Building a comparable corpus and a benchmark for Spanish medical text simplification

Construcción de un corpus comparable y un recurso de referencia para la simplificación de textos médicos en español

Leonardo Campillos-Llanos,[1] **Ana R. Terroba Reinares,**[2]
Sofía Zakhir Puig,[1] **Ana Valverde-Mateos**[3], **Adrián Capllonch-Carrión**[4]
[1]ILLA - Consejo Superior de Investigaciones Científicas (CSIC)
[2]Fundación Rioja Salud
[3]Unidad de Terminología Médica, Real Academia Nacional de Medicina de España
[4]Centro de Salud Retiro, Hospital General Universitario Gregorio Marañón
{leonardo.campillos,sofia.zakhir}@csic.es, arterroba@riojasalud.es,
avalverde@ranm.es, adrian.capllonch@salud.madrid.org

Abstract: We report the collection of the CLARA-MeD comparable corpus, which is made up of 24 298 pairs of professional and simplified texts in the medical domain for the Spanish language (>96M tokens). Texts types range from drug leaflets and summaries of product characteristics (10 211 pairs of texts, >82M words), abstracts of systematic reviews (8138 pairs of texts, >9M words), cancer-related information summaries (201 pairs of texts, >3M tokens) and clinical trials announcements (5748 pairs of texts, 451 690 words). We also report the alignment of professional and simplified sentences, conducted manually by pairs of annotators. A subset of 3800 sentence pairs (149 862 tokens) has been aligned each by 2 experts, with an average inter-annotator agreement *kappa* score of 0.839 (±0.076). The data are available in the community and contributes with a new benchmark to develop and evaluate automatic medical text simplification systems.
Keywords: Comparable corpus. Medical text simplification. Biomedical natural language processing.

Resumen: Se describe la recogida del corpus comparable CLARA-MeD, formado por 24 298 pares de textos profesionales y simplificados de dominio médico en lengua española (>96M palabras). Los tipos de textos varían desde prospectos médicos y fichas técnicas de medicamentos (10 211 pares de textos, >82M palabras), resúmenes de revisiones sistemáticas (8138 pares de textos, >9M palabras), resúmenes de información sobre el cáncer (201 pares de textos, >3M palabras) y anuncios de ensayos clínicos (5748 pares de textos, 451 690 palabras). También presentamos el alineamiento de frases técnicas y simplificadas, realizado a mano por pares de anotadores. Un subconjunto de 3800 pares de frases (149 862 tokens) se han emparejado, con un acuerdo medio entre anotadores con valor $kappa = 0.839$ (±0.076). Los datos están disponibles en la comunidad y este nuevo recurso permite desarrollar y evaluar sistemas de simplificación automática de textos médicos.
Palabras clave: Corpus comparable. Simplificación de textos médicos. Procesamiento de lenguaje natural biomédico.

1 Introduction

Text simplification is the task of *transforming a text into an equivalent which is more understandable* (Saggion et al., 2011). The application of natural language processing (NLP) techniques makes it possible to automate the simplication of texts across domains and tasks, ranging from legal and administrative texts (Scarton et al., 2018), language learning (Petersen and Ostendorf, 2007), users with special reading needs (Barbu et al., 2015) or health literacy (Kindig et al., 2004).

Corpus data are required for analysing text simplification strategies, developing and testing NLP systems. This work introduces

a new resource made up of documents from the medical domain, which is available at: https://digital.csic.es/handle/10261/269887.

2 Background

Text simplification approaches are commonly conceived as a translation task—from the technical to laymen's register. Simplification involves operations at multiple linguistic levels: grammar (e.g. simpler word order, passive to active voice), discourse (e.g. split long sentences) or lexis (e.g. replace complex words with clearer synonyms). Some automatic simplification approaches rely on rule-based or lexicon-based modules to address each linguistic level. In contrast, data-driven approaches may use machine translation of monolingual (professional/simplified) corpora—at present, mostly via deep-learning-based methods (Van den Bercken et al., 2019; Sakakini et al., 2020; Devaraj et al., 2021; Martin et al., 2021).

Regardless of the approach, dedicated corpora are need: comparable resources (professional and simplified versions of a text) or, ideally, parallel corpora (texts with almost identical content in different registers).

In addition to corpora for the English language (Van den Bercken et al., 2019; Sakakini et al., 2020), simplification text collections exist for Brazilian Portuguese (Caseli et al., 2009), German (Klaper et al., 2013), Italian (Tonelli et al., 2016) or French (Grabar and Cardon, 2018; Gala et al., 2020). Other multilingual resources have been released in challenges for complex word identification (Yimam et al., 2017)

For Spanish, there is the EASIER corpus,[1] with different characteristics compared to the CLARA-MeD text collection. First, the EASIER corpus is a general domain resource; even though some sentences are related to health topics, the CLARA-MeD resource focuses only on the medical domain. Second, the EASIER corpus gathers 3977 sentences annotated with 8155 complex words, and 3396 sentences labeled with 7892 suggested synonyms. The CLARA-MeD corpus is not annotated with these data, but features a sentence-level alignment of 3800 pairs of technical and simplified sentences, following specific criteria (§3.6). Lastly, besides including parallel data, the CLARA-MeD corpus is larger in size. These data can be used to enlarge the number of aligned sentences or can be annotated in more detail in future versions.

Several methods have been applied to collect such type of corpora. Wikipedia and Simple Wikipedia have been aligned to obtain a parallel corpus in the general domain (Zhu et al., 2010). However, the correspondence of content between technical/simplified versions are often deficient (Xu et al., 2015). Moreover, this method can not be directly extended to languages without a simplified Wikipedia. In these cases, some teams (Palmero Aprosio et al., 2019; Rauf et al., 2020) have created synthetic corpora via translations from SimpleWikipedia or similar sources such as WikiLarge (Grabar and Cardon, 2018). Another method is manual simplification by domain or task specialists (Gala et al., 2020), which assures linguistic quality but requires an adequate team and is more time-consuming. Hybrid methods have also been conducted for medical English (Van den Bercken et al., 2019; Moramarco et al., 2021).

3 Methods and Sources

Figure 1 summarizes the workflow applied to create the corpus. The first stage involved collecting a comparable resource. We gathered medical texts with two versions (for professionals and laymen readers) from sources recently reported (Moreno-Sandoval et al., 2019). At the current stage, we have not used articles from the Medicine category in Wikipedia, given that there is not a Spanish version from Simple Wikipedia.

The second stage implied matching professional and simplified sentences from each comparable subcorpus. In the following, we detail the data sources to create this resource (§ 3.1-§ 3.4), and Section 3.6 explains the criteria and methods to align sentences.

3.1 Drug leaflets and summaries of product characteristics

The Medicine Online Information Center (Centro de Información de Medicamentos, CIMA) [2] is a drug-related service and knowledge database maintained by the Spanish Agency of Medicines and Medical Devices

[1] shorturl.at/lvCJU

[2] https://cima.aemps.es

Figure 1: Methods and stages to compile the corpus.

(AEMPS). CIMA provides all the information related to drugs prescribed in Spain through a search engine, the Nomenclator resource (a rich database of medical drugs), and pharmaceutical/pharmacovigilance reports. In addition, the information about each medical drug (indication, medical brand, dosage, unit of presentation, etc.) is provided in two types of documents: summaries of product characteristics (written in a professional register and aimed at healthcare professionals) and drug leaflets (with simpler structures or terms, and aimed at patients). We downloaded both types of data, and release only cleaned, noise-free texts with both versions available (10 211 pairs of texts, 82 907 317 words).[3]

3.2 Systematic reviews

The Cochrane Library[4] is an updated database of systematic reviews and metanalyses. This is the main collection of medical evidence (Sackett et al., 1996), mostly from publications reporting results of clinical trials. Healthcare professionals use this database to keep up to date with the latest evidence to apply in their clinical practice. The Cochrane Library is a multilingual resource, although not all reviews are available in all languages. Each review presents an abstract of the full text and also a summary in plain language, which is aimed at a non-specialist readership. We collected a total of 8138 pairs of documents (9 618 698 words).

3.3 Cancer-related information summaries

The National Cancer Institute (NCI) website presents a large volume of bilingual (English and Spanish) information.[5] Contents revolve around cancer types, disorders and symptoms, oncological therapies, pharmacological substances, genetics, screening, prevention, palliative care or patient-oriented counseling. Noteworthy, *Physician Data Queries* (*PDQ*) articles gather updated and evidence-based information about essential aspects of cancer in two versions: for professionals and patients. We collected 201 pairs of PDQ texts, and removed noisy information from the web pages (e.g. URLs, content menus, tables, etc). The cleaned texts amount up to 3 044 461 words.

3.4 Clinical trials announcements

The European Clinical Trials register (EudraCT)[6] gathers public data about all clinical trials conducted in the European Union, either at national or multinational level. Clinical trials announcements are published both in English and the European language corresponding to the countries involved in the trial. Trial announcements describe the trial protocols, patients and participants, interventions, indications and expected outcomes of the trial, among others. Two sections are written with equivalent contents between scientific (aimed at healthcare professionals) and popularization levels (aimed at patients or laymen users): the public and scientific title of the trial, and the public and scientific indication. To gather these data, we reused 700 texts from the Clinical Trials for Evidence Based Medicine in Spanish (CT-EBM-SP) corpus (Campillos-Llanos et al., 2021). We also downloaded more than 7500 announcements from the website, and selected only those with both a public and scientific version of the title and/or indication. After filtering out noisy or redundant data, we gathered 5784 pairs of texts (451 690 words).

3.5 Descriptive statistics

Table 1 includes excerpts of professional and simplified versions of each data source. Table 2 shows the word count of the comparable corpus, and Table 3, the average of sentences per text and average words per sentence (with standard deviation, SD). Texts

[3] The data from the Cochrane Library cannot be released without permission.

[4] https://www.cochranelibrary.com/

[5] https://www.cancer.gov

[6] https://www.clinicaltrialsregister.eu

Source	Professional version	Simplified version
CIMA	*La administración concomitante de metamizol con metotrexato u otros antineoplásicos puede aumentar la toxicidad sanguínea de los antineoplásicos particularmente en pacientes de edad avanzada.* 'Concomitant administration of metamizole with methotrexate or other antineoplastics may increase the blood toxicity of antineoplastics particularly in elderly patients.'	*Si se administra conjuntamente con metotrexato u otros medicamentos para el tratamiento de los tumores (antineoplásicos), puede potenciar los efectos tóxicos en sangre de los antineoplásicos, sobre todo en pacientes de edad avanzada.* 'If co-administered with methotrexate or other drugs for the treatment of tumors (antineoplastics), it may potentiate the toxic effects of antineoplastics in the blood, especially in elderly patients.'
Cochrane	*La administración de suplementos de vitamina D podría disminuir la necesidad de ventilación mecánica invasiva, pero la evidencia es incierta (evidencia de certeza baja).* 'Vitamin D supplementation may decrease need for invasive mechanical ventilation, but the evidence is uncertain (low-certainty evidence).'	*La vitamina D podría reducir la necesidad de conectar a los pacientes a un respirador para ayudarles a respirar, pero se desconoce la evidencia.* 'Vitamin D may reduce the need for patients to be put on a ventilator to help them breathe, but the evidence is uncertain.'
EudraCT	*Ensayo clínico aleatorizado, doble ciego, controlado con placebo, para evaluar la eficacia y seguridad de la vacuna COMIRNATY (vacuna COVID-19 ARNm, Pfizer-BioNTech) en personas con COVID persistente* 'Randomized, double-blind, placebo-controlled clinical trial to evaluate the efficacy and safety of the COMIRNATY vaccine (COVID-19 mRNA vaccine, Pfizer-BioNTech) in people with long COVID'	*El objetivo del estudio es analizar si la administración de una vacuna contra la infección COVID19 puede hacer disminuir los síntomas de COVID persistente.* 'The aim of the study is to analyze whether the administration of a vaccine against COVID19 infection can reduce the symptoms of long COVID.'
NCI	*El LH que se diagnostica durante el primer trimestre del embarazo no constituye un indicador absoluto de la necesidad de un aborto terapéutico.* 'HL that is diagnosed in the first trimester of pregnancy does not constitute an absolute indication for therapeutic abortion.'	*Cuando el linfoma de Hodgkin se diagnostica durante el primer trimestre del embarazo, no siempre significa que se aconsejará a la mujer que interrumpa el embarazo.* 'When Hodgkin lymphoma is diagnosed in the first trimester of pregnancy, it does not necessarily mean that the woman will be advised to end the pregnancy.'

Table 1: Samples of professional and simplified versions of texts from different data sources.

from CIMA (drug leaflets and summaries of product characteristics) and NCI (cancer-related summaries) are longer. Professional texts from all sources tend to be longer than simplified texts. Likewise, the average sentence length (number of words per sentence) is generally longer in the professional version of almost all sources (except for CIMA).

3.6 Parallel text alignment

We aligned 3800 professional-laymen sentences extracted from the CLARA-MeD corpus. Pairs of annotators with varied backgrounds (a computational linguist, a medical doctor and medical terminologists) matched scientific and simplified versions of a subset of sentences.

Gathering parallel sentences from the EudraCT announcements was straightforward. Each clinical trial announcement contains two versions (for patients/laymen users and healthcare professionals) of specific sections: the public and scientific title of the trial, and a public and scientific indication. We gathered the data from both versions, which yielded a preliminary noisy alignment of 5784 sentence pairs. Similar sentences were rejected, and two annotators per data batch conducted a manual revision. We followed a set of linguistic criteria to accept a sentence pair as adequate equivalences (§ 3.6.2). Problematic pairs of sentences were discussed to achieve a consensus.

For the other sources, we automated the alignment by extracting, for each professional sentence, the most similar simplified version (§ 3.6.1). After the semi-automatic alignment, we followed the same methodology for the manual revision, and two annotators per data batch checked the sentence pairs. We thereby filtered out the most reliable sentence pairs and assessed the inter-annotator agreement.

3.6.1 Semi-automatic alignment

To gather aligned pairs of sentences, we should combine, for each pair of texts, all professional sentences with all simplified ones. Nonetheless, the amount of candidate pairs collected in this way is unaffordable to be re-

Source	Text pairs	Professional	Simplified	Total
CIMA	10 211	55 463 410	27 443 907	82 907 317
Cochrane abstracts	8138	6 235 454	3 383 244	9 618 698
EudraCT	5748	255 902	195 788	451 690
NCI	201	2 093 569	955 480	3 049 049
Total	**24 298**	**64 048 335**	**31 978 419**	**96 022 166**

Table 2: Word count per source data of the comparable corpus.

Source		Professional	Simplified
CIMA	Avg sentences per text (±SD)	431.72 (±234.47)	210.03 (±71.96)
	Avg words per sentence (±SD)	12.36 (±2.48)	12.76 (±1.61)
Cochrane abstracts	Avg sentences per text (±SD)	34.06 (±9.05)	19.57 (±11.70)
	Avg words per sentence (±SD)	22.08 (±3.94)	22.02 (±4.39)
EudraCT	Avg sentences per text (±SD)	1.77 (±0.65)	1.75 (±0.62)
	Avg words per sentence (±SD)	26.47 (±11.20)	19.73 (±8.74)
NCI	Avg sentences per text (±SD)	505.53 (±492.57)	309.39 (±177.61)
	Avg words per sentence (±SD)	20.53 (±3.75)	15.47 (±1.49)

Table 3: Average (avg) sentences per text and average words per sentence (±standard deviation).

vised, and only very few pairs will be adequate alignments.

Previous work (Cardon and Grabar, 2020) has reduced this *search space* by leveraging parsing information and developing a machine-learning classifier using features such as sentence length, the Levenshtein distance or number of shared words between each version, among others. The alignment may also be automated by means of tools such as CATS (Štajner et al., 2018), which extracts similar sentences according to character n-grams, the average of word embeddings (WAVG) in the sentence or the continuous word alignment-based similarity analysis model. Another tool is MASSAlign (Paetzold et al., 2017), which aligns paragraphs or sentences by computing a TF-IDF-based similarity matrix of items between each pair of texts, and a vicinity procedure to complete the alignment iteratively.

Nonetheless, we experimented with a state-of-the-art procedure, BERT-based Sentence Embeddings (Reimers and Gurevych, 2019). With this method, we obtained an embedding representation of each sentence, and compute the cosine similarity between the professional and simplified version. We applied a threshold set empirically on the cosine similarity score ($cosine > 0.6$). We discarded sentences that were not similar enough, to obtain a subset to be revised manually later. We provide a companion python jupyter notebook to reproduce our method.[7]

3.6.2 Alignment criteria

We adopted the guidelines from Grabar and Cardon (2018) to align pairs of sentences (e.g. identical pairs are not used), and we added new rules. Table 4 summarizes our criteria.

We followed these criteria and rejected those sentences that each pair of annotators per data batch judged as bad alignments. Disagreements were discussed to achieve a consensus, and the medical practitioner solved specific questions about medical aspects of the contents. We aligned a total of 3800 sentence pairs (149 862 words). The average inter-annotator agreement between experts was of $kappa = 0.839$ (±0.076), which represents an *almost perfect agreement*.

4 Conclusions and future work

This work has presented the CLARA-MeD corpus of comparable (professional/laymen) medical texts in Spanish, a new contribution for analysing text simplification strategies and conduct experiments on text simplification tasks. A limitation of our work is the scarce data obtained to train and test data-intensive methods (e.g. deep-learning). More

[7] https://github.com/lcampillos/CLARA-MeD/

1. We prioritize aligning one-to-one sentences; however, in some cases, one simplified sentence needs to be aligned with two professional ones, and vice versa:

P: *Linfoma folicular recidivante/resistente* ('Relapsed/refractory follicular lymphoma')
S: *El linfoma folicular es un cáncer que afecta a los glóbulos blancos llamados linfocitos. El término recaída o refractaria indica una enfermedad que vuelve a crecer o no responde al tratamiento* ('Follicular lymphoma is a cancer that affects white blood cells called lymphocytes. The term relapsed or refractory indicates disease that grows again or does not respond to the treatment.')

2. Sentence pairs that only differ in punctuation or functional words (e.g. prepositions or adverbs) are not aligned if the simplified version does not have a simpler structure.

3. Simplified sentences that have unintelligible acronyms without their explanation or expansion are not used (we except widely-used acronyms: e.g. *SIDA*, 'AIDS'):

P: *Cancer colorectal (CCR)* ('Colorectal cancer (CRC)')
S: *El CCR es el desarrollo del cáncer desde el colon o el recto* (Not aligned)
('CRC is the development of cancer from the colon or rectum')

4. Professional and simplified sentences are not aligned if the simplified version presents a large loss of essential information that is present in the professional version:

P: *Tratamiento del síndrome de Hunter y deterioro cognitivo*
('Treatment of Hunter syndrome and cognitive impairment')
S: *Síndrome de Hunter: deficiencia de la enzima iduronato 2-sulfatasa* (Not aligned)
('Hunter syndrome-Iduronate-2-Sulfatase enzyme deficiency')

5. We do not align sentence pairs with incoherent data, imprecise information or contradictions between the professional and the simplified version:

P: *Diabetes mellitus tipo 1* ('Type 1 Diabetes Mellitus')
S: *Altos niveles de azucar (glucosa) en sangre* (Not aligned)
('High levels of sugar (glucose) in the blood')

6. Sentences consisting of paraphrases, definitions or explanations of technical terms are considered adequate simplified versions and can be aligned with the professional version:

P: *Colitis ulcerosa* ('Ulcerative Colitis')
S: *La colitis ulcerosa es una enfermedad inflamatoria intestinal que provoca inflamación en el revestimiento del intestino grueso con irritación e hinchazón* ('Ulcerative Colitis is a type of inflammatory bowel disease that causes the lining of the large intestine (colon) to become inflamed (irritated and swollen)') (Aligned)

7. Aligned sentences may have redundant information or elliptic words (e.g. prepositions), minor spelling, grammar or typographic errors, provided that the meaning is not distorted:

P: *Neumonía por SARS-CoV-2* ('SARS-CoV-2 infected patients with pneumonia')
S: *Neumonía COVID* ('COVID-Pneumonia') (Aligned)

Table 4: Alignment criteria with examples (*P*: professional; *S*: simplified).

drug-related documents from CIMA need to be cleaned and revised to be used. Moreover, more types of comparable data could be obtained from medical websites with two versions (professional-oriented and patient-oriented contents). In addition to this, the methodology applied by van der Bercken et al. (2019) could be useful to widen the coverage of Spanish medical texts from Wikipedia, and thus include this source in the cor-

pus. Sometimes, the sentence-level alignment might not be satisfactory because there is not always a one-to-one correspondence. In such cases, a paragraph-level alignment is a more suitable option (Devaraj et al., 2021). Lastly, the corpus is not annotated with complex words, which would be useful in complex word identification (CWI) tasks.

Even so, this is the first version of a benchmark for medical text simplification in the Spanish language. The high inter-annotator agreement values show a fine sentence-level alignment, which was assessed by linguists, a lexicographer and with the advice of a health professional. This ensures that these data are a quality benchmark for developing and testing medical text simplification systems.

Acknowledgements

We thank the reviewers for their valuable comments to improve this work. Project CLARA-MED (PID2020-116001RA-C33) funded by MCIN/AEI/ 10.13039/501100011033/, in project call: "Proyectos I+D+i Retos Investigación".

References

Barbu, E., Martín-Valdivia, M. T., Martinez-Camara, E., and Urena-López, L. A. (2015). Language technologies applied to document simplification for helping autistic people. *Expert Systems with Applications*, 42(12):5076–5086.

Campillos-Llanos, L., Valverde-Mateos, A., Capllonch-Carrión, A., and Moreno-Sandoval, A. (2021). A clinical trials corpus annotated with UMLS entities to enhance the access to evidence-based medicine. *BMC medical informatics and decision making*, 21(1):1–19.

Cardon, R. and Grabar, N. (2020). Construction d'un corpus parallèle à partir de corpus comparables pour la simplification de textes médicaux en français. *Traitement Automatique des Langues*, 61(2):15–39.

Caseli, H. M., Pereira, T. F., Specia, L., Pardo, T. A., Gasperin, C., and Aluísio, S. M. (2009). Building a Brazilian Portuguese parallel corpus of original and simplified texts. *Proc. of 10th CICLing*, 41:59–70.

Devaraj, A., Marshall, I., Wallace, B., and Li, J. J. (2021). Paragraph-level simplification of medical texts. In *Proc. of the NAACL 2021*, pages 4972–4984.

Gala, N., Tack, A., Javourey-Drevet, L., François, T., and Ziegler, J. C. (2020). Alector: A parallel corpus of simplified French texts with alignments of misreadings by poor and dyslexic readers. In *Proc. of LREC 2020*, page 1353–1361.

Grabar, N. and Cardon, R. (2018). CLEAR - Simple corpus for medical French. In *Proc. of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9.

Kindig, D. A., Panzer, A. M., Nielsen-Bohlman, L., et al. (2004). *Health literacy: a prescription to end confusion*. Washington (DC): National Academies Press.

Klaper, D., Ebling, S., and Volk, M. (2013). Building a German/simple German parallel corpus for automatic text simplification. In *Proc. of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR 2013), Sofia, Bulgaria*.

Martin, L., Fan, A., de la Clergerie, É., Bordes, A., and Sagot, B. (2021). Muss: Multilingual unsupervised sentence simplification by mining paraphrases. *arXiv preprint arXiv:2005.00352*.

Moramarco, F., Juric, D., Savkov, A., Flann, J., Lehl, M., Boda, K., Grafen, T., Zhelezniak, V., Gohil, S., Korfiatis, A. P., et al. (2021). Towards more patient friendly clinical notes through language models and ontologies. In *Proc. of the AMIA Annual Symposium*, pages 881–890.

Moreno-Sandoval, A., Torre-Toledano, D., Valverde-Mateos, A., and Campillos-Llanos, L. (2019). Estudio sobre documentos reutilizables como recursos lingüísticos en el marco del desarrollo del plan de impulso de las tecnologías del lenguaje. *Procesamiento del Lenguaje Natural*, 63:167–170.

Paetzold, G., Alva-Manchego, F., and Specia, L. (2017). Massalign: Alignment and annotation of comparable documents. In *Proceedings of the IJCNLP 2017, System Demonstrations*, pages 1–4.

Palmero Aprosio, A., Tonelli, S., Turchi, M., Negri, M., and Di Gangi Mattia, A.

(2019). Neural text simplification in low-resource conditions using weak supervision. In *Workshop on Methods for Optimizing and Evaluating Neural Language Generation (NeuralGen)*, pages 37–44.

Petersen, S. E. and Ostendorf, M. (2007). Text simplification for language learners: a corpus analysis. In *Workshop on speech and language technology in education*. Citeseer.

Rauf, S. A., Ligozat, A.-L., Yvon, F., Illouz, G., and Hamon, T. (2020). Simplification automatique de texte dans un contexte de faibles ressources. In *Actes 6e conférence Traitement Automatique des Langues Naturelles (TALN), vol. 2*, pages 332–341.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992.

Sackett, D. L., Rosenberg, W. M., Gray, J. M., Haynes, R. B., and Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't. *British Medical Journal*, 312(7023):71–72.

Saggion, H., Gómez-Martínez, E., Etayo, E., Anula, A., and Bourg, L. (2011). Text simplification in simplext: Making texts more accessible. *Procesamiento del lenguaje natural*, (47):341–342.

Sakakini, T., Lee, J. Y., Duri, A., Azevedo, R. F., Sadauskas, V., Gu, K., Bhat, S., Morrow, D., Graumlich, J., Walayat, S., et al. (2020). Context-aware automatic text simplification of health materials in low-resource domains. In *Proc. of the 11th LOUHI Workshop*, pages 115–126.

Scarton, C., Paetzold, G., and Specia, L. (2018). Simpa: A sentence-level simplification corpus for the public administration domain. In *Proc. of LREC 2018*, pages 4333–4338.

Štajner, S., Franco-Salvador, M., Rosso, P., and Ponzetto, S. P. (2018). CATS: A tool for customized alignment of text simplification corpora. In *Proc. of LREC 2018*, pages 3895–3903.

Tonelli, S., Aprosio, A. P., and Saltori, F. (2016). SIMPITIKI: a Simplification corpus for Italian. In *CLiC-it/EVALITA*, pages 4333–4338.

Van den Bercken, L., Sips, R.-J., and Lofi, C. (2019). Evaluating neural text simplification in the medical domain. In *Proc. of the World Wide Web Conference*, pages 3286–3292.

Xu, W., Callison-Burch, C., and Napoles, C. (2015). Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Yimam, S. M., Štajner, S., Riedl, M., and Biemann, C. (2017). Multilingual and cross-lingual complex word identification. In *Proc. of the Int. Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 813–822.

Zhu, Z., Bernhard, D., and Gurevych, I. (2010). A monolingual tree-based translation model for sentence simplification. In *Proc. of the 23rd Intern. Conference on Computational Linguistics (COLING 2010)*, pages 1353–1361, Beijing, China.