Procesamiento del Lenguaje Natural, Revista nº 74, marzo de 2025, pp. 95-108 recibido 03-12-2024 revisado 22-01-2025 aceptado 03-02-2025

Complex Word Identification for Lexical Simplification in Spanish Texts for Patients

Identificación de palabras complejas para simplificación léxica de textos para pacientes en español

Federico Ortega-Riba,[1] **Leonardo Campillos-Llanos,**[2] **Doaa Samy**[3]
[1]University of Colorado at Boulder, Departments of Linguistics and Computer Science
[2]ILLA - Consejo Superior de Investigaciones Científicas
[3]Facultad de Filosofía y Letras, Universidad Complutense de Madrid
federico.ortegariba@colorado.edu, leonardo.campillos@csic.es, dkhalil@ucm.es

Abstract: This work describes the task of complex word identification (CWI) in Spanish medical texts for patients. Identifying complex words is the first step in lexical simplification, which aims to overcome the language gap between patients and healthcare professionals, enable access to information, and ensure unambiguous terminology for effective and clear communication. As part of the task, we created a medical complex words annotation guideline and compiled a corpus consisting of 225 texts (162575 tokens). A total of 18203 complex words (single and multi-words) were manually labeled, each text being annotated by two linguists with high inter-annotator agreement (F1 = 84.42%). The corpus was utilized to train two machine learning classifiers (Support Vector Machines and Logistic Regression) as baselines, in addition to seven deep learning transformer models. The models were selected by considering two factors: language (Spanish and multilingual) and domain (general or medical). The final results on the test set achieve an overall average F1 score of 79.02 (±0.65) for the transformer model with the best performance.
Keywords: Automatic Text Simplification, Language Resources, Corpora.

Resumen: Este artículo describe la tarea de identificación de palabras complejas en textos médicos en español para pacientes. Este es el primer paso para la simplificación léxica, cuyo objetivo es superar la barrera lingüística entre pacientes y profesionales sanitarios, permitir el acceso a la información y garantizar una terminología sin ambigüedades y una comunicación clara y eficaz. Se ha creado una guía de anotación y se ha compilado un corpus de 225 textos (162575 tokens). Se anotaron 18203 palabras complejas (entidades simples como multipalabra), siendo cada texto revisados por dos lingüistas, y alcanzando un alto valor de acuerdo entre anotadores (F1 = 84.42%). El corpus se ha empleado para entrenar modelos de aprendizaje automático (máquinas de soporte vectorial y regresión logística) como referencia, y siete modelos de aprendizaje profundo basados en transformers. Estos modelos fueron seleccionados considerando dos factores: idioma (español o multilingüe) y dominio (general o médico). Los experimentos finales muestran una puntuación F1 de 79.02 (±0.65) para el modelo transformer con mejores resultados.
Palabras clave: Simplificación Automática de Textos, Recursos Lingüísticos, Corpus.

1 Introduction

Lexical simplification (LS) focuses on improving accessibility to textual information by replacing complex words with simpler alternatives (Saggion, 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 and Friedman, 2005). An in-depth analysis of the fluctuations of medical discourse as well as the individual's reading comprehension (Shahid et al., 2022) can reveal how language shapes patient relationships and knowledge communication in

ISSN 1135-5948 DOI 10.26342/2025-74-7

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 instructions to prevent stomach cancer in a leaflet. The shifts in tone, terminology or structure would provide enough insights to understand communication effectiveness with laymen patients (Bellés-Fortuño, 2016; Montalt and García-Izquierdo, 2016).

Despite the efforts by healthcare providers to ensure clinical descriptions of symptoms and diagnoses, time constraints and limited consultations times are an important aspect that could hinder communication in this respect (Jabour, 2020). These issues, alongside low health literacy levels, interfere in patients' ability to understand relevant information (Roter and Hall, 2006).

In order to overcome these barriers, a Natural Language Processing (NLP) application could be used to detect medical complex words for patients before providing them with an explanation or simpler synonym. This process may be automated with deep learning or machine learning (ML) models that can be integrated in an end-user application. This article reports our efforts to address the first step in the LS approach, i.e. the annotation and complex word identification task.

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 and Erjavec, 2017). 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 et al., 2018; Qiang et al., 2020).

In this work, we address these tasks in Spanish texts for patients, in line with the research goals of the CLARA-MeD project.[1] We provide an annotated corpus of medical complex words as a gold standard. The annotated corpus was used to fine-tune state-of-the-art pre-trained transformer models specifically for this task (Vaswani et al., 2017). Transformer models have demonstrated excellent performance for a wide range of NLP tasks. 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. For the purposes of this research, seven transformer-based models were selected, paying special attention to two parameters (language and domain), as two key factors to test the performance after fine-tuning. In addition to these models, we compared classical ML models as baselines: Support Vector Machines (SVM) and Logistic Regression (LR) classifiers.

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: Clinical trials (CT), Consent Forms (CF) and Patient information documents (PID).

In brief, our contributions are as follows:

1. A corpus of 225 texts (162575 tokens), which was doubly-annotated with 18203 complex words and high agreement values (F1 = 84.42%). The dataset and code used in this study are publicly accessible on: `https://digital.csic.es/handle/10261/373675`.

2. A guideline with criteria for annotating complex words in Spanish medical texts.

3. A set of CWI classifiers with seven transformer-based models and two baseline machine classifiers (SVM and LR). The code is available on: `https://github.com/fede-ortega/LS-CWI-ES/` and the fine-tuned models are available on: `https://huggingface.co/CLARA-MeD`.

The article is structured with a Background section (§2) and a description of the methods used (§3). Our results are shown in §4. Then, a discussion (§5) is established to interpret the key findings of the study and compare them with previous research in the medical field, highlighting similarities and differences, before the conclusions (§6).

[1] `https://clara-nlp.uned.es/home/med/`(Retrieved 2025/2/10)

2 Background

2.1 Dichotomy of Medical Discourse

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", which may be a mono- or a multi-word expression.

Provided that both paradigmatic and syntagmatic variations play a major role in knowledge representation (Faber et al., 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 cognition, syntax and pragmatics. These aspects explain the intersection between general and specialized discourse, as they cannot be separated into two water-tight compartments. Understanding the interplay between patients' general discourse and healthcare providers' specialized discourse is essential for effective communication.

2.2 Texts for patients

Real life doctor-patient interaction has the advantage of rephrasing, asking questions and answering in a short time. Nonetheless, laymen patients often lack the specialized knowledge required to understand a doctor's report or laboratory test on their own.

This explains the relevance of health literacy and the importance of providing health services targeting patients care. Health literacy is the individual's capability to effectively understand and use health information Ratzan (2001). Interpreting complex medical terminology in healthcare settings is fundamental for patients. In the context of NLP, ML or deep learning applications must be designed to assist patients to accurately recognize medical complex words to provide clear, comprehensible information which allows LS.

Many countries, including the USA, Canada, Australia, China and countries from the European Union, have prioritized health literacy in their policies and practices. In line with Juvinyà-Canal, Bertran-Noguer, and Suñer-Soler (2018), specifically in Europe, health literacy is seen as a vital component of the European health strategy.

The Spanish healthcare policies highlight patients' awareness since 2019, when the Spanish Health Literacy Network started collaborating in research projects. Members of the association have focused on heart disease (Falcón et al., 2022) or COVID-19 (Martínez et al., 2022), and started other campaigns such as *Health without doubts* (Fernández, Juvinyà, and Suñer, 2021b) or *Always ask three questions* (Fernández, Juvinyà, and Suñer, 2021a).[2]

These initiatives help patients make informed decisions about their health, adhere to medical advice and engage with healthcare providers, owing to the following reasons:

1. Clarity and precision: language experts ensure that medical terms are clear, reduce ambiguity and lower the risk of misinterpretation. 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.

2. Standardization: linguists contribute to the standardization of medical knowledge, and facilitate consistent terminology across languages and regions, which helps global health communication.

2.3 Terminological resources

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

A crucial part of the complex word 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:

[2] `https://shorturl.at/jrjtu` (Accessed 2024/8)

1. The International Classification of Diseases (ICD) vs 10 (World Health Organization, 2004).[3]
2. The Medical Subject Headings (MeSH) (Lipscomb, 2000).[4]
3. SNOMED Clinical Terms (SNOMED CT) (Donnelly and others, 2006).[5]
4. The Medical Dictionary for Regulatory Activities (MedDRA) (Brown, Wood, and Wood, 1999).[6]
5. The Anatomical Therapeutic Chemical (ATC) Classification (World Health Organization, 2019).[7]

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. In the annotation part of this work, the UMLS (Bodenreider, 2004) 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 in Figure 1.

In the NLP field, these terminological resources allow the expansion of complex words with synonyms for the same 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 paraphrases.

2.4 Lexical simplification

Understanding medical texts, such as our own health records or scientific findings related to our medical conditions, is crucial for everyone. However, medical texts often use specialized complex words and abbreviations derived from Latin or Greek. This makes medical texts hard to understand (Keselman and Smith, 2012), and it is not a current issue, as the need for making changes has been studied for decades.

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
ecocardiografia	A5556982	SCTSPA	PT	40701008
ecocardiografia (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
Ecocardiografia	A9214335	MSHSPA	MH	D004452
Ecocardiografia 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

Figure 1: Example of search in the UMLS for *echocardiography* (CUI: C0013516).

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, the International Plain Language Federation[8] is the institution that sets an ISO standard so that readers can easily find and understand what they need. Our work presented in this research article 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.

Alarcon 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, Calixto, and Saggion, 2015), which poses a significant challenge when working with languages that have limited annotated

[3] `https://shorturl.at/kK6Nz` (Retrieved 2024/8)
[4] `https://shorturl.at/qAR7V` (Retrieved 2024/8)
[5] `https://www.snomed.org/` (Retrieved 2024/8)
[6] `https://www.meddra.org/search` (Accessed 2024/8)
[7] `https://shorturl.at/9j7JA` (Retrieved 2024/8)
[8] `https://www.iplfederation.org/` (Accessed 2024/8)

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: CWI, 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.

Specifically, in this article, we define complex word as any lexical item that hinders the understanding of the text contents for a non-specialized reader. Complex words may be medical terms that convey specialized knowledge (e.g., jargon or acronyms), or general domain words that are rare in everyday usage, which include mono-, multi-words and abbreviations. Given the variety of profiles with which we could associate the means of this task, we focus on individuals with functional literacy as to using the technical resources required to access a medical text understanding tool.

In the past decade, research groups have extensively explored CWI and several shared tasks have been organized (Yimam et al., 2017; Yimam et al., 2018; Ortiz-Zambranoa and Montejo-Ráez, 2020; Saggion et al., 2023). Some teams have used lexicon-based approaches over the past few years (Sulayes, 2020; Deléger and Zweigenbaum, 2009), which demands creating datasets with candidate complex words, often using specialized medical sources (Elhadad and Sutaria, 2007); a list of resources is enumerated in (Paetzold and Specia, 2017).

Another approach is based on word frequency thresholds (Bott et al., 2012; Leroy et al., 2013), which may nonetheless lack adequate performance in real use. Other groups have addressed CWI by means of machine learning (Shardlow, 2013) or formulating it as a sequence labeling task (Yimam et al., 2017; Gooding and Kochmar, 2019). Recently, standard logistic regression models have been applied across languages to alleviate the data bottleneck (Finnimore et al., 2019). The predominant trend has been characterized by word embeddings, and, in recent times, deep learning (De Hertog and Tack, 2018), including recurrent neural networks (RNN) (Pylieva et al., 2019) and current Large Language Models (LLMs) (Smădu et al., 2024). For more details on CWI, we refer to a recent survey (North, Zampieri, and Shardlow, 2023).

3 Methods

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 models for lexical simplification, as well as other traditional ML algorithms. For this purpose, three collections of 75 open-source texts have been manually annotated and peer-revised to achieve a gold standard. Note that these texts do not contain personal data from patients. These sets of texts belong to three different typologies:

1. Consent forms (CFs):[9] 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 Fundación Rioja Salud[10] and accredited websites, such as Consejería de Salud y Consumo de la Junta de Andalucía.[11]

2. Clinical trial announcements (CTAs): 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).[12]

3. Patient information documents (PIDs): texts of informative nature targeting patients and general audience. Topics range from transplants to several 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[13] and the Spanish National Transplant Organization.[14]

The corpus files were divided into three sets of training (60%), development (20%) and test (20%). The statistics of each collection of texts can be seen in Table 1.

[9]These consent forms were provided by Ana Rosa Terroba, a collaborator in the CLARA-MeD project, in which this research was conducted.

[10]`https://shorturl.at/rMN4M` (Retrieved 2024/1)

[11]`https://shorturl.at/xgkkk` (Retrieved 2024/1)

[12]`https://shorturl.at/Yn8IW` (Retrieved 2024/1)

[13]`https://shorturl.at/sNqsr` (Retrieved 2024/1)

[14]`https://www.ont.es/` (Retrieved 2024/1)

Set	#Texts	#Sentences	#Tokens	#Complex Words
CFs	51/9/15	1798/429/528	33473/7338/9525	3354/699/851
CTAs	51/9/15	1699/324/374	32755/8006/6901	5105/1190/1114
PIDs	51/9/15	2064/281/680	43161/7014/14402	4043/678/1169
Total	153/27/45	5561/1034/1582	109389/22358/30828	12502/2567/3134

Table 1: Corpus statistics; counts in each split are separated by /, i.e., (train/dev/test).

3.2 Annotation process

All texts have been pre-processed using the CLARA-MeD tool (Campillos-Llanos et al., 2024),[15] a dictionary-based system that detects difficult-to-understand complex words according to dedicated patient-oriented lexicons and a word frequency list with a threshold of 5000 (Figure 2). The detected words are underlined and are possible candidates to be annotated as (`complex_word`, `CW`). The resulting text with the highlighted complex words is then uploaded to the BRAT annotation tool (Stenetorp et al., 2012) in `.ann` format. Lastly, linguists manually validate if the automatically detected candidates are complex words (or whose definition or synonyms need to be improved), or label more complex words that the CLARA-MeD tool did not detect based on difficulties overcome during the annotation process.

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.

3.3.1 Nested complex words

Nested complex words are not annotated. Only the more specific `complex_word` (or that with the longest span) is labeled. For example (id of the text is given between brackets): *oclusión tubárica* instead of *tubárica* and *oclusión tubárica* (2022-000422-16).

3.3.2 Frequent complex words

Frequent complex words that are commonly used by patients are not annotated; for example, *cirugía*, *diabetes* or *cáncer*. When there is doubt on whether to annotate or not a complex word, we decide if the word can be further simplified. For example,*intervención quirúrgica* can be simplified to *operación* (2022-002680-30). Therefore, we annotate *intervención quirúrgica*, but not *operación*.

[15] `http://claramed.csic.es/demo` (Accessed 2024/8)

3.3.3 Discontinuous complex words

Discontinuous entities are not annotated; the full span between the discontinuous entities (including the words between them) is annotated instead. For example, we annotate *tumor (T) 4b*, instead of *tumor* and *4b* separately. This is also the case for multi-word elements only discontinued by a typographic symbol annotated as one complex word. For example, we annotated *VIH 1/2* instead of *VIH 1* and *2* separately (2022-003594-33).

3.3.4 Measure units

Measure units will not be annotated when they are of general use (e.g., *mg.*). However, we annotated those that are infrequent to a layman reader and are needed to understand the text. For example, μg ('microgram') in: $\geq$2 μg (2021-001396-16).

3.3.5 Foreign words

Foreign words will not be annotated if there is a translated equivalent in the text. Example: *Diagnóstico de enfermedad de Crohn (Crohn's disease)* (2021-003314-39).

However, we annotate any foreign word if it is only used without an Spanish translation. Example: *RSV* which stands for 'Respiratory Syncytial Virus' (2022-003124-41).

3.3.6 Names of genes

Genes will not be annotated, except for those which are highly relevant or associated with a disease. Example: *BRCA*: gene associated with breast cancer (2022-003594-33).

3.3.7 Names of clinical trials

Names of clinical trials are not be annotated; e.g., *CLOU064A2301* (2022-001034-11).

3.3.8 Synonyms

We annotate synonyms of the same complex words; e.g., *diagnóstico temprano*, synonym of *diagnóstico precoz* (aula_cyl_erc_3).

3.3.9 Adjectives

Adjectives which do not belong to a phraseological unit will not be annotated. Example: *valvulopatía estenótica grave* in which *grave*

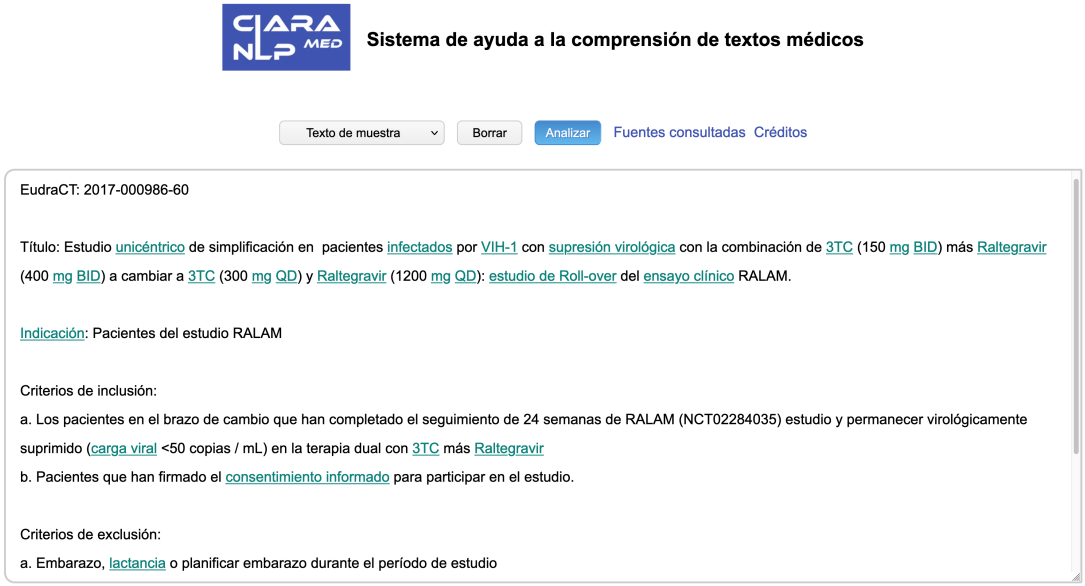

Figure 2: Example of the CLARA-MeD tool and a sample CT text.

is not included as a phraseological unit (2020-003312-27). However, the adjectives often create a complete `complex_word` with a concept identifier in the UMLS Metathesaurus. Thus, we annotate such cases; for example: *insuficiencia cardiaca refractaria* (ci_sec_1).

3.3.10 Organizations

We annotated acronyms of medical organizations; e.g., *NYHA* (2020-003312-27).

3.4 Inter-annotator agreement

Four linguists revised the 225 texts (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 BRATEval library[16] in java. Once the texts were annotated, the .ann files were extracted. These .ann files were then converted to .conll format. As the last step, the .conll format files were converted to .json to adapt to the requirements of models in Hugging Face.

3.5 Model Training

Transformer-based models available on the Hugging Face hub (Wolf et al., 2020) were utilized for the CWI task. The choice of transformers was based on their state-of-the-art performance for NLP tasks. The models were fine-tuned for our task by adapting the notebook on token classification[17] available in the Hugging Face's GitHub. The following are the seven models used in this work:

1. BETO – Spanish BERT (Cañete et al., 2020): a monolingual, general domain model with 110M parameters, trained on the Spanish subsets of Wikipedia and the Open Parallel Corpus.

2. Biomedical Language model for Spanish (RoBERTa EHR) (Carrino et al., 2022): a monolingual and domain-specific model, trained on Spanish medical texts (125M parameters).

3. RoBERTa-EHR-CT (Campillos-Llanos et al., 2021): a monolingual and domain-specific pretrained model from RoBERTa EHR, with the same number of parameters, but fine-tuned on 1200 texts about clinical trials in Spanish.

4. mBERT base model (Devlin et al., 2018): a multilingual, general domain model with 110M parameters, pretrained on Wikipedia for 104 languages.

5. Medical mT5 (García-Ferrero et al., 2024): a multilingual, domain-specific model, which was pre-trained on large-scale medical data sources in English, French, Italian and Spanish; we used the large version (738M parameters).

6. mDeBERTa v3 (He, Gao, and Chen, 2021): a multilingual, general domain

[16] `https://shorturl.at/tlVSq` (Retrieved 2024/4)
[17] `https://shorturl.at/Y1vRP` (Retrieved 2024/8)

model, pre-trained on the CommonCrawl corpus for 100 languages; we used the large version (190M parameters).

7. RoBERTa-BNE-large (Gutiérrez-Fandiño et al., 2022): the large version (355M parameters) of the monolingual, general-domain model trained on data from the Spanish National Library.

Each transformer 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, and we provide the average and standard deviation of the three experimental rounds for each model. All experiments were run in Google Colab Pro.

In addition to the transformer-based models, we have tested other classical ML models as baselines, using Scikit-learn (Pedregosa et al., 2011): SVMs and LR classifiers. SVM classifiers work by finding the hyperplane that best separates data points of different classes with the maximum margin. SVMs use kernel functions to transform the input space, allowing for complex decision boundaries. For its part, LR is a binary classification model that uses the logistic function (sigmoid). The classifier models the input features using optimization techniques and outputs values between 0 and 1.

For ML models (SVM and LR), we report the average across 5-fold cross-validation on the train and development sets, and the final scores on the test set. We employed features focusing on lexical classes and positional characteristics of tokens. Namely, we used the lowercase form of the token, the preceding word, and the subsequent word to capture contextual information. Additional binary features were incorporated to identify if the token was in uppercase, title case, or consisted entirely of digits. Furthermore, we utilized the part-of-speech (POS) tag of each token, obtained using the spaCy library (Honnibal et al., 2017), to enrich the syntactic understanding of the text.

The performance of the models was measured using common evaluation metrics (Hripcsak and Rothschild, 2005), and post-evaluation of results was conducted to analyze false positives and false negatives. The evaluation metrics were precision, recall, F1-score and accuracy for all approaches. Given the high classification imbalance—i.e., most tokens do not belong to the `complex_word` class—we report the micro-average F1 and we use the `class_weight='balanced'` parameter for the Scikit-learn classifiers.

4 Results

The transformer model with the best performance in the F1 measure was mDeBERTaV3, followed by BETO and Medical mT5; whereas for other ML classifiers, SVM succeeded in the CWI task compared to Logistic Regression (Table 2). In the 5-fold cross-validation with the training and development sets, the SVM obtained an average F1 of 90.29, but the F1 decreased when applied on the test set (F1 = 76.6). However, the F1 value was still below a random classifier.

The results from mDeBERTaV3 suggest that it captured a high number of correctly labeled complex words. When it comes to recall, the MarIA model scored the best results with 82.98, which indicates that it had 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 appeared to be the most balanced one, with the highest F1 score. Furthermore, it had a strong precision and a similar recall. The BETO model excelled in recall as well, 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 scored the highest accuracy, which indicates good performance, yet less favorable F1 outcomes than the vast majority of the models. Overall, the RoBERTa EHR model performed the weakest among transformers, with the lowest metrics' scores, specifically an average F1 of 69.62.

In terms of performance across classes, we observed that the inside class (`I-CW`) achieved lower scores compared to the begin class (`B-CW`). For example, the `B-CW` class had an F1 score of ∼57% and ∼59% on the test set with the LR and SVM models, respectively; whereas the F1 score of the `I-CW` class decreased to ∼26% and ∼25%, respectively. Figure 3 shows the confusion matrix of the gold standard and the predictions by the mDeBERTa model. In a shallow error analysis of this model, most errors seem to appear in the `B-CW` class, which is mislabelled as the `O` class, or vice versa. A detailed error analy-

	Precision	Recall	F1	Accuracy
dccuchile-bert-base-spanish-wwm-uncased	75.01 (±1.11)	82.98 (±0.60)	78.78 (±0.34)	93.49 (±0.13)
roberta-bne-large	64.50 (±11.50)	**83.40 (±0.54)**	72.32 (±7.47)	92.07 (±2.97)
bert-base-multilingual-cased	76.23 (±0.78)	77.03 (±1.74)	76.63 (±1.19)	92.10 (±0.48)
microsoft-mdeberta-v3-base	**79.05 (±1.39)**	79.01 (±0.70)	**79.02 (±0.65)**	94.86 (±0.22)
RoBERTa-bsc-bio-ehr-es	62.58 (±3.34)	78.54 (±0.31)	69.62 (±2.03)	94.21 (±0.62)
RoBERTa-es-clinical-trials-ner	70.44 (±1.07)	78.82 (±1.32)	74.39 (±0.98)	95.21 (±0.07)
Medical-mt5-large	74.94 (±1.16)	82.07 (±0.40)	78.34 (±0.77)	94.72 (±0.13)
LinearSVM	76.60	76.60	76.60	76.60
Logistic Regression	75.24	75.24	75.24	75.24

Table 2: Results on the test set (for transformer models, we report the average of 3 experiments and the ± standard deviation).

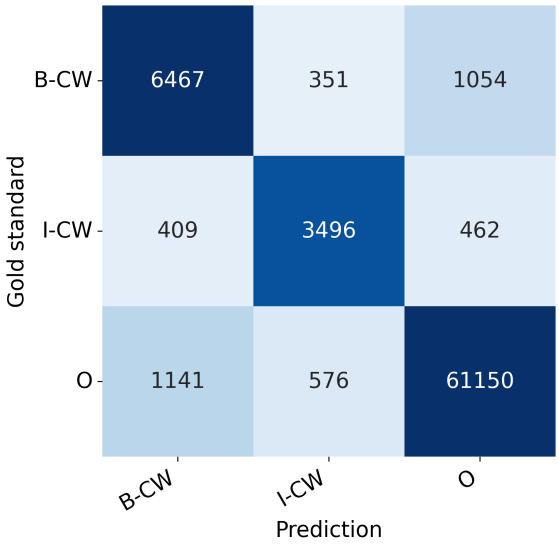

Figure 3: Confusion matrix for the test results with the mDeBERTa model compared to our gold standard.

sis could shed more light on the most frequent words that are misclassified. Due to the small size of our annotated corpus, a thorough error analysis is left for future work, when more data is collected and annotated.

5 Discussion

Comparing these findings with previous studies, we observe certain similarities with other classifiers for the CWI task. For example, Alarcon et al. (2019) reported an F1 score of 74.97 using an SVM classifier, achieving slightly lower results. In another article by Alarcon, 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. In our research, the ability of transformer-based models to capture semantic relationships and contextual information make them dynamic, especially for multiword CWs. 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.

Surprisingly, the general-purpose models achieved the best scores, compared to the domain-specific models like mT5, which was pre-trained using medical texts from sources like ClinicalTrials or PubMed. This could mean some of the annotated complex words

in the training data do not belong necessarily to the medical domain or are polysemous words. Some examples might be:

1. *revocar*, *consentimiento* or *reintervención* (examples from several texts).
2. *exploración*: when used as examination to find a pathology instead of going out to explore a new place (ci_sed_9).
3. *coma*: used as the state of profound unconsciousnes instead of the third person of the verb *comer* in Spanish (aula_cyl_diabetes_5).
4. *instrumental*: as the medical equipment necessary in a surgery instead of the music-related meaning (ci_86).
5. *progresar*: with a negative meaning as 'worsening' rather than a positive one (ci_ser_4).

In spite of the favorable results obtained, we are aware of certain challenges, given the nature of the problem and due to the fact that developing the corpus and annotating it was a crucial starting point. Addressing the identified limitations is feasible by adopting strategies such as: 1) Increasing the corpus size, as this would allow for further generalization by the transformers models. 2) Introducing embeddings as features for the ML model. 3) Performing a thorough error analysis on a subset of predictions, to provide further insights to improve the identification and classification of complex words. As observed in our results, the poorer performance on the inner class (`I-CW`) could reflect that models are less robust when predicting whether modifiers or adjectives should be considered part of the complex word (e.g., *aguda* in *apendicitis aguda*, 'acute appendicitis'). Indeed, annotators often hesitated to annotate a wide or short span of multi-word CWs. However, understanding where the models fail is still unclear. 4) Using metadata regarding the thematic sub-domains within the corpus to establish a difference between complex words belonging to different specialties, e.g. complex words in ophthalmology or in oncology. 5) Reexamining our corpus to establish a difference between complex words belonging to the general domain and those complex words belonging exclusively to the medical domain. In doing so, we might ascertain an explanation of performances using general and domain-specific models.

After considering our limitations, further factors should be taken into account if the tested classifiers are to be integrated in real-time services or solutions. If so, the feasibility of integrating ML-based or transformer-based classifiers would vary. Transformer classifiers provide better results but are less transparent and require high computational capacities. On the other hand, ML classifiers provide less accuracy, but they are not computationally demanding.

6 Conclusions

In this study, we evaluated standard ML models (SVMs and Logistic Regression) and fine-tuned transformer-based models for CWI in Spanish texts for patients. Our results showed that transformer-based models tend to achieve similar F1 scores and perform better than traditional ML classifiers. Our 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 225 texts and 18203 complex words, which may limit the training of our models. Therefore, future research could explore extending the annotations and enlarge the corpus size for better generalizations, and reassess which complex words might belong to the general discourse. By continuing to refine and expand upon these methods, we can make significant strides towards more inclusive communication between Spanish health providers and patients.

Acknowledgements

We greatly thank Ana Valverde-Mateos and Ana Rosa Terroba-Reinares for revising the annotation of some texts of the experimental corpus. Federico Ortega-Riba was funded by a CSIC JAE Intro 2023 scholarship. This work was conducted within the framework of the CLARA-MeD project (PID2020-116001RA-C33), funded by MICIU/AEI/10.13039/501100011033/ (call: "Proyectos I+D+i Retos Investigación").

References

Alarcon, R., L. Moreno, and P. Martínez. 2021. Lexical simplification system to improve web accessibility. *IEEE Access*, 9:58755–58767.

Alarcon, R., L. Moreno, I. Segura-Bedmar, and P. Martínez. 2019. Lexical simplification approach using easy-to-read resources. *Procesamiento del Lenguaje Natural*, 63:95–102.

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

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

Bott, S., L. Rello, B. Drndarević, and H. Saggion. 2012. Can Spanish be simpler? LexSiS: Lexical simplification for Spanish. In *Proceedings of COLING 2012*, pages 357–374.

Brown, E. G., L. Wood, and S. Wood. 1999. The medical dictionary for regulatory activities (MedDRA). *Drug safety*, 20(2):109–117.

Cabré, M. T. 2003. Theories of terminology: Their description, prescription and explanation. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 9(2):163–199.

Campillos-Llanos, L., F. Ortega-Riba, A. R. Terroba-Reinares, A. Valverde-Mateos, and A. Capllonch-Carrión. 2024. CLARA-MeD Tool – A System to Help Patients Understand Clinical Trial Announcements and Consent Forms in Spanish. *Studies in Health Technology and Informatics*, pages 95–99.

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 medical informatics and decision making*, 21:1–19.

Carrino, C. P., J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estapé, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre, and M. Villegas. 2022. Pretrained biomedical language models for clinical NLP in Spanish. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 193–199, Dublin, Ireland, May. Association for Computational Linguistics.

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

Cimino, J. J. 1998. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of information in medicine*, 37(04/05):394–403.

De Hertog, D. and A. Tack. 2018. Deep learning architecture for complex word identification. In *Proceedings of the 13th workshop on innovative use of NLP for building educational applications*, pages 328–334.

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 (BUCC)*, pages 2–10.

Devlin, J., M. Chang, K. Lee, and K. Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Donnelly, K. et al. 2006. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279.

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.

Faber, P., M. Tercedor, S. Montero Martínez, P. Araúz, J. A. Prieto Velasco, C. Lopez-Rodriguez, A. Reimerink, C. Linares, M. De Quesada, J. Gómez-Moreno, and A. San Martín. 2012. *A Cognitive Linguistics View of Terminology and Specialized Language*. Applications of cognitive linguistics. De Gruyter Mouton.

Falcón, M., A. Torres, E. Hernández, I. Del Arco, P. Bas, and M. Fernández.

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.

Fernández, M., D. Juvinyà, and R. Suñer. 2021a. Salud sin dudas – salut sense dubtes.

Fernández, M., D. Juvinyà, and R. Suñer. 2021b. "fes sempre tres preguntes" (haz siempre tres preguntas).

Finlayson, M. A. and T. Erjavec. 2017. Overview of annotation creation: Processes and tools. *Handbook of Linguistic Annotation*, pages 167–191.

Finnimore, P., E. Fritzsch, D. King, A. Sneyd, A. U. Rehman, F. Alva-Manchego, and A. Vlachos. 2019. Strong baselines for complex word identification across multiple languages.

García-Ferrero, I., R. Agerri, A. A. Salazar, E. Cabrio, I. de la Iglesia, A. Lavelli, B. Magnini, B. Molinet, J. Ramirez-Romero, G. Rigau, J. M. Villa-Gonzalez, S. Villata, and A. Zaninello. 2024. Medical mT5: An Open-Source Multilingual Text-to-Text LLM for The Medical Domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11165–11177.

Gooding, S. and E. Kochmar. 2019. Complex word identification as a sequence labelling task. In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1148–1153, Florence, Italy, July. Association for Computational Linguistics.

Gutiérrez-Fandiño, A., J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, A. Gonzalez-Agirre, C. Armentano-Oller, C. Rodriguez-Penagos, and M. Villegas. 2022. MarIA: Spanish language models. *Procesamiento del lenguaje natural*, 68:39–60.

He, P., J. Gao, and W. Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *arXiv preprint arXiv:2111.09543*.

Honnibal, M., I. Montani, S. Van Landeghem, and A. Boyd. 2017. spacy: Industrial-strength natural language processing in python. *Journal of Artificial Intelligence Research*, 60:549–593.

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

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

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

Keselman, A. and C. A. Smith. 2012. A classification of errors in lay comprehension of medical documents. *Journal of biomedical informatics*, 45(6):1151–1163.

Leroy, G., J. E. Endicott, D. Kauchak, O. Mouradi, M. Just, et al. 2013. User evaluation of the effects of a text simplification algorithm using term familiarity on perception, understanding, learning, and information retention. *Journal of Medical Internet Research*, 15(7):e2569.

Lipscomb, C. E. 2000. Medical subject headings (MeSH). *Bull Med Lib Assoc*, 88(3):265.

Makaryus, A. N. and E. A. Friedman. 2005. Patients' understanding of their treatment plans and diagnosis at discharge. In *Mayo clinic proceedings*, volume 80, pages 991–994. Elsevier.

Martínez, E., M. Falcón, A. B. Maldonado, G. Ruiz, and O. Monteagudo-Piqueras. 2022. Monitorización del comportamiento y las actitudes de la población relacionadas con la covid-19 en la región de murcia 2020-2022. cosmo- carm: Estudio oms.

Moen, H., L.-M. Peltonen, M. Koivumäki, H. Suhonen, T. Salakoski, F. Ginter, and S. Salanterä. 2018. Improving layman readability of clinical narratives with unsupervised synonym replacement.

In *Building Continents of Knowledge in Oceans of Data: The Future of Co-Created eHealth*. IOS Press, pages 725–729.

Montalt, V. and I. García-Izquierdo. 2016. Exploring the link between the oral and the written in patient-doctor communication. *Ordóñez-López, Pilar & Nuria Edo-Marzá (eds.)*.

North, K., M. Zampieri, and M. Shardlow. 2023. Lexical complexity prediction: An overview. *ACM Computing Surveys*, 55(9):1–42.

Ortiz-Zambranoa, J. A. and A. Montejo-Ráez. 2020. Overview of alexs 2020: First workshop on lexical analysis at SEPLN. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*, volume 2664, pages 1–6.

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

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12:2825–2830.

Pylieva, H., A. Chernodub, N. Grabar, and T. Hamon. 2019. Rnn embeddings for identifying difficult to understand medical words. In *Proceedings of the 18th BioNLP workshop and shared task*, pages 97–104.

Qiang, J., Y. Li, Y. Zhu, Y. Yuan, and X. Wu. 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.

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

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

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

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*.

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

Shardlow, M. 2013. A comparison of techniques to automatically identify complex words. In *51st annual meeting of the association for computational linguistics proceedings of the student research workshop*, pages 103–109.

Smădu, R.-A., D.-G. Ion, D.-C. Cercel, F. Pop, and M.-C. Cercel. 2024. Investigating large language models for complex word identification in multilingual and multidomain setups.

Štajner, S., I. Calixto, and H. Saggion. 2015. Automatic text simplification for Spanish: Comparative evaluation of various simplification strategies. In *Proceedings of the international conference recent advances in natural language processing*, pages 618–626.

Stenetorp, P., S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. In *Proc. of the Demonstrations at the 13th Conference of the EACL*, pages 102–107.

Sulayes, A. R. 2020. General lexicon-based complex word identification extended with stem n-grams and morphological engines. In *IberLEF@SEPLN*.

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

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V.

Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Williams, M., R. Parker, D. Baker, N. Parikh, K. Pitkin, W. Coates, and J. Nurss. 1995. Inadequate functional health literacy among patients at two public hospitals. *JAMA: The Journal of the American Medical Association*, 274(21):1677–1682.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

World Health Organization. 2004. *International Statistical Classification of Diseases and Related Health Problems vs. 10*. World Health Organization.

World Health Organization. 2019. *Anatomical Therapeutic Chemical classification*. Uppsala: Nordic Council on Medicines.

Yimam, S. M., C. Biemann, S. Malmasi, G. Paetzold, L. Specia, S. Štajner, A. Tack, and M. Zampieri. 2018. A Report on the Complex Word Identification Shared Task 2018. In J. Tetreault, J. Burstein, E. Kochmar, C. Leacock, and H. Yannakoudakis, editors, *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana, June. Association for Computational Linguistics.

Yimam, S. M., S. Stajner, M. Riedl, and C. Biemann. 2017. Multilingual and cross-lingual complex word identification. In *RANLP*, pages 813–822.