LexComSpaL2: A Lexical Complexity Corpus for Spanish as a Foreign Language

Jasper Degraeuwe, Patrick Goethals
Ghent University (LT^3 / MULTIPLES research groups)
Groot-Brittanniëlaan, 9000 Ghent, Belgium
Jasper.Degraeuwe, Patrick.Goethals@UGent.be

Abstract

We present LexComSpaL2, a novel corpus which can be employed to train personalised word-level difficulty classifiers for learners of Spanish as a foreign/second language (L2). The dataset contains 2,240 in-context target words with the corresponding difficulty judgements of 26 Dutch-speaking students who are learning Spanish as an L2, resulting in a total of 58,240 annotations. The target words are divided over 200 sentences from 4 different domains (economics, health, law, and migration) and have been selected based on their suitability to be included in L2 learning materials. As our annotation scheme, we use a customised version of the 5-point lexical complexity prediction scale (Shardlow et al., 2020), tailored to the vocabulary knowledge continuum (which ranges from no knowledge over receptive mastery to productive mastery; Schmitt, 2019). With LexComSpaL2, we aim to address the lack of relevant data for multi-category difficult prediction at word level for L2 learners of other languages than English.

Keywords: Lexical Complexity Prediction, Personalisation, Spanish as a Foreign Language

1. Introduction

The presence of unknown words can prevent people from fully understanding the meaning of a text. Some studies translated this observation into lexical thresholds, claiming that 95 to 98% of the words in a text should be known for optimal comprehension and successful inference of unknown words (Laufer and Ravenhorst-Kalovski, 2010). Although the existence of such concrete thresholds has been questioned, it is generally agreed that text comprehension and vocabulary knowledge are positively correlated: the more words someone knows in a given text, the better that person will understand the text (Laufer and Ravenhorst-Kalovski, 2010; Schmitt et al., 2011). These principles also apply to the productive aspects of language: the more words people know, the more content they will be able to convey (Milton, 2013).

Conversely, this implies that a lack of vocabulary knowledge can be an obstacle, a situation which frequently occurs when the language at hand is not one's native language (L1). In search of a (partial) answer to this issue, research within the domains of foreign language learning (FLL) and Computer-Assisted Language Learning (CALL) has devoted considerable attention to the creation of dedicated tools (e.g., reading and writing assistants; Akhlaghi et al., 2019) and resources (e.g., vocabulary lists organised into different proficiency levels; Dang et al., 2017) for foreign/second language (L2) learners.

During the development of those tools and resources, the identification of possibly difficult words (or word uses) in running text can play a pivotal role. As an alternative for the labour-intensive process of identifying difficult words by hand, which is still common practice within FLL and CALL (Tack, 2021), computational linguistic methods proved to be a path worth exploring, especially after the introduction of neural networks and large language models opened a whole new range of opportunities (Alfter, 2021; Tack, 2021). Methods exploiting computer-readable resources in which words are linked to difficulty levels (or frequency bands, since frequency is known to correlate with difficulty; Schmitt, 2010b) constitute a first option, as they can automatically assign words in digital(ised) texts to their corresponding difficulty/frequency label. However, apart from having limited coverage (only the words included in the resources will be assigned a label), this approach does not take into account individual differences between learners.

To overcome these limitations, machine learning (ML) systems can be designed, which offer much more flexibility: in theory, they can classify any text, sentence or word into any set of difficulty levels, and tailor predictions to individual learner profiles. However, the performance and applicability of ML systems heavily depend on the annotated data they are trained on: annotations from L1 speakers will differ from those of L2 learners, for example, and annotations according to a binary format (i.e. with "difficult" or "non-difficult" as the two possible labels) will lead to other output values than annotations based on a multi-category classification system such as the Common European Framework of Reference for Languages (CEFR), which includes six different labels ranging from A1 to C2.

The present study aims to make a contribution to the domain of word-level difficulty prediction for FLL purposes by presenting LexComSpaL2 (**Lex**ical

LREC-COLING 2024, pages 10432–10447
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

Complexity for **Spa**nish **L2**)[1], a novel corpus with three distinctive features. First, the difficulty judgements included in the dataset are based on a customised annotation format which combines insights from FLL and computational linguistics. The second distinctive characteristic is the learner-centred perspective of the difficulty judgements: instead of relying on graded resources (e.g., François and De Cock, 2018) or frequency lists (e.g., Davies and Hayward Davies, 2018), the corpus represents individual learner judgements which can be used to train personalised models. Finally, by taking Spanish as the target language we expand the coverage of the field, since, to the best of our knowledge, there does not yet exist any Spanish dataset for difficulty prediction with FLL as the target setting.

The paper is structured as follows: Section 2 provides an overview of the related research, both from a linguistic (Section 2.1) and computational linguistic (Section 2.2) perspective. Next, the compilation of the LexComSpaL2 corpus, its annotation, and its statistics are presented in Section 3. Finally, Section 4 includes a discussion of the dataset, after which concluding remarks together with possible directions for future work are provided in Section 5.

2. Related Research

2.1. Difficulty/complexity in FLL/CALL

Having an extensive vocabulary knowledge is usually considered as an indispensable requisite to be able to function in a foreign language (Milton, 2013; Schmitt, 2010a), with a combination of implicit and explicit vocabulary activities being generally recognised as the go-to learning method (Nation, 2019). Explicit vocabulary learning activities (e.g., fill-in-the-blanks exercises) require paying deliberate attention to vocabulary items, while in implicit activities the increase in vocabulary knowledge is achieved as a secondary effect, because the main goal of the activity is the successful completion of an authentic task, such as understanding the plot of a book (Ellis, 1994; Krashen, 1989). In both of these strands, it is essential to at least have an indication about which words might be difficult to understand or produce for the target learner. As for the implicit approach, Krashen's (1989) Input Hypothesis states that learners acquire language/vocabulary when the input they are exposed to is comprehensible but slightly beyond their current knowledge. This implies that, to create the activities, it should be known which parts of the input are comprehensible and which are not. In a similar vein, for explicit learning it has to be decided which words to include in and exclude from the activities, a task which becomes considerably easier if it is known which words are (un)known by the target learner.

In linguistics, "word difficulty/complexity" is a multi-faceted concept. As a starting point, we take the notion of "linguistic complexity", which can be subdivided into different categories/dichotomies (Kortmann and Szmrecsanyi, 2012). A first dichotomy refers to global (i.e. the complexity of a language as such) versus local complexity (i.e. at a phonological, morphological, syntactic, lexical, semantic or pragmatic level). A second distinction concerns absolute (or objective) versus relative (or agent-related/cognitive) complexity. The former type refers to complexity as established by the linguistic properties of words, ranging from the number of morphemes over the number of vowels and diphthongs to the homonymous and/or polysemous character of words (i.e. the number of different meanings/senses they have). Especially the last feature plays an important role in an L2 context. Often, lexically ambiguous words have one high-frequency "easy" meaning and one or several low-frequency, specialised "difficult" meanings, making them more challenging to process and learn than single-meaning words (Bensoussan and Laufer, 1984; Degani and Tokowicz, 2010).

As opposed to absolute complexity, relative complexity corresponds to the complexity as perceived by a particular language learner, bringing psycholinguistic factors and world knowledge into the equation (Kortmann and Szmrecsanyi, 2012; North et al., 2023). In an L2 setting, an additional crucial factor in determining this agent-related complexity is the influence of one's L1. False friends (e.g., ES *gracioso* ['funny'] - NL *gracieus* ['graceful']) and cognates (e.g., ES *proyecto* - NL *project* - EN *project*), for example, are L1-related phenomena which can have a considerable impact on the degree of complexity as perceived by L2 learners. As relative complexity has also been referred to as "difficulty", in the remainder of this paper we will use "complexity" for absolute complexity and "difficulty" for relative complexity.

Another crucial aspect are the categories words can be assigned to. The most straightforward option would be a binary categorisation of complex/difficult versus non-complex/non-difficult. However, vocabulary knowledge is usually conceptualised as a continuum from no knowledge over receptive (passive) mastery to productive (active) mastery (Laufer, 1998; Nation, 2019; Schmitt, 2019), meaning that a continuous categorisation is likely to be a more suitable solution. A well-known conceptualisation of this continuum is the self-report Vocabulary Knowledge Scale (Wesche and Paribakht, 1996; see also Section 3.2).

In any case, both in binary and in scale-like cate-

[1]The corpus is made available through a GitHub repository, and a sample is provided in Table 5 at the end of the paper.

gorisations, the more specific the reference point in mind, the more informative the assigned labels will be. Suppose that a teacher wants to classify all words in a given text for a heterogeneous group of L2 students: when taking "L2 learners" in general as the reference point, the output of the classification process will be very generic; when analysing the text three times with "beginner", "intermediate", and "advanced" as the reference points in mind, the output will be more informative; when classifying the words for each student individually, the output will be most informative.

Finally, to measure complexity, a large series of complexity/readability metrics have been developed, which are based on local complexity features and usually operate at text level (e.g., Flesch Reading Ease; Flesch, 1951). These measures and their L2 variants have been widely applied in FLL and integrated into CALL environments, but it remains questionable whether their threshold-based composition (e.g., classifying a word as complex when it has *n* or more syllables) can accurately identify complexity at word level, let alone identify which words might be perceived as difficult by a specific learner. As a result, the classification of textual input into difficulty levels has usually been performed manually in the fields of FLL and CALL (Tack, 2021), with teachers marking complex/difficult words in reading materials as a prototypical example. It should be noted, though, that several resources can be consulted to facilitate this manual process, from graded vocabulary lists (Dang et al., 2017) to frequency lists (Davies and Hayward Davies, 2018). For Spanish in particular, the *Plan Curricular* developed by the Instituto Cervantes constitutes a valuable resource, as it includes sections in which vocabulary items are linked to different CEFR levels.

2.2. Difficulty/complexity in Computational Linguistics

Computer-driven methods which identify complex/difficult words have been designed for a wide range of target audiences, ranging from children (Kajiwara et al., 2013) to people suffering from dyslexia (Rello et al., 2013). In this subsection we will provide a brief overview of the main word-level concepts and methods for L2 learners as the specific target audience.

In computational linguistics, a task-oriented component is added to the linguistic concept of difficulty/complexity: we need data on which the automatic classifiers can be trained, which means that we need to link concrete words to concrete difficulty/complexity labels in an "inventory". One common approach to building such inventory is exploiting computer-readable versions of the same resources as used for manual consultation (cf. supra). Graded course books (e.g., based on CEFR levels) also serve this purpose, as they allow words to be assigned to the level at which the words first occur in the books (Alfter, 2021). Another approach is to collect human annotations, which can be done through platforms such as Amazon Mechanical Turk (Shardlow et al., 2021) or by means of specific research experiments (Tack, 2021).

This inventory as such already provides enough information to build a straightforward classifier which simply assigns all words in a given input text to their label in the inventory, an approach often adopted in vocabulary profiling (Finlayson et al., 2023). However, this method has one major drawback: the coverage of the classifier will always be limited to the words included in the inventory. To overcome this limitation, a set of "features" can be gathered for the set of target words. These features tend to be quantifiable variables that can be computed automatically, such as frequency, word length, cognateness, number of syllables and, more recently, word embedding values obtained from large language models such as BERT (Devlin et al., 2019). Based on these features, it is possible to train full-fledged ML systems which are able to generalise and make predictions for unseen words.

2.2.1. CWI

In complex word identification (CWI), the goal is to label words as either complex or non-complex (see "Label (CWI)" in Table 1 for an example). It should be noted that the term "complex" in CWI combines elements from both "complexity" and "difficulty" (Section 2.1), as it refers to the difficulty an individual may experience in understanding a given word as a result of both the word's linguistic properties and factors belonging to the individual (North et al., 2023). CWI has been integrated into many

Sentence		
La sala de espera de pediatría está repleta de niños que moquean. ('The paediatric waiting room is full of children sniffling.')		
Target word	**Label (CWI)**	**Label (LCP)**
sala	0	1
espera	0	2
pediatría	1	4
repleta	1	3
niños	0	1
moquean	1	5

Table 1: Example of annotation using binary CWI labels compared to continuous LCP labels.

applications (e.g., lexical simplification pipelines), but its binary nature has shown to be prone to low inter-annotator agreement (Zampieri et al., 2017).

The 2018 shared task on CWI (Yimam et al., 2018) showed that ML-assisted strategies provide extensive coverage and obtain the best performance on the CWI task. Early ML-oriented studies include the work of Paetzold and Specia (2016), who addressed CWI as a part of their lexical simplification approach for non-native speakers of English as the target audience. More recently, Tack (2021) gathered binary difficulty judgements of L2 learners of French to train neural networks which are able to make contextualised and personalised predictions. This last aspect in particular is highly important, as personalising CWI models has shown to lead to the best performance (Gooding and Tragut, 2022).

As far as CWI datasets are concerned, the 2018 shared task (Yimam et al., 2018) provided a considerable contribution for English, German, Spanish and French as target languages. However, the annotators were L1 speakers who were explicitly instructed to assume a broad target audience ranging from children over L2 learners to people with reading impairments (Yimam et al., 2018, p. 67). Next, the CLexIS2 dataset (Ortiz Zambrano and Montejo-Ráez, 2021) aims to contribute to CWI and lexical simplification for Spanish in an educational setting, but the complex word annotations in the dataset come from L1 speakers of Spanish in computing studies, again not from L2 learners of Spanish.

2.2.2. LCP

In lexical complexity prediction (LCP), a word's complexity is evaluated by assigning a value from a 5-point scale instead of providing a binary complex versus non-complex judgement (see "Label (LCP)" in Table 1 above for an example). As was the case with CWI, "complexity" in LCP should be interpreted as an amalgam of the concepts of complexity and difficulty. Importantly, in contrast with CWI and its binary character, LCP enables making predictions based on "comparative complexity", i.e. determining whether a target word is more or less complex than another target word (North et al., 2023).

As appears from the datasets released (Shardlow et al., 2020) and shared tasks organised (Shardlow et al., 2021) over the course of the past few years, LCP has been attracting more and more attention. Nevertheless, studies and datasets with L2 learners as the specific target audience remain scarce, especially on word level. A rare example can be found in the work of Lee and Yeung (2018), who had Japanese learners of English rate a set of 12,000 English words on a 5-point scale in order to develop personalised lexical simplification models. As was the case in the domain of CWI, adopting a learner-centred and personalised perspective has been identified as an important avenue for future research within LCP (North et al., 2023).

Finally, the research being conducted into difficulty/complexity classifiers which predict CEFR levels should also be highlighted, as their scale-like nature bears much resemblance with the concepts behind LCP. Alfter (2021), for instance, trained feature-based ML algorithms based on CEFR-labelled resources as input (e.g., ELELex; François and De Cock, 2018). In a similar vein, Aleksandrova and Pouliot (2023) present a lexical complexity classifier (based on a support vector classifier algorithm) which predicts contextually-aware CEFR-based labels for both single words and multiword expressions in English as well as French.

3. Dataset

3.1. Data Collection

From the related research it can be concluded that computer-driven difficulty/complexity prediction for L2 learners is still relatively unexplored terrain, especially for languages other than English. Moreover, adopting a learner-centred approach has been identified as an important aspect, and it has been shown that multi-category rating methods such as LCP open up a wider range of applications than the binary CWI method. Therefore, in this study we will adapt the principles of LCP to the "no knowledge – receptive mastery – productive mastery" continuum of vocabulary knowledge and have L2 learners of Spanish make annotations according to this adapted LCP rating scale.

To build a representative dataset (i.e. including data which can be used in L2 Spanish materials), we select sentences from 4 different domain-specific newspaper article corpora (on economics, health, law, and migration)[2]. We adopt this approach for two main reasons: first, domain-specific words represent specialised knowledge that is crucial to learning a particular topic (Webb and Nation, 2017). Second, domain-specific vocabulary consists of both high- and low-frequency words, which should lead to a diverse dataset with understandable as well as challenging vocabulary for all proficiency levels.

The selection procedure of the sentences consists of the following series of steps (which is repeated for each of the 4 domain-specific corpora): first, we build a keyword list based on Odds Ratio (Pojanapunya and Watson Todd, 2018) as the "key-

[2]The corpora are available within the Spanish Corpus Annotation Project (Goethals, 2018), which offers an Intelligent CALL (ICALL) environment for L2 Spanish teachers and students with data from a wide range of uniformly tokenised, POS-tagged, lemmatised and parsed corpora (Appendix A).

Rating	Original LCP description	VKS	Adapted description
1	Very easy: this word is very familiar to me	I can use this word in a sentence: _____.	I know this word and its meaning, and I also use it actively in speaking/writing.
2	Easy: I am aware of the meaning of this word	I know this word. It means _____. (synonym or translation)	I know this word and its meaning, but I might not be able to use it on the top of my head in an oral/written conversation. When I have some time to think, however, I do think I would use it naturally.
3	Neutral: this word is neither difficult nor easy	I have seen this word before, and I think it means _____. (synonym or translation)	I have heard/seen this word before and given the context I think that I more or less know what it means, but I do not see myself using this word actively.
4	Difficult: the meaning of this word is unclear to me, but I may be able to infer it from the sentence	I have seen this word before, but I don't know what it means.	This word sounds vaguely familiar and based on the context I could make an educated guess about its meaning, but I would still need a dictionary to be able to understand its exact meaning.
5	Very difficult: I have never seen this word before / this word is very unclear to me	I don't remember having seen this word before.	This word does not sound familiar at all to me, and even based on the context I do not know what it means, so I would definitely need a dictionary to get to know its meaning.

Table 2: Comparison between original LCP scale descriptions, Vocabulary Knowledge Scale (VKS) descriptions (Wesche and Paribakht, 1996), and our own customised descriptions (engrafted onto the "no knowledge – receptive mastery – productive mastery" continuum of vocabulary knowledge; Schmitt, 2019). The descriptions presented to the participants are the adapted LCP descriptions.

ness" metric (Gabrielatos, 2018)[3]. Next, we take the first 50 keywords from that list and select, for each keyword, 1 sentence from the corpus. The selection of the sentence is realised by means of an adapted version of the selection method proposed by Pilán et al. (2016), which is specifically designed to extract pedagogically suitable sentences from corpora (Appendix B).

Finally, all selected sentences across the 4 domains are joined together to form the final dataset. Every content word in the sentences (except for adverbs) represents a target word to be annotated, with the maximum number of instances of the same lemma being limited to 5 (Shardlow et al., 2021). In this version of the dataset, only single words are considered.

In summary, our LexComSpaL2 corpus aims to be representative (by including various domains, which echoes the often thematic structure of vocabulary classes and materials), contextualised (by preserving sentence contexts, which enables in-context presentation of target words during an-

Proficiency level (PL)	Details	Number of students
PL1	2^{nd} bachelor ($\approx$B1 level on CEFR scale)	10
PL2	3^{rd} bachelor ($\approx$B2 level)	8
PL3	Master's degree ($\approx$C1 level)	8

Table 3: Overview of participant data. All participants are enrolled in the Applied Linguistics career at Ghent University.

[3]To obtain the keyword list, we calculate the Odds Ratio values for all nouns by comparing the lemma frequency in the study corpus (i.e. the domain-specific corpus) to the lemma frequency in a reference corpus (also available within SCAP). Only candidate items with a statistically significant effect size according to the Bayesian Information Criterion (values $\geq$2; Wilson, 2013) and a keyness value higher than 1 (i.e. items which are more key to the study corpus than to the reference corpus) are maintained, after which the remaining items are ranked from highest to lowest keyness.

notation) and pedagogically suitable (by adopting a dedicated selection method). An overview of the dataset statistics is presented in Section 3.3, which will also formulate an answer to our assumptions that (1) domain-specific vocabulary is a mix of high- and low-frequency items, and that (2) this mix should on its turn lead to a mix of easy and more difficult target words.

3.2. Data Labelling

As our goal is to arrive at a learner-centred dataset, 26 students of L2 Spanish are recruited as participants. We assign a unique ID to each participant and collect information on their L1 (in this case, all participants are Dutch-speaking natives), their proficiency level (measured by the stage of the university career they are currently in, see Table 3) and the number of years they have been studying Spanish (usually equal to their proficiency level). These data can then be used as additional variables in the ML models trained on the dataset, next to other features such as word embeddings, frequency/dispersion, cognateness, word length, and number of syllables (see also Section 2.2). In the end, the model should be able to output personalised difficulty predictions for any sequence of words it receives as input.

All target words are presented in their original sentence context to the participants, meaning that the corpus can also be used to analyse how the in-context usage of words affects their complexity (Shardlow et al., 2020). Each participant is asked to annotate all 200 sentences according to a customised annotation scheme, inspired by the LCP scale and tailored to the "no knowledge – receptive mastery – productive mastery" continuum of vocabulary knowledge (Schmitt, 2019). As mentioned before, the Vocabulary Knowledge Scale (VKS; Section 2.1) undertakes a similar effort (see column "VKS" in Table 2 above). However, performing VKS-based annotations is a time-consuming task, as both passive and active knowledge are tested explicitly (by asking a synonym, translation or usage example). Therefore, we choose to make our scale fully self-perceived, but not without taking a series of measures to make the self-report judgements as qualitative and reliable as possible. First of all, we organise the annotations as on-site sessions without any time constraints, allowing us to provide guidance and answer questions whenever necessary. For their annotation work, the participants also receive a financial compensation, serving as an additional incentive for them to complete the classification task diligently. Thirdly, we provide more elaborate and explicit descriptions of the different LCP labels compared to the regular ones (see column "Adapted description" in Table 2).

Prior to starting the experiment, participants were given a written document including the instructions (Appendix C), which were discussed orally with one of the researchers involved in the study. Apart from highlighting that participants should base their annotations primarily on their intuitions and needs as L2 learners, the instructions also emphasised that it was the in-context meaning of lexically ambiguous target words which should be evaluated, rendering the current version of the LexComSpaL2 dataset "implicitly word-sensed". To make the dataset "explicitly word-sensed", the output of a word sense disambiguation system (WSD) could be used to link the difficulty judgements of ambiguous words to specific word sense labels. As a hypothetical example, let us suppose that the WSD system is applied to sentence 1_1 in Table 5, which contains the ambiguous word *celebrar* ('to party' / 'to hold, to organise'). A performant WSD system would assign the word *celebrado* to the 'to hold, to organise' sense, meaning that the difficulty judgements for this particular instance of *celebrar* can be linked to the concept 'to hold, to organise' instead of to the word form *celebrado* or to the lemma *celebrar*. In other words, this extra dataset layer would enable

Sentences		Target words		Frequency target words	
Total (per domain)	Average length (SD)	Total (unique)	Average per sentence (SD)	Frequency range	Percentage
200 (50)	28.85 (2.98)	2,240 (1,863)	11.2 (2.14)	1 - 1,000	0.24
				1,001 - 2,000	0.14
				2,001 - 3,000	0.09
				3,001 - 4,000	0.07
				4,001 - 5,000	0.05
				>5,000	0.41

Table 4: The statistics for LexComSpaL2 (on sentences and target words). Standard deviation is abbreviated as SD, and the frequency ranges are based on Davies and Hayward Davies (2018).

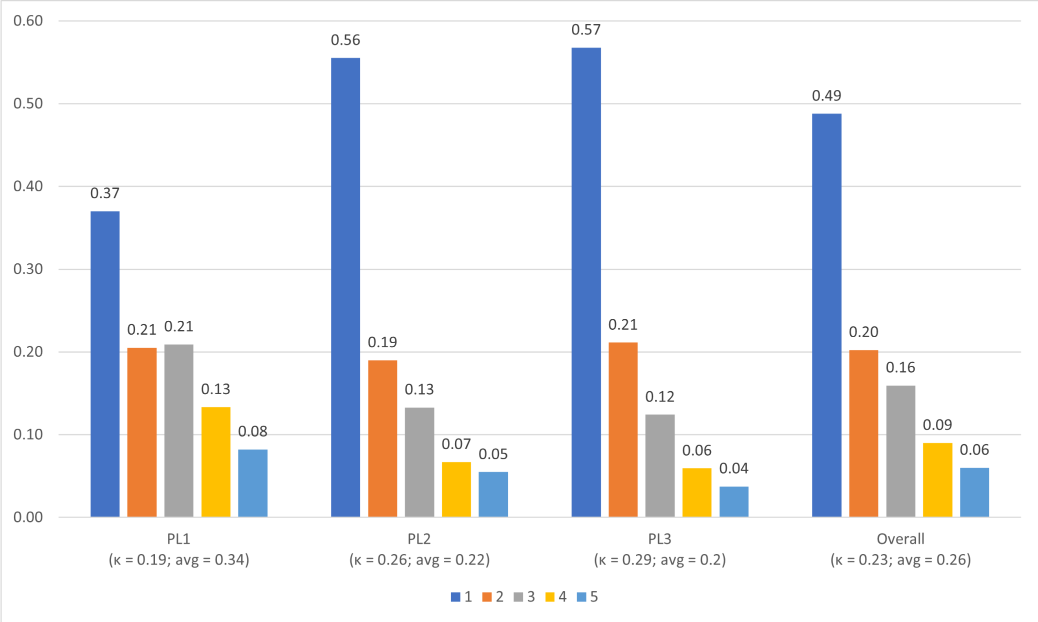

Figure 1: The statistics for LexComSpaL2 (on annotations). Distributions for the 5 LCP labels (see Table 2) are reported both per proficiency level (PL; see also Table 3 above) and overall. Inter-annotator agreement and average difficulty (avg; normalised in the range 0-1) are added between parentheses. Inter-annotator agreement is measured by Fleiss' kappa (K), with scores below 0 indicating less agreement than could be expected by chance and a value of 1 indicating full agreement.

ML classifiers trained on LexComSpaL2 to yield more fine-grained predictions.

3.3. Statistics

In summary, the LexComSpaL2 corpus includes 2,240 target words (1,863 unique lemmas), distributed over 200 different sentences (50 per domain). All target words are evaluated by each of the 26 participants, resulting in a total of 58,240 observations. A sample taken from the dataset is provided in Table 5 at the end of the paper, and a comprehensive overview of the dataset statistics is presented in Table 4 (details on the sentences and target words) and Figure 1 (details on the annotations). Below, we will briefly discuss the most important aspects.

First, as shown in Table 4, the distribution of the target words according to the frequency ranges proposed by Davies and Hayward Davies (2018) suggests that the corpus obtains a fairly good balance between frequent and less frequent words (a little over 40% of the words does not belong to the top 5,000 most frequent words in Spanish). This finding confirms our assumption that choosing sentences from domain-specific corpora would lead to a representative mix of high- and low-frequency words (Section 3.1).

The subsequent assumption that this mix would then also result in a diverse dataset containing both understandable and potentially more challenging vocabulary items (regardless of the students' proficiency levels), is corroborated by the statistics presented in Figure 1. To obtain the average scores, the annotations were normalised in the range 0-1 ($1 \rightarrow 0$, $2 \rightarrow 0.25$, $3 \rightarrow 0.5$, $4 \rightarrow 0.75$, $5 \rightarrow 1$). Although, overall, most words are known (fairly) well (69% for labels 1 and 2 combined, resulting in a relatively low normalised average difficulty score of 0.26), a considerable number of words is only known passively (16% for label 3), vaguely (9% for label 4), or not at all (6% for label 5). The statistics at the level of the individual groups show that there is a considerable difference between PL1 and PL2 (normalised average difficulty of 0.34 compared to 0.22), but no so much between PL2 and PL3 (0.22 compared to 0.2).

A second crucial observation to be made concerns the inter-annotator agreement scores included in Figure 1. In fact, the statistics reveal a relatively low inter-annotator agreement (K=0.23 overall), even within the individual proficiency level groups (values between 0.19 and 0.29). This finding underpins the need for individual predictions, as aggregating the scores for each word cannot be said to adequately represent the judgements of most of the learners. Therefore, instead of providing one single difficulty value for each target word, the LexComSpaL2 corpus includes all individual annotations, which further distinguishes our dataset from existing LCP corpora. To be complete, average scores (both overall and per proficiency level) are also added to the dataset, as they can still serve as (distant) approximations of difficulty.

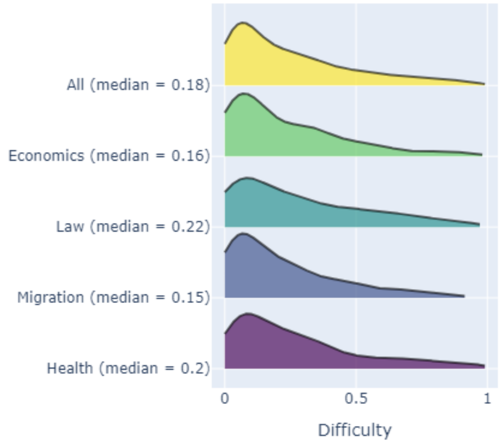

Figure 2: Ridge line plot presenting the probability density function of the individual domains included in the dataset, as well as the entire dataset ("All"). Scores are normalised in the range 0 - 1 (Shardlow et al., 2020).

Next, in Figure 2 we present a ridge line plot of the average word difficulty scores, grouped per domain. This plot allows us to visualise any major differences between the different domains included in the dataset. However, the plot indicates that the domains follow approximately the same distribution, meaning that the difficult words are spread relatively well across the 4 domains.

Finally, in anticipation of the LexComSpaL2 corpus being used in the future, we propose a fixed dataset split into training, validation, and test sets[4]. This should allow for a fair comparison between models being trained on the corpus. To enable cross-validation, we provide 10 different 80/10/10 splits for the training/validation/test sets. The sets are constructed at sentence level (to enable training ML models which take into account the context, such as neural networks with BiLSTM layers), and the different domains are always equally distributed within each set.

4. Discussion

With our LexComSpaL2 corpus, we aim to make a substantial contribution to the field of automatic word-level difficulty prediction for (Spanish) L2 learners. The sentences and target words included in the corpus come from 4 different domains and were deliberately selected based on their pedagogical suitability. With an average difficulty of 0.26, the words in our corpus fall towards the easier end of the LCP scale. Selecting the sentences and words based on frequency bands (Shardlow et al., 2020) could have led to a more balanced dataset, but would have jeopardised its representativeness. As the 1000 most frequent words in Spanish alone account for about 80% of the words in written text and 88% in spoken text (Davies, 2005), it is safe to say that potential L2 materials will always contain a large proportion of frequent words, so difficulty classifiers should be able to handle them appropriately (i.e. learn that they are more likely to be perceived as easy, especially by upper-intermediate and advanced learners).

Apart from its representativeness, another distinctive characteristic of LexComSpaL2 are the individual annotations, gathered based on a customised annotation scale. The annotations can be linked to the participant features (unique ID, proficiency level, and years of experience) and used to train personalised models, which have shown to lead to the best performance (Gooding and Tragut, 2022; Tack, 2021). The models could then be employed to create customised L2 Spanish materials tailored to the individual needs of students, both for implicit activities (e.g., scanning reading materials to select only those with less than *n* % of vaguely known and unknown words) and explicit ones (e.g., creating fill-in-the-blanks exercises to practice words which are known passively but not yet actively).

Regarding the limitations of the dataset in its current format, it should first of all be noted that caution is required when using the LexComSpaL2 dataset for setups in which the targeted learners do not have Dutch as their L1. Due to factors such as false friends/cognates (see also Section 2.1), cultural significance, and academic curriculum design in the home country, it is to be expected that groups of, say, L1 Chinese or Arabic students learning Spanish will display considerably different vocabulary difficulty profiles compared to L1 Dutch students. Secondly, as brought forward in Section 3.2, the instructions urged the students to base their annotations for lexically ambiguous words on the in-context meaning of the target word rather than on the isolated word form, but these annotations are not yet "explicitly" linked to word sense labels.

5. Conclusion and Future Work

In this article we have presented the LexComSpaL2 corpus, which can be used to train word-level difficulty classifiers for (Spanish) L2 learners as the target audience. The dataset contains data from 4 different sources (newspaper corpora on economics, health, law, and migration) and totals 58,240 difficulty judgements provided by 26 L2 Spanish learn-

[4]See the GitHub repository for full details.

ers of different proficiency levels. As our annotation scheme, we tailored the lexical complexity prediction scale to the vocabulary knowledge continuum.

As for future work, a first important avenue would be the collection of difficulty judgements from L2 Spanish students with other L1s than Dutch. Secondly, we plan to release an ML classifier trained on LexComSpaL2 in the near future, which can then serve as a baseline model. Thirdly, there still remain several opportunities to further enrich the dataset. The inclusion of larger contexts (surrounding sentences or entire paragraph) for training large language models on the dataset would be such dataset update worth exploring, as would be the addition of extra participant features, such as the results on a proficiency test (e.g., a cloze test; Marcos Miguel, 2020).

Finally, to avoid that new learners need to annotate all 200 sentences before they can get personalised predictions, we will perform an item analysis to identify the most "valuable" sentences. Based on this reduced dataset, it would also become possible to use an annotation scheme which explicitly gauges productive knowledge, instead of the fully self-perceived scale used in the present study.

6. Acknowledgements

This research has been carried out as part of a PhD fellowship on the IVESS project (file number 11D3921N), funded by the Research Foundation - Flanders (FWO). Additionally, we want to express our sincere gratitude to the reviewers for their valuable feedback and suggestions.

7. Bibliographical References

Elham Akhlaghi, Branislav Bédi, Matthias Butterweck, Cathy Chua, Johanna Gerlach, Hanieh Habibi, Junta Ikeda, Manny Rayner, Sabina Sestigiani, and Ghil'ad Zuckermann. 2019. Demonstration of LARA: A Learning and Reading Assistant. In *Proc. 8th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2019)*, pages 37–38.

Desislava Aleksandrova and Vincent Pouliot. 2023. CEFR-based Contextual Lexical Complexity Classifier in English and French. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 518–527, Toronto, Canada. Association for Computational Linguistics.

Sentence ID	Sentence text	Target word	Average judgement	Individual judgements
1_1	El directivo, que ha celebrado un almuerzo de Navidad con la prensa, ha asegurado que [. . .] ('The manager, who has held a Christmas lunch with the press, has assured that [. . .]')	directivo	{PL1: 0.3, PL2: 0.34, PL3: 0.22, overall: 0.29}	{PARTP1: 3, . . . , PARTP26: 1}
		celebrado	{PL1: 0.13, PL2: 0, PL3: 0.06, overall: 0.07}	{PARTP1: 2, . . . , PARTP26: 1}
		...		
...				
4_50	Las investigaciones sobre atención primaria, neurología, oncología médica y microbiología van después, [. . .] ('Research into primary care, neurology, medical oncology and microbiology comes after, [. . .]')	investigaciones	{PL1: 0.28, PL2: 0.03, PL3: 0.06, overall: 0.13}	{PARTP1: 1, . . . , PARTP26: 1}
		atención	{PL1: 0.2, PL2: 0.03, PL3: 0.03, overall: 0.1}	{PARTP1: 2, . . . , PARTP26: 1}
		...		

Table 5: Examples from the LexComSpaL2 corpus, with target words being underlined.

David Alfter. 2021. *Exploring natural language processing for single-word and multi-word lexical complexity from a second language learner perspective*. Göteborgs Universitet, Göteburg. OCLC: 1292463715.

Marsha Bensoussan and Batia Laufer. 1984. Lexical Guessing in Context in EFL Reading Comprehension. *Journal of Research in Reading*, 7(1):15–32.

Thi Ngoc Yen Dang, Averil Coxhead, and Stuart Webb. 2017. The Academic Spoken Word List. *Language Learning*, 67(4):959–997.

Mark Davies. 2005. Vocabulary range and text coverage: Insights from the forthcoming Routledge frequency dictionary of Spanish. In *Selected proceedings of the 7th Hispanic Linguistics Symposium*, pages 106–115. Citeseer.

Mark Davies and Kathy Hayward Davies. 2018. *A frequency dictionary of Spanish: core vocabulary for learners*, second edition edition. Routledge frequency dictionaries. Routledge, London ; New York.

Tamar Degani and Natasha Tokowicz. 2010. Ambiguous words are harder to learn. *Bilingualism: Language and Cognition*, 13(3):299–314.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nick Ellis. 1994. Vocabulary acquisition: The implicit ins and outs of explicit cognitive mediation. In Nick Ellis, editor, *Implicit and explicit learning of languages*, pages 211–282. Academic Press, London.

Natalie Finlayson, Emma Marsden, and Laurence Anthony. 2023. Introducing MultilingProfiler: An adaptable tool for analysing the vocabulary in French, German, and Spanish texts. *System*, 118:103122.

Rudolf Flesch. 1951. *How to test readability*. Harper.

Thomas François and Barbara De Cock. 2018. ELELex: a CEFR-graded lexical resource for Spanish as a foreign language. Louvain-la-Neuve.

Costas Gabrielatos. 2018. Keyness analysis: Nature, metrics and techniques. In C. Taylor and A. Marchi, editors, *Corpus Approaches To Discourse: A critical review*, pages 225–258. Routledge, Oxford.

Patrick Goethals. 2018. Customizing vocabulary learning for advanced learners of Spanish. In *Technological innovation for specialized linguistic domains : languages for digital lives and cultures, proceedings of TISLID'18*, pages 229–240. Éditions Universitaires Européennes. Event-place: Gent, Belgium.

Sian Gooding and Manuel Tragut. 2022. One Size Does Not Fit All: The Case for Personalised Word Complexity Models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 353–365, Seattle, United States. Association for Computational Linguistics.

Tomoyuki Kajiwara, Hiroshi Matsumoto, and Kazuhide Yamamoto. 2013. Selecting Proper Lexical Paraphrase for Children. In *Proceedings of the 25th Conference on Computational Linguistics and Speech Processing (ROCLING 2013)*, pages 59–73, Kaohsiung, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Bernd Kortmann and Benedikt Szmrecsanyi, editors. 2012. *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*. DE GRUYTER.

Stephen Krashen. 1989. We Acquire Vocabulary and Spelling by Reading: Additional Evidence for the Input Hypothesis. *The Modern Language Journal*, 73(4):440–464.

Batia Laufer. 1998. The Development of Passive and Active Vocabulary in a Second Language: Same or Different? *Applied Linguistics*, 19(2):255–271.

Batia Laufer and Geke C. Ravenhorst-Kalovski. 2010. Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, 22(1):15–30.

John Lee and Chak Yan Yeung. 2018. Personalizing Lexical Simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 224–232, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Nausica Marcos Miguel. 2020. Analyzing morphology-related strategies in Spanish L2 lexical inferencing: how do suffixes matter? *International Review of Applied Linguistics in Language Teaching*, 58(3):351–377.

James Milton. 2013. Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In Camilla Bardel, Christina Lindqvist, and Batia Laufer, editors, *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis*, number 2 in Eurosla Monographs Series, pages 57–78.

Paul Nation. 2019. The Different Aspects of Vocabulary Knowledge. In Stuart Webb, editor, *The Routledge Handbook of Vocabulary Studies*, pages 15–29. Routledge, London.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2023. Lexical Complexity Prediction: An Overview. *ACM Computing Surveys*, 55(9):1–42.

Jenny A. Ortiz Zambrano and Arturo Montejo-Ráez. 2021. CLexIS2: A New Corpus for Complex Word Identification Research in Computing Studies. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1075–1083, Held Online. INCOMA Ltd.

Gustavo Paetzold and Lucia Specia. 2016. Unsupervised Lexical Simplification for Non-Native Speakers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

Ildikó Pilán, Elena Volodina, and Lars Borin. 2016. Candidate sentence selection for language learning exercises: From a comprehensive framework to an empirical evaluation. *Revue Traitement Automatique Des Langues*, 57(3):67–91.

Punjaporn Pojanapunya and Richard Watson Todd. 2018. Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory*, 14(1):133–167.

Luz Rello, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. 2013. Frequent Words Improve Readability and Short Words Improve Understandability for People with Dyslexia. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Paula Kotzé, Gary Marsden, Gitte Lindgaard, Janet Wesson, and Marco Winckler, editors, *Human-Computer Interaction – INTERACT 2013*, volume 8120, pages 203–219. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Lecture Notes in Computer Science.

Norbert Schmitt. 2010a. Chapter 3. Key Issues in Teaching and Learning Vocabulary. In Rubén Chacón-Beltrán, Christian Abello-Contesse, and María Del Mar Torreblanca-López, editors, *Insights into Non-native Vocabulary Teaching and Learning*, pages 28–40. Multilingual Matters.

Norbert Schmitt. 2010b. *Researching Vocabulary*. Palgrave Macmillan UK, London.

Norbert Schmitt. 2019. Understanding vocabulary acquisition, instruction, and assessment: A research agenda. *Language Teaching*, 52(02):261–274.

Norbert Schmitt, Xiangying Jiang, and William Grabe. 2011. The Percentage of Words Known in a Text and Reading Comprehension. *The Modern Language Journal*, 95(1):26–43.

Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. CompLex — A New Corpus for Lexical Complexity Prediction from Likert Scale Data. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 57–62, Marseille, France. European Language Resources Association.

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. SemEval-2021 Task 1: Lexical Complexity Prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online. Association for Computational Linguistics.

Anaïs Tack. 2021. *Mark My Words! On the Automated Prediction of Lexical Difficulty for Foreign Language Readers*. PhD thesis, UCLouvain & KU Leuven, Louvain-la-Neuve, Belgium.

Stuart Webb and Paul Nation. 2017. *How vocabulary is learned*. Oxford Handbooks for Language teachers. Oxford University Press, Oxford.

Marjorie Wesche and T. Sima Paribakht. 1996. Assessing Second Language Vocabulary Knowledge: Depth Versus Breadth. *The Canadian Modern Language Review*, 53(1):13–40.

Andrew Wilson. 2013. Embracing Bayes factors for key item analysis in corpus linguistics. In Markus Bieswanger and Amei Koll-Stobbe, editors, *New Approaches to the Study of Linguistic Variability*, number 4 in Language Competence and Language Awareness in Europe, pages 3–11. Peter Lang, Frankfurt.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A Report on the Complex Word Identification

Shared Task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Gustavo Paetzold, and Lucia Specia. 2017. Complex Word Identification: Challenges in Data Annotation and System Performance. In *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2017)*, pages 59–63, Taipei, Taiwan. Asian Federation of Natural Language Processing.

8. Appendices

A. List of Corpora

In Table 6 below the corpora used in this study are listed. All corpora have been compiled within the SCAP initiative (Goethals, 2018). The specialised corpora with IDs 1 to 4 were used as the domain-specific corpora. The reference corpus combines the contents of all corpora in the table, resulting in one large 111 million-word corpus. When we use the reference corpus for keyness calculations, we first check if there is an overlap between the texts from the study corpus and those of the reference corpus. If so, the corresponding texts are removed from the reference corpus before performing the calculations.

B. Sentence Selection Method

In Table 7 below we include details on the method we used to select pedagogically suitable sentences from the four domain-specific corpora. The method is based on Pilán et al. (2016) and adapted to Spanish.

C. Annotation Instructions

In Figure 3 and Figure 4 below we include the written instructions provided to the participants during the data labelling process.

Corpus ID	Medium	Genre	Topic area	Source(s)	Samples	Words
Specialised corpora						
1	Written	News-papers	Economics	Cinco Días	19,251	10,787,219
2	Written	News-papers	Law	El País	41,351	24,695,247
3	Written	News-papers	Migration	El País	9,198	7,728,019
4	Written	News-papers	Health	El País	16,972	9,088,971
5	Written	News-papers	Tourism	El País (section "El Viajero")	10,839	7,589,762
General corpora						
1	Written	Fiction	Adult prose	Various Spanish novels (>year 2000)	531	43,661,672
2	Written	Fiction	Youth literature	Various Spanish novels (>year 2000)	104	7,528,422

Table 6: List of corpora used in the study.

Nr.	Criterion	Value
Search term		
1	Absence of search term	False
2	Number of matches	1
3	Position of search term	Anywhere
Well-formedness		
4	Dependency root	True
5	Ellipsis	False
6	Incompleteness	False
7	Non-lemmatised tokens	0
8	Non-alphabetical tokens	0
NEW	Explicit subject	True
Context independence		
9	Structural connective in isolation	False
10	Pronominal anaphora	0
11	Adverbial anaphora	0
12	L2 complexity in CEFR level	Unlimited
Additional structural criteria		
13	Negative formulations	0
14	Interrogative speech	False
15	Direct speech	False
16	Answer to closed questions	False
17	Modal verbs	$\leq$1
18	Sentence length	25 - 35 tokens
Additional lexical criteria		
19	Difficult vocabulary	Unlimited
20	Word frequency	Unlimited
21	Out-of-vocabulary words	Unlimited
22	Sensitive vocabulary	False
23	Typicality	None
24	Proper names	0
25	Abbreviations	0

Table 7: Details of the parameters we applied to select pedagogically suitable sentences from the domain-specific corpora. Criteria not included in Pilán et al. (2016) are marked as "NEW".

Predicting the difficulty level of words in Spanish

Example

El directivo , que ha celebrado un almuerzo de Navidad con la prensa , ha asegurado que ninguna teleco en el país le ha mostrado su preocupación .											
directivo (directivo)	1	2	3	4	5	asegurado (asegurar)	1	2	3	4	5
celebrado (celebrar)	1	2	3	4	5	teleco (teleco)	1	2	3	4	5
almuerzo (almuerzo)	1	2	3	4	5	país (país)	1	2	3	4	5
Navidad (navidad)	1	2	3	4	5	mostrado (mostrar)	1	2	3	4	5
prensa (prensa)	1	2	3	4	5	preocupación (preocupación)	1	2	3	4	5

Context

With this research experiment, we want to develop a system that automatically predicts the difficulty level of words in Spanish, taking into account the profile of the end user. The dataset on which we will train the system consists of four sets of fifty sentences, which you will assess in a minute. The sentences will each come from texts around a particular domain: **economy/business**, **law**, **social themes/migration**, and **health.** In this way, we want to arrive at a representative dataset, since vocabulary classes (and the corresponding exercises and learning materials) are often organised per domain.

1

Figure 3: Annotation instructions provided to the participants (translated from Dutch into English), page 1.

Instructions

- Read the sentence carefully
- Assign each of the listed words from the sentence to one of the categories described below by marking the box with an "x" or circling the number.

1	I know this word and its meaning, and I also use it actively in speaking/writing.
2	I know this word and its meaning, but I might not be able to use it on the top of my head in an oral/written conversation. When I have some time to think, however, I do think I would use it naturally.
3	I have heard/seen this word before and given the context I think that I more or less know what it means, but I do not see myself using this word actively.
4	This word sounds vaguely familiar and based on the context I could make an educated guess about its meaning, but I would still need a dictionary to be able to understand its exact meaning.
5	This word does not sound familiar at all to me, and even based on the context I do not know what it means, so I would definitely need a dictionary to get to know its meaning.

Additional observations

- It is **not** allowed to consult any resources during the experiment. All that matters is your judgement/intuition as a student of Spanish.
- There is a real chance that (many) words will be unknown to you. This is absolutely normal, and part of the design of the experiment.
- If the target word is a word with multiple meanings, base your annotation on the meaning in which the word appears in the specific context of the sentence. In other words: if you know the word, but not in the meaning in which it appears in the sentence, rate it as a word you do not know.
- Take your time to think about which category you assign to each word; it is crucial that the annotations represent your judgement/intuition. Although the idea is also not to spend minutes thinking about one single word, naturally.
- The experiment is relatively intensive from a cognitive point of view, so be sure to take enough small breaks in between (at least after having completed a set).

2

Figure 4: Annotation instructions provided to the participants (translated from Dutch into English), page 2.