and Adam Kilgarriff. 2004. The SENSEVAL-3 English lexical sample task. In *SENSEVAL-3*, pages 25–28, Barcelona.

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

Martha Palmer, Hoa Trang Dang, and Christiane Fellbaum. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, 13(2):137–163.

SE Petersen and M Ostendorf. 2007. Text simplification for language learners: A corpus analysis. In *Speech and Language Technology for Education Workshop (SLaTE)*, pages 69–72.

DL Sackett, WM Rosenberg, JA Gray, RB Haynes, and WS Richardson. 1996. Evidence based medicine: what it is and what it isn't. *BMJ*, 312(7023):71–2.

Advaith Siddharthan. 2006. Syntactic simplification and text cohesion. *Research on Language & Computation*, 4(1):77–109.

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

L Specia, SK Jauhar, and R Mihalcea. 2012. Semeval-2012 task 1: English lexical simplification. In **SEM 2012*, pages 347–355.

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

Sowmya Vajjala and Detmar Meurers. 2015. Readability-based sentence ranking for evaluating text simplification. Technical report, Iowa State University.

Jean Véronis. 1998. A study of polysemy judgments and inter-annotator agreement. In *SENSEVAL-1*, Herstmonceux Castle, England.

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.

Sanja Štajner and Maja Popović. 2016. Can text simplification help machine translation? *Baltic J. Modern Computing*, 4(2):230–242.

Tu Thanh Vu, Giang Binh Tran, and Son Bao Pham. 2014. Learning to simplify children stories with limited data. In *Intelligent Information and Database Systems*, pages 31–41.

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.

Mark Yatskar, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2010. For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia. In *NAACL*, pages 365–368.

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