Procesamiento del Lenguaje Natural, Revista nº 74, marzo de 2025, pp. 221-239 recibido 05-12-2024 revisado 21-01-2025 aceptado 10-02-2025

Exploring Linguistic Features in a New Readability Corpus for Spanish

Exploración de características lingüísticas en un nuevo corpus de lecturabilidad en español

Sandra Rodríguez Rey, André Bernárdez Braña, Marcos Garcia
Centro Singular de Investigación en Tecnoloxías Intelixentes (CITIUS)
Universidade de Santiago de Compostela
{sandrarodriguez.rey,andre.bernardez.brana,marcos.garcia.gonzalez}@usc.gal

Abstract: The reading difficulty of a given text has traditionally been calculated using readability formulas, which measure some linguistic properties of texts and provide a score. Current methods for automatic readability assessment are mostly based on supervised models which use manually defined linguistic features learned from texts classified by readability levels. While reference corpora are available for various languages, existing resources for Spanish are often limited in genre diversity, and primarily designed for tasks like text simplification or teaching Spanish as a foreign language, making them less suitable for training classifiers. This paper presents a new readability corpus for Spanish, which contains 2,563 texts from 11 categories and 68 subcategories, manually classified into four levels of readability. Its compilation and topic selection was specifically defined for adult readers, with a focus on automatic classification tasks. This study also analyzes the most relevant linguistic properties regarding each of the levels, and explores the use of language models' surprisal as a readability predictor, whose correlation with the levels indicates its usefulness for training automatic classifiers.
Keywords: Readability, Text classification, Complexity features, Adult Learning.

Resumen: El nivel de complejidad para la lectura de un texto se calculaba tradicionalmente mediante las fórmulas de lecturabilidad, que miden algunas propiedades lingüísticas de los textos y proporcionan una puntuación. Los métodos actuales de evaluación automática de la lecturabilidad se basan en modelos supervisados que utilizan características lingüísticas definidas manualmente y aprendidas a partir de textos clasificados por niveles de complejidad. Aunque existen corpus de referencia para varios idiomas, los recursos existentes para el español suelen ser limitados en cuanto a diversidad de géneros y están diseñados principalmente para tareas como la simplificación de textos o el aprendizaje del español como lengua extranjera, lo que los hace menos adecuados para el entrenamiento de clasificadores. Este artículo presenta un nuevo corpus de lecturabilidad en español, que contiene 2.563 textos de 11 géneros y 68 subgéneros, clasificados manualmente en cuatro niveles de lecturabilidad. Su compilación y selección temática se definió específicamente para lectores adultos, con especial atención en las tareas de clasificación automática. Este estudio también analiza las propiedades lingüísticas más relevantes en relación con cada uno de los niveles y explora el uso de la *surprisal* de los modelos de lengua como predictor de la lecturabilidad, cuya correlación con los niveles indica su utilidad para el entrenamiento de clasificadores automáticos.
Palabras clave: Lecturabilidad, clasificación de textos, características de complejidad, aprendizaje para adultos.

ISSN 1135-5948 DOI 10.26342/2025-74-15

1 Introduction

Readability refers to how easily a text can be read and understood (Campos Saavedra et al., 2014). Knowing the degree of readability of texts is useful in several areas, such as language learning, the development of tools for people with reading difficulties, the creation of content, and the implementation of accessibility policies. It can facilitate tasks such as selecting appropriate reading materials or adapting the content of texts to make them clearer, more accessible, and easier to understand (Vajjala, 2022).

The traditional calculation method is the application of readability formulas such as Flesch (Flesch, 1948) or SMOG (Harry and Laughlin, 1969), that measure some characteristics of the texts such as word and sentence length, or whether the vocabulary used is familiar or unfamiliar to the reader (Campos Saavedra et al., 2014). These formulas have been gradually replaced by machine learning models based on linguistic features defined by experts, whose results are better than those obtained with readability formulas (François, 2015). The most widely used models use classical machine learning algorithms and, more recently, methods based on deep learning are being explored (Madrazo Azpiazu and Pera, 2019). As most models are supervised, they need corpora annotated according to readability levels.

The literature collects reference corpora for different languages, e.g., English (Schwarm and Ostendorf, 2005), French (Wilkens et al., 2022), or Portuguese (Ribeiro, Mamede, and Baptista, 2024). Available resources for Spanish include the Coh-Metrix-Esp (Quispesaravia et al., 2016), Newsela (Xu, Callison-Burch, and Napoles, 2015), CAES (Rojo and M. Palacios, 2016), CEDEL2 (Lozano, 2022), Simplext (Saggion et al., 2011), kwiziq and Hablacultura (Vásquez-Rodríguez et al., 2022). However, existing resources for Spanish are limited, and some are restricted due to privacy licenses or access difficulties (Vásquez-Rodríguez et al., 2022). Moreover, they tend to lack a variety of genres, and were designed for second language learning or text simplification, being less useful for developing automatic readability assessment tools.

Taking the above into account, this paper introduces a new readability corpus for Spanish, which contains 2,563 documents manually classified into four levels. The corpus compilation has been designed with adult readers with low literacy skills in mind, aimed at developing automatic tools to improve their reading skills. Thus, the texts come from 11 categories (and 68 subcategories), and include a wide variety of topics. In addition to the design, compilation, and classification processes, this study analyzes the distribution of various linguistic features across levels, reinforcing previous research that indicate that length-based features or lexical diversity correlate with readability. Additionally, we also explore the use of language models' (LMs) surprisal as a readability predictor, suggesting that average surprisal from both encoder and decoder transformer models might be used for classification tasks.

In summary, this paper contributes a new corpus specifically designed for automatic classification tasks, and an analysis of linguistic features that affect text complexity, together with the exploration of LMs' surprisal as a readability predictor. This new resource is available at Zenodo and is freely distributed and re-usable according to CC BY-NC-ND 4.0 license for research purposes.[1] It is, to the best of our knowledge, the largest readability corpus available for Spanish.

Besides this introduction, the paper is organized as follows: Section 2 introduces the most relevant readability corpora, with a special focus on Spanish resources. Then, Section 3 outlines the corpus creation process and its key characteristics, while Section 4 presents a detailed analysis of linguistic features across levels. Then, we explore the use of LMs surprisal as a readability predictor (Section 5), and conclude this study and draw ideas for future research in Section 6.

2 An overview of existing readability corpora

In recent years, a great deal of research has focused on the creation of corpora of texts labeled by levels of complexity, both in Spanish and in other languages. This section presents the most relevant ones for Spanish, English and some related Romance languages. First, the existing resources for Spanish are described, followed by a selection of English cor-

[1] The corpus is included in the multilingual iRead4Skills Dataset 1: corpora by complexity level for FR, PT and SP, available at `https://zenodo.org/records/13127399`.

pora, and some similar datasets for Romance languages as well as multilingual resources.

Among those datasets for Spanish, as stated by Vásquez-Rodríguez et al. (2022), there are corpora classified by levels of complexity that are freely available, and others with privacy licenses permitting their use, such as Newsela or Simplext. We describe these corpora, as well as the Coh-Metrix-Esp, CAES, CEDEL2, kwiziq and HablaCultura.

The Coh-Metrix-Esp corpus (Quispesaravia et al., 2016) is composed of 100 texts of two levels: 50 literary texts labeled as easy, which are mainly fables for children, and 50 labeled as difficult, which are mainly stories for adults (Vásquez-Rodríguez et al., 2022). In addition to being used in the development of the Coh-Metrix-Esp tool, this corpus has also been used in the creation of MultiAzterTest, a tool for analyzing textual stylometry and evaluating the readability of texts (Bengoetxea and Gonzalez-Dios, 2021).

Newsela (Xu, Callison-Burch, and Napoles, 2015) is a corpus of 1,130 news articles (1,301,767 tokens) rewritten by translators to match four complexity levels based on school grades (Xu, Callison-Burch, and Napoles, 2015). The texts are classified following the Lexile readability score, a metric that defines the readability of a text based on syntactic and semantic features (Lennon and Burdick, 2014).

The Corpus de Aprendices de Español (CAES) (Rojo and M. Palacios, 2016) consists of texts written by learners of Spanish at different levels according to the Common European Framework of Reference (CEFR). The learners come from diverse countries and have different native languages. The current version (2.1) consists of 6,561 texts (1,045,097 language units), which reflect the most common topics and text types for each level. It is worth noting that texts may contain errors as they were written by foreign students.

In addition to these three corpora, there are other more recent resources, such as Simplext (Saggion et al., 2011; Saggion et al., 2015), which contains texts at two difficulty levels: 200 journalistic texts and their adaptations with manual simplifications made by the DILES research group of the Universidad Autónoma de Madrid (Saggion et al., 2011) or those extracted from kwiziq and HablaCultura (Vásquez-Rodríguez et al., 2022), which contain free articles from these websites tagged according to the CEFR.

There are several corpora with similar characteristics in English, with different annotations related to text readability: based on the recommended reading age and categorized into three readability levels. First, the Weekly Reader corpus (Schwarm and Ostendorf, 2005) consists of 2,400 texts from an American educational news magazine, classified based on the recommended reading age, with labels indicating the readability level (Jian, Xiang, and Le, 2022). Second, WeeBit (Vajjala and Meurers, 2012) includes part of the Weekly Reader corpus and texts from the BBC Bitesize website (Jian, Xiang, and Le, 2022). Third, TextEvaluator (Sheehan, Flor, and Napolitano, 2013) contains 934 texts that are classified by level of complexity (from 8 to 18 years old) and by genre (informational, literary, and mixed) and subgenre, making it a representative dataset of the full range of text types considered by teachers and students in U.S. classrooms (Sheehan, Flor, and Napolitano, 2013). In addition, the CommonLit Ease of Readability (CLEAR) corpus (Heintz et al., 2022), which consists of approximately 5,000 texts categorized by grade level (from 8 to 18 years old), is similar to the previous one in terms of genre and subgenre classification. Finally, the OneStopEnglish corpus contains 189 newspaper articles collected from The Guardian in three versions rewritten by teachers, and is classified into three levels according to their readability (Vajjala and Lučić, 2018).

Other datasets with similar characteristics for other languages such as French, Italian, Portuguese, Basque or Catalan should also be considered, including multilingual corpora. Recent resources in French include the FLE-CORP, FLM-CORP and FSW corpora, and the Ljl, bibebook and JeLisLibre datasets. FLE-CORP is a corpus of 2,734 texts, classified according to CEFR levels, extracted from French as a foreign language books, including texts of 8 text types (Wilkens et al., 2022). The FLM-CORP (Wilkens et al., 2022) consists of 334 texts classified in 9 categories according to school years (from 9 to 18 years old) and extracted from Belgian French, history and science textbooks, belonging to four text types. The other corpora contain only literary texts. On the one hand, the three datasets of Hernandez, Oulbaz, and Faine (2022), Ljl (746 books), bibebook (207

books) and JeLisLibre (44 books), contain corpora of children's and adult books with free licenses, classified by complexity level: based on StoryWeaver levels (Ljl); in three levels according to recommended reading age (bibebook); and by school year (JeLisLibre). On the other hand, FSW (Ngo and Parmentier, 2023) comprises 1,228 texts from StoryWeaver, categorized into five readability levels based on the website's level description.

For Italian, the READ-IT corpus (Dell'Orletta, Montemagni, and Venturi, 2014), consists of 638 journalistic texts (original and easy-to-read versions), while the CELI corpus (Grego Bolli, Rini, and Spina, 2017) contains 213 annotated texts from B2 and C2 exams. In Portuguese, the dataset created by Ribeiro, Mamede, and Baptista (2024) contains 598 texts extracted from the Portuguese exams of the Instituto Camões and classified according to CEFR levels from A1 to C1. In Basque, there is the Leveled Basque Science Popularization Corpus (Gonzalez-Dios et al., 2014), which includes 400 scientific articles of two levels (simple and complex) (Bengoetxea and Gonzalez-Dios, 2021).

Finally, there are two multilingual corpora that deserve to be mentioned. VikiWiki (Madrazo Azpiazu and Pera, 2020) includes about 5,400 texts in Spanish, English, Catalan, Basque, French and Italian, randomly selected from Vikidia, an online and public encyclopedia for people with low reading comprehension that adapts Wikipedia articles, and the corresponding Wikipedia texts. CEDEL2 (Lozano, 2022) contains 4,399 documents, 1,112 of them written in Spanish by native speakers, and the rest by Spanish learners, which can be filtered by Spanish proficiency in 6 levels: beginner, intermediate, and advanced (each with lower and upper sublevels). It also includes texts in English, Portuguese, Greek, Japanese and more.

So far, we have identified six methods for classifying corpora. The first approach distinguishes two difficulty levels based on the source and genre, such as the Vikiwiki, LBSPC, READ-IT and Coh-Metrix-Esp corpora. The second one is by CEFR levels, using texts already classified by experts, such as the Instituto Camões dataset and the CELI, FLE-CORP, kwiziq, HablaCultura, CAES and CEDEL2 corpora. The third is based on the levels from the resource they came from (FSW or Ljl). Corpora can also be classified by recommended reading age, as seen in Bibebook, TextEvaluator, Weekly Reader, and WeeBit corpora. Some resources are classified by grade level (JeLisLibre, FLM-CORP, CLEAR, and Newsela). Finally, classification can be defined by experts, as in OneStopEnglish (classified by teachers) and Simplext (classified by researchers).

However, most of these classifications may not be suitable for readability classification tasks due to the following reasons: (i) two levels are insufficient for indicating reading complexity; (ii) the CEFR levels are designed for foreign language learners, who have different difficulties than native speakers; (iii) the ad-hoc classifications mentioned above include literature for children and adolescents, who have different reading preferences than adults; (iv) the reading age is usually based on school years, from childhood to adulthood (usually up to 18 years), and the same is true of (v) grade level classifications, as most adults with low literacy skills do not continue their education beyond this age; (vi) levels defined by experts may be accurate if they are designed for readability classification tasks rather than simplification tasks, and a standardized classification is needed.

Existing resources for Spanish include a very limited variety of text genres for classification tasks, as they come from the literary, journalistic and educational fields. Moreover, some of them are designed for text simplification, as in the case of Newsela and Simplext, and others are designed for teaching and learning Spanish as a foreign language, as in the case of the CAES, CEDEL2, kwiziq and HablaCultura corpora, all of which are labeled according to the CEFR. Therefore, the Spanish corpus presented here is a valuable resource for research in readability, offering broad utility due to its wide variety of genres and topics. To the best of our knowledge, it is the largest readability corpus available for Spanish, featuring its own classification developed with input from experts in adult education and readability research.

3 Description of the corpus creation process

This section describes the main characteristics of the corpus, starting by the design guidelines, its compilation process as well as the classification and validation of the doc-

uments. The corpus is representative of the standard Castilian variety of Spanish.

3.1 Corpus creation guidelines

Readability levels: We started by defining the levels of the corpus according to their degree of readability. We defined 3 levels: L1 (very easy), L2 (easy), and L3 (plain), together with an additional fourth level (L4, including documents with higher degrees of complexity, and compiled after the validation process) which serves as a control category to capture texts that do not clearly align with the primary three readability levels, and therefore helping to delineate the boundaries between the target levels (L1-L3). The three target levels have been defined by experts according to a set of lexical and conceptual, verbal, syntactic and cohesion characteristics (see Appendix A for details).[2]

Genres and domains: Texts of different genres and topics are included, representing the most common text genres and topics of interest to an adult reader (Correia et al., 2024), from news or travel guides to encyclopedia entries, cooking recipes, or self-help books. In total, we have collected documents from 11 categories and 68 subcategories. A detailed list of categories and subcategories is provided in Appendix B. An attempt was made to compile ten texts per level and subcategory from different authors, although this was not fully achieved after the validation process, since some text genres, due to their characteristics, correspond to one level of readability and there is a low volume of written texts of more or less readability within these genres.[3]

Corpus size: Regarding the size of the corpus, we attempted to collect at least 2,000 documents, which would be sufficient for training automatic classification tools. At the end of the compilation and validation process (see below), we reached 2,563 texts (660 of level 1, 660 of level 2, 889 of level 3, and 354 classified as more complex, L4). Concerning the length of the documents, we selected texts between 250 and 500 words.

Metadata and format: For each compiled document we incorporate its metadata in a spreadsheet which includes information such as author, translator, language, publisher, place and date of publication, ISBN, ISSN, URL, DOI, etc. Regarding the format, we include two versions of each document: the original format (e.g., HTML or PDF) together with a TXT file.

3.2 Compilation and validation

The methodology of the corpus creation can be grouped in 4 main steps: (i) source and document selection; (ii) excerpt selection and revision; (iii) initial classification and validation; and (iv) final compilation.

Source and document selection: A large number of sources of different types were used in the selection process. From many sources, texts were collected for only one specific subcategory and in many cases only one text per source. The types of sources used are varied, including: blogs, newspapers and magazines, institutional websites, social networks, online store websites, academic databases, books, educational resources, scientific data sources, government data sources, corporate databases, political party websites, legislative portals, and websites of religious institutions.

Excerpt selection and revision: After the document search phase, text excerpts were selected based on the defined length guidelines. The texts were revised, and typing errors were corrected. Metadata for each document was also collected in this phase.

Initial classification and validation: The process of compiling, classifying and validating the texts was carried out by two linguists with different but complementary perceptions: One being a native speaker of Spanish with language teaching experience, and the other a foreign language speaker with a good command of the target variety. An initial classification was performed using a set of complexity descriptors initially defined for each level (detailed in Appendix A) as a guide. This appendix contains, among others, lexical and conceptual, syntactic, verbal and cohesion descriptors by level of readability. These descriptors are very detailed and many of them cannot be automatically extracted (e.g., the proportion of concrete concepts, pronoun anaphora and ellipsis, specialized concepts, or linear temporal relations). Therefore, they were applied at a surface

[2] An up-to-date version of the Complexity Levels defined in the iRead4Skills project can be found at `https://zenodo.org/records/10459090`.

[3] For example, it is almost impossible to find administrative documents belonging to L1.

level following the experts' intuitions in accordance with the objectives of the corpus.

After compilation, each linguist validated the documents compiled by the other, ensuring accuracy, consistency, and adherence to corpus criteria. The final level of each text was determined by this validation, as the pre-classified level was available and already judged as adequate or inadequate. In some cases of doubt, other experts were consulted, and a meeting was held to determine the final readability level.

The validation process by the two linguists allowed us to calculate the agreement between them: the observed agreement between the initial classification and the validation was of 0.45, while the Cohen's κ was of 0.25.[4] As previous research, these results indicate that both the complexity descriptors and the classification task is challenging and involves a certain degree of subjectivity. When analyzing the validation individually, we have observed that the native linguist modified more documents initially classified by the foreign expert than vice-versa ($\kappa = 0.11$ vs. $\kappa = 0.30$), suggesting that the differences in criteria may also be influenced by a different perspective on text complexity depending on the native or non-native background of the expert.

Final compilation: After the collection and validation phases, 1,076 texts were reclassified according to the results of the validation process. Due to the reclassification, there was a notable imbalance in the number of texts among the three levels, with very few texts remaining in level 1 and the emergence of a new category of texts more complex than those originally considered (L4). Therefore, a new phase of compilation for L1 and L2 was initiated to obtain 590 additional texts. It is worth noting that for this final stage we relied on sources with an already defined classification (e.g., learning materials), which together with the validation by experts ensures a higher quality.

4 Characteristics of the texts according to readability levels

The corpus has been processed with NLP tools and enriched with information from external resources. Then, we observed the distribution of features studied in previous work as potential descriptors of different levels of readability, including length-based features, distance between syntactic dependencies, percentages of PoS-tags, or features based on lexical properties (Sheehan, Flor, and Napolitano, 2013; Curto, Mamede, and Baptista, 2015; Quispesaravia et al., 2016; Heintz et al., 2022; Wilkens et al., 2022).

4.1 Statistical extraction method

To identify linguistic properties and to compute their statistics per readability level, we processed the corpus with Spanish modules of Stanza (Qi et al., 2020), converting it into CoNLL-U format.[5] This step includes automatic sentence splitting, lemmatization, PoS-tagging and morphological features, and dependency parsing using the *universal dependencies* format.

Additionally, we used external resources such as specialized lexicons to automatically identify other linguistic features. Thus, the words were also grouped based on the following lexicons: (i) according to their level of complexity, using the CEFR levels, defined in the Spanish lexicon provided by Blanco Escoda (2024)[6]; (ii) by the presence of affixes, prefixes, and suffixes listed in the index extracted from the RAE dictionary (Pérez, Alameda, and Cuetos, 2003); and (iii) by the frequency and repetition of the words. A semantic analysis was also performed by calculating the percentage of polysemous words appearing in the SAW database (Fraga et al., 2017) and in the lexicon of Haro et al. (2017), as well as the frequency and orthographic neighborhood (i.e., how a word is connected to others in the vocabulary) of the words collected in the index of Pérez, Alameda, and Cuetos (2003).

Based on the extracted information, simple metrics (e.g., length or syllable count) and advanced metrics requiring external data (e.g., orthographic neighborhood or polysemy) were calculated using custom scripts.

4.2 Results

In general, the results confirm previous research in this field, as some phenomena show an association with respect to the defined lev-

[4] Observed agreement refers to the proportion of instances where all annotators agree.

[5] `https://universaldependencies.org/format.html`

[6] The lexicon is available at `https://zenodo.org/records/10889986`.

els. In this section, we perform a qualitative analysis of some of these phenomena, whose results are shown in Table 1 (the statistics for all features can be seen in Appendix C).

Average length of texts, sentences and words: On average, the length of the texts is directly proportional to their difficulty, i.e. the longer the text, the more difficult it is to read, as longer texts tend to contain more content and detail, making them more difficult to understand (Curto, Mamede, and Baptista, 2015).[7] Thus, texts belonging to readability level 3 have, on average, twice as many words as texts belonging to level 1. This can also be seen in the figures for the average number of lemmas per text. Sentence length is a factor that reflects the syntactic difficulty of a text (Henry, 1980). This can be observed in the corpus, in which the average sentence length increases from level 1 to level 3. This is also reflected in the average sentence length in tokens. In terms of punctuation, easier texts have shorter sentences and more punctuation, although the number of commas and semicolons (which increase the complexity of the texts) increases with the level (see Table 7 in Appendix C). As for the average word length, it is slightly longer at the higher levels. These numbers support the idea that word length is directly related to their level of difficulty (Henry, 1980).

Word type distribution: The percentage of adpositions, adjectives and determiners increases progressively as the level increases. The opposite is true for the percentage of verbs, numerals and punctuation elements. There are more of these elements at the lowest level than at the highest. In addition to verbs, statistics were extracted on infinitives, gerunds, participles, and finite verbs, as it has been shown that some of these verb forms are considered relevant for text classification models in terms of readability (Madrazo Azpiazu and Pera, 2020). The figures show that the percentage of infinitives and finite verbs is higher at simpler levels, while the number of participles and gerunds is higher at more complex levels.

[7]It should be noted that, except for a generic definition of the length of the documents (minimum and maximum words), during the compilation and validation of texts, its size has not been taken into account for their classification.

Lexical diversity: The number of unique words in a text affects its readability, as the higher the number, the more complex the text is to read (Islam et al., 2012). This measure, known as hapax legomena, has been used as a proxy for lexical diversity in Portuguese (Curto, Mamede, and Baptista, 2015). According to Kornai (2008), the usual percentage of unique words in a corpus is between 40 and 60% (Islam, Mehler, and Rahman, 2012). This corpus has 71%. As the level increases, more unique words are used, and fewer repeated words are found. Similarly, the number of lemmas is lower at level 1 and increases progressively. Redundancy in texts, which can be interpreted here in terms of the number of repeated words, lemmas and different tokens, facilitates comprehension (Henry, 1980). The combination of these factors leads to the conclusion that the lexical diversity of texts increases in line with an increase in the readability level.

Average dependency distance: The average distance between dependencies increases slightly as the level rises. As this distance often indicates more complex sentence structures, it may contribute to higher text complexity (von Glasersfeld, 1970). In our corpus, the results show a progressive increase in dependency distance between levels 1 and 3, following the mentioned trend.

Percentages of lexical occurrences by readability levels and orthographic neighbors: The percentage of words classified in the corpus as belonging to the very easy, easy and plain levels confirms that the level of readability is generally low, since the use of lexical items belonging to the very easy level predominates and the number of items belonging to the plain level is very low. The statistical analysis according to the level shows that the proportion of very easy vocabulary is greater than that of more complex levels, while the proportion of plain lexical items is lower. Furthermore, it can be observed that there are more A1 lexical items in level 1, more A2 lexical items in level 2, and more B1 lexical items in level 3 than in the other levels.

There is a relationship between lexical retrieval, the number of lexical neighbors, and their frequency (Andrews, 1997). Wilkens et al. (2022) offer a proposal for measuring this relationship: to determine the number of

	L1	L2	L3
Avg w/text	181.94	332.30	399.90
Avg token/text	219.49	382.86	457.34
Avg lemmas/text	101.72	162.36	193.98
Avg dif. token/text	119.25	192.39	226.44
Avg w/sentence	13.71	22.80	25.46
Avg charact./w	4.59	4.84	4.97
Avg dependency dist.	3.25	3.47	3.63
% PUNCT	15.97	12.56	11.99
% ADP	11.47	13.56	14.48
% ADJ	4.42	6.17	6.69
% DET	11.41	12.66	13.51
% VERB	9.15	9.08	8.32
% NUM	2.14	1.35	1.25
% participles	8.03	13.01	15.47
% finite verbs	60.94	57.33	55.44
% infinitives	20.53	20.12	19.4
% gerunds	1.48	3.11	3.53
% repeated words	28.85	28.76	27.98
% hapax legomena	71.15	71.24	72.02
% A1 words	84.86	79.69	78.59
% A2 words	11.24	13.79	13.43
% B1 words	3.90	6.52	7.98
Avg orthog. neighbors	5.49	4.93	4.74

Table 1: Statistical results across levels. Underlined phenomena exhibit significant variation (one-way ANOVA, $p < 0.05$).

neighbors and their average and cumulative frequency in a reference corpus. In this corpus, there is a slight decrease in the number of orthographic neighbors as the readability level increases.

Dialogue: Finally, the presence of dashes and question, exclamation and quotation marks can indicate the presence of dialogue in the text (Henry, 1980). Several authors (e.g., Dolch, Henry, François) argue that the presence of dialogue in the text is an indicator of greater simplicity (Henry, 1980) and therefore greater readability. The greater number of question and exclamation marks and dashes in level 1 and their progressive decrease in higher levels (see Appendix C) indicate that dialogue is a more prominent feature in the simpler levels, which could influence the level of readability of the texts.

In sum, this section shows that, with only minor exceptions, the linguistic features of our corpus follow the same tendencies as previous research in other languages, reinforcing its value as resource for fostering readability research in Castilian Spanish.

5 Surprisal as a Readability Predictor

Despite the recent advances in natural language processing, it is still not clear to what extent current LMs improve previous approaches for automatic readability assessment. On the one hand, some studies show that the performance of fine-tuned transformers largely vary across corpora, being sometimes surpassed by traditional machine learning classifiers (Deutsch, Jasbi, and Shieber, 2020), while more recent studies advocate for hybrids models that integrate linguistic features into embedding vectors extracted from transformers models (Li, Ziyang, and Wu, 2022; Wilkens et al., 2024).

Surprisal from language models: On unsupervised scenarios, Martinc, Pollak, and Robnik-Šikonja (2021) evaluated the usefulness of language models perplexity to rank a set of documents according to their complexity. Inspired by this approach, we explore the use of LMs *surprisal* as a readability feature. Surprisal (Hale, 2001; Levy, 2008) is the negative log probability assigned

by a LM to a given word in context, and has been used as a strong predictor for reading times (Demberg and Keller, 2008; Wilcox et al., 2023), where the higher the surprisal the more slow the word to read, and vice-versa. On sentence-based tasks, surprisal values have been also used as grammatical acceptability measures (Lee and Vu, 2024). Following *Hypothesis 1* of Martinc, Pollak, and Robnik-Šikonja (2021), we predict that language models trained on general texts (newspapers, wikipedia, etc.) that fall in the middle of the readability spectrum will exhibit lower surprisal for easy texts (L1), and higher surprisal for more complex documents (L3, L4). As a result, their surprisal values may correlate with the four readability levels of the corpus.

Models and methods: For our experiments, we obtain sentence surprisals for all documents in the dataset using `minicons`[8] (Misra, 2022). For generative models, we computed the surprisal of a given word w_t as $\text{Surprisal}(w_t) = -\log P(w_t \mid w_{<t})$ (where $w_{<t}$ is the preceding context). For bidirectional models, we compared the 'pseudo-log-likelihood' proposed by Salazar et al. (2020) to the variation of Kauf and Ivanova (2023), which takes into account out-of-vocabulary and multi-token words.[9] Given a document and a LM, we first compute the surprisal of each sentence as the sum of the negative log probabilities of each of the tokens in each of the words in the sentence, and then obtain its average and median as proxies for its linguistic acceptability with respect to the training data.[10] We compare both encoder and decoders models from different architectures, and monolingual and multilingual ones. For encoders, we used the monolingual BETO (Cañete et al., 2020), RoBERTa-BNE (base and large) (Fandiño et al., 2022), and Bertin-RoBERTa (De la Rosa et al., 2022), and the multilingual XLM-RoBERTa (base and large) (Conneau et al., 2020). As auto-regressive models, we compared the monolingual Bertin-GPT-J-6B (De la Rosa and Fernández, 2022) to Llama 3.2 1B[11] and to XGLM-1.7B (Lin et al., 2022). Experimets were performed on a standard PC with a NVIDIA Titan Xp GPU (12gb).

Method	Feature	ρ
Length-based	ASL	**0.547**
BETO-base	Mean	0.501
	Median	0.483
Bertin-RoBERTa-base	Mean	<u>0.556</u>
	Median	0.538
RoBERTa-base BNE	Mean	0.455
	Median	0.448
RoBERTa-large BNE	Mean	0.462
	Median	0.457
XLM-RoBERTa-base	Mean	0.456
	Median	0.417
XLM-RoBERTa-large	Mean	0.465
	Median	0.434
Llama 3.2 1B	Mean	<u>0.579</u>
	Median	**0.558**
XGLM-1.7B	Mean	**0.581**
	Median	<u>0.553</u>
Bertin GPT-J-6B	Mean	<u>0.556</u>
	Median	0.534

Table 2: Spearman ρ correlations (all of them significative) between surprisal values and the 4 levels of the corpora. *ASL* is the average sentence length in number of tokens. Numbers in bold are the best results per type (mean/median), while underscores correlations are those surprisals which improve ASL.

Results: Surprisal values are compared to average sentence length (ASL), which according to previous research is one of the most predictive features from those not needing external resources (Wilkens et al., 2022). The results on Table 2 show that the surprisals of auto-regressive models, and those of Bertin-RoBERTa-base surpass the ASL feature, indicating that surprisal of some models might be a powerful readability predictor. The results also suggest that, as expected, the training corpus has a strong influence on the models' behavior, as models of different sizes trained with similar data (e.g., both RoBERTa-BNE, both XLM-RoBERTa, or both Bertin –GPT and RoBERTa) have comparable results.

[8] `https://github.com/kanishkamisra/minicons`

[9] We report the results of the latter approach, which show 0.02 better correlation on average.

[10] We have also analyzed the maximum and minimum surprisals (not reported here): The former produced similar results than the median, while minimum seems not to be a good predictor, with average correlations of 0.28.

[11] `https://huggingface.co/meta-llama/Llama-3.2-1B`

In sum, this experiment provides additional evidence that LMs' surprisal, as suggested by previous studies using perplexity, are correlated with readability, and therefore they could be incorporated as new features for both supervised and unsupervised classification approaches.

6 Conclusions and future work

This paper presents a new readability corpus for Spanish designed for automatic classification tasks, together with an analysis of linguistic features that affect text readability and an investigation into the potential of LMs surprisal as a readability predictor.

This new corpus is presented as a comprehensive resource that addresses a crucial need in the research field, namely the diversity of genres and textual topics, as well as the necessity for corpora designed for readability classification tasks. The existing resources for Spanish are limited in scope, representing only genres from the literary, journalistic and educational domains. Moreover, their design is oriented towards text simplification and the teaching of Spanish as a foreign language. This new corpus includes texts of all genres and topics that are commonly read and of interest to adults. The texts are labeled and classified according to their level of readability, following a taxonomy based on the research work of experts in the fields of adult learning and readability. In the process of creating the corpus, we encountered challenges related to the number of texts per subcategory and level, the availability of texts at the most basic level, and the manual classification of texts.

The statistical analysis, when considered alongside the existing literature, enables us to draw certain conclusions regarding the relationship between specific linguistic phenomena and the readability level of the text. It can thus be concluded that longer texts, sentences, and words, as well as more participles, gerunds, and lexical diversity, are linked to higher levels of readability. Conversely, more infinitives and finite verbs are linked to lower levels. The findings also suggest that a greater distance between dependencies, a greater number of specific PoS-tags and more B1 items are associated with higher readability levels. However, more orthographic neighbors are associated with lower levels.

Furthermore, we also explored the suitability of the LMs surprisals as a readability predictor, by analyzing their correlation with the levels of the corpora. A detailed comparison of state-of-the-art models, both encoders and generative, suggest that the average surprisal is a good predictor that could be incorporated into automatic classifiers of readability levels.

New lines of research emerge from this work related to the influence of certain linguistic features on the readability of texts (e.g., distance between dependencies, orthographic neighbors or derived words), to the use of LMs surprisal as a predictor (e. g., by exploring it with other models and datasets) or to the creation of new corpora for other languages, using the one presented here as a model. Finally, it is worth mentioning that the release of the corpus will contribute to foster research in readability in Spanish, both from a theoretical point of view (e.g., investigating what makes a document more or less readable) and from an applied perspective such as using it to train and evaluate readability classifiers.

Acknowledgments

This work has received financial support from the European Commission (Project: iRead4Skills, Grant number: 1010094837, Topic: HORIZON-CL2-2022-TRANSFORMATIONS-01-07, DOI: 10.3030/101094837), from *Xunta de Galicia - Consellería de Cultura, Educación, Formación Profesional e Universidades* (Centro de investigación de Galicia accreditation 2024-2027 ED431G-2023/04), and the European Union (European Regional Development Fund - ERDF), and by MCIN/AEI/10.13039/501100011033, (PID2022-142843NB-I00) and by a Ramón y Cajal grant (RYC2019-028473-I).

References

Andrews, S. 1997. The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin & Review*, 4(4):439–461, December.

Bengoetxea, K. and I. Gonzalez-Dios. 2021. MultiAzterTest: a Multilingual Analyzer on Multiple Levels of Language for Readability Assessment.

Blanco Escoda, X. 2024. Léxico mejorado

del español para los niveles del MCER A1, A2 y B1, y nociones metalingüísticas de base. *Onomázein: Revista de lingüística, filología y traducción de la Pontificia Universidad Católica de Chile*, Extra 14:150–167.

Campos Saavedra, D., P. Contreras Carmona, B. Riffo Ocares, M. Véliz, and A. Reyes Reyes. 2014. Complejidad textual, lecturabilidad y rendimiento lector en una prueba de comprensión en escolares adolescentes. *Universitas Psychologica*, 13(3), jan.

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*.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July. Association for Computational Linguistics.

Correia, S., R. Amaro, M. Ricardo, and X. Blanco Escoda. 2024. iread4skills - reading skills survey, July.

Curto, P., N. Mamede, and J. Baptista. 2015. Automatic text difficulty classifier - assisting the selection of adequate reading materials for european portuguese teaching. In *Proceedings of the 7th International Conference on Computer Supported Education CSEDU*, volume 1, pages 36–44, 01.

De la Rosa, J. and A. Fernández. 2022. Zero-shot reading comprehension and reasoning for spanish with BERTIN GPT-J-6B. In M. M. y Gómez, J. Gonzalo, F. Rangel, M. Casavantes, M. Ángel Álvarez Carmona, G. Bel-Enguix, H. J. Escalante, L. Freitas, A. Miranda-Escalada, F. Rodríguez-Sánchez, A. Rosá, M. A. Sobrevilla-Cabezudo, M. Taulé, and R. Valencia-García, editors, *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)*. CEUR Workshop Proceedings.

De la Rosa, J., E. G. Ponferrada, M. Romero, P. Villegas, P. G. de Prado Salas, and M. Grandury. 2022. Bertin: Efficient pre-training of a spanish language model using perplexity sampling. *Procesamiento del Lenguaje Natural*, 68(0):13–23.

Dell'Orletta, F., S. Montemagni, and G. Venturi. 2014. Assessing document and sentence readability in less resourced languages and across textual genres. *ITL - International Journal of Applied Linguistics*, 165:163–193, 12.

Demberg, V. and F. Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210.

Deutsch, T., M. Jasbi, and S. Shieber. 2020. Linguistic features for readability assessment. In J. Burstein, E. Kochmar, C. Leacock, N. Madnani, I. Pilán, H. Yannakoudakis, and T. Zesch, editors, *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17, Seattle, WA, USA → Online, July. Association for Computational Linguistics.

Fandiño, A. G., J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, and M. Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68.

Flesch, R. 1948. A readability formula in practice. *Elementary English*, 25(6):344–351.

Fraga, I., I. Padrón, M. Perea, and M. Comesaña. 2017. I saw this somewhere else: The spanish ambiguous words (saw) database. *Lingua*, 185:1–10.

François, T. 2015. When readability meets computational linguistics: a new paradigm in readability. *Revue française de linguistique appliquée*, 20(2):79–97.

Gonzalez-Dios, I., M. J. Aranzabe, A. Díaz de Ilarraza, and H. Salaberri. 2014. Simple or complex? assessing the readability of Basque texts. In J. Tsujii and J. Hajic, editors, *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 334–344, Dublin, Ireland, August.

Dublin City University and Association for Computational Linguistics.

Grego Bolli, G., D. Rini, and S. Spina. 2017. Predicting readability of texts for italian l2 students: A preliminary study. In *ALTE (2017). Learning and Assessment: Making the Connections – Proceedings of the ALTE 6th International Conference, 3-5 May 2017*, pages 272–278. Association of Language Testers in Europe, may.

Hale, J. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Haro, J., P. Ferré, R. Boada, and J. Demestre. 2017. Semantic ambiguity norms for 530 spanish words. *Applied Psycholinguistics*, 38(2):457–475.

Harry, G. and M. Laughlin. 1969. Smog grading - a new readability formula. *The Journal of Reading*.

Heintz, A., J. S. Choi, J. Batchelor, M. Karimi, and A. Malatinszky. 2022. A large-scaled corpus for assessing text readability. *Behavior Research Methods*, 55, 03.

Henry, G. 1980. Lisibilité et compréhension. *Communication & Langages*, 45(1):7–16.

Hernandez, N., N. Oulbaz, and T. Faine. 2022. Open corpora and toolkit for assessing text readability in French. In R. Wilkens, D. Alfter, R. Cardon, and N. Gala, editors, *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*, pages 54–61, Marseille, France, June. European Language Resources Association.

Islam, Z., A. Mehler, and R. Rahman. 2012. Text readability classification of textbooks of a low-resource language. In R. Manurung and F. Bond, editors, *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 545–553, Bali, Indonesia, November. Faculty of Computer Science, Universitas Indonesia.

Jian, L., H. Xiang, and G. Le. 2022. English text readability measurement based on convolutional neural network: A hybrid network model. *Computational Intelligence and Neuroscience*, 2022:1–9, 03.

Kauf, C. and A. Ivanova. 2023. A better way to do masked language model scoring. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 925–935, Toronto, Canada, July. Association for Computational Linguistics.

Kornai, A. 2008. *Mathematical Linguistics*. Advanced Information and Knowledge Processing. Springer London, 1 edition. Hardcover published: 10 November 2007, Softcover published: 22 October 2010, eBook published: 16 December 2007.

Lee, S. Y. and M. H. Vu. 2024. The effects of distance on NPI illusive effects in BERT. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9443–9457, Miami, Florida, USA, November. Association for Computational Linguistics.

Lennon, C. and H. Burdick. 2014. The lexile® framework as an approach for reading measurement and success. Accessed: June 11, 2024.

Levy, R. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Li, W., W. Ziyang, and Y. Wu. 2022. A unified neural network model for readability assessment with feature projection and length-balanced loss. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7446–7457, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Lin, X. V., T. Mihaylov, M. Artetxe, T. Wang, S. Chen, D. Simig, M. Ott, N. Goyal, S. Bhosale, J. Du, R. Pasunuru, S. Shleifer, P. S. Koura, V. Chaudhary, B. O'Horo, J. Wang, L. Zettlemoyer, Z. Kozareva, M. Diab, V. Stoyanov, and

X. Li. 2022. Few-shot learning with multilingual generative language models. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Lozano, C. 2022. Cedel2: Design, compilation and web interface of an online corpus for l2 spanish acquisition research. *Second Language Research*, 38(4):965–983.

Madrazo Azpiazu, I. and M. S. Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

Madrazo Azpiazu, I. and M. S. Pera. 2020. Is cross-lingual readability assessment possible? *Journal of the Association for Information Science and Technology*, 71(6):644–656.

Martinc, M., S. Pollak, and M. Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179, March.

Misra, K. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

Ngo, D. V. and Y. Parmentier. 2023. Towards sentence-level text readability assessment for French. In S. Štajner, H. Saggio, M. Shardlow, and F. Alva-Manchego, editors, *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 78–84, Varna, Bulgaria, September. INCOMA Ltd., Shoumen, Bulgaria.

Pérez, M., J. Alameda, and F. Cuetos. 2003. Frecuencia, longitud y vecindad ortográfica de las palabras de 3 a 16 letras del diccionario de la lengua española (rae, 1992). *REMA, ISSN 1135-6855, Vol. 8, Nº. 2, 2003, pags. 1-10*, 8, 01.

Qi, P., Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Quispesaravia, A., W. Perez, M. Sobrevilla Cabezudo, and F. Alva-Manchego. 2016. Coh-Metrix-Esp: A complexity analysis tool for documents written in Spanish. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4694–4698, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Ribeiro, E., N. Mamede, and J. Baptista. 2024. Automatic text readability assessment in European Portuguese. In P. Gamallo, D. Claro, A. Teixeira, L. Real, M. Garcia, H. G. Oliveira, and R. Amaro, editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 97–107, Santiago de Compostela, Galicia/Spain, March. Association for Computational Lingustics.

Rojo, G. and I. M. Palacios, 2016. *Learner Spanish on computer: Current trends and future perspectives*, pages 55–87. John Benjamins Publishing Company, 12.

Saggion, H., E. Gómez-Martínez, E. Etayo, A. Anula, and L. Bourg. 2011. Text simplification in simplext. making text more accessible. *Procesamiento del Lenguaje Natural*, 47:341–342, 09.

Saggion, H., S. Štajner, S. Bott, S. Mille, L. Rello, and B. Drndarevic. 2015. Making it simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Trans. Access. Comput.*, 6(4), May.

Salazar, J., D. Liang, T. Q. Nguyen, and K. Kirchhoff. 2020. Masked language model scoring. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online, July. Association for Computational Linguistics.

Schwarm, S. and M. Ostendorf. 2005. Reading level assessment using support vector

machines and statistical language models. In K. Knight, H. T. Ng, and K. Oflazer, editors, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 523–530, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Sheehan, K. M., M. Flor, and D. Napolitano. 2013. A two-stage approach for generating unbiased estimates of text complexity. In L. Rello, H. Saggion, and R. Baeza-Yates, editors, *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*, pages 49–58, Atlanta, Georgia, June. Association for Computational Linguistics.

Vajjala, S. 2022. Trends, limitations and open challenges in automatic readability assessment research. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5366–5377, Marseille, France, June. European Language Resources Association.

Vajjala, S. and I. Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. 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 297–304, New Orleans, Louisiana, June. Association for Computational Linguistics.

Vajjala, S. and D. Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163—-173, 06.

Vásquez-Rodríguez, L., P.-M. Cuenca-Jiménez, S. Morales-Esquivel, and F. Alva-Manchego. 2022. A benchmark for neural readability assessment of texts in Spanish. In S. Štajner, H. Saggion, D. Ferrés, M. Shardlow, K. C. Sheang, K. North, M. Zampieri, and W. Xu, editors, *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 188–198, Abu Dhabi, United Arab Emirates (Virtual), December. Association for Computational Linguistics.

von Glasersfeld, E. 1970. The problem of syntactic complexity in reading and readability. *Journal of Reading Behavior*, 3(2):1–14.

Wilcox, E. G., T. Pimentel, C. Meister, R. Cotterell, and R. P. Levy. 2023. Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Wilkens, R., D. Alfter, X. Wang, A. Pintard, A. Tack, K. P. Yancey, and T. François. 2022. FABRA: French aggregator-based readability assessment toolkit. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1217–1233, Marseille, France, June. European Language Resources Association.

Wilkens, R., P. Watrin, R. Cardon, A. Pintard, I. Gribomont, and T. François. 2024. Exploring hybrid approaches to readability: experiments on the complementarity between linguistic features and transformers. In Y. Graham and M. Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2316–2331, St. Julian's, Malta, March. Association for Computational Linguistics.

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

A Appendix: Readability Level Characteristics

Level 1» A1 (light/transparent (easy)/ more dependent)
Short and simple texts for the purpose of performing familiar tasks or short and simple texts introducing new information (e.g., didactic texts). Typically, simple, everyday concepts. It is assumed that the speaker has limited access to communication domains. Ideally, the topic is presented initially. Basic communication contexts: basic day-to-day (transport schedules / lists, menus / general instructions, item price information); family/personal communications; simple/basic information. Absence of figures of speech. Basic lexicon (active lexicon) – known words and simple expressions memorized = used in everyday matters (e.g., transport, food, family, work). Frequent and concrete main and copulative verbs and frequent and concrete nouns, that is, concepts/ideas with a higher level of concreteness than abstraction. Rare affixation; except frequent affixes such as -mente, -ción, re-. Short periods, with simple conjunctions and in direct order (Subject Verb Object). Rare auxiliary verbs (except for copulative verbs) and few anaphoric references: the referential chain is complete and does not occur in an elliptical way (e.g., 'A Juan no le cae bien porque no es nada simpática' vs. 'A Juan no le cae bien debido a su completa falta de simpatía'.). Coordination structures (Noun Phrase & Noun Phrase, Adjective & Adjective...) with copulative, disjunctive and adversative conjunctions are admitted. Some frequently used subordination structures (subordinate temporal adverbials, for example), except for those less frequent (reduced adverbials of Infinitive) are admitted. Occurrence of some periphrastic constructions, more usual and therefore more "decipherable" (e.g., estar + Gerund /empezar/ir + a + Infinitive; dejar/acabar + de + Infinitive). No compound tenses. Use of Indicative verb tenses. Personal Infinitive and Gerund are admitted. Simple temporal location. Temporal cohesion is given by means of temporal adverbs or connectors (hoy, mañana, antes, después...) and not by verb tenses. (e.g., 'Ella se fue antes de que llegara Juan.' vs. 'Juan llegó y ella ya se había ido')

Features	
lexical-conceptual	basic lexicon (absence of foreign words and less of 15% of non-basic vocabulary) mostly concrete nouns (concrete concepts) daily and familiar concepts (personal life) and new simple concepts rare affixation (-izar; -ción, -mente, etc)
verbs	simple tenses in Infinitive, Indicative, Gerund mainly main verbs (no auxiliary) copulative verbs (ser, estar, seguir, continuar, andar, parecer) frequent multi-word verbs (estar/ir a; poder/querer) in Indicative mode + Inf.
syntactic structures	simple coordination (y, o, pero, ni...) simple subordination (subject relatives, temporal - antes de, después de, cuando; causative - porque; conditional (si) + Indicative
cohesion	few pronoun anaphora and ellipsis one or two temporal location(s)

Table 3: Level 1 features.

Level 2 » A2 (mild/grey/clear/less dependent)
Short texts that are interesting for the reader to inform himself or in moments of leisure or with the purpose of carrying out tasks. Some presence of abstract concepts (such as feelings, states of mind, religiosity, qualities, and defects, etc.). Concepts linked to the world known personally and professionally to the reader. Usual communication contexts: work context (specific instructions); Media (news of interest, e.g., sports); Commercial Communication (ads). Basic lexicon (active lexicon) and expanded passive lexicon with frequent words. Main and copulative verbs and frequent nouns in different domains where the reader routinely interacts or is interested in. Some affixation, frequent prefixes and suffixes (e.g., -aje, -ción, -ero, -mente, -ísimo, -izar, -sub, super-, in-). Short periods, with coordinated conjunctions and most of the subordinate conjunctions, both in direct order (Subject Verb Object) and in other possibilities. Admits subject relative subordinate clauses, but not object clauses (e.g., 'El niño que abrazó a su madre' vs. 'El niño al que su madre abrazó.') Admits subordinates with Indicative, Subjunctive, and Infinitive. Verbs in simple tenses. Some periphrastic constructions such as the passive voice (especially in the Indicative). Compound tenses are present (e.g., pretérito pluscuamperfecto: había comido) More complex temporal reference, in linear sequence. Temporal cohesion can be given via verb tenses. The reader can link different parts of the text and make a global sense of them.

Features	
lexical-conceptual	presence of more domain-specific lexicon (with explanation) less concrete nouns and ideas more diverse domain concepts (e.g., professional context) some affixation (-aje, -ción, -ero, -mente, -ísimo, -izar, sub-, super-, in-)
verbs	Infinitive, Indicative, Gerund and Subjunctive; some compound tenses; passive form in Indicative main and auxiliary verbs copulative verbs (all others + permanecer) less frequent multi-word verbs (e.g., modal verbs) + Subjunctive
syntactic structures	coordination (no solo... sino también, tanto como, o…o, bien…bien, ni...ni, además) subordination (subject relatives, temporal (tan pronto como que, antes que) causative (dado que, por), conditional (si, en caso de que), others (conforme, según, como, a medida que... + subjunctive)
cohesion	presence of pronoun anaphora and ellipsis more than one temporal reference, in linear sequence

Table 4: Level 2 features.

Level 3 » B1 (heavy/intense/opaque/dark/independent)
Texts of different sizes and on varied topics of interest to the reader for information or leisure. Varied concepts. Readers are able to step out of their comfort zone. More contact with the online world. The reader is able to infer at a more complex level (e.g., infer opinions from opinion texts), including at a multimodal level, with texts in less common formats (e.g., infographics). Leisure (stories; travel diaries, fiction); Professional (theoretical articles); Media (reportage, opinion articles); Online (forums) communication contexts. Varied lexicon to express subjects in any of the communication domains. Less passive lexicon, due to diverse contact. Polysemy of certain words. Occurrence of frequent foreign words (e.g., timing, hobby, show). Nouns that express both concrete and abstract concepts to describe situations, reactions, emotions, thoughts, etc. Some frequent domain-specific verbs and nouns, describing trendy or situations known to the reader. Main and copulative verbs, frequent and some domain-specific. Occurrence of most affixations, with the exception of erudite and less frequent affixes. The reader is able to infer the meaning of derived words. Longer periods, with simple and compound sentences and a greater variety of conjunctions and syntactic order. Some high-frequency irregular verbs are admitted. Presence of modal verbs, with uses and meanings in common expressions and in unusual contexts. Indicative, Subjunctive, Imperative and Conditional moods, both in the active voice and in the passive voice. Passives with -se (e.g., 'El trabajo se hace bien o no se hace'). Complex temporal reference, in non-linear sequence. The reader can detect information (albeit basic) that is not explicit in the text.

Features	
lexical-conceptual	presence of domain-specific terminology (without explanation) domain-specific and abstract nouns and ideas (infographics) abstract or domain-specific concepts (communication contexts other than personal or professional contexts) frequent affixation
verbs	all moods and tenses (including conditional) frequent multi-word verbs and compound tenses, including more unusual modal verbs
syntactic structures	coordination and subordination; non- linear syntactic order all passive forms (in all modes, passives with 'se')
cohesion	frequent anaphora and ellipsis diverse temporal localization with non-linear relations

Table 5: Level 3 features.

B Appendix: Text categories and subcategories

Categories	**Subcategories**
Personal communication	ticket, personal letter, diary, SMS / online chat, list / agenda
Institutional / professional communication	minute, letter, internal release, press release, report, instructions, newsletter, web page
Social media	editorial, news, reportage, interview, opinion article, scientific dissemination article, biography, horoscope, obituary, weather report
Commercial communication / dissemination	ad, flyer, institutional social media messages, menu, label, user manual, medicine leaflet
Didactic book	textbook, encyclopedia, cookbook, glossary
Fiction book	short story, fable, epic, novel, poetry, drama
Non-fiction book	biography, chronicle, essay, diary, preface / prologue, dedicatory, self-help book, travel diary, memoirs, letter, travel guide
Academic	article, project, abstract, critical review, report, thesis, essay
Political	speech, motion / report, program
Legal	law, contract, notification letter / public notice
Religious	prayer, scriptures, homilies, catechism

Table 6: Categories and subcategories in the corpus.

C Appendix: Complete characteristics of corpus texts by level

	Total	**L1**	**L2**	**L3**	**L4**
No. sentences	44 957	9 900	12 035	16 867	6 155
Avg No. sentences/text	17.54	15.00	18.23	18.97	17.36
No. words	843 218	120 044	219 429	355 570	148 175
Avg No. w/text	328.95	**181.94**	**332.30**	**399.90**	**418.29**
No. tokens	972 279	144 833	252 827	406 640	167 979
Avg No. tokens/text	379.30	**219.49**	**382.86**	**457.34**	**474.15**
No lemmas	418 690	67 108	107 190	172 456	71 936
Avg No. lemmas/text	163.34	**101.72**	**162.36**	**193.98**	**203.10**
No. dif. tokens	489 761	78 678	127 027	201 326	82 730
Avg No. dif. tokens/text	191.07	**119.25**	**192.39**	**226.44**	**233.59**
Avg No. w/sentence	22.01	**13.71**	**22.80**	**25.46**	**27.39**
Avg No. token/sentence	25.21	16.28	26.16	28.89	30.86
Avg No. charact./w	4.86	**4.59**	**4.84**	**4.97**	**5.15**
Avg No. charact./token	4.37	4.04	4.36	4.50	4.68
Avg No. dot/text	15.06	12.34	15.67	16.44	15.53
Avg No. comma/text	20.13	**11.0**	**20.41**	**24.66**	**25.21**
Avg No. semicolon/text	0.57	**0.23**	**0.57**	**0.74**	**0.76**
Avg dependency distance	3.49	**3.25**	**3.47**	**3.63**	**3.63**
% PUNCT	12.62	**15.97**	**12.56**	**11.99**	**11.24**
% PROPN	5.13	5.13	4.93	5.06	5.57
% ADP	13.95	**11.47**	**13.56**	**14.48**	**15.44**
% PRON	5.06	4.59	5.62	5.13	4.44
% ADJ	6.37	**4.42**	**6.17**	**6.69**	**7.63**
% CCONJ	3.38	3.44	3.5	3.34	3.23
% DET	13.05	**11.41**	**12.66**	**13.51**	**13.99**
% NOUN	19.57	18.3	18.92	19.89	20.91
% ADV	2.79	2.96	3.18	2.71	2.28
% AUX	2.35	2.49	2.63	2.26	2.05
% SCONJ	1.8	1.58	1.98	1.86	1.57
% VERB	8.42	**9.15**	**9.08**	**8.32**	**7.03**
% NUM	1.43	**2.14**	**1.35**	**1.25**	**1.35**
% INTJ	0.59	1.31	0.45	0.51	0.36
% X	0.03	0.05	0.02	0.03	0.03
% PART	0.03	0.01	0.02	0.03	0.03
% SYM	0.09	0.05	0.1	0.1	0.09
% participles	14.07	**8.03**	**13.01**	**15.47**	**18.62**
% finite verbs	56.49	**60.94**	**57.33**	**55.44**	**53.05**
% infinitives	19.59	**20.53**	**20.12**	**19.4**	**18.11**
% gerunds	3.08	**1.48**	**3.11**	**3.53**	**3.54**
% repeated words	28.29	**28.85**	**28.76**	**27.98**	**27.12**
% hapax legomena	71.71	**71.15**	**71.24**	**72.02**	**72.88**
% A1 words	80.41	**84.86**	**79.69**	**78.59**	**77.99**
% A2 words	12.95	11.24	13.79	13.43	13.43
% B1 words	6.64	**3.90**	**6.52**	**7.98**	**8.58**
% polysemous words	3.45	2.97	3.67	3.62	3.50
Avg No. orthog. neighbors	4.86	**5.49**	**4.93**	**4.74**	**4.49**
Avg No. ?/text	2.12	**2.9**	**2.35**	**1.85**	**0.91**
Avg No. !/text	3.05	3.59	2.95	3.03	2.28
Avg No. quotation marks/text	2.65	1.13	2.82	3.42	3.22
Avg No. -/text	1.37	**2.12**	**1.19**	**1.21**	**0.71**
% prefixes	0.80	0.80	0.84	0.78	0.77
% affixes	30.13	33.08	29.89	28.80	28.41
% suffixes	29.33	32.28	29.05	28.03	27.64

Table 7: Characteristics of the texts in the corpus according to their level of readability. Numbers in bold are those with an apparent relationship to at least the three target levels (L1-L3).