Procesamiento del Lenguaje Natural, Revista nº 72, marzo de 2024, pp. 31-43 recibido 30-11-2023 revisado 18-01-2024 aceptado 01-02-2024

Enhancing the understanding of clinical trials with a sentence-level simplification dataset

Mejora de la comprensión de ensayos clínicos con un conjunto de datos simplificados a nivel de frase

Leonardo Campillos-Llanos,[1] **Rocío Bartolomé,**[2] **Ana R. Terroba Reinares**[3]
[1]ILLA (CSIC) [2]Fac. Filosofía y Letras (UAM) [3]Fund. Rioja Salud
leonardo.campillos@csic.es, rocio.bartolome@uam.es, arterroba@riojasalud.es

Abstract: We introduce a dataset with 1200 manually simplified sentences (144 019 tokens) from clinical trials in Spanish. A total of 1040 announcements from the European Clinical Trials Register (EudraCT) were analyzed to select sentences with ambiguities or exceeding 25 words. Simplification criteria were devised in an annotation guideline, which is released publicly along with the dataset. We obtained two versions: syntactically simplified sentences, and sentences with syntactic and lexical simplification. We report a quantitative, a qualitative and a human evaluation, in which three independent evaluators assessed the grammaticality/fluency, semantic adequacy and overall simplification. Results show that the resource is suitable for advancing research on automatic simplification of medical texts.
Keywords: Text simplification, Medical language processing, Clinical trials.

Resumen: Se presenta un conjunto de 1200 frases de ensayos clínicos en español simplificadas manualmente (144 019 tokens). Se analizaron 1040 anuncios del Registro Europeo de Ensayos Clínicos (EudraCT), seleccionando frases con ambigüedades o con más de 25 palabras. Se elaboraron criterios de simplificación recogidos en una guía distribuida públicamente con el conjunto de datos. Se obtuvieron dos versiones: oraciones simplificadas sintácticamente, y oraciones con simplificación léxica y sintáctica. Se presenta una evaluación cuantitativa, cualitativa y por tres evaluadores independientes sobre la gramaticalidad/fluidez, adecuación semántica y simplificación. Los resultados muestran que el recurso es adecuado para avanzar en la investigación en simplificación automática de textos médicos.
Palabras clave: Simplificación de textos, PLN médico, Ensayos clínicos.

1 Introduction

Achieving a plain language version of medical documents helps patients to enhance their understanding of health-related information and their adherence to treatment (Ondov, Attal, and Demner-Fushman, 2022). Potential participants in clinical trials might find eligibility criteria grammatically complex and rife with medical jargon (Wu et al., 2016), which hinders patients from taking part in a study. Automatic text simplification (Shardlow, 2014; Saggion, 2017), complemented with human supervision, has been shown to produce more understandable texts for patients (Lalor, Woolf, and Yu, 2019) and clinical researchers (Fang et al., 2021). Indeed, simplification also enhances (bio)medical language processing, given that such pre-processing makes it easier to parse coordinated or relative clauses (Peng et al., 2012) or complex compound phrases (Wei, Leaman, and Lu, 2014) before text mining.

To develop simplification systems for medical texts in Spanish, we created a dataset of 1200 manually simplified sentences from trial announcements. We release publicly a guideline and the resource in two versions: simplified sentences at the syntax-level, and with lexical and syntactical simplification.[1]

Figure 1 shows a sample of the original version of a trial announcement and its syntactical simplification. Long sentences in the technical version are shortened or split in the simplified version. Some nominalizations are changed to a verb or adjective form, which are easier to understand: e.g. *capacidad del*

[1]https://digital.csic.es/handle/10261/346579

ISSN 1135-5948 DOI 10.26342/2024-72-2

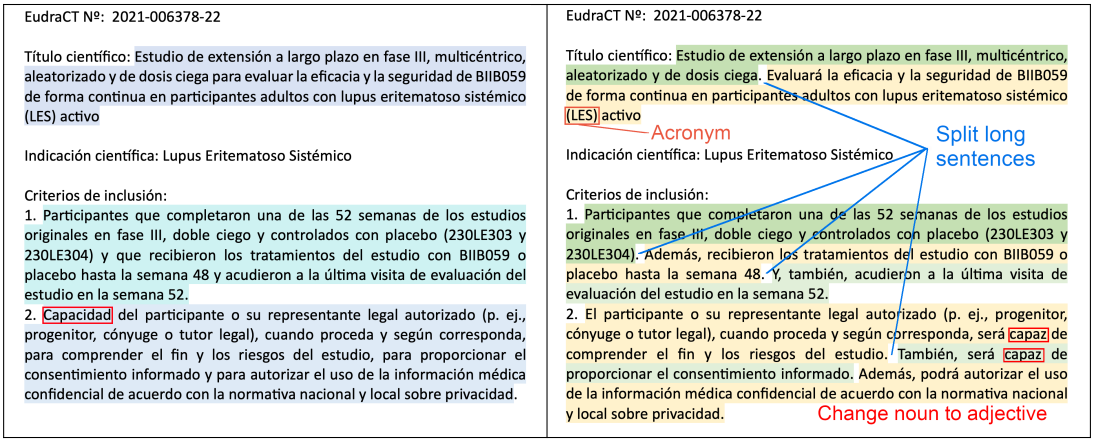

Figure 1: An unsimplified trial announcement (left) and its syntactic simplification (right).

participante... ('ability of the participant...') → *el participante será capaz...* ('the participant will be able to...'). Still, acronyms (*LES*) and technical terms (*multicéntrico*) require lexical simplification. The next sections report the background (§2), the methods (§3) and the evaluation (§4).

2 Background

Text simplification involves operations at all linguistic levels (lexis, syntax and discourse).

Methods for lexical simplification (Paetzold and Specia, 2017) generally rely on curated lexicons with technical and simplified words (Grabar and Hamon, 2016), paraphrase extraction (Elhadad and Sutaria, 2007; Deléger and Zweigenbaum, 2009), or machine learning-based approaches (Shardlow, 2013). Currently, deep learning methods are gaining ground through word-embeddings, prompt-based methods and large language models (LLMs), as explained in a recent survey (North et al., 2023). Lexical simplification has been addressed in the recent TSAR challenge (Saggion et al., 2023).

Syntactic simplification requires arranging words to achieve a word order with unambiguous references, split long sentences, change passive to active voice or rewrite nominalization structures to verb or adjective forms. Several works have used rules learned from corpora in order to apply simplification operations (Siddharthan, 2006; Peng et al., 2012; Seretan, 2012; Collados, 2013; Brouwers et al., 2014; Mukherjee et al., 2017). Most rules rely on dependency or part-of-speech tagging to derive simplification rules; for example, by parsing parallel sentences from technical and simplified texts (Szep et al., 2019). In contrast, other methods propose detecting syntactic simplification cues that do not rely on heavy syntactic analysis (Evans and Orăsan, 2019).

Discourse phenomena also require syntactic operations to simplify structures beyond the sentence and abridge long paragraphs. In addition, anaphora and co-reference might cause ambiguities to understand the content (Wilkens, Oberle, and Todirascu, 2020).

Lastly, texts may be simplified at all levels using transfer learning techniques (Menta and García-Serrano, 2022; Trienes et al., 2022; Alarcón, Martínez, and Moreno, 2023).

Evaluating simplification may be subjective (Grabar and Saggion, 2022), but standardized methods exist. However, quantitative approaches, such as readability formulae (Flesch, 1948), are not always adequate for medical texts (Zeng-Treitler et al., 2007). Moreover, metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) or SARI (Xu et al., 2016) are limited when assessing simplification, since they may correlate negatively with simplicity (Sulem, Abend, and Rappoport, 2018) or do not assess simplification operations thoroughly (Alva-Manchego, Scarton, and Specia, 2021). Human assessment of simplifications is thus beneficial.

Simplification tasks rely on lexicons or parallel (technical/simplified) corpora, which are scarce for Spanish (Segura-Bedmar and Martínez, 2017; Ferrés and Saggion, 2022; Alarcon, Moreno, and Martínez, 2023). Some were created in multilingual projects but are small (Xu, Callison-Burch, and Napoles, 2015; Martin et al., 2021; Joseph et al., 2023). We introduce a dataset to develop and test simplification tools. Table 1 shows samples.

Original	*Ensayo clínico para establecer los efectos de las dosis bajas de rtPA y los efectos de la reducción intensiva de la presión arterial en pacientes con accidente cerebrovascular isquémico agudo.* (2014-002823-86) 'Clinical trial to establish the effects of low-dose rtPA and the effects of intensive blood pressure lowering in patients with acute cerebrovascular accident'
Syntactic simplification	*Ensayo clínico para establecer los efectos de las dosis bajas de* ***rtPA*** *y los efectos de* ***reducir intensamente*** *la presión arterial. Se estudiará en pacientes con accidente cerebrovascular isquémico agudo.* 'Clinical trial to establish the effects of low-dose rtPA and the effects of lowering blood pressure intensively. This will be studied in patients with acute cerebrovascular accident'
Lexical and syntactic simplification	*Ensayo clínico para establecer los efectos de las dosis bajas de rtPA y los efectos de reducir intensamente la presión arterial.* ***rTPA es el activador recombinante del plasminógeno tisular****, un medicamento que ayuda a disolver los coágulos de sangre. Se estudiará en pacientes con accidente cerebrovascular isquémico agudo (****ictus****).* 'Clinical trial to establish the effects of low-dose rtPA and the effects of lowering blood pressure intensively. rtPA stands for recombinant tissue plasminogen activator, a medical drug that helps to dissolve blood clots. This will be studied in patients with acute cerebrovascular accident (stroke)'
Original	*Mujeres en tratamiento de TRA que reciban embriones propios o donados que presenten un desarrollo endometrial inferior a 5 mm a pesar de haber recibido un tratamiento con estrogenoterapia.* (2016-001716-38) 'Women in ART treatment that receive own or donated embryos and presenting an endometrial development less than 5 mm despite having received treatment with estrogen therapy.'
Syntactic simplification	*Mujeres en tratamiento de* ***TRA*** *que reciban embriones propios o donados. Las mujeres presentarán un desarrollo* ***endometrial*** *inferior a 5 mm a pesar de haber recibido un tratamiento con* ***estrogenoterapia****.* 'Women in ART treatment that receive own or donated embryos. These women will have an endometrial development less than 5 mm despite having received treatment with estrogen therapy.'
Lexical and syntactic simplification	*Mujeres en* ***tratamiento de reproducción asistida*** *que reciban embriones propios o donados. Las mujeres presentarán un desarrollo del* ***endometrio (capa del útero)*** *inferior a 5 mm a pesar de haber recibido una* ***terapia de estrógenos (hormonas)****.* 'Women in assisted reproductive treatment that receive own or donated embryos. These women will have a development of endometrium (the innermost layer of the uterus) less than 5 mm despite having received treatment with estrogens (hormones).'

Table 1: Samples of technical sentences and manually simplified (EudraCT id in brackets).

3 Methods

3.1 Data preparation

Three linguists analyzed trial announcements from EudraCT.[2] A set of 700 texts come from (Campillos-Llanos et al., 2021) and cover the period 2009-2020; and another set contains 340 texts (issued in the years 2020-2022). In total, we analyzed 1040 texts. However, we only used 510 trials (49.04%), because we discarded texts that were too long (above 1500 tokens), had lists with more than 10 lab values or had sentences that could not be simplified syntactically. Sentences with co-reference ambiguities, digressions or exceeding 25 words were selected (we followed a criterion supported by experts in Plain Language (da Cunha, 2022)). The criteria are detailed in §3.2 and Tables 2 and 3.

[2] `https://www.clinicaltrialsregister.eu`

APPO	**Appositive phrases**
Orig	*Sujetos, **varones y mujeres**, con diagnóstico de insuficiencia renal.* ('Subjects, men and women, diagnosed with renal failure.') (2014-001296-32)
Simp	*Sujetos con diagnóstico de insuficiencia renal.* ('Subjects diagnosed with renal failure.')
CONJ	**Conjunctions (coordination and subordination)**
Orig	*Diagnóstico por la imagen mediante fármacos radiactivos **con el objetivo de** localizar glándulas paratiroideas anómalas cuando las pruebas de imagen convencionales son negativas **y así** poder planificar de forma óptima el tratamiento quirúrgico.* ('Diagnostic imaging using radioactive pharmaceuticals to locate abnormal parathyroid glands when conventional imaging tests are negative, as a necessary condition for planning an optimal surgical treatment.') (2019-002729-31)
Simp	*Diagnóstico por la imagen mediante fármacos radiactivos **para** localizar glándulas paratiroideas anómalas cuando las pruebas de imagen convencionales son negativas. **Así** se podrá planificar de forma óptima el tratamiento quirúrgico.* ('Diagnostic imaging using radioactive pharmaceuticals to locate abnormal parathyroid glands when conventional imaging tests are negative. Surgical treatment can then be optimally planned.')
COREF	**Co-reference and anaphora**
Orig	*Ensayo clínico para la identificación de biomarcadores basados en técnicas ómicas (..), y su variabilidad inter e intraindividual **que** permitan la mejora en la individualización del tratamiento.* ('Clinical trial for the identification of biomarkers based on omics techniques (..), and their inter and intra-individual variability that allow the improvement in the individualization of treatment.') (2019-002795-13)
Simp	*Ensayo clínico para identificar biomarcadores basados en técnicas ómicas (..), y su variabilidad inter e intraindividual. **Estas técnicas** permitirían la mejora en la individualización del tratamiento.* ('Clinical trial to identify biomarkers based on omics techniques (..), and their inter- and intra-individual variability. These techniques would allow the improvement in the individualization of treatment.')
LEN	**Long sentences**
Orig	*Las pacientes fértiles deberán obtener resultado negativo en una prueba de embarazo en orina en las 24 horas previas a la primera dosis del fármaco del estudio.* ('Female subjects of childbearing potential must have a negative urine pregnancy test within 24 hours prior to the first dose of study drug.') (2019-001565-33)
Simp	*Las pacientes fértiles deberán obtener resultado negativo en una prueba de embarazo en orina. Se hará en las 24 horas previas a la primera dosis del fármaco del estudio.* ('Female subjects of childbearing potential must have a negative urine pregnancy test. This will be performed within 24 hours prior to the first dose of study drug.')
NEG	**Negation**
Orig	***No más de 1 año** antes de la fecha de inclusión.* ('No more than 1 year prior to enrollment.') (2015-003759-23)
Simp	***Un año o menos** antes de la fecha de inclusión.* ('One year or less prior to enrollment.')

Table 2: Syntactic simplification aspects according to linguistic criteria.

NOM	**Change nouns/adjectives to verb form**
Orig:	*Paracetamol en el **tratamiento** del dolor.* ('Paracetamol in the treatment of pain.') (2015-004482-88)
Simp:	*Paracetamol para **tratar** el dolor.* ('Paracetamol to treat pain.')
PAS	**Passive to active voice**
Orig:	*4 semanas previas a dosificación, o más si **es requerido** por las regulaciones locales.* ('4 weeks before dosage, or more if it is required by local regulations') (2016-001227-31)
Simp:	*4 semanas previas a dosificación, o más si las regulaciones locales lo **requieren**.* ('4 weeks before dosage, or more if local regulations require it.')
REDUN	**Redundancies**
Orig	*Se debe consultar al monitor médico antes de que el participante del **estudio** se incorpore al **estudio** AS0014.* ('The medical monitor must be consulted prior to the study participant's entry into the AS0014 study.') (2019-004163-47)
Simp	*Se debe consultar al monitor médico antes de que el participante se incorpore al **estudio** AS0014.* ('The medical monitor must be consulted before the participant enters into the AS0014 study.')
OVERS	**Oversimplification**
Orig	*En participantes sintomáticos, uno de los criterios para el diagnóstico de posible demencia frontotemporal de variante conductual o de subtipo semántico o de afasia progresiva primaria.* ('In symptomatic patients, one of the criteria for the diagnosis of probable behavioral variant FTD or FTD-semantic subtype or FTD-Progressive Nonfluent Aphasia.') (2019-004066-18)
Simp	*En participantes sintomáticos, **que tengan** uno de los criterios para el diagnóstico de posible demencia frontotemporal de variante conductual. **También, que tengan** posible demencia frontotemporal de subtipo semántico o de afasia progresiva primaria.* ('In symptomatic patients, participants who have one of the criteria for the diagnosis of possible behavioral variant of frontotemporal dementia. Also, participants who have possible frontotemporal dementia of semantic subtype or primary progressive aphasia.')
OTHER	**Other:** This label gathers aspects related to style, punctuation or grammar that enhance the clarity of the sentence or avoid ambiguities; these operations include fixing number or gender disagreement, preposition errors or unnatural word order.
Orig	*Aborto recurrente, preeclampsia previa o enfermedades hematológicas. Uso de **fármacos vasoactivos**: Fundamentalmente **relacionadas** con la hipertensión.* ('Recurrent miscarriage, previous preeclampsia or hematologic diseases. Use of vasoactive drugs: Fundamentally related to hypertension.') (2017-001878-42)
Simp	*Aborto recurrente, preeclampsia previa o enfermedades hematológicas. Uso de fármacos vasoactivos fundamentalmente relacionados con la hipertensión.* ('Recurrent miscarriage, previous preeclampsia or hematologic diseases. Use of vasoactive drugs mainly related to hypertension.')

Table 2: Syntactic simplification aspects according to linguistic criteria (cont.).

3.2 Simplification criteria

We followed the works by experts in plain language (da Cunha, 2022), the recommendations of the International Plain Language Federation,[3] the guideline prepared by the European Commission (European-Commission, 2016) and lexical simplification analyses (Koptient, Cardon, and Grabar, 2019; Carbajo and Moreno-Sandoval, 2023). We also applied the criteria defined in former work (Campillos-Llanos et al., 2022).

We provide two versions: syntactically simplified sentences, and sentences with syntactic and lexical simplifications. The version without lexical simplification is intended for research on syntactic simplification (e.g. development of a dedicated tool). The fully simplified one is provided for end-to-end systems that simplify sentences at all levels. A guideline gathers the simplification criteria.[4] Tables 2 and 3 show all simplification aspects.

3.3 Analysis and evaluation

To understand the distribution of topics across sentences, we used Medical Subject Heading (MeSH) Tree Entry Terms from the corresponding source text. Each EudraCT trial announcement has a MeSH descriptor (section E.1.1.2) of the therapeutic area. Nonetheless, these are not always accurate, and our topic distribution is only illustrative. We also counted the most frequent medical concepts in the sentences. Although the distributed dataset is not normalized to Concept Unique Identifiers from the Unified Medical Language System (Bodenreider, 2004), we used a lexicon (Campillos-Llanos, 2023) for the normalization used in this analysis.

To measure the quality of our simplifications, we conducted quantitative and qualitative measurements. First, we compared the word count, the number of syllables per sentence, the count of polysyllable words (with at least 3 syllables) and of monosyllable words in original and simplified sentences. We used `Textstat` (Bansal and Aggarwal, 2021). Simplified sentences should be shorter, have less syllables or less polysyllable words. We also compared the dependency tree height. This is a measure of structural complexity, given that more complex sentences have deeper syntactic dependency trees, as other teams showed (Alva-Manchego et al., 2020; Martin et al., 2020). The dependency tree depth should be shorter in simplified sentences. We computed this value with the Spacy `es_core_news_sm` model (vs. 3.3.3). For example, the following unsimplified sentence (with an apposition) has a token count of 7 and a dependency tree height value of 3: *Subjects, men and women, with renal insufficiency.* In contrast, the simplified version (without the apposition) has less tokens (4) and a shallower dependency tree height (2): *Subjects with renal insufficiency.* Figure 2 shows the dependency parsing of both sentences (obtained with Spacy).

To compare the lexical diversity of each simplification version, we computed the type-token ratio (TTR), a measure that has been used to describe other corpora (Trienes et al., 2022). The TTR is the proportion between unique tokens (*types*) and all tokens in a corpus. The higher the TTR value, the more lexically diverse a text is, and presumably more complex. We used a Python script.[5]

As a proxy for readability, we computed the average Inflesz score for each version (original, syntactic simplification and lexical and syntactic simplification). This is a perspicuity-based measure to estimate how clear and comprehensible a text is, according to the count of words, syllables and sentences. The Inflesz value was validated in Spanish health texts (Barrio-Cantalejo et al., 2008). The higher the score, the more readable the text is. We used a Python implementation.[6]

We did not use BLEU nor SARI since we did not compare the output of any simplification method with the human simplifications.

For a qualitative evaluation, three subjects (one linguist and two documentalists who were not involved in the simplification) assessed 100 random simplified sentences (50 with syntax simplification, and 50 with both the syntax and lexical simplification). They evaluated grammaticality and fluency (G/F), semantic coherence and meaning adequacy (M), and overall simplification (S), in line with previous work (Saggion et al., 2015; Koptient and Grabar, 2020). A 5-point Likert scale questionnaire was distributed (5 was the highest score, and 1 the lowest). We modified instructions originally prepared by other teams (Yamaguchi et al., 2023) to fit the Spanish language.

[3] `https://www.iplfederation.org/`

[4] `https://digital.csic.es/handle/10261/346579`

[5] Available at: `https://acortar.link/N49259`

[6] Available at: `https://acortar.link/9i8yF0`

ABBR	**Expanding abbreviations/acronyms**	
Orig:	*Tratamiento para el* ***MDE****.* (‘Treatment for MDE.’)	(2019-002704-41)
Simp:	*Tratamiento para el* ***episodio depresivo mayor****.* (‘Treatment for mayor depressive episode.’)	
ADD-LEX	**Adding a lexeme**	
Orig:	*Tolerabilidad de macitentan.* (‘Tolerability of macitentan.’)	(2013-003822-96)
Simp:	*Tolerabilidad del* ***medicamento*** *macitentan* (‘Tolerability of macitentan medical drug.’)	
DEL-LEX	**Deleting a lexeme**	
Orig	*Elvitegravir* ***(EVG)*** *administrado junto a darunavir.* (‘Elvitegravir (EVG) administered with darunavir.’)	(2013-001476-37)
Simp	*Elvitegravir administrado junto a darunavir.* (‘Elvitegravir administered with darunavir.’)	
HYP	**Replacement with a hypernym**	
Orig	*Ensayo clínico, simple ciego, aleatorizado, controlado y prospectivo.* (‘Single blind, randomized, controlled prospective clinical trial.’)	(2012-005571-14)
Simp	***Ensayo clínico de investigación.*** (‘Clinical research trial.’)	
PAR	**Paraphrase or definition**	
Orig	*Tratamiento con amikacina* ***intravenosa****.* (‘Treatment with intravenous amikacine.’)	(2014-001296-32)
Simp	*Tratamiento con amikacina* ***administrada en vena****.* (‘Treatment with amikacine administered into the vein.’)	
SYN	**Simpler synonym**	
Orig	***Profilaxis*** *habitual.* (‘Usual prophylaxis.’)	(2019-002233-11)
Simp	***Prevención*** *habitual.* (‘Usual prevention.’)	
TRANS	**Translation**	
Orig	*Test de embarazo de la visita de* ***screening****.* (‘Pregnancy test at the screening visit.’)	(2020-001901-22)
Simp	*Test de embarazo de la visita de* ***selección****.*	

Table 3: Lexical simplification aspects according to linguistic criteria.

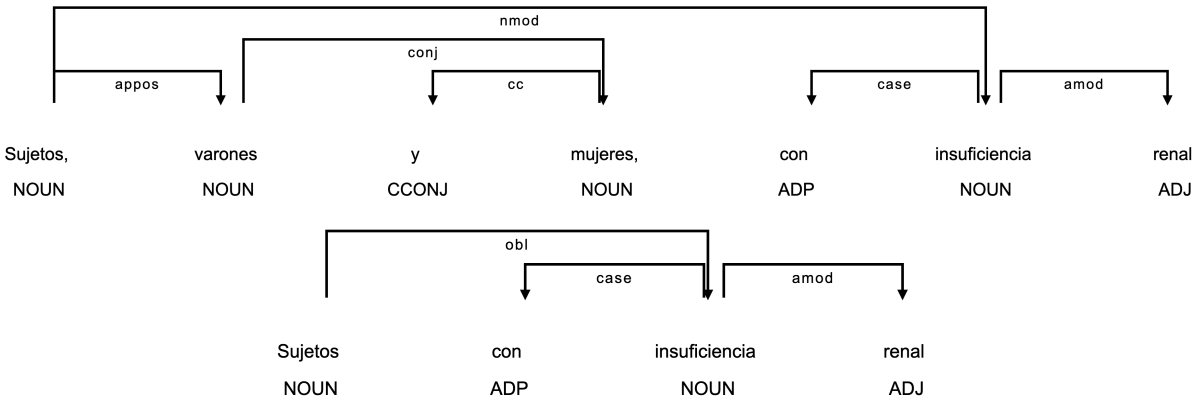

Figure 2: Dependency parsing of an unsimplified sentence with an apposition (*appos*) above and the simplified sentence (without the apposition) below; *ADP*: ‘adposition’ (∼preposition).

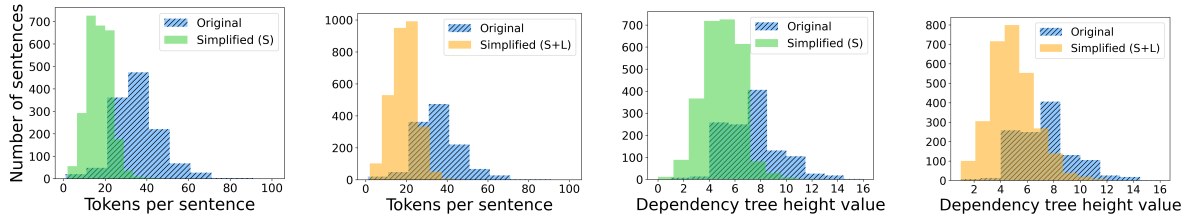

Figure 3: Tokens per sentence and dependency tree height values of original sentences, sentences with syntactic simplification (S) and with syntactic and lexical simplification (S+L).

	Original	**S**	**S+L**
Tokens (tk)	43 229	45 013	55 777
Types	5625	5669	5764
TTR	0.13	0.10	0.12
Avg tk/st	35.66 (±10.51)	17.31 (±3.91)	19.16 (±3.93)
Avg st	1.02 (±0.16)	2.21 (±0.62)	2.47 (±0.85)
Avg syl/st	66.55 (±19.45)	31.50 (±10.57)	34.48 (±11.97)
Avg mon/st	19.40 (±6.64)	9.03 (±3.87)	9.89 (±4.13)
Avg pol/st	8.63 (±2.65)	4.34 (±2.12)	4.61 (±2.37)

Table 4: Counts; *S*: syntactic simplification; *S+L*: syntactic and lexical simplification; *Avg*: average; *TTR*: type-token ratio; *st*: sentence; *syl*: syllables; *mon*: monosyllable words; *pol*: polysyllable.

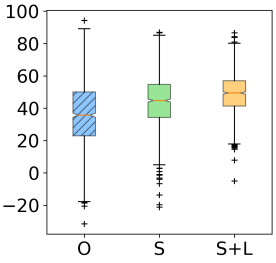

Figure 4: Inflesz scores.

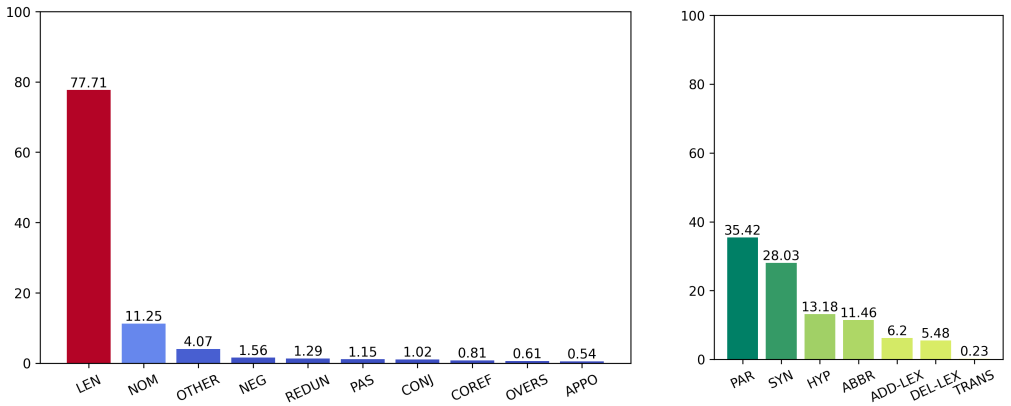

Figure 5: Distribution (%) of syntactic (left) and lexical simplification operations (right).

4 Results

4.1 Descriptive statistics

Table 4 shows descriptive statistics. The syntactically simplified version has shorter sentences and contain less syllables and polysyllable words. However, monosyllable words are more abundant in the original version. The number of simplified sentences tends to be slightly superior to the original version; indeed, many simplifications involved splitting long sentences. The fully-simplified version contains marginally longer sentences, more syllables and polysyllable words compared to the version with only syntactic simplification. Similarly, the TTR scores were lower in the syntactically simplified version. The original and the fully-simplified versions were more lexically diverse, possibly because they have more jargon or paraphrases, respectively.

With regard to readability (Figure 4), the average Inflesz score of the original version was of 35.69 (±19.72), which is interpreted as *Very difficult*. The syntactically simplified version has a higher score (44.19±15.42); and sentences with both syntactic and lexical simplification have a higher score (48.99±11.75). Scores of both simplified versions are considered *Somewhat difficult* in Inflesz.

Regarding the dependency tree height, the average value was of 7.13 in the original sentences; 4.80 in the syntactically-simplified

ones; and 5.06 in the sentences both syntactically and lexically simplified. Figure 3 shows the distribution of word count and dependency tree height values across versions. Statistical tests of these values, count of syllables, monosyllable and polysyllable words across versions showed statistically significant differences (Kruskall-Wallis, $p<0.0001$).

Overall, lexical aspects needed more simplification. Figure 5 shows the distribution of simplification operations. A total of 1476 syntactic aspects were simplified (an average of 1.23 operations per sentence). Shortening sentence length and changing nominal/adjective structures to verbs were the most frequent operations in our data. In turn, 2208 lexical aspects were simplified (an average of 1.84). Altogether, semantic-related lexical changes (paraphrasing, synonym and hypernym replacement) are estimated to represent up to 76.63% of lexical operations. Abbreviations/acronyms account for a 11.46%.

4.2 Health topics and concepts

Table 5 shows the 15 most frequent UMLS Concept Unique Identifiers (CUIs). Most refer to research tasks (*clinical trials, evaluation, randomization*), participants (*patients, study subjects*) and general entities about conditions or procedures (*disease, pharmaceutical preparations (prep.)*, *medicament*).

Freq	CUI and preferred term
2604	C0008976; Clinical Trials
2122	C0008972; Clinical Research
1875	C2603343; Study
1719	C0030705; Patients
1372	C0087111; Therapeutic procedure
857	C0220825; Evaluation
529	C0681850; Study Subject
490	C0034656; Randomization
438	C0008976; Clinical Trials
412	C0012634; Disease
353	C0013227; Pharmaceutical prep.
350	C0221423; Illness
348	C1510438; Assay
337	C0456386; Medicament
337	C0304228; Proprietary drug

Table 5: The most frequent CUIs.

Figure 6 plots the topic distribution of Medical Subject Headings (MeSH) in the EudraCT source texts (15 most frequent topics). Most sentences come from clinical trial announcements about cancer (19.69%), virus diseases (11.42%), nervous system diseases (9.25%) and cardiovascular diseases (7.48%).

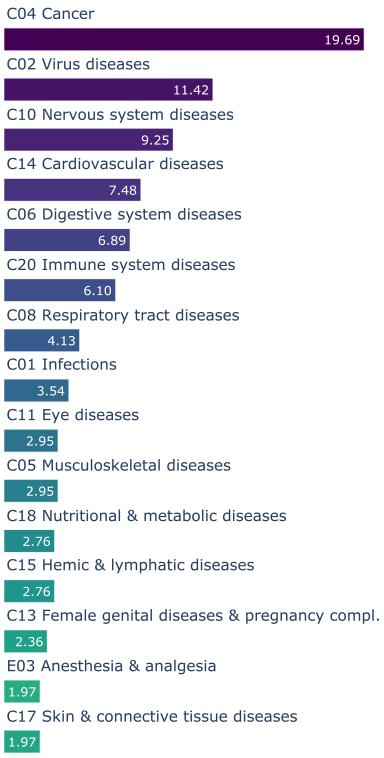

Figure 6: The 15 most frequent MeSH topics (%) in the EudraCT source texts.

	G/F	M	Sim	Avg
S	4.9	4.9	3.6	4.5
S+L	4.8	4.9	4.3	4.7

Table 6: Human evaluation; *G/F*: grammaticaly/fluency; *M*: meaning; *Sim*: simplification; *S*: syntactically simplified; *S+L*: syntactically and lexically simplified; *Avg*: average.

4.3 Human evaluation

Table 6 includes the evaluation results; we also include the average of the three aspects as in (Maddela, Alva-Manchego, and Xu, 2021). The average simplification scores (*Sim*) were slightly lower in the version with only syntactical simplification. Grammaticality and fluency aspects (*G/F*) were moderately similar. Still, some sentences from the version with both lexical and syntactic simplification were penalized due to many explanations in brackets that decreased the perceived readability. However, with regard

O: *Se estudiará* ***en el tratamiento de*** *pacientes* ((PRO) (V)) (PP ((PREP) (NP ((DET)(N) (PP ((PREP) (NP (N))))))	(NOM)
S: *Se estudiará* ***para tratar*** *pacientes* ((PRO) (V)) (PP (PREP) ((V) (NP (NOUN))))	
O: *Sujetos****, varones y mujeres,*** *con insuficiencia renal* (NP ((N) (PUNCT) (NP (N) (CCONJ) (N)) (PUNCT) (PP (PREP) (NP (N) (AP (ADJ))))))	(APPO)
S: *Sujetos con insuficiencia renal* (NP ((N) (PP (PREP) (NP (N) (AP (ADJ))))))	

Table 7: Samples of rules to change original (O) to simplified sentences (S). ADJ: 'adjective'; AP: 'adjective phrase'; DET: 'determiner'; N: 'noun'; NP: 'noun phrase'; PRO: 'pronoun'; PP: 'prepositional phrase'; $PREP$: 'preposition'; $PUNCT$: 'punctuation'; V: 'verb'.

to simplification, the fully simplified version received higher scores. This implies that both syntactic and lexical aspects achieved the best overall simplification, although we need to improve sentence fluency.

5 Discussion

Readers can not always understand medical documents due to long sentences, terminology and opaque acronyms. Simplifying medical texts needs to address syntactic and lexical aspects. However, any simplification task poses the challenge of guaranteeing to transmit the meaning of the text with precision.

We present a manually-simplified dataset for automatic simplification of medical texts. Our quantitative evaluation showed that sentence length, average tokens, syllables and polysyllable words per sentence, and dependency tree height values were lower in simplified sentences—i.e., these are less complex. Inflesz readability scores showed that simplified sentences are less difficult; still, according to this scale, they are *Somewhat difficult*. This is in line with our human evaluation, in which the overall simplification was rated in a 5-point Likert scale with lower scores (compared to fluency or semantic adequacy). All in all, there is still room for improvement, but the version with syntactic and lexical simplifications was rated better on average.

Our work has several limitations. First, the dataset size is small, which makes it difficult to train data-intensive approaches. More sentences need to be simplified by humans, which is a labor-intensive task. Second, the human evaluation could be subjective and is not strong (only 3 subjects assessed 100 simplified sentences, due to time constraints). Third, we did not test any syntactic simplification system. Some tools are only available for English (Mukherjee et al., 2017; Scarton et al., 2017; Chatterjee and Agarwal, 2021). Other tools for Spanish are not openly accessible (Ferrés et al., 2016). Creating (or re-adapting) a system for the Spanish language is out the scope of the present work. Lastly, although we described linguistic simplification aspects, we did not annotate abstract operations (e.g. `delete`, `add`, `move` or `replace`), as in other works (Bott and Saggion, 2011; Cardon et al., 2022).

On the whole, this is one of the few available resources for medical text simplification in Spanish. The dataset can be used to derive sentence-level simplification rules. Part-of-speech tagging the original and simplified sentences allows linguists to extract rules across registers, as other teams did (Seretan, 2012; Szep et al., 2019). Table 7 illustrates some samples of simplification rules.

6 Conclusion

We presented a dataset of 1200 sentences from clinical trial announcements in Spanish. Three experts simplified them manually according to criteria recorded in a guideline, which is shared publicly along with the dataset. We distribute a syntactically simplified version, and another with lexical and syntactic simplification. We reported descriptive statistics, an analysis of health topics and concepts, and a quantitative evaluation. A human evaluation showed that the simplified sentences are adequate, and the fully simplified version was assessed better. The main limitations are the small dataset size and the limited human evaluation.

Acknowledgements

This work was conducted in project CLARA-MED (PID2020-116001RA-C33) funded by MCIN/AEI/10.13039/501100011033/ (call: "Proyectos I+D+i Retos Investigación").

References

Alarcón, R., P. Martínez, and L. Moreno. 2023. Tuning bart models to simplify spanish health-related content. *Procesamiento del Lenguaje Natural*, 70:111–122.

Alarcon, R., L. Moreno, and P. Martínez. 2023. EASIER corpus: A lexical simplification resource for people with cognitive impairments. *Plos one*, 18(4):e0283622.

Alva-Manchego, F., L. Martin, A. Bordes, C. Scarton, B. Sagot, and L. Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proc. of the 58th ACL*, page 4668–4679.

Alva-Manchego, F., C. Scarton, and L. Specia. 2021. The (un) suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

Bansal, S. and C. Aggarwal. 2021. Textstat. `https://pypi.org/project/textstat/`.

Barrio-Cantalejo, I. M., P. Simón-Lorda, M. Melguizo, I. Escalona, M. I. Marijuán, and P. Hernando. 2008. Validación de la Escala INFLESZ para evaluar la legibilidad de los textos dirigidos a pacientes. 31(2):135–152.

Bodenreider, O. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Bott, S. M. and H. Saggion. 2011. Spanish text simplification: An exploratory study. *Procesamiento del Lenguaje Natural*, 47:87–95.

Brouwers, L., D. Bernhard, A.-L. Ligozat, and T. François. 2014. Syntactic sentence simplification for French. In *Proc. of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 47–56.

Campillos-Llanos, L., A. R. Terroba Reinares, S. Zakhir Puig, A. Valverde-Mateos, and A. Capllonch-Carrión. 2022. Building a comparable corpus and a benchmark for Spanish medical text simplification. *Procesamiento del lenguaje natural*, pages 189–196.

Campillos-Llanos, L. 2023. MedLexSp–a medical lexicon for Spanish medical natural language processing. *Journal of Biomedical Semantics*, 14(1):1–23.

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

Carbajo, B. and A. Moreno-Sandoval. 2023. Financial concepts extraction and lexical simplification in spanish. *(Under review)*.

Cardon, R., A. Bibal, R. Wilkens, D. Alfter, M. Norré, A. Müller, W. Patrick, and T. François. 2022. Linguistic corpus annotation for automatic text simplification evaluation. In *Proc. of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1842–1866.

Chatterjee, N. and R. Agarwal. 2021. DEPSYM: A Lightweight Syntactic Text Simplification Approach using Dependency Trees. In *CTTS@ SEPLN*, pages 42–56.

Collados, J. C. 2013. Splitting complex sentences for natural language processing applications: Building a simplified Spanish corpus. *Procedia-Social and Behavioral Sciences*, 95:464–472.

da Cunha, I. 2022. Un redactor asistido para adaptar textos administrativos a lenguaje claro. *Procesamiento del Lenguaje Natural*, 69:39–49.

Deléger, L. and P. Zweigenbaum. 2009. Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. In *Proc. of the 2nd Workshop on Building and Using Comparable Corpora*, pages 2–10.

Elhadad, N. and K. Sutaria. 2007. Mining a lexicon of technical terms and lay equivalents. In *Biological, translational, and clinical language processing*, pages 49–56.

European-Commission. 2016. *Cómo escribir con claridad*. Brussels: Directorate-General for Translation, Publications Office.

Evans, R. and C. Orăsan. 2019. Identifying signs of syntactic complexity for rule-based sentence simplification. *Natural Language Engineering*, 25(1):69–119.

Fang, Y., J. H. Kim, B. R. S. Idnay, R. A. Garcia, C. E. Castillo, Y. Sun, H. Liu, C. Liu, C. Yuan, and C. Weng. 2021. Participatory design of a clinical trial eligibility criteria simplification method. In *Medical Informatics Europe*, pages 984–988.

Ferrés, D., M. Marimon, H. Saggion, and A. AbuRa'ed. 2016. YATS: yet another text simplifier. In *Proc. of the 21st Int. Conf. on Applications of Natural Language to Information Systems, NLDB 2016*, pages 335–342. Springer.

Ferrés, D. and H. Saggion. 2022. ALEXSIS: a dataset for lexical simplification in Spanish. In *Proc. of LREC 2022*, pages 3582–94, Marseille, France.

Flesch, R. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Grabar, N. and T. Hamon. 2016. A large rated lexicon with French medical words. In *Proc. of LREC 2016*, pages 2643–2648, Portorož, Slovenia.

Grabar, N. and H. Saggion. 2022. Evaluation of automatic text simplification: Where are we now, where should we go from here. In *Actes de la 29e Conférence TALN*, pages 453–463.

Joseph, S., K. Kazanas, K. Reina, V. J. Ramanathan, W. Xu, B. C. Wallace, and J. J. Li. 2023. Multilingual simplification of medical texts. *arXiv preprint arXiv:2305.12532*.

Koptient, A., R. Cardon, and N. Grabar. 2019. Simplification-induced transformations: typology and some characteristics. In *BioNLP 2019*, page 309–318.

Koptient, A. and N. Grabar. 2020. Fine-grained text simplification in French: steps towards a better grammaticality. In P. Bath, P. Jokela, and L. Sbaffi, editors, *Proc. of Int. Symp. on Health Information Management Research*.

Lalor, J. P., B. Woolf, and H. Yu. 2019. Improving electronic health record note comprehension with NoteAid: randomized trial of electronic health record note comprehension interventions with crowdsourced workers. *Journal of medical Internet research*, 21(1):e10793.

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proc. of Workshop on Text Summarization of ACL*, pages 74–81, Barcelona, Spain.

Maddela, M., F. Alva-Manchego, and W. Xu. 2021. Controllable text simplification with explicit paraphrasing. In *Proc. of NAACL*, pages 3536–3553.

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

Martin, L., B. Sagot, E. de la Clergerie, and A. Bordes. 2020. Controllable sentence simplification. In *Proc. of LREC 2020*, pages 4689–4698, Marseille, France.

Menta, A. and A. García-Serrano. 2022. Controllable sentence simplification using transfer learning. *Proc. of the Working Notes of CLEF*.

Mukherjee, P., G. Leroy, D. Kauchak, S. Rajanarayanan, D. Y. R. Diaz, N. P. Yuan, T. G. Pritchard, and S. Colina. 2017. NegAIT: A new parser for medical text simplification using morphological, sentential and double negation. *Journal of biomedical informatics*, 69:55–62.

North, K., T. Ranasinghe, M. Shardlow, and M. Zampieri. 2023. Deep learning approaches to lexical simplification: A survey. *arXiv preprint arXiv:2305.12000*.

Ondov, B., K. Attal, and D. Demner-Fushman. 2022. A survey of automated methods for biomedical text simplification. *Journal of the American Medical Informatics Association*, 29(11):1976–1988.

Paetzold, G. H. and L. Specia. 2017. A survey on lexical simplification. *Journal of Artificial Intelligence Research*, 60:549–593.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. of the 40th Annual Meeting of the ACL*, pages 311–318.

Peng, Y., C. O. Tudor, M. Torii, C. H. Wu, and K. Vijay-Shanker. 2012. iSimp: A sentence simplification system for biomedical text. In *2012 IEEE International Conference on Bioinformatics and Biomedicine*, pages 1–6. IEEE.

Saggion, H. 2017. *Automatic text simplification*, volume 32. Synthesis Lectures on Human Language Technologies, Springer.

Saggion, H., S. Štajner, S. Bott, S. Mille, L. Rello, and B. Drndarevic. 2015. Making it Simplext: Implementation and evaluation of a text simplification system for Spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):1–36.

Saggion, H., S. Štajner, D. Ferrés, K. C. Sheang, M. Shardlow, K. North, and M. Zampieri. 2023. Findings of the TSAR-2022 shared task on multilingual lexical simplification. *arXiv preprint arXiv:2302.02888*.

Scarton, C., A. P. Aprosio, S. Tonelli, T. M. Wanton, and L. Specia. 2017. MUSST: A multilingual syntactic simplification tool. In *Proc. of the IJCNLP 2017, System Demonstrations*, pages 25–28.

Segura-Bedmar, I. and P. Martínez. 2017. Simplifying drug package leaflets written in Spanish by using word embedding. *Journal of biomedical semantics*, 8(1):1–9.

Seretan, V. 2012. Acquisition of Syntactic Simplification Rules for French. In *Proc. of LREC*, pages 4019–4026.

Shardlow, M. 2013. A comparison of techniques to automatically identify complex words. In *Proc. of the 51st annual meeting of the Association for Computational Linguistics*, pages 103–109.

Shardlow, M. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70.

Siddharthan, A. 2006. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4:77–109.

Sulem, E., O. Abend, and A. Rappoport. 2018. BLEU is not suitable for the evaluation of text simplification. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, editors, *Proc. of the 2018 EMNLP Conference*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Szep, A., M. Szep, G. Leroy, D. Kauchak, N. Kloehn, D. Revere, and M. Just. 2019. Algorithmic generation of grammar simplification rules using large corpora. *AMIA Summits on Translational Science Proceedings*, 2019:72–81.

Trienes, J., J. Schlötterer, H.-U. Schildhaus, and C. Seifert. 2022. Patient-friendly clinical notes: towards a new text simplification dataset. In *Proc. of the TSAR-2022 Workshop*, pages 19–27.

Wei, C.-H., R. Leaman, and Z. Lu. 2014. Simconcept: A hybrid approach for simplifying composite named entities in biomedicine. In *Proc. of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 138–146.

Wilkens, R., B. Oberle, and A. Todirascu. 2020. Coreference-based text simplification. In *Proc. of the 1st READI Workshop*, pages 93–100.

Wu, D. T., D. A. Hanauer, Q. Mei, P. M. Clark, L. C. An, J. Proulx, Q. T. Zeng, V. V. Vydiswaran, K. Collins-Thompson, and K. Zheng. 2016. Assessing the readability of clinicaltrials.gov. *Journal of the American Medical Informatics Association*, 23(2):269–275.

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

Xu, W., C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Yamaguchi, D., R. Miyata, S. Shimada, and S. Sato. 2023. Gauging the gap between human and machine text simplification through analytical evaluation of simplification strategies and errors. In *Findings of EACL 2023*, pages 359–375.

Zeng-Treitler, Q., H. Kim, S. Goryachev, A. Keselman, L. Slaughter, and C.-A. Smith. 2007. Text characteristics of clinical reports and their implications for the readability of personal health records. *Studies in health technology and informatics*, 129(2):1117.