UNIVERSIDAD
COMPLUTENSE
MADRID

MÁSTER EN LINGÜÍSTICA Y TECNOLOGÍAS
ESPECIALIDAD LINGÜÍSTICA COMPUTACIONAL

TRABAJO DE FIN DE MÁSTER
Curso académico 2023-2024

Lexical Simplification in Spanish Texts for Patients: The Complex Word Identification Task

APELLIDOS Y NOMBRE: Federico Ortega Riba

DNI: 77440656R

TUTOR: Dr. Leonardo Campillos Llanos

COTUTORA: Dra. Doaa Samy Khalil Shawer

CONVOCATORIA: Septiembre

DECLARACIÓN DE INTEGRIDAD ACADÉMICA

El abajo firmante, D. Federico Ortega Riba con D.N.I. nº 77440656R

Declaro que el presente trabajo, titulado *Lexical Simplification in Spanish Texts for Patients: The Complex Word Identification Task* entregado en el tiempo y forma previstos como Trabajo de Fin de Máster del *Máster en Lingüística y Tecnologías* es fruto de mi investigación y trabajo personal, y que en él no constan contenidos ni materiales cuyas fuentes no estén claramente identificadas y citadas en el cuerpo del texto o en la bibliografía.

Entiendo, por tanto, que incurriré en plagio si se diera, entre otras, alguna de las siguientes circunstancias:

- Entregar un trabajo ajeno como si fuera propio.
- Copiar un texto literalmente o parafrasearlo sin citar la fuente consultada.
- Entregar un trabajo copiado, en todo o en parte, de fuentes impresas o electrónicas.

Comprendo que el plagio es una grave ofensa académica que puede tener consecuencias negativas en la calificación de mi TFM.

Y para que conste, firmo esta declaración en Boulder, a 8 de septiembre de 2024

Firmado: Federico Ortega Riba

ORTEGA
RIBA
FEDERICO -
77440656R

Digitally signed
by ORTEGA RIBA
FEDERICO -
77440656R
Date: 2024.09.08
10:37:07 -06'00'

To my family, for always supporting me during my education, even though they still
don’t understand what I do

To Ger, for encouraging me to dream big

To Leo, for believing in a translator

Abstract

Medical texts are often filled with technical jargon, which can be challenging to understand, especially for patients with low health literacy levels. In a healthcare system where time limitations lead to difficulties for understanding techniques, pathologies and medical instructions, lexical simplification (LS) can help to allow information accessibility and increase informed decision-making. Additionally, in a highly dynamic and context-dependent framework as medical discourse, linguists are crucial for the standardization of knowledge to ensure terms are clear and unambiguous. Our research focuses in the automatic complex word identification (CWI) task in Spanish text for patients. This is the first step in the pipeline of current lexical simplification methods and the starting point to bridge the language gap between healthcare providers and patients. Identifying complex words enables simplification and substitute generation, used to tailor content to the reader's comprehension level. To develop the task, an annotation scheme for medical texts is consolidated to compile a single and multi-word corpus which will serve as our dataset for CWI. The corpus for the fine-tuning task (train and development splits) consists of 60 clinical trials (CTs), 60 consent forms (CFs) and 60 patient information documents (PIDs). To examine language and domain variation, we tested both monolingual and multilingual, as well as general domain and domain-specific models. The tests were conducted with 15 new CTAs, 15 CFs, and 15 PIDs, achieving an overall average F1 score of 76.13, and an F1 score of 79.02 for the model with the best performance.

Keywords: *Natural language processing, computational linguistics, lexical simplification, Spanish language, data annotation, medical terminology*

Resumen

Los textos médicos suelen presentar términos que pueden resultar difíciles de entender, sobre todo para aquellos pacientes con niveles bajos de alfabetización sanitaria. En un sistema en el que las restricciones de tiempo dificultan la comprensión de procedimientos, patologías e indicaciones médicas, la simplificación léxica puede ayudar a facilitar el acceso a la información. Además, en un entorno tan dinámico como el discurso médico, los lingüistas juegan un papel fundamental a la hora de ofrecer una estandarización del conocimiento y garantizar una terminología sin ambigüedades. Nuestra investigación se centra en la tarea de identificación automática de palabras complejas (CWI, en inglés) en textos en español para pacientes. Este es el primer paso en las investigaciones actuales sobre simplificación léxica, así como el punto de partida para eliminar la brecha lingüística entre los profesionales sanitarios y los pacientes. La CWI permite la simplificación y la generación de sustitutos, utilizados como herramientas para adaptar los contenidos al nivel de educación sanitaria del paciente. Para el desarrollo de la tarea, se ha consolidado una guía de anotación de textos médicos con el objetivo de compilar un corpus compuesto por términos simples y multipalabra. El conjunto de datos para la parte del ajuste consta de 60 ensayos clínicos (CT, en inglés), 60 consentimientos informados (CF, en inglés) y 60 documentos con información para pacientes (PID, en inglés). Para analizar la variación en cuanto a idioma y grado de especialización de los modelos utilizados, se han comparado modelos monolingües y multilingües, así como de dominio general y de dominio médico. Las pruebas finales se realizaron con 15 CTA, 15 FC y 15 PID nuevos, alcanzando una puntuación F1 con una media global de 76.13, y con una puntuación de 79.02 para el modelo con mejores resultados.

Palabras clave: *Procesamiento del lenguaje natural*, *lingüística computacional*, *simplificación léxica*, *español*, *anotación de datos*, *terminología médica*

Contents

1. Introduction

Lexical simplification focuses on improving accessibility to textual information by replacing complex words with simpler alternatives (Saggion & Hirst, 2017). When it comes to healthcare information, this process becomes even more important, since patients sometimes have limited vocabulary and difficulties understanding their diagnoses (Williams et al, 1995; Makaryus & Friedman, 2005). Medical discourse analysis involves examining the ways in which doctors and patients take part in different contexts and purposes. An in-depth analysis on the medical discourse's fluctuations as well as the individual's reading comprehension (Shahid et al., 2022) can reveal how language shapes patient relationships and the communication of knowledge in the medical field. To illustrate how medical conditions and treatments are described in professional texts, we could consider the differences between a consent form from a surgical stomach procedure and the directions to prevent stomach cancer in a leaflet. The clear shifts in tone, terminology, structure and other language features would provide enough insights to understand communication effectiveness with laymen patients reading both texts (Bellés-Fortuño, 2016; Montalt & García-Izquierdo, 2016).

Despite the efforts from healthcare providers to ensure clinical descriptions of symptoms and diagnoses, time constraints and limited consultations times are an important aspect that could hinder the communication in this respect (Jabour, 2020). In many countries across Europe, healthcare appointments are brief and doctors might feel rushed to complete their schedule on time. The time pressure also contributes to incomplete explanations, creating a gap in communication. These issues, alongside low health literacy levels, hinder patients' ability to understand relevant information (Roter, 2006).

In order to overcome these barriers, the first step in the lexical simplification approach presented in our paper consists of the annotation of complex words. The annotation task may be quite challenging even for linguists, and the use of terminologies and ontologies in lexical simplification within the field is crucial (Finlayson & Erjavec, 2017). These types of resources serve as a powerful tool for precise definitions of medical concepts and are essential to ensure that simplified terms convey the original semantics of a complex word. Similarly, terminologies and ontologies help to categorize and structure specialized information, enabling an easier identification of key elements in a medical text.

Once the annotation of complex words has been conducted following guidelines from standardized terminological sources, the computational process of lexical simplification starts. It typically involves the automatic identification of words, a proposal for simpler synonyms and the selection of the most appropriate word based on context, semantics and sentence readability (Moen, 2018; Qiang et al., 2020). In this research work, we focus on the first task: CWI in Spanish texts for patients. With aligned tokens and the annotations reaching a gold standard, a corpus of medical complex words is compiled. The alignment of complex words ensures that each token of a sentence corresponds to a complex word or a simple word.

The corpus files are divided into three sets of training, validation and test. Subsequently, the annotated corpus was used to fine-tune a pretrained transformer model specifically for this task (Vaswani et al., 2017). Transformer models have demonstrated superior performance for a wide range of NLP tasks, unlike other machine learning (ML) models, such as support vector machines (SVM) and conditional random fields (CRF). The self-attention mechanisms of transformer models allow them to capture complex relationships in the dataset, and have also outperformed other ML methods on benchmark datasets (Wolf et al., 2020). Moreover, these state-of-the-art models are robust to variations in real-world medical data, learning faster to ignore noise and focus on relevant linguistic patterns. The transformers models subject to fine-tuning are selected and the parameters adapted to our task. For the purposes of this research, six suitable transformer-based models were selected, paying special attention to two parameters: language and domain. These two factors have been crucial to test the performance after fine-tuning with our corpus. Therefore, our work will be useful to evaluate linguistic nuances and multilingual capabilities of the selected models, as well as whether a lack of original exposure to terminology and contextual usage may affect how a model detects complex words in Spanish texts. During this process, the models were tested on CWI within three different text typologies: CTs, CFs and PIDs. The performance of the models was measured using common evaluation metrics (Hripcsak, 2005), and post-evaluation of results was conducted to analyze false positives, false negatives and assess model performance. A discussion is established afterwards to interpret the key findings of the study and compare them with previous research in the medical field, highlighting similarities and differences.

The structure of the work is as follows. The second section is a literature background where linguistic features in medical discourse are analyzed, and the CWI task is introduced. In the third section, our methods for the fine-tuning of transformer-based models are described, alongside the annotation guidelines followed to compile our corpus. Lastly, our results are shown in the fourth section, followed by a discussion on future work and our conclusions.

2. Background

2.1. Medical discourse's dichotomy

One of the main difficulties in a medical context is defining the variabilities of terminology used between professionals and patients. According to the ISO 1087 *Terminology work and terminology science – Vocabulary* (2019) a term is a "designation that represents a general concept by linguistics means". This term may be a single word or a multiword expression, which cannot be divided without losing its meaning. For instance, *give birth* is an English multiword and non-separable expression which means *to have a baby*[1]. In Spanish, personal pronouns may be added to a word, resulting in words fusion such as *extirpárselo* ('remove it from them') or *dializarle* ('dialyze them').

Additionally, Latinisms, Greek and Latin roots are frequently used as well, as they are really productive in medical language: some of the most used prefixes are *cardio-* for heart diseases, *auto-* for *self* or *neuro-*/*neur-*/*nerv-* related to the nervous system. Provided that both paradigmatic and syntagmatic variations play a major role in knowledge representation (Faber, 2012), linguists can facilitate specialized communication and knowledge transfer between specialized users and patients. According to Cabré (2003), terminological units' membership to a general or specialized domain depends on three main aspects: cognition, syntax and pragmatics. This explains the intersection between general and specialized discourse: these cannot be separated in two water-tight compartments, in which both are multidimensional, as a specialized discourse may contain another one. Building on this idea, it is evident that understanding the interplay between general discourse among patients and specialized discourse among healthcare providers is essential for effective communication.

Let us think of three medical interactive settings to illustrate how these aspects can help distinguish domains. Taking into account that finding literal conversations from publicly available sources is quite challenging due to privacy policies and the specific nature of such dialogues, the examples provided are not extracted from a real transcription. They have been AI-generated with ChatGPT (OpenAI, version of 2024) and manually modified following the analysis of medical discourse in literature instead, such as those presented in Wodak (2006), Abdramanova (2023) and De Belder (2013).

[1] https://www.merriam-webster.com/dictionary/give%20birth (Retrieved June 15, 2024)

A. Patient-doctor communication

Conversation example	Insights of medical discourse features
Dr. A: Good morning, what can I do for you? Patient A: Good morning, doctor, I have been feeling dizzy since I woke up, and I have a stomach ache. Dr. A: I'm going to do a physical examination now to get a better idea of what's going on. Please lie down on the examination table and lift your shirt a bit. *[after the examination]* Dr. A: I'm going to ask you to sit up now. Based on your symptoms and my examination, there are a few possibilities we need to consider. To get a clearer picture, I'd like to order some blood tests and possibly an ultrasound of your abdomen.	At first, this seems like a general discourse interaction. However, if this patient is finally diagnosed with *diverticulitis*, they may need an explanation of the pathology. The patient may not know that medical terminology often uses the suffix -*itis* to indicate inflammation (Abramanova, 2023), but in all likelihood, they might guess a health problem is involved, and they will need to know what a *diverticulum* is. A successful therapeutic conversation would mean the patient knows they have an inflammation in the wall of the large intestine, creating a semi-specialized language context in which *stomach ache* and *diverticulitis* are involved. In terms of semantic properties, one is more precise than the other, the term *diverticulitis* helps to describe a connection between the event of abdominal pain and the nature of the pathology.

Table 1. Conversation example and insights of medical discourse features in a patient-doctor dialogue.

B. Gastroenterologist doctor-doctor informal communication

Conversation example	Insights of medical discourse features
Dr. A: hello, Dr. B, what do we have today? Dr. B: So, Patient A has CD. He's been experiencing chronic inflammation in the gastrointestinal tract, particularly affecting the ileum and parts of the colon. His symptoms include abdominal pain, severe diarrhea, fatigue, and weight loss. We've started him on a biologic therapy to help manage the inflammation, and we're closely monitoring his response to treatment.	In this example, general and specialized discourse start to combine. There are certain elements that characterize informal language in this extract: • Conciseness with clarity: the use of a straightforward and simple statement reflects the informal setting. There is no attempt to use overly technical jargon or unnatural formal medical terminology. • Conversational tone: the use of contractions and listing symptoms in a straightforward manner makes the information accessible and easy to understand. Direct address: the use of direct and personal language such as "Patient A" and "he" creates a sense of immediacy and personal involvement in the patient's care.

Table 2. Conversation example and insights of medical discourse features in a casual conversation between two gastroenterologists.

C. Ophthalmologists doctor-doctor formal communication

Conversation example	Insights of medical discourse features
Dr. A: In our recent study involving 116 patients diagnosed with retinitis pigmentosa at the University of Bonn, we utilized next-generation sequencing to analyze the genetic underpinnings of the disease. This method has significantly improved our diagnostic yield. Specifically, we identified pathogenic mutations in 70% of our cohort, a notable increase compared to previous methodologies. We found 110 different mutations across 30 genes, with 46 of these being novel at the time of diagnosis. Dr. B: That's impressive. How did you ensure the accuracy of the NGS data?	This dialogue uses precise scientific terminology to ensure clarity and specificity. It is highly specialized and few connotations of general discourse are depicted, like Dr. B's reaction to the topic of Dr. A. The formality of the text is conveyed through the following structures: • Terminology: *next-generation sequencing* (*NGS*), *pathogenic mutations*, and *in silico prediction tools* are specific to genetics and molecular biology, ensuring that both doctors understand the methods and implications. • Structured and planned discourse: descriptions of sequencing technologies provide a clear picture of the procedures followed. Doctor A emphasizes the robustness and accuracy of the results. Based opinion: phrases like *diagnostic yield* and *novel mutations* convey the advancements made through the use of NGS, underlining the practical benefits and future potential of these technologies.

Table 3. Conversation example and insights of medical discourse features in a formal conversation between two ophthalmologists.

In Table 1, it is shown that, in a conversation between a patient and a doctor, the latter should use their interpersonal and communication skills to enable understanding and give therapeutic instructions. In this sense, a layman patient can refer to a doctor that they are suffering from a pathology in simple words. However, when a doctor-doctor dialogue takes place (Tables 2 and 3), it is possible to encounter grammatical structures which are more obscure for laymen patients, yet are on the same level of communication for both agents taking part in the conversation. This is the case for the second and third examples, where technical terminology is not simplified or explained, such as *CD* or *NGS* owing to the fact that same-type healthcare specialists are involved in a conversation. Furthermore, in an academic context of medicine, those abbreviations are also used, yet a higher register is preferred.

In addition to these three contexts studied above, diseases like diabetes mellitus, HIV, and COVID-19 have become widely understood by the general public due to their high prevalence, significant impact on public health, and extensive media coverage. For instance, according to a report from the World Health Organization (2023), diabetes affects over 422 million people globally, and has prompted widespread educational initiatives about its management and prevention. On the other hand, HIV and AIDS, since their identification in the early 1980s, have been the focus of major international health campaigns, which have resulted in significant public knowledge about transmission and prevention, as the UNAIDS report of 2023 claims. Last but not least, the COVID-19 pandemic, due to its rapid global spread, has dominated public discourse since 2019, with continuous updates and guidelines from health authorities like the WHO and the CDC.

In contrast, rarer conditions, listed in the National Organization for Rare Disorders, such as acromegaly[2] or amyloidosis[3] do not receive the same level of public attention. Acromegaly, a hormonal disorder caused by excess growth hormone, and amyloidosis, a group of diseases where abnormal protein deposits build up in organs and tissues, affect fewer people. Consequently, they lack the widespread public media coverage that drives general understanding. These diseases are often only discussed within specialized medical communities or among those directly affected, resulting in lower public awareness.

After reviewing three examples of medical discourse variation in medical and oral settings, it is shown that the best way to study specialized knowledge units is by studying their behavior in different scenarios. For the purpose of this work, only the three dimensions described above have been studied, nevertheless, there are other scenarios in which the combinations of general and specialized discourse may vary: a doctor-resident conversation, an ophthalmologist-gastroenterologist conversation, a parent-doctor conversation or a student-doctor conversation. Whether it is a text or speech, some terms, phrases or sentences do not belong to only one specialized domain, they can be present in more than one and behave in different ways. As Faber (2012) states:

> Understanding a terminology-rich text requires knowledge of the domain, the concepts within it, the propositional relations within the text, as well as the conceptual relations between concepts within the domain. (p.8).

2.2. Texts for patients

Real life doctor-patient interaction has the advantage of rephrasing, asking questions and providing answers in a short period of time. Nonetheless, laymen patients often lack the

[2] https://rarediseases.org/rare-diseases/acromegaly/ (Retrieved June 20, 2024)

[3] https://rarediseases.org/rare-diseases/amyloidosis/ (Retrieved June 20, 2024)

specialized knowledge required to understand a doctor’s report or laboratory tests on their own, and medical terminology might create a linguistic barrier for non-specialized users who sometimes require clarification of complex topics. In addition to these difficulties, doctors often lack enough time to facilitate their patients’ health understanding about medication, pathologies or techniques, in or outside of a consultation. This explains the relevance of health literacy and the importance of providing health services targeting patients care. In such use cases, complex word identification tasks applying NLP techniques, as the task studied in this dissertation, could play an essential role in use cases targeting patients care.

Described as the capability to effectively understand and use health information, health literacy is fundamental for patients to interpret complex medical terminology that is often encountered in healthcare settings. According to Ratzan (2001), the term *health literacy* was coined in 1974 and focuses on an individual's ability to understand the complex demands of health maintenance in modern society. Over the last two decades, the idea has gained attention due to its advantages for both individual and public health. In light of the increasing prevalence associated with non-communicable diseases, such as cancer or heart diseases (WHO, 2023), there is an existing need for laymen patients to take greater responsibility for managing their health. Chen Liu et al. (2020) define the risks of low health literacy as follows:

> Inadequate health literacy is associated with difficulties in comprehension of health information, limited knowledge of diseases and lower medication adherence, which contribute to poor health, high risk of mortality, insufficient and ineffective use of healthcare, increased costs, and health disparities. (p. 1).

Notwithstanding the relevance of a healthcare professional to identify pathologies, a medium health literacy level empowers patients to better understand their diagnoses, treatment plans, and medication instructions, which is crucial for their overall health management. In the context of NLP, this means that deep learning applications designed to assist patients must be capable of accurately recognizing complex medical terms to provide clear, comprehensible information which allows lexical simplification. Enhancing health literacy is therefore pivotal for developing NLP applications that can bridge the gap between complex medical language and patient understanding, leading to improved health outcomes and patient empowerment.

Consequently, many countries, including the USA, Canada, Australia, China and countries from the European Union, have prioritized health literacy in their policies and practices. The World Health Organization also endorses health literacy as one of the main agents in achieving sustainable development goals. In line with Juvinyá-Canal et al. (2018), specifically in Europe, health literacy is seen as a vital component of the European health strategy. This strategy requires the involvement of patients all over the world due to the increasing complexity of health information and its accessibility, particularly via the internet. As specified in the Eurobarometer report (European Union, 2017), a majority of Europeans now turn to the internet when seeking answers to health-related questions, and they believe these answers to come from trust-worthy sites. Therefore, policies promoting digitalization and e-health must also consider the digital health literacy levels of both the general population and healthcare professionals to ensure the reliability of online health information. The Spanish healthcare policies highlight patients' awareness since 2019, when the Spanish Health Literacy Network started collaborating in research projects. Communication between health professionals, patients and caregivers have been of first importance ever since, and members of the association have focused on topics such as heart disease (Falcón et al., 2022b) or COVID-19 (Falcón et al., 2022a), as well as other campaigns such as *Health without doubts* (Fernández et al., 2021b) or *Always ask three questions* (Fernández et al., 2021a), all of them available at the Spanish Health Literacy Network's website[4].

The role of linguistics in simplifying medical texts is vital as it bridges the gap between complex information and general understanding for laymen patients, enhancing health outcomes. Making terminology accessible to a diverse population ensures that health information is tailored for the varying literacy levels of patients. The idea that linguists can notably contribute to technical fields has been underpinned for decades. Sager et al. (1980, p. 13) already covered the principles and practices of special languages, including medical terminology, and claim that "it is essential for the individual to have access to complex fields of knowledge and science if he is to take a rational part in societal development". The way in which linguists help patients make informed decisions about their health, adhere to medical advice and engage with healthcare providers can be explained based on the following reasons:

[4] https://www.alfabetizacionsalud.com/que-hacemos/investigacion/ (Retrieved August 14, 2024)

1. Clarity and precision: language experts ensure that medical terms are clear, reduce ambiguity and lower the risk of misinterpretation in medical communication. Cimino (1998) proposed an example of context-sensitive ambiguity and context-independent ambiguity by explaining the concept of *myocardial infarction*. This concept could mean *right ventricular infarction* or *left ventricular infarction*; however, the pathophysiologic process does not vary. This does not apply to the term *diabetes*, which can also be specified by adding *mellitus*, *gestational*, *neonatal*, *type 1 or* 2, with a different pathological process. This example shows that, even with a fine-grained concept like this, there will always be some finer-grained ones. On top of this, it is possible to encounter abbreviations like *MI*, which are clearly ambiguous as they could mean *myocardial infarction* or *mitral insufficiency*. Whereas ambiguity can be allowed in the vocabulary for specialized physicians, as seen in the example of Cimino, language experts can have a major role in reducing the unequivocal meaning of an abbreviation based on context.
2. Standardization: linguists contribute to the standardization of medical knowledge, and facilitate consistent terminology across different languages and regions, which is essential for global health communication. The Unified Medical Language System, UMLS, (Bodenreider, 2004)[5] is a clear example of how providing a consistent categorization of concepts helps in information retrieval tasks within a robust framework validated by the scientific community. This topic will be further analyzed in section 2.3.

2.3. Terminological resources

Terminology extraction for text annotation requires an extensive overview of various medical thesauri, classification systems, and standards crucial for the registration of medical information. Since no single terminology serves all purposes, the resources reviewed for the task presented in this article include detailed descriptions and applications of each system within clinical contexts. This section summarizes the main terminological knowledge studied for this task.

As observed in 2.2., linguistic variation with acronyms, synonyms or hypernyms requires standardization for a normalized representation of information. The main objective is to preserve the integrity of the message in its sense and interpretation. In order

[5] https://uts.nlm.nih.gov/uts/umls/home (Retrieved August 14, 2024)

to understand how terminological or terminographical resources work, a difference between three categories should be made:

1. Controlled vocabularies and thesauri: thesauri are controlled vocabularies that include preferred terms, synonyms, brief definitions, and descriptors for indexing, or information retrieval. Ontologies, which go a step further, incorporate relationships between concepts.
2. Hierarchies and classifications: multiple classifications, such as the Medical Subject Headings (MeSH)[6], organize terms in various hierarchical structures. For example, *pulmonary abscess* can be found under both *bacterial infections and mycoses* and *respiratory tract diseases* categories.
3. Monaxial vs. multiaxial terminologies: monaxial terminologies describe a single type of aspect, whereas multiaxial terminologies describe multiple aspects on different axes, allowing for post-coordination and qualifiers.

A crucial part of the terminology annotation task of this job consisted in finding key medical classification systems and using their standardized terminology to identify complex words. Some of the most useful have been:

1. The International Classification of Diseases (ICD)[7].
2. The Medical Subject Headings (MeSH).
3. The SNOMED Clinical Terms (SNOMED CT)[8].
4. The Medical Dictionary for Regulatory Activities (MedDRA)[9].
5. The Anatomical Therapeutic Chemical (ATC) Classification[10].

Each resource fulfills a different purpose, from biomedical and health-related information to pharmacovigilance in drug regulation. All these tools are integrated in the UMLS metathesaurus, which provides a comprehensive framework to bring together the biomedical ontologies. Campillos-Llanos (2023) provides more information on thesauri and ontologies for Spanish medical language processing in case a further explanation is needed.

[6] https://www.nlm.nih.gov/mesh/meshhome.html (Retrieved August 14, 2024)

[7] https://www.who.int/standards/classifications/classification-of-diseases (Retrieved August 14, 2024)

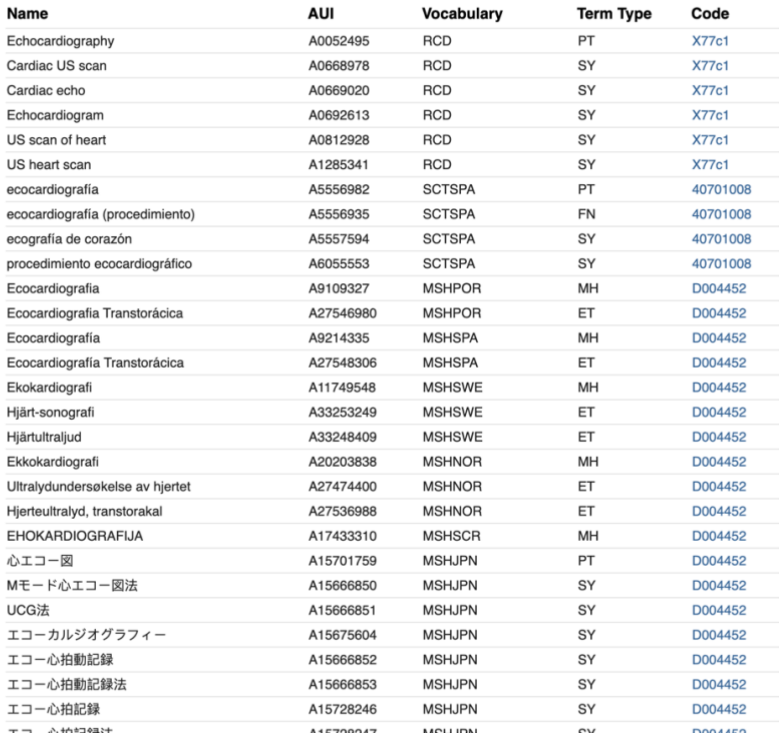
[8] https://www.snomed.org/ (Retrieved August 14, 2024)

[9] https://www.meddra.org/search (Retrieved August 14, 2024)

[10] https://www.who.int/tools/atc-ddd-toolkit/atc-classification (Retrieved August 14, 2024)

In addition to all these resources, each contributing significantly to the depth and breadth of the study, the insights gained from the UMLS provided a foundational understanding and guided the research direction. The UMLS consists of three principal components: a metathesaurus which integrates various terminologies and ontologies already mentioned, a semantic network that provides a categorization of all concepts, and a specialist lexicon of biomedical terms with syntactic, morphological, and orthographic information supporting NLP applications.

It is important to bear in mind that the UMLS metathesaurus is not a final user application and needs to be adapted to the task the user is conducting. In the annotation part of this work, the UMLS was of paramount importance, as it allowed to map concepts between languages. For example, when some widespread English abbreviations appeared in clinical texts like *echo* ('echocardiography'), which is not the equivalent of *eco* in Spanish, since it is the abbreviation for *ecografía* ('US' or 'ultrasound' in English). The search for *echo* can be seen as follows in Figure 1.



Name	AUI	Vocabulary	Term Type	Code
Echocardiography	A0052495	RCD	PT	X77c1
Cardiac US scan	A0668978	RCD	SY	X77c1
Cardiac echo	A0669020	RCD	SY	X77c1
Echocardiogram	A0692613	RCD	SY	X77c1
US scan of heart	A0812928	RCD	SY	X77c1
US heart scan	A1285341	RCD	SY	X77c1
ecocardiografía	A5556982	SCTSPA	PT	40701008
ecocardiografía (procedimiento)	A5556935	SCTSPA	FN	40701008
ecografía de corazón	A5557594	SCTSPA	SY	40701008
procedimiento ecocardiográfico	A6055553	SCTSPA	SY	40701008
Ecocardiografia	A9109327	MSHPOR	MH	D004452
Ecocardiografia Transtorácica	A27546980	MSHPOR	ET	D004452
Ecocardiografía	A9214335	MSHSPA	MH	D004452
Ecocardiografía Transtorácica	A27548306	MSHSPA	ET	D004452
Ekokardiografi	A11749548	MSHSWE	MH	D004452
Hjärt-sonografi	A33253249	MSHSWE	ET	D004452
Hjärtultraljud	A33248409	MSHSWE	ET	D004452
Ekkokardiografi	A20203838	MSHNOR	MH	D004452
Ultralydundersøkelse av hjertet	A27474400	MSHNOR	ET	D004452
Hjerteultralyd, transtorakal	A27536988	MSHNOR	ET	D004452
EHOKARDIOGRAFIJA	A17433310	MSHSCR	MH	D004452
心エコー図	A15701759	MSHJPN	PT	D004452
Mモード心エコー図法	A15666850	MSHJPN	SY	D004452
UCG法	A15666851	MSHJPN	SY	D004452
エコーカルジオグラフィー	A15675604	MSHJPN	SY	D004452
エコー心拍動記録	A15666852	MSHJPN	SY	D004452
エコー心拍動記録法	A15666853	MSHJPN	SY	D004452
エコー心拍記録	A15728246	MSHJPN	SY	D004452
エコー心拍記録法	A15728247	MSHJPN	SY	D004452

Figure 1. Example of a search display using UMLS for *echocardiography* (UMLS CUI: C0013516).

This overview highlights the complexity and necessity of standardized medical terminologies and classification systems for data management. In the NLP field, these terminological resources allow the expansion of terms with synonyms for an equal concept. Nonetheless, exact synonyms hardly occur, and each specialty has its own connotations. Considering that there is no perfect relation between natural language expressions and concepts of a domain, these tools help to reduce variation and lack of consensus, organize polysemic words and register paraphrases. Furthermore, in the texts used for this work, we have not been confronted with semantic voids, as the topics of each text are appreciably clear and are usually within the title of the texts, e.g., in *Evaluación del efecto de la Anestesia Libre de Opioides* (2022-001027-33). However, when machine annotation took place, the software used sometimes opted for the common-level or *familiar* meaning for some words. This issue will be addressed and further analyzed in the results section with significant examples.

These resources are not only helpful for the healthcare sector, they are an excellent tool for linguists. The UMLS broad range of medicine specialties has been a fundamental aspect for this work, since its terminologies such as SNOMED, MeSH, ICD or LOINC included in the metathesaurus facilitated semantic interoperability. With regards to semantics, the UMLS expanded the concepts with synonyms, which was an essential part for understanding the annotation task that was taking place and how it would be automatically processed. As Dalianis (2018) contends, "the identification of the semantic meaning can support directly the understanding of the clinical text, but the semantic tagging can also be used as features for machine learning". (p. 42). This is valuable for identifying and expanding abbreviations and acronyms in clinical texts but also for mapping concepts to other terminologies.

2.4. Lexical simplification

Understanding medical texts, such as our own health records or scientific findings related to our medical issues, is crucial for everyone. However, medical texts often use specialized terms and abbreviations derived from Latin or Greek, as seen in section 1.1. This makes medical texts hard to understand (Keselman & Smith, 2012). Comprehending these texts can be particularly difficult for laymen who are not accustomed to looking up unfamiliar terms. As the need to make information accessible to everyone increases, previous research has shown that replacing difficult words with simpler synonyms can make texts easier to understand (Abrahamsson et al., 2014). This is not a current fashion, as the need for making changes has been studied for decades. Blum and Levenston (1978) maintain that:

> The ways in which lexical items in natural languages can substitute for each other in specific contexts give rise to the strategies of lexical simplification common to these diverse linguistic contexts. The universal principles involved are probably based on systemic relationships between lexical items such as hyponymy, synonymy, antonymy and converseness. The awareness of these relationships, together with the ability to use circumlocution and paraphrase, is part of every speaker's semantic competence and enables him, when the need arises, to express complex meanings by indirect means. (p. 400).

What is more, a difference should be made between plain language and easy-to-read language. The latter does not only involve simplification, but improvement over the visualization of the text, such as using bullet points and short enumerations or adjusting each sentence of a text to a certain number of characters. Conversely, for the former it is the International Plain Language Federation[11] the institution that sets an ISO standard so that readers can easily find and understand what they need. According to this Federation, the standard is founded on a globally recognized definition of plain language. It was created by an international team of experts using ISO's esteemed consensus model and is supported by empirical evidence. It consolidates the information utilized and may provide new information the reader was not previously aware of. This makes it a valuable tool for quickly evaluating one's work, ensuring no steps were missed when writing or editing a document.

As Zilio et al. describe (2020), in NLP, the task of text simplification involves rewriting a text, adding definitions or other supplementary information, or removing unnecessary information to reduce the text's complexity while ensuring that the simplified text's meaning remains largely unchanged and that the new version reads naturally and smoothly for the reader (Siddharthan, 2002; Siddharthan, 2014). Our work presented in this dissertation is more focused on plain language enhancement to improve professional practices rather than entirely changing the structures of the studied texts. Our simplification task does not aim to automatically replace complex phrases, instead, it is intended to assist laymen patients in understanding more complex texts.

Alarcón et al. (2019) briefly reviewed the various methods that exist to achieve this goal for the Spanish language, including supervised, unsupervised and hybrid techniques. Supervised methods require annotated datasets to fulfill their purpose (Štajner et al. 2015), which poses a significant challenge when working with languages that have

[11] https://www.iplfederation.org/ (Retrieved August 20, 2024)

limited annotated corpora for text simplification (Saggion et al., 2011). Regarding methodological strategies, Paetzold and Specia (2017) suggest that lexical simplification should be carried out in four stages: Complex Word Identification, Generation of Substitutes (GS), selection of substitutes, and substitutes ranking. Our work adheres to this methodology, but we will focus exclusively on the CWI task. CWI involves identifying the words in a sentence that need simplification, meaning it determines which words are complex in a given text. Substitute generation involves creating potential synonyms for the identified complex words. Most research utilizes existing dictionaries, with WordNet being the most commonly used (Lal & Ruger, 2002). In the substitute selection phase, the most appropriate synonym is chosen from the set of generated synonyms based on factors like simplicity and context.

Given the variety of profiles with which we could associate the means of this task, the ideal target audience would be defined as individuals with functional literacy as to using the technical resources required to access the tool. This concept of *functional literacy* is described by UNESCO[12] as “the capacity of a person to engage in all those activities in which literacy is required for effective function of his or her group and community and also for enabling him or her to continue to use reading, writing and calculation for his or her own and the community’s development”. In this sense, the range of usage is moderately limited but not restrictive enough to take a population group out of the equation.

[12] https://uis.unesco.org/en/glossary-term/functional-literacy (Retrieved July 28, 2024)

3. Methodology

3.1. Dataset statistics

After considering which approach our task should follow, the main objective of this work is to test the performance of existing transformer-based model for lexical simplification. Our work may serve as the starting point of a future LS tool for functionally literate patients who are not familiar with specific terminological expressions, and this may result in a doctor-patient gap which cannot be overcome. In order to do so, three collections of 60 open-source texts have been manually annotated and peer-revised to achieve a gold standard.

These sets of texts belong to three different typologies:

1. Consent forms: this first typology refers to the form which patients willingly complete in order to undergo a clinical intervention or for accepting participation in a clinical experiment. These texts come from websites such as Sociedad Española de Reumatología (SER)[13], Sociedad Española de Anestesiología y Reanimación (SEDAR)[14], Sociedad Española de Cardiología[15], Sociedad Española de Cardiología Pediátrica y Cardiopatías Congénitas[16], Consejería de Salud y Consumo de la Junta de Andalucía[17], among others[18]. The majority of texts are listed under the identification *ci* followed by a number, e.g., *ci_32*.

2. Clinical trial announcements: the public information about any controlled study assessing the safety and efficacy of a therapeutic agent involving consenting human subjects. This set was extracted from the European Union Drug Regulating Authorities Clinical Trials Database (EudraCT)[19] and the texts of this collection maintain their identification number from the original database, such as *2018-001167-23*.

[13] https://www.ser.es/profesionales/que-hacemos/investigacion/herramientas/hojas-informativas-y-consentimiento/ (Retrieved January 24, 2024)
[14] https://sedar.es/index.php/cientifico/consentimientos-informados (Retrieved January 24, 2024)
[15] https://secardiologia.es/arritmias/cientifico/consentimiento-informado (Retrieved January 24, 2024)
[16] https://secardioped.org/wp-content/uploads/2019/10/c.pdf (Retrieved January 24, 2024)
[17] https://www.juntadeandalucia.es/organismos/saludyconsumo/areas/sistema-sanitario/derechos-garantias/paginas/ci-oncologia.html (Retrieved January 24, 2024)
[18] These consent forms were provided by Ana Rosa Terroba, a collaborator in the CLARA-MeD project, in which this dissertation was conducted.
[19] https://eudract.ema.europa.eu/ (Retrieved January 24, 2024)

3. Other collection of informative texts or patient information documents: unifies informative texts about topics ranging from transplants to different types of cancer, pain or diseases. This set was primarily extracted from the public patient portal of the Spanish Autonomous Region of Castilla y León[20] and the Spanish National Transplant Organization[21]. Those texts belonging to the former are listed with the id *aula_cyl*, such as *aula_cyl_cancercolon1*; while the texts from the latter are grouped following the id *ont*, as in *ont_1_historia*.

The dataset used in this study is freely accessible to the public and can be found on https://github.com/fede-ortega/LS-CWI-ES.

The statistics of each collection of texts can be seen in Table 4.

Set	Avg. Sentences	Avg. Tokens	No. CW	Avg. CW
CTs	33.71	679.35	6295	104.91
CFs	37.11	680.18	4053	67.55
PIDs	39.08	836.25	4721	78.6

Table 4. statistics of each dataset of our corpus (avg.: average; CW: complex words).

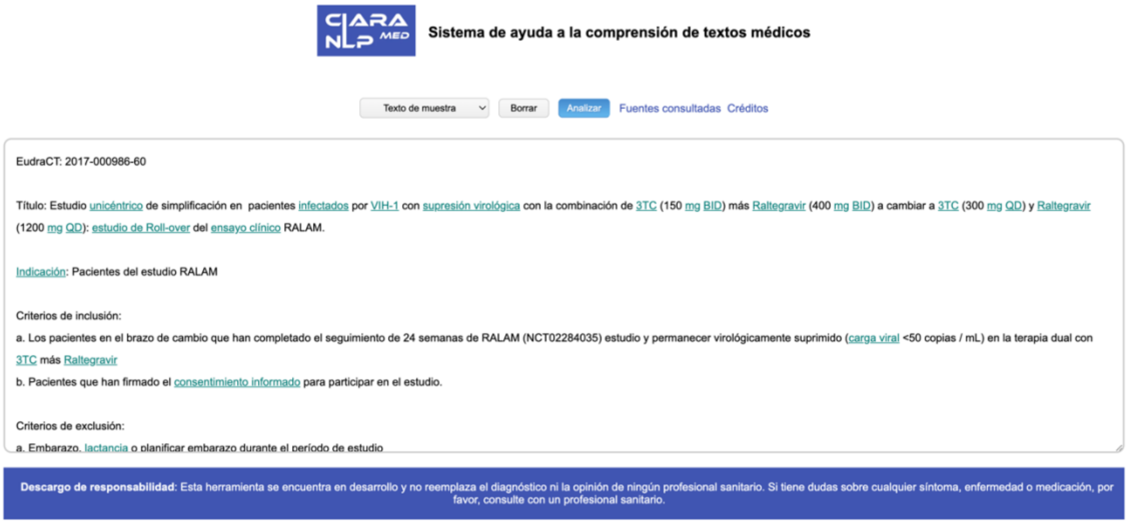
3.2. Annotation process

All texts have been annotated using the BRAT tool (Stenertop et al. 2012)[22], which is distributed as open source. In order to evaluate the terms that the system cannot detect or whose definition or synonyms need to improve, the following steps are taken.

First, each text is copied in the online tool[23] (Campillos-Llanos et al., 2024). Once the text is copied, by clicking on *Analizar* the detected words will be highlighted and underlined (Figure 2). This is the previous step to click *Descargar* to download the .ann file.

The annotation tool BRAT uses this .ann format, which needs the same name as the original .txt document. For example, the text *aula_cyl_parkinson.txt* will be aligned with the text *aula_cyl_parkinson.ann*. All data has been annotated in a local server.

[20] https://www.saludcastillayleon.es/AulaPacientes/es/enfermedades (Retrieved January 24, 2024)
[21] https://www.ont.es/ (Retrieved January 24, 2024)
[22] http://brat.nlplab.org/ (Retrieved January 24, 2024)
[23] http://claramed.csic.es/demo (Retrieved August 20, 2024)



Sistema de ayuda a la comprensión de textos médicos

EudraCT: 2017-000986-60

Título: Estudio unicéntrico de simplificación en pacientes infectados por VIH-1 con supresión virológica con la combinación de 3TC (150 mg BID) más Raltegravir (400 mg BID) a cambiar a 3TC (300 mg QD) y Raltegravir (1200 mg QD): estudio de Roll-over del ensayo clínico RALAM.

Indicación: Pacientes del estudio RALAM

Criterios de inclusión:
a. Los pacientes en el brazo de cambio que han completado el seguimiento de 24 semanas de RALAM (NCT02284035) estudio y permanecer virológicamente suprimido (carga viral <50 copias / mL) en la terapia dual con 3TC más Raltegravir
b. Pacientes que han firmado el consentimiento informado para participar en el estudio.

Criterios de exclusión:
a. Embarazo, lactancia o planificar embarazo durante el período de estudio

Descargo de responsabilidad: Esta herramienta se encuentra en desarrollo y no reemplaza el diagnóstico ni la opinión de ningún profesional sanitario. Si tiene dudas sobre cualquier síntoma, enfermedad o medicación, por favor, consulte con un profesional sanitario.

Figure 2. Example of the ClaraMeD tool using the available sample text of a CT.

The task consists of revising complex word entities (Complex_Word, CW) that the system detects automatically:

- Add terms or entities which have not been annotated (false negatives). For example, if the system has not detected *hemoderivados* as Complex_Word, it is selected and annotated.
- Delete terms or entities incorrectly annotated (false positives). For example, if the system annotates *vómitos* as Complex_Word, the annotation is deleted using the menu.
- Correct the span of the annotated entities. For example, if the system detected *punción*, but the correct Complex_Word is *punción accidental*, this span can be changed with the menu.

3.3. Annotation criteria

As the annotation task progressed, more annotation criteria were added, which are listed as follows with the appropriate examples and the texts in which they were present:

A. Nested entities

Nested entities are not annotated. Only the more specific term or the complex_word with the longest range will be included. Examples (id of the text is given between brackets):

(1) *oclusión tubárica* instead of *oclusión* and *tubárica* as separate complex_word (2022-000422-16).

(2) *bloqueo neurolítico del ganglio celíaco* instead of *bloqueo neurolítico* and *ganglio celíaco* as separate complex_word (ci_sed_19).

(3) *espesor central de la córnea* instead of *espesor central* and *córnea* as separate complex_word (2021-006456-14).

B. Frequent or widespread entities

Frequent entities that are now used by patients without medical knowledge. If the system does not return an annotation, it will be left unannotated. Example: *cirugía*, *diabetes* or *cáncer*.

However, when there are problematic terms, and there is doubt on whether to annotate the entity or not, we must decide if the word can be further simplified. Examples:

(1) *intervención quirúrgica* can be simplified with a more general synonym as *operación* (2022-002680-30). Therefore, *operación* is not annotated.

(2) *glucemia* is a complex word which can be explained as *azúcar en la sangre* (aula_cyl_diabetes_5). Therefore, *azúcar en la sangre* is not annotated.

C. Discontinuous entities

Discontinuous entities will not be annotated, and multiword entities which are separated by one or more irrelevant words will take these considerations:

(1) Annotate only the term which presents difficulty for understanding. Example:

- *tumor (T) 4b* (2018-001167-23).

(2) Annotate all components of the entity. Examples:

- *VIH 1/2* (2022-003594-33).
- *cáncer de mama ipsi o contralateral* (2021-002346-33).

D. Measure units

Measure units will not be annotated when they are of general use (e.g., *mg*.), but can be annotated those that are necessary to understand the text. Example:

(1) ≥2 *μg* (2021-001396-16).

E. Foreign words

Foreign words will not be annotated if there is a translated equivalent in the text. Example:

(1) Diagnóstico de enfermedad de Crohn (*Crohn's disease*) (2021-003314-39).

This is not the case if the foreign word is only used to refer to the disease. Example:

(2) *RSV* which stands for 'Respiratory Syncytial Virus' (2022-003124-41).

F. Names of genes

Names of genes will not be annotated, except for those which are highly relevant or associated with a disease. Example:

(1) *BRCA*: gene associated with breast cancer (2022-003594-33).
(2) *HER2*: gene used as tumor indicator (2022-003594-33).
(3) *EGFR*: gene of the epidermal growth factor receptor (2022-003016-87).

G. Names of clinical trials

Names of clinical trials will not be annotated. Example:

(1) *CLOU064A2301* (2022-001034-11).
(2) *COMET-ICE* (2021-000724-35).

H. Synonyms

Synonyms of the same complex words can be annotated. Example:

(1) *diagnóstico temprano*, synonym of *diagnóstico precoz* (aula_cyl_erc_3).

I. Adjectives

Adjectives which do not belong to a phraseological unit will not be annotated. Examples:

(1) *valvulopatía estenótica grave* in which *grave* is not included as a phraseological unit (2020-003312-27).

However, sometimes the adjectives create a complete entity. Examples:

(2) *insuficiencia cardiaca refractaria*, which is a complete entity (ci_sec_1).
(3) *rigidez matutina*, which is a complete phraseological unit (aula_cyl_dolormuscesq_2).

Additionally, if the adjective is an entity but it can accompany other words, it can be annotated individually, for example *maniobras de reanimación neonatal* in which *maniobras de reanimación* and *neonatal* are separate entities.

J. Organizations

Organizations and general guides for patients are also annotated. Example:

(1) *NYHA* (2020-003312-27).

(2) *Federación Española de Parkinson* (aula_cyl_parkinson).

(3) *Guía de práctica clínica* (aula_cyl_diabetes_4).

3.4. Preliminary evaluation

The 180 texts were revised by four linguists (each text was revised by two annotators), and the inter-annotator agreement (IAA) was calculated using the F-measure. The strict IAA was 84.42% and the relaxed IAA was 91.58%, which was quantified using the library BRATEval[24] in java.

Once the texts were annotated, the .ann files were extracted. This format consists of four columns:

- The first one indicates the number of the term.
- The second one provides the annotation.
- The third and fourth columns are the offsets of the CW in the given text.
- The fifth one is the annotated CW.

To illustrate this, a real example from ci_carped_2 is shown:

T1	CW 10 16	ductus
T2	CW 28 34	C.I.V.
T3	CW 113 119	ductus
T4	CW 131 137	C.I.V.
T5	CW 451 475	consentimiento informado
T6	CW 598 623	comunicaciones congénitas
T7	CW 633 638	aorta
T8	CW 644 660	arteria pulmonar

[24] https://perso.limsi.fr/pz/blah2015 (Retrieved May 3, 2024)

The .ann files were then converted to .conll format. The .conll files consist of two columns in which the word appears in the first column and the class appears in the second column (Sang & De Meulder, 2003). Our label list included three types of features, following the BIO tagging scheme, which encodes the span of each entity:

- O (out): if the entity token is a non-complex word.
- B (begin): if the detected token is the beginning of the entity. In our case, *B-CW* for single complex words or the first word of a multi-word complex term.
- I (inside): if the detected token is inside of the entity. In our task, *I-CW* for the rest of the words in a multi-word term.

For instance, in the sentence *Se permiten sujetos que recibieron quimioterapia adyuvante tras cistectomía con intención curativa* (2018-001167-23), each token is annotated as follows:

```
Se      O
permiten        O
sujetos O
que     O
recibieron      O
quimioterapia   B-CW
adyuvante       I-CW
tras    O
cistectomia     B-CW
con     O
intención       O
curativa        O
```

The .conll format files were converted to .json to adapt to the requirements of models in Hugging Face. The .json format consists of a set for each sentence of the text which is divided into a list of tokens, with the words of the sentence, and a list of ner_tags, with the annotation provided in the .colln format for each token. To illustrate this, we could analyze the sentence *Metástasis conocida en el SNC o meningitis carcinomatosa*. (2018-001167-23) in the .json format:

```
{

"tokens": ["Metástasis", "conocida", "en", "el", "SNC", "o", "meningitis", "carcinomatosa", "."],

"ner_tags": ["B-CW", "O", "O", "O", "B-CW", "O", "B-CW", "I-CW", "O"]

}.
```

3.5. Model training

Transformer-based models available on the Hugging Face hub (Wolf et al., 2020) were utilized for the task of lexical simplification. The choice of transformers was based on their state-of-the-art performance for NLP tasks, and these models were fine-tuned for our classification task. The six models were fine-tuned following the notebook on token classification available in the Hugging Face's GitHub[25]. The six models used in this task are listed as follows:

(1) BETO – Spanish BERT[26] (Cañete et al., 2020): monolingual and general model trained on a big Spanish annotated corpus.

(2) Biomedical Language model for Spanish (RoBERTa EHR)[27] (Carrino et al., 2022): monolingual and domain-specific pretrained model on a corpus of medical texts and clinical reports in Spanish with more than 963 millions of tokens.

(3) RoBERTa-EHR-CT[28] (Campillos-Llanos et al., 2021): monolingual and domain-specific pretrained model from RoBERTa EHR, which was fine-tuned on texts about clinical trials to detect four types of semantic groups from the UMLS: anatomy, pharmacological substances, pathologies and procedures.

(4) mBERT base model[29] (Devlin et al., 2018): multilingual and general pretrained model on 104 languages.

(All references below were retrieved August 2, 2024)

[25] https://github.com/huggingface/notebooks/blob/main/examples/token_classification.ipynb

[26] https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased

[27] https://huggingface.co/PlanTL-GOB-ES/bsc-bio-es

[28] https://huggingface.co/lcampillos/roberta-es-clinical-trials-ner

[29] https://huggingface.co/google-bert/bert-base-multilingual-cased

(5) Medical mT5[30] (García-Ferrero et al., 2024): multilingual and domain-specific pretrained model on English, Spanish, French and Italian corpora.

(6) mDeBERTaV3[31] (He et al., 2021): multilingual and general pretrained model which improves BERT with disentangled attention and enhanced masked encoder.

Each model was fine-tuned with a batch size of 16, trained with 30 epochs, an early stopping of 5, a learning rate of 2e-5 and the Adam optimizer. Models were run with three seeds (100, 500 and 1000), and we provide the average and standard deviation of the three experimental rounds for each model. All experiments were run in Google Colab Pro[32]. The evaluation metrics were precision, recall, F1-score and accuracy.

3.6. Metrics

In this section, we will succinctly explain how the true positives (TP), false positives (FP) and false negatives (FN) in the tests are used to compute the evaluation metrics, which are precision, recall and F1, defined as follows:

(1) Precision: measures the accuracy to correctly identify complex words, and it is useful when the cost of FP is high. It is defined as the ratio of TP predictions to the total of TP and FP.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

(2) Recall: measures the ability to capture the maximum number of complex words in tasks where the cost of FN is high. It is defined as the ratio of TP to the total number of TP and FN.

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

(3) F1-Score: measures the overall performance of a classification model. It is defined as the harmonic mean of precision and recall.

$$F1 = 2 \text{ x } \frac{\text{Precision x Recall}}{\text{Precision} + \text{Recall}}$$

(All references below were retrieved August 2, 2024)

[30] https://huggingface.co/HiTZ/Medical-mT5-xl

[31] https://huggingface.co/microsoft/mdeberta-v3-base

[32] https://colab.research.google.com/

4. Results

For each model, we report the average precision, recall, F1 and accuracy of the three seeds in our test dataset, and the standard deviation is specified between brackets. The model that showed the best performance in the F1 measure was mDeBERTaV3, followed by BETO and Medical mT5 (Table 5).

Model	P	R	F1	Acc.
BETO	75.01 (±1.11)	**82.98 (±0.60)**	78.78 (±0.34)	93.49 (±0.13)
RoBERTa EHR	62.58 (±3.34)	78.54 (±0.31)	69.62 (±2.03)	94.21 (±0.62)
RoBERTa-EHR-CT	70.44 (±1.07)	78.82 (±1.32)	74.39 (±0.98)	**95.21 (±0.07)**
mBERT	76.23 (±0.78)	77.03 (±1.74)	76.63 (±1.19)	92.10 (±0.48)
Medical mT5	74.94 (±1.16)	82.07 (±0.40)	78.34 (±0.77)	94.72 (±0.13)
mDeBERTaV3	**79.05 (±1.39)**	79.01 (±0.70)	**79.02 (±0.65)**	94.86 (±0.22)

Table 5. Evaluation of CWI results in the test dataset (P: Precision; R: Recall; acc.: Accuracy).

In terms of precision, mDeBERTaV3 achieved the highest score with 79.05, suggesting that it captures a high number of correctly labeled complex words. When it comes to recall, BETO scored the best results with 82.98, which indicates that it has a better performance in identifying the maximum number of correctly labeled complex words in all observations of the actual class. Among all the models considered, mDeBERTaV3 appears to be the most balanced one, with the highest F1 score. Furthermore, it has a strong precision and a similar recall. The BETO model excels in recall, which makes it a good choice if the task is minimizing missed complex words or false negatives; however, it may include more false positives compared to mDeBERTaV3. The RoBERTa-EHR-CT model scores the highest accuracy, which indicates good performance, yet less favorable F1 outcomes than the vast majority of the models. Overall, the RoBERTa EHR model performs the weakest, with the lowest metrics' scores, specifically a 69.62 F1-score.

Surprisingly, the general-purpose models achieved the best scores, compared to the results for domain-specific models, which could mean some of the annotated complex words in the training dataset do not belong necessarily to the medical domain or are polysemous words. Some examples to illustrate this might be:

(1) *revocar*, *consentimiento*, *axila*, *heces*, *nauseas* or *reintervención* (examples on several texts).

(2) *exploración*: when used as examination to find a pathology instead of going out to explore a new place (ci_sed_9).

(1) *coma*: used as the state of profound unconsciousness instead of the third person of the verb *comer* in Spanish (aula_cyl_diabetes_5).

(2) *instrumental*: as the medical equipment necessary in a surgery instead of the music-related meaning (ci_86).

(3) *progresar*: with a negative meaning as ‘worsening’ rather than a positive one (ci_ser_4).

(4) *saciedad*: recognized in the Spanish idiomatic expression *hasta la saciedad* (ont_2_tipos_1).

5. Discussion

Comparing these findings with previous studies, we observe certain similarities with other classifiers for lexical simplification. For example, Alarcón et al. (2019) reported an F1 score of 74.97 using a SVM classifier, achieving slightly lower results. In another article by Alarcón, Moreno and Martínez (2021), their research resulted in an F1 score of 72.7 using BERT independently, indicating a similar trend with their previous work. On the other hand, Truică, Stan and Apostol (2023), obtained better precision and recall results with a multilayer perceptron; however, their overall accuracy did not fall within the range of our reported values, with a 15% difference. The results from Truică using SVM, random forest or extra randomized trees presented discrepancies compared to those of the perceptron, suggesting an interesting comparison between classifiers. Nonetheless, drawing a direct comparison between models might be challenging due to the differences in our dataset from the ones presented in previous research.

The implications of these findings are relevant because they indicate that transformers demonstrate good performance in lexical simplification tasks. Having said that, the CWI task in Spanish texts is yet to be tested using other datasets and traditional ML classifiers, such as SVM or CRF. In our research, transformer-based models' ability to capture semantic relationships and contextual information make them dynamic, especially for multi-word terms. This capability enables them to be competent in grasping nuanced meanings in which words are used, and to be fine-tuned on domain-specific datasets to test their performance.

In spite of our optimal results, the main limitation of our work is the size of our corpus. The transformers used would benefit from a more extensive corpus of more than 180 texts, since it would allow them to generalize better to unseen medical-related complex words, increasing versatility and accuracy. For future work, it would also be recommended to reexamine our corpus of medical texts and establish a difference between complex words belonging to the general domain and those complex words belonging exclusively to the medical domain. In doing so, it would be possible to ascertain an explanation of better performances between general domain and domain-specific models.

6. Conclusions

In this study, transformer-based models were fine-tuned and evaluated for CWI in Spanish texts for patients. Our results demonstrated that the use of transformer-based models, while similar, tends to achieve a slightly higher F1 score compared to other previous studies which used mostly SVM. These findings might indicate that language and domain are not the most relevant factors in the CWI task with our data, since the mDeBERTaV3, BETO and Medical mT5 models were the ones with the best performance, although their scores were relatively similar. Even so, our outcomes deserve to be confirmed with additional experiments. The simplification system was evaluated with a corpus consisting of 180 texts and 15069 complex words, which may limit the training of our models. Therefore, future research could explore extending the annotations of our corpus for better generalizations, reassess which complex words might belong to general discourse, and testing performance on other traditional ML models. By continuing to refine and expand upon these methods, we can make significant strides towards more inclusive communication between Spanish health providers and patients.

7. References

Abdramanova, D. (2023). The Specialized Word Formation Processes in Legal and Medical Discourse and Their Impact on Professional Communication. *Центральноазиатский журнал образования и инноваций*. *2*(11). 51–58. https://www.in-academy.uz/index.php/cajei/article/view/22767

Abrahamsson, E., Forni, T., Skeppstedt, M., & Kvist, M. (2014). Medical text simplification using synonym replacement: Adapting assessment of word difficulty to a compounding language. *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, 57–65.

Acromegaly. (2021). In *National Organization for Rare Disorders*. Retrieved June 20, 2024, from https://rarediseases.org/rare-diseases/acromegaly/

Alarcón, R., Moreno, L., & Martínez, P. (2021). Lexical simplification system to improve web accessibility. *IEEE Access*, *9*, 58755-58767.

Alarcón, R., Moreno, L., Segura-Bedmar, I., & Martínez, P. (2019). Lexical simplification approach using easy-to-read resources. *Procesamiento del Lenguaje Natural*, *63*, 95-102.

Amyloidosis. (2023). In *National Organization for Rare Disorders*. Retrieved June 20, 2024, from https://rarediseases.org/rare-diseases/amyloidosis/

Bas-Sarmiento, P., Lamas-Toranzo, M. J., Fernández-Gutiérrez, M., & Poza-Méndez, M. (2022). Health Literacy, Misinformation, Self-Perceived Risk and Fear, and Preventive Measures Related to COVID-19 in Spanish University Students. *Int. J. Environ. Res. Public Health*, *19*, 15370. https://doi.org/10.3390/ijerph192215370

Bellés-Fortuño, B. (2016). Popular science articles vs. scientific articles: a tool for medical education. *Medical discourse in professional, academic and popular settings*, 55-78.

Blum, S., & Levenston, E. A. (1978). Universals of Lexical Simplification. *Language Learning*, *28*(2), 399–415.

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

Brown, E. G., Wood, L., & Wood, S. (1999). The medical dictionary for regulatory activities (MedDRA). *Drug safety*. *20*(2), 109–117. https://doi.org/10.2165/00002018-199920020-00002

Cabré, M. T. (2003). Theories of terminology: their description, prescription and explanation. *Terminology*, *9*(2): 163-199.

Campillos-Llanos, L. (2023). MedLexSp–a medical lexicon for Spanish medical natural language processing. *Journal of Biomedical Semantics*, *14*(1), 2. https://doi.org/10.1186/s13326-022-00281-5

Campillos-Llanos, L., Ortega-Riba, F., Terroba, A. R., Valverde-Mateos, A., & Capllonch-Carrión, A. (2024). CLARA-MeD Tool - A System to Help Patients Understand Clinical Trial Announcements and Consent Forms in Spanish. *Studies in health technology and informatics*, *316*, 95–99. https://doi.org/10.3233/SHTI240354

Campillos-Llanos, L., Valverde-Mateos, A., Capllonch-Carrión, A., & 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*, 1-19.

Cañete, J., Chaperon, G., Fuentes, R., Ho, J-H., Kang, H., & Pérez, J. (2020). Spanish Pre-Trained BERT Model and Evaluation Data. *PML4DC at ICLR 2020*.

Carrino, C. P., Llop, J., Pàmies, M., Gutiérrez-Fandiño, A., Armengol-Estapé, J., Silveira-Ocampo, J., Valencia, A., González-Aguirre, A., & Villegas, M. (2022). Pretrained biomedical language models for clinical NLP in Spanish. *Proceedings of the 21st Workshop on Biomedical Language Processing*, 193-199.

Cimino, J. J. (1998). Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of information in medicine*, *37*(4-5), 394–403.

Dalianis, H. (2018). Medical classifications and terminologies. *Clinical Text Mining: Secondary Use of Electronic Patient Records*, *5*, 35-43. Berlin: Springer.

De Belder, Z. (2013) Power and Discourse Comparing the Power of Doctor in Two Contrasting Interactive Encounters. *Innervate*. *5*, 106–121.

Devlin, J., Ch., M-W, Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR.*

dyalize. (2011). In *Merriam-Webster.com*. Retrieved June 15, 2024, from https://www.merriam-webster.com/dictionary/dyalize.

El-Sappagh, S., Franda, F., Ali, F., & Kwak, K. S. (2018). SNOMED CT standard ontology based on the ontology for general medical science. *BMC medical informatics and decision making*, *18*, 1-19.

European Commission. (2017). *Standard Eurobarometer*. https://europa.eu/eurobarometer/surveys/detail/2142.

European Union. (Version of 2024). *European Union Drug Regulating Authorities Clinical Trials Database (EudraCT)*. Retrieved January 24, 2024, from https://eudract.ema.europa.eu/

Faber, P. B. (Ed.). (2012). *A cognitive linguistics view of terminology and specialized language*. Berlin, Boston: De Gruyter Mouton.

Falcón, M., Bas, P., & Fernández, M. (2022). *Monitorización del comportamiento y las actitudes de la población relacionadas con la COVID-19 en España*. Instituto de Salud Carlos III.

Falcón, M., Torres, A., Hernández, E., Del Arco, I., Bas, P., & Fernández, M. (2022). *Desarrollo y efectividad de una intervención de mHealth en la mejora de la Alfabetización en Salud y autogestión del paciente pluripatológico con insuficiencia cardíaca: un ensayo controlado aleatorizado*. INiBICA.

Fernández-Gutiérrez, M., Bas-Sarmiento, P., Marín-Paz, A. J., Castro-Yuste, C., Sánchez-Sánchez, E., Hernández-Encuentra, E., Vinolo-Gil, M. J.., Carmona-Barrientos, I., & Poza-Méndez, M. (2023). Self-management in heart failure using mHealth: A content validation. *International Journal of Medical Informatics, 171*, 104985. https://doi.org/10.1016/j.ijmedinf.2023.104986

Fernández, M., Juvinyà, D., & Suñer, R. (2021). *"Fes sempre tres preguntes" (Haz siempre tres preguntas)*. Xarxa HPH Catalunya.

Fernández, M., Juvinyà, D., & Suñer, R. (2021). *Salud sin dudas – Salut sense dubtes*. Xarxa HPH Catalunya.

Finlayson, M. A., & Erjavec, T. (2017). Overview of annotation creation: Processes and tools. *Handbook of Linguistic Annotation*, 167-191.

García-Ferrero, I., Agerri, R., Salazar, A. A., Cabrio, E., de la Iglesia, I., Lavelli, A., Magnini, B., Molinet, B., Ramírez-Romero, J., Rigau, G., Villa-González, J. M., Villata, S., & Zaninello, A. (2024). Medical mT5: an open-source multilingual text-to-text LLM for the medical domain. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 11165-11177.

give birth. (2011). In *Merriam-Webster.com*. Retrieved June 15, 2024, from https://www.merriam-webster.com/dictionary/give%20birth.

Google. (2024). *Google Colaboratory*. Retrieved August 20, 2024, from https://colab.research.google.com/

He, P., Gao, J., & Chen, W. (2021). DeBERTaV3: Improving DeBERTa using electra-style pre-training with gradient-disentangled embedding sharing. *Natural Language Processing*, 9019-9052.

Hripcsak, G., & Rothschild, A. S. (2005). Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, *12*(3), 296-298.

International Plain Language Association. (n.d.) *What is plain language?* Retrieved August. 20, 2024, from https://www.iplfederation.org/plain-language/

International Standard Organization. (2019). *Terminology work and terminology science – Vocabulary* (ISO Standard n.º 1087:2019). https://www.iso.org/standard/62330.html.

International Standard Organization. (2023). *Plain language Part 1: Governing principles and guidelines* (ISO Standard No. 24495-1:2023). https://www.iso.org/standard/78907.html.

Jabour, A. M. (2020). The impact of longer consultation time: A simulation-based approach. *Applied Clinical Informatics*, *11*(05), 857-864.

Junta de Andalucía. (2014). *Consentimiento informado. Especialidad en oncología: Listado de consentimientos informados disponibles en la especialidad de Oncología*. Retrieved January 24, 2024, from https://www.juntadeandalucia.es/organismos/saludyconsumo/areas/sistema-sanitario/derechos-garantias/paginas/ci-oncologia.html.

Junta de Castilla y León – Aula de Pacientes. (2018). *Enfermedades*. Retrieved January 24, 2024 from https://www.saludcastillayleon.es/AulaPacientes/es/enfermedades.

Juvinyà-Canal, D., Bertran-Noguer, C., & Suñer-Soler, R. (2018). Alfabetización para la salud, más que información. *Gaceta sanitaria*, *32*, 8-10.

Keselman, A., & Smith, C. (2012). A classification of errors in lay comprehension of medical documents. *Journal of Biomedical Informatics*, *45*(6), 1151–1163.

Lal, P., & Ruger, S. (2002). Extract-based summarization with simplification. *Proc. ACL*, *10*, 3-10.

Liu, C., Wang, D., Liu, C., Jiang, J., Wang, X., Chen, H., Ju, X., & Zhang, X. (2020). What is the meaning of health literacy? A systematic review and qualitative synthesis. *Family medicine and community health*, *8*(2), e000351. https://doi.org/10.1136/fmch-2020-000351

Makaryus, A. N., & Friedman, E. A. (2005). Patients' understanding of their treatment plans and diagnosis at discharge. *Mayo clinic proceedings*, *80*(8), 991-994.

Ministerio de Sanidad – Organización Nacional de Transplantes (ONT). (2023). *Información al ciudadano*. Retrieved January 24, 2024, from https://www.ont.es/

Moen, H., Peltonen, L. M., Koivumäki, M., Suhonen, H., Salakoski, T., Ginter, F., & Salanterä, S. (2018). Improving Layman Readability of Clinical Narratives with Unsupervised Synonym Replacement. *Studies in health technology and informatics*, *247*, 725–729.

Montalt, V., & García-Izquierdo, I. (2016). Exploring the link between the oral and the written in patient-doctor communication. *Medical discourse in professional, academic and popular settings*, 103-125.

OpenAI. (2024). *ChatGPT* (June 30 version) [Large language model]. https://chat.openai.com/chat.

Paetzold, G. H., & Specia, L. (2017). A survey on lexical simplification. *Journal of Artificial Intelligence Research*, *60*, 549-593.

Qiang, J., Li, Y., Zhu, Y., Yuan, Y., & Wu, X. (2020). LSBert: A Simple Framework for Lexical Simplification. *ArXiv, abs/2006.14939*.

Ratzan, S. C. (2001). Health literacy: communication for the public good. *Health promotion international*, *16*(2), 207–214. https://doi.org/10.1093/heapro/16.2.207.

Real Academia Nacional de Medicina de España (RANME). (2010). *Diccionario de términos médicos*. Madrid: Editorial Panamericana. https://dtme.ranm.es/

Real Academia Nacional de Medicina de España (RANME). (2023). *Diccionario panhispánico de términos médicos*. https://dptm.es/

Rogers, F. B. (1963). Medical subject headings. *Bull Med Libr Assoc*. *51*(1), 114–6.

Roter, D., & Hall, J. A. (2006). *Doctors talking with patients/patients talking with doctors*. Bloomsbury Publishing.

Sager, J. C. (1980). *English special languages: principles and practice in science and technology* (1. Aufl). Brandstetter.

Saggion, H., & Hirst, G. (2017). Lexical Simplification. *Automatic text simplification*, 32, 21-31. Morgan & Claypool.

Saggion, H., Gómez-Martínez, E., Etayo, E., Anula, A., & Bourg, L. (2011). Text simplification in simplext: Making texts more accessible. *Proces. Leng. Nat.*, *47*, 341-342.

Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 142–147.

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

Shahid, R., Shoker, M., Chu, L. M., Frehlick, R., Ward, H., & Pahwa, P. (2022). Impact of low health literacy on patients' health outcomes: a multicenter cohort study. *BMC health services research*, *22*(1), 1148.

Siddharthan, A. (2002). An architecture for a text simplification system. *Language Engineering Conference*. 64–71.

Siddharthan, A. (2014). A survey of research on text simplification. *ITL-International Journal of Applied Linguistics. Special Issue on Readability and Text Simplification.* Peeters Publishers, Belgium.

Sociedad Española de Anestesiología, Reanimación y Terapeútica del Dolor (SEDAR). (n.d.). *Consentimientos informados*. Retrieved January 24, 2024, from https://sedar.es/index.php/cientifico/consentimientos-informados.

Sociedad Española de Cardiología Pediátrica y Cardiopatías Congénitas (SECPCC). (2011). *Formularios de consentimientos informados*. Retrieved January 24, 2024, from https://secardioped.org/wp-content/uploads/2019/10/c.pdf.

Sociedad Española de Cardiología. (2014). *Consentimiento informado*. Retrieved January 24, 2024, from https://secardiologia.es/arritmias/cientifico/consentimiento-informado.

Sociedad Española de Reumatología (SER). (n.d.). *Hojas informativas y de consentimiento.* Retrieved January 24, 2024, from https://www.ser.es/profesionales/que-hacemos/investigacion/herramientas/hojas-informativas-y-consentimiento/.

Spanish Health Literacy Network. (n.d.). *Investigación*. Retrieved August 14, 2024, from https://www.alfabetizacionsalud.com/que-hacemos/investigacion/.

Štajner, S., Calixto, I., & Saggion, H. (2015). Automatic text simplification for Spanish: Comparative evaluation of various simplification strategies. *Proc. Int. Conf. Recent Adv. Natural Lang. Process*, 618-626.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. I. (2012). BRAT: a web-based tool for NLP-assisted text annotation. *Proc. of the Demonstrations at the 13th Conference of the European Chapter of the Assoc. for Computational Linguistics*, 102-107.

The path that ends AIDS: UNAIDS Global AIDS Update 2023. (2023). Geneva: Joint United Nations Programme on HIV/AIDS.

Truică, C. O., Stan, A. I., & Apostol, E. S. (2023). SimpLex: a lexical text simplification architecture. *Neural Computing and Applications*, *35*(8), 6265-6280.

United Nations Educational, Scientific, and Cultural Organization. (2021). *Functional literacy*. Retrieved July 28, 2024, from https://uis.unesco.org/en/glossary-term/functional-literacy.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L, & Polosukhin, I. (2017). Attention is All You Need. *Advances in Neural Information Processing Systems (NeurIPS)*.

Williams, M. V., Parker, R. M., Baker, D. W., Parikh, N. S., Pitkin, K., Coates, W. C., & Nurss, J. R. (1995). Inadequate functional health literacy among patients at two public hospitals. *Jama*, *274*(21), 1677-1682.

Wodak, R. (2006). Medical Discourse: Doctor-Patient Communication. *Encyclopedia of Language & Linguistics*, Second Edition, 7, 681-687. Oxford: Elsevier.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, … & Rush, A. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 38-45.

World Health Organization (WHO). (1996). *W.H.O.'s Collaborating centre for drug statistics methodology: Guidelines for ATC classification and DDD assignment*.

World Health Organization (WHO). (2019). *International statistical classification of diseases and related health problems* (11th ed.). Retrieved August 14, 2024, from https://icd.who.int/

World Health Organization. (2023). *Diabetes*. Retrieved August 14, 2024, from https://www.who.int/news-room/fact-sheets/detail/diabetes

World Health Organization. (2023). *Noncommunicable diseases*. Retrieved August 14, 2024, from https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases

Zilio, L., Paraguassu, L. B., Hercules, L. A. L., Ponomarenko, G., Berwanger, L., & Finatto, M. J. B. (2020). A lexical simplification tool for promoting health literacy. *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*. 70-76.