You Shall Know a Word's Difficulty by the Family It Keeps: Word Family Features in Personalised Word Difficulty Classifiers for L2 Spanish

Jasper Degraeuwe
Ghent University (LT^3 / MULTIPLES research groups)
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
Jasper.Degraeuwe@UGent.be

Abstract

Designing vocabulary learning activities for foreign/second language (L2) learners highly depends on the successful identification of difficult words. In this paper, we present a novel personalised word difficulty classifier for L2 Spanish, using the LexComSpaL2 corpus as training data and a BiLSTM model as the architecture. We train a *base* version (using the original LexComSpaL2 data) and a *word family* version of the classifier (adding word family knowledge as an extra feature). The base version obtains reasonably good performance (F1 = 0.53) and shows weak positive predictive power (ϕ = 0.32), underlining the potential of automated methods in determining vocabulary difficulty for individual L2 learners. The "word family classifier" is able to further push performance (F1 = 0.62 and ϕ = 0.45), highlighting the value of well-chosen linguistic features in developing word difficulty classifiers.

1 Introduction

In the rapidly evolving digital era of the 21st century, language and technology are growing closer together than ever. Language technology tools – especially those driven by large language models (LLMs) – have become very adept at performing a wide range of tasks, ranging from summarising documents to translating texts from one language into another. In some cases, their output has even shown to be virtually indistinguishable from human-written materials (Else, 2023).

At the same time, despite the wealth of technological assistance, being able to understand and speak a foreign/second language (L2) can be said to remain an indispensable skill for anyone who wants to fully engage in foreign cultures, as building sustainable intercultural relationships involves tasks that are much more difficult to achieve by means of technological tools alone, such as interpreting humour and facial expressions (Godwin-Jones, 2019). What language technology tools do possess, however, is the ability to play the role of a valuable assistant in the L2 learning process.

This interface between second language acquisition (SLA) and computer assistance has commonly been referred to as Computer-Assisted Language Learning or CALL. Recently, the field of CALL has witnessed a growing interest in the specific subdomain of Intelligent CALL (ICALL). With techniques coming from the field of Artificial Intelligence (AI) and its subdomain of Natural Language Processing (NLP) as their source of "intelligence", ICALL environments aim to — among other goals — facilitate and/or (partially) automate the creation of language learning materials. Although the origins of ICALL can be traced back as far as to the 1980s (Nyns, 1989), it was not until the advent of static (Mikolov et al., 2013) and contextualised word embeddings (Devlin et al., 2019) followed by full-fledged generative LLMs that a true paradigm shift from CALL towards ICALL has started to take shape.

ICALL platforms can foster virtually any language skill, but in this paper we will specifically focus on ICALL for *vocabulary learning* purposes. With a large body of studies showing that text comprehension and vocabulary knowledge are positively correlated (e.g., Laufer and Ravenhorst-Kalovski, 2010; Schmitt et al., 2011), we know that a wide vocabulary is a fundamental requisite to be able to function in a language. Or, in the words of Wilkins (1972, p. 111): "without grammar very little can be conveyed, without vocabulary *nothing* can be conveyed". To help learners expand their L2 vocabulary, a combination of implicit and explicit learning activities is to be recommended (Nation, 2019; Schmitt, 2010a). 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 by-product,

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 312–325

because the main goal of the activity is the successful completion of an authentic task such as understanding the plot of a book.

In both of these strands, knowing which words might be difficult for target learners to understand or produce is a valuable source of information. In the case of the implicit approach, one of the key notions is that learners acquire vocabulary when they are exposed to input that is comprehensible but slightly beyond their current knowledge (Lichtman and VanPatten, 2021), implying that it should be known which parts of the input are comprehensible and which are not. For explicit learning, on the other hand, informed decisions need to be made on which words to in- and exclude from the activities, a task that becomes considerably easier if we know which words are (un)known by the learners.

In other words, identifying difficult words (or word uses) plays a pivotal role during the development of vocabulary learning materials. As an alternative to the labour-intensive process of identifying these words by hand, research within the field of ICALL has explored NLP-driven approaches to perform this task. Methods exploiting computer-readable resources in which words are linked to difficulty levels (or frequency bands, since frequency correlates with difficulty[1]; Schmitt, 2010b) constitute a first option, as they can automatically assign words in digital texts to their corresponding difficulty/frequency label (Finlayson et al., 2023).

However, apart from having limited coverage (only words included in the resources will be assigned a label), this approach does not take into account individual differences among learners. To overcome these limitations, machine learning systems can be designed, which offer 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 (Tack, 2021).

In this paper, we present a first-of-its-kind individualised word difficulty classifier for L2 Spanish. As our training data, we make use of LexComSpaL2 (**Lex**ical **Com**plexity for **Spa**nish **L2**; Degraeuwe and Goethals, 2024), a publicly available dataset[2] containing 2,240 in-context target words with the corresponding difficulty judgements of 26 L2 Spanish students. We compare the results of training two different versions of the classifier: a *base* version (only using the original LexComSpaL2 data) and a *word family* version (adding word family knowledge as an extra feature).

[1]It should be noted that this difficulty - frequency correlation does not mean that word difficulty *equals* word frequency. As shown in previous research (Pintard and François, 2020), word difficulty cannot be predicted by frequency values alone.

[2]The dataset is made available through a GitHub repository and was released under the ODC-By license, which grants the right to freely use and adapt the data as long as any use of the dataset is adequately attributed.

2 Related Research

This literature overview discusses lexical difficulty/complexity as defined in the field of linguistics in Section 2.1, the technique of lexical complexity prediction (i.e. the approach adopted to build LexComSpaL2) in Section 2.2, individualised learning in Section 2.3, and a detailed account of word families (used to create a "linguistically enriched" version of LexComSpaL2) in Section 2.4.

2.1 Difficulty and Complexity in SLA

In linguistics, the concept of "word difficulty/complexity" is usually subdivided into several dimensions, often dichotomous in nature. One of the most prominent distinctions is the one between *absolute* (or *objective*) and *relative* (or *agent-related*) complexity (Kortmann and Szmrecsanyi, 2012). In the former type, complexity is understood in terms of the linguistic properties of words, ranging from their length over the number of vowels and diphthongs they contain to their homonymic and/or polysemous character (i.e. the number of different meanings/senses they have). Especially the last feature plays an important role in an SLA setting, as lexically ambiguous items have shown to be more challenging to process and learn than single-meaning words (Bensoussan and Laufer, 1984).

Relative complexity (also denominated "difficulty"; Bulté et al., 2025), on the other hand, corresponds to the complexity as perceived by a particular language learner, meaning that psycholinguistic factors and world knowledge can come into play (North et al., 2023; Kortmann and Szmrecsanyi, 2012). In an L2 setting, an additional crucial factor in determining agent-related complexity is L1 influence, which can manifest itself through false friends (e.g., ES *listo* ['ready'] - NL *list* ['ruse, trick']) or cognates (e.g., ES *individuo* - NL *individu* - EN *individual*).

2.2 Lexical Complexity Prediction

Computational approaches to identifying difficult/complex words focus on "operationalising" the abovementioned linguistic concepts. A crucial aspect of this operationalisation is the presence of

some kind of "inventory" in which words are linked to discrete difficulty/complexity labels. One possible way to build such inventories is exploiting computer-readable versions of graded vocabulary lists (Dang et al., 2017), frequency lists (Davies and Hayward Davies, 2018), or graded coursebooks (in which case words are assigned to the level at which they first occur; Alfter, 2021). Another approach is to collect human annotations, either through online (crowdsourcing) platforms (Shardlow et al., 2021) or by means of dedicated research experiments (Tack, 2021).

The dataset used in this study, LexComSpaL2, falls in the last category (for more details on the corpus, see Section 3.1.1). The LexComSpaL2 annotations were gathered according to the principles of lexical complexity prediction (LCP; see Table 1 for an example), a relatively new strand within the field of NLP that provides an alternative to the binary complex word identification (CWI) approach (which labels words as either complex or non-complex; Yimam et al., 2018).

By using a five-point scale going from "very easy" to "very difficult" (for the full descriptors, see Section 3.1.1), LCP not only yields more fine-grained judgements than the binary CWI labels, it also enables making predictions based on "comparative complexity" (i.e. whether a word is more or less complex than another target word; North et al., 2023). Importantly, the term "complexity" as used in the field of LCP represents an amalgam of the concepts of complexity and difficulty described in Section 2.1, as it refers to the difficulty an individual may experience in understanding a given word as a result of both their personal knowledge and a word's linguistic properties (North et al., 2023). In this paper we adopt the same comprehensive definition but will give preference to the term "difficulty" instead of "complexity", since the individualisation of the predictions puts slightly more emphasis on the (personal knowledge of the) learner than on the linguistic properties of the target words.

Sentence	
They do hold elections, but candidates have to be endorsed by the conservative clergy, so dissenters are by definition excluded.	
Target word	**LCP Label**
do	1
hold	2
elections	3
candidates	1
have	1
be	1
endorsed	4
conservative	2
clergy	5
dissenters	5
definition	1
excluded	2

Table 1: Fictitious example of LCP annotations. The target sentence is taken from the CompLex dataset, the first LCP corpus ever created (Shardlow et al., 2020). In line with the LexComSpaL2 corpus, only nouns, verbs, and adjectives are considered in the example.

To the best of our knowledge, LexComSpaL2 is the only available LCP dataset that (1) specifically targets L2 learners and (2) enables training personalised word difficulty classifiers. Other LCP datasets are mostly constructed for the purpose of training models that can be integrated in a lexical simplification pipeline (Paetzold and Specia, 2017). A comprehensive overview of existing LCP datasets can be found in Shardlow et al. (2024).

Regarding the features used in LCP classifiers, recent research has revealed that a hybrid approach combining linguistic features (see Section 2.1) and LLM embeddings (e.g., BERT embeddings; Devlin et al., 2019) results in the highest performance (Ortiz-Zambrano et al., 2025). Earlier research, however, showed that also with static word embeddings good performance levels can be achieved (Tack, 2021). In this paper, we build on this line of research by combining static word embeddings with linguistic information on word families. By focusing on word families we aim to gain new insights into the value of linguistic features in automated word difficulty prediction, as previous research has mainly paid attention to lexical features related to the word itself (e.g., word length and number of syllables) and semantic features taken from resources such as WordNet (Fellbaum, 1998; e.g., number of synonyms, hypernyms, and/or hyponyms of a given target word).

2.3 Individualised Learning

As already touched upon above, another core and unique aspect of our word difficulty classifier is that it outputs *personalised* predictions. This way, we aim to integrate findings from the literature on indi-

vidual differences in SLA[3]. In brief, research in this domain has demonstrated that a variety of factors related to the individual can impact the learning process and learning outcomes. As mentioned in Section 2.1, a first crucial dimension of individual differences is the linguistic background of the learner, particularly (proficiency in) their L1 and experience with (learning) other languages (Degani and Goldberg, 2019). Other individual differences that have a considerable impact on the vocabulary learning process of L2 learners include cognitive factors such as memory capacity (Martin and Ellis, 2012) and the degree of out-of-school exposure to the L2 (De Wilde et al., 2022).

The domains of ICALL and NLP started to devote increasingly more attention to individualising system outputs. The most comprehensive approach to personalising the L2 learning process can be found in Intelligent Language Tutoring/Teaching Systems (ILTSs), which tailor learning materials to the specific needs of individual users on a macro (selecting and sequencing activities) and/or micro level (providing scaffolded feedback) (Meurers et al., 2019; Ruiz et al., 2023). Regarding word difficulty prediction, both in the domain of CWI (Gooding and Tragut, 2022; Tack, 2021) and LCP (Degraeuwe and Goethals, 2024; North et al., 2023) efforts have been undertaken to adopt a learner-centred and personalised perspective. In the present study, we aim to continue this line of research.

2.4 Word Families

Finally, we briefly discuss the concept of word families (Bauer and Nation, 1993), based on which we expanded the LexComSpaL2 dataset and trained a separate version of the classifier. As defined by Webb (2021, p. 941), "[w]ord families are made up of a headword, its inflections, and derivations". For the headword *address*, for example, this means that the word family consists of both the nominal (*addresses*) and verbal inflections (*addresses*, *addressed*, *addressing*), as well as derivations of the two (e.g., *addressee*, *readdress*, *unaddressed*) plus their inflected forms (e.g., *addressees*, *readdresses*, *readdressed*). Supported by empirical evidence from cognitive linguistics (Zhang and Lin, 2021), one of the main arguments in favour of using word family information in an L2 learning setting is that, once learners have acquired knowledge of the form-meaning connection of a given family member (e.g., *legal*), they can use their knowledge of the morphological system to infer the meaning of other members of the family (e.g., *legally*, *illegal*) (Nation and Webb, 2011; Nation, 2016).

[3]For extensive overviews of this domain, we refer to Dörnyei (2014) or Skehan (1991).

3 Methodology

The methodology consists of two main steps: (1) the preparation of the dataset on which the different versions of the classifier should be trained (Section 3.1) and (2) the actual development and training of the classifier (Section 3.2).

3.1 Dataset Preparation

3.1.1 Original LexComSpaL2 Dataset

To train the base version of the personalised word difficulty classifier, we used the LexComSpaL2 dataset in its original format (Degraeuwe and Goethals, 2024; see Table 2 for a dataset sample). LexComSpaL2 includes 2,240 target words distributed over 200 sentences coming from four different domains (economics, health, law, and migration). The sentences were selected from L1 newspaper corpora using a dedicated method specifically designed to extract pedagogically suitable sentences from corpus data (Pilán et al., 2016). Regarding the annotations, 26 L2 Spanish learners (from different proficiency levels but all L1 Dutch) were asked to rate the (in-context) difficulty of all nouns, verbs, and adjectives in the 200 sentences according to the five-point LCP scale. Importantly, Degraeuwe and Goethals (2024) tailored the original LCP descriptors to L2 learners as the target audience by projecting the LCP labels onto the vocabulary knowledge continuum (Schmitt, 2019), which conceptualises vocabulary knowledge as a construct that gradually moves from "no knowledge" over "receptive mastery" to "productive mastery" (see Table 3 for the adapted scale).

In summary, the 58,240 self-perceived judgements of word difficulty included in LexComSpaL2 constitute relevant and representative data to train personalised word difficulty classifiers for L2 learners, as the annotations were (1) provided by actual L2 learners and (2) taken from pedagogically suitable sentences that were selected in an attempt to mimic the often thematic organisation of real-life vocabulary learning courses and materials.

3.1.2 Word Family-Enriched Dataset

To enrich the original LexComSpaL2 dataset with word family information, we considered the following three word family levels: the word's **token**

Sentence ID	Sentence text	Target word	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	{P1: 3, P2: 2, P3: 2, [...], P24: 3, P25: 1, P26: 1}
		celebrado	{P1: 2, P2: 1, P3: 1, [...], P24: 2, P25: 1, P26: 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	{P1: 1, P2: 1, P3: 1, [...], P24: 1, P25: 1, P26: 1}
		atención	{P1: 2, P2: 1, P3: 1, [...], P24: 1, P25: 1, P26: 1}
		...	

Table 2: Sample from the LexComSpaL2 corpus that was also presented in Degraeuwe and Goethals (2024). Aggregated judgements (per proficiency level and overall) were omitted from the sample, since we only used the individual judgements to train the classifier. Target words are underlined and "P" stands for participant.

Rating	Original LCP description	Adapted description
1	Very easy: this word is very familiar to me	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 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 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	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	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 3: Original LCP descriptions compared to adapted descriptions proposed by Degraeuwe and Goethals (2024). The adapted descriptions are based on the vocabulary knowledge continuum (Schmitt, 2019).

form (also called the "type"), the word's **lemma**, and the **source** from which the word's lemma is derived (i.e. the "parent" of the lemma in the "family tree")[4]. As an illustration, let us consider the word *desaparecido* ('disappeared'): the **token** level consists of all occurrences of this exact word form in the LexComSpaL2 dataset; the **lemma** level corresponds to the lemma of *desaparecido* (i.e. the infinitive *desaparecer* - 'to disappear') and includes all of its other conjugated forms (e.g., *desaparezco*, *desaparecieron*); the **source** level corresponds to the "parent" of *desaparecer* (i.e. *aparecer* - 'to appear') and encompasses all inflected forms of this parent (e.g., *aparezco*, *aparecía*).

Next, we applied the following procedure to every target word in the LexComSpaL2 dataset:

1. Check if the exact **token** of the target word occurs more than once in the corpus. If so, we (1) calculate if there is a statistically significant difference ($p \leq 0.05$) between the annotations and (2) gather, for all participants individually, the lowest and highest annotated LCP value for the token in question.

2. Check if the **lemma** of the target word occurs more than once in the corpus. If so, we repeat the process described in the first step, but in this case for the target word's lemma.

3. Check if the **source** of the target word's lemma occurs more than once in the corpus. If so, we repeat the process described in the first

[4]The *token* and *lemma* levels correspond to, respectively, Level 1 and Level 2 of the Bauer and Nation (1993) taxonomy. The *source* level is specific to this study.

step for the source lemma. If the target word's lemma is a headword, this step is skipped.

All words were disambiguated for part of speech using Stanza[5], meaning that words such as *humano* ('human'), which can both be a noun and adjective, constitute two different tokens. To compute statistical significance, we used the non-parametric Kruskall-Wallis H-test (Kruskal and Wallis, 1952), which can be applied to two or more samples and does not assume normally distributed data. To look up the source lemma of a given target word, we used the publicly available word family resource[6] developed within the Spanish Corpus Annotation Project (Goethals, 2018).

We added the data to the original LexComSpaL2 corpus by creating three new versions of the corpus (one per word family level), in which we added four extra columns: one indicating if the target word occurs multiple times (*True* or *False* as value), one indicating if the annotations differ significantly (*True*, *False*, or *N/A* if the target word only occurs once), and two columns including the lowest and highest annotation per participant (or again *N/A*). This "word family-enriched" version of the dataset are made available as a part of the original LexComSpaL2 GitHub repository[7].

The descriptive statistics of the word family enrichment are presented in Table 4. For the token and lemma levels, *#candidates* refers to the number of, respectively, unique tokens and lemmas that occur more than once in the corpus. For the source level, *#candidates* refers to the number of unique tokens whose source lemma also occurs in the corpus. The *#statSignDiff* column indicates for how many of those candidates the learners' annotations differed significantly. Although these numbers are a by-product of the research and should also be interpreted as such, it does seem opportune to highlight that they seem to confirm as well as contradict the assumption that knowledge of one word family member means that learners also know the meaning of other family members (Section 2.4). At the token and lemma levels, the low number of statistically significant differences reveals that the annotations (and therefore also the degree of word knowledge) were consistent across the different occurrences. For the lemma level, this means that if learners acquired a certain degree of knowledge for one inflected form, they are highly likely to have the same degree of knowledge for other inflected forms. At the source level, however, we see that this conclusion does not hold, as in 106 of the 248 cases there was a statistically significant difference in the learners' difficulty judgements. The results clearly indicate that knowledge acquired at the token level is not necessarily transferred to the source level in the family tree (or vice versa). Returning to the example above, this means that if learners know *desaparecido* it does not necessarily imply that they also know *aparecer* (or vice versa).

Level	#candidates	#statSignDiff	#enriched
Token	159	2	355 / 2,240
Lemma	273	6	632 / 2,240
Source	248	106	297 / 2,240

Table 4: Descriptive statistics of word family enrichment of LexComSpaL2. The *#enriched* column contains the number of target words for which information other than *N/A* was added.

3.2 Classifier Training

3.2.1 Base Classifier

As the architecture for the base classifier, we used a Bidirectional Long Short-Term Memory (BiLSTM) model that follows a similar design as the CWI classifier presented in Tack (2021), who found that this type of neural network is able to personalise difficulty predictions. For each observation (i.e. a word linked to a learner's LCP annotation), the base model takes the following features as input: a character embedding[8], the word's fastText embedding (Cañete, 2019), and the participant information (unique identifier, proficiency level, years of experience, and L1[9]). Based on a softmax activation function, the output layer yields a probability distribution for the different classes, with the class for which the highest probability is obtained being selected as the predicted difficulty level for the word in question. A simplified visualisation of the model's architecture is presented in Figure 1 (for a full visualisation, see Appendix A) and the underlying code is made available in a GitHub repository[10].

[5] https://stanfordnlp.github.io/stanza/
[6] https://scap.ugent.be/overview-resources/
[7] https://github.com/JasperD-UGent/LexComSpaL2

[8] Randomly initialised and trained with a convolutional neural network (De Hertog and Tack, 2018).
[9] It should be noted that, since all annotators in the LexComSpaL2 corpus have Dutch as their L1, this feature will not contribute to personalising the predictions.
[10] https://github.com/JasperD-UGent/personalised-word-difficulty-classifier

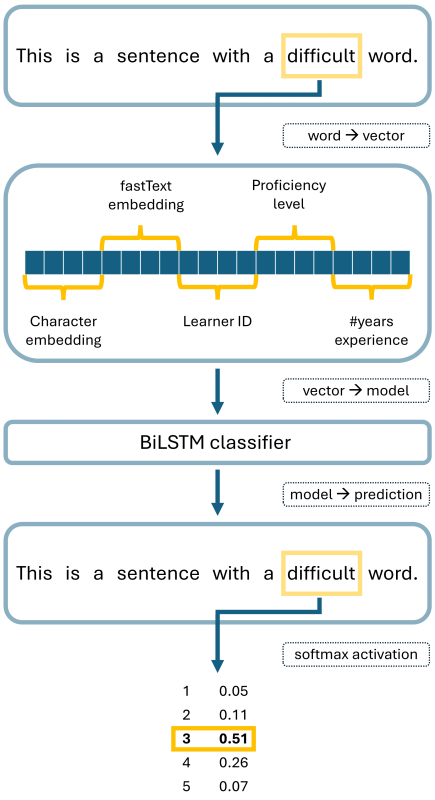

Figure 1: Simplified representation of BiLSTM word difficulty classifier.

To evaluate the model, a tenfold cross-validation setup was adopted. As our dataset split, we used the split added by Degraeuwe and Goethals (2024) to the LexComSpaL2 repository in anticipation of the corpus being used to train future machine learning models. This sentence-level dataset split includes ten different "folds" of the LexComSpaL2 data into a training (160 sentences or 80%), validation, and test set (20 sentences or 10% each). The complete overview of the number of training instances per fold is included in Appendix B.

Regarding training parameters, we set the number of epochs to 50, the batch size to 64, and the loss strategy to *sparse_categorical_crossentropy*. *Adam* was used as the optimiser and an early stopping monitor on the validation loss (with a patience of 10) was added to the training process. In each cross-validation run, the weights for the training samples were calculated (Appendix C) and used for weighting the loss function. A mask was set to all "non-target words" (i.e. all tokens which are not a noun, verb, or adjective and thus did not receive a label during the data collection) and their input vectors and sample weight were set to 0. This way, the sentence context was still correctly represented but the masked tokens were ignored during training. Finally, zero-padding (to the maximum sentence length of 35) was applied to all inputs and outputs.

3.2.2 Word Family-Enriched Classifier

The word family version of the classifier was built based on the exact same architecture as the base version. The only difference is that one additional feature was added to the word vector, containing the content of the four columns that were added to the dataset (Section 3.1.2). A simplified visualisation of this updated word vector is presented in Figure 2. To gain insights into the impact of each word family level, we trained the classifier based on (1) only the *token* level information as extra data, (2) only the *lemma* level data, (3) only the *source* level data, and (4) all three levels combined (*combi*). The word vector for *combi* was obtained by concatenating the "word family feature" from the three individual levels (i.e. the light-coloured part on the right-hand side of the vector visualisation in Figure 2) and appending these values to the "base features" (i.e. the dark-coloured part on the left-hand side of the vector in Figure 2).

4 Results and Analysis

The performance scores are presented in Table 5. We compare the results against a naïve most frequent label (MFL) baseline, which always predicts the most frequent difficulty label in the dataset (i.e. label 1). For evaluation, we first calculate two measures that are insensitive towards changes in class distribution: (1) the D' coefficient, which determines the degree of certainty in the predictions (Smith et al., 2021), and (2) the Matthews correlation coefficient (abbreviated as MCC and denoted as ϕ), which determines the quality of the predictions by estimating the strength of association between the true and predicted classes (Matthews, 1975). Changes in the values of these

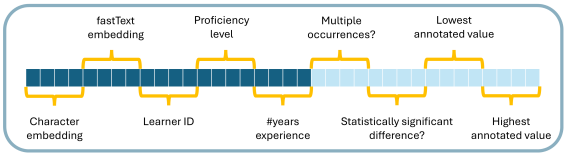

Figure 2: Simplified representation of word vector enriched with word family information. Word family values are different for each of the three possible word family levels (*token*, *lemma*, and *source*). Highest and lowest annotated value are provided per participant.

Classifier type	D' ↑	MCC ↑	F1 ↑	MSE ↓	RMSE ↓	Accuracy ↑
MFL baseline	0	0	0.32	2.61	1.62	0.49
Base	0.18 (± 0.01)	0.32 (± 0.02)	0.53 (± 0.02)	1.32 (± 0.1)	1.15 (± 0.04)	0.56 (± 0.02)
Word family (token)	0.23 (± 0.01)	0.37 (± 0.02)	0.56 (± 0.01)	1.25 (± 0.07)	1.12 (± 0.03)	0.59 (± 0.02)
Word family (lemma)	0.26 (± 0.01)	0.4 (± 0.02)	0.59 (± 0.02)	1.18 (± 0.08)	1.09 (± 0.04)	0.61 (± 0.02)
Word family (source)	0.23 (± 0.01)	0.38 (± 0.02)	0.57 (± 0.02)	1.24 (± 0.11)	1.11 (± 0.05)	0.59 (± 0.02)
Word family (combi)	**0.32** (± 0.01)	**0.45** (± 0.02)	**0.62** (± 0.02)	**1.11** (± 0.1)	**1.05** (± 0.05)	**0.63** (± 0.02)

Table 5: Performance of the different personalised word difficulty classifiers. We report the mean score across the ten cross-validation runs for each of the six performance metrics. Standard deviation values are included between parentheses and the top score per metric is presented in bold.

metrics can be fully attributed to changes in the model, and not to aspects inherent to the data such as class imbalance (Tack, 2021). MCC values can go from -1 (inverse prediction) over 0 (average random prediction) to 1 (perfect prediction), while the D' coefficient ranges between 0 (no discriminative power) and 1 (full discriminative power).

In addition, we also report three commonly used metrics in machine learning: weighted F1 (i.e. the harmonic mean of precision and recall), mean squared error or MSE (i.e. the average squared difference between the true and predicted values), and root mean squared error or RMSE (which converts MSE values to the same units as the dependent variable, in our case the 1-5 LCP scale). Finally, we also include the intuitive accuracy metric (i.e. the number of correct predictions divided by the total number of predictions).

Since, to the best of our knowledge, this is the first study which specifically analyses the potential of including word family information as an input feature, our research provides valuable new insights into the added value of linguistic features in LCP-based classifiers. The results in Table 5 unequivocally show that word family information has a noticeable positive impact on model performance, with the top-performing *combi* classifier achieving a large increase on all metrics in comparison to both the MFL baseline ($+0.32$ for D'; $+0.45$ for MCC; -0.57 for RMSE) and the base classifier ($+0.14$ for D'; $+0.13$ for MCC; -0.1 for RMSE).

When breaking down the results per type of classifier, a first finding to be highlighted is that, though leaving ample room for improvement, the **base classifier** already achieves reasonably good performance. The mean D' and MCC values (0.18 and 0.32, respectively) suggest that the model has acquired weak positive discriminatory and predictive power, while the RMSE score reveals that – on average and including penalisation – the model's prediction is only 1.15 away from the true label. Importantly, the base model also outperforms the MFL baseline by a large margin (e.g., MCC of 0 compared to 0.32, weighted F1 of 0.53 compared to 0.32, and RMSE of 1.15 compared to 1.62).

When comparing the base to the **word family-enriched classifier**, the results clearly show that any type of word family information is helpful for the model, as all subtypes outperform the base version on every metric. The increase in performance is most notable at the lemma level ($+0.08$ for MCC; -0.06 for RMSE), suggesting that the model successfully leveraged information on the knowledge a given learner has acquired for one inflected form of a lemma to predict the label of other inflected forms of that lemma. However, it should be noted that the lemma level is also the level at which most instances in the dataset were enriched (Table 4), which may have played an important role in this particular subtype obtaining the largest increase. Regarding the results for the source subtype, it should be highlighted that – next to the lower number of enriched instances – the 106 statistically significant differences in annotations between the target word and its source (see Section 3.1.2) might be a second reason for the smaller increase at the source level compared to the lemma level.

In summary, the findings of our study provide strong evidence in favour of integrating word family features (and well-chosen linguistic features in general) into personalised word difficulty classifiers. Particularly, with the *combi* classifier obtaining the highest scores, the take-home message is that the more relevant data on word family knowl-

edge are added, the better the classifier's predictions of word difficulty become.

5 Conclusion

In this paper, we presented a personalised word difficulty classifier for L2 Spanish, trained on the LexComSpaL2 dataset (Degraeuwe and Goethals, 2024). Based on a straightforward BiLSTM architecture with a softmax activation function, the classifier can take any Spanish target sentence as input and will predict a difficulty label ranging from 1 to 5 for every content word in the sentence. Moreover, thanks to the inclusion of learner-specific features in the training process (e.g., proficiency level and years of experience), the model attempts to tailor its output to the unique profile of every learner individually. In doing so, the classifier goes beyond the generic, one-size-fits-all difficulty levels often used in L2 vocabulary learning resources (e.g., based on the Common European Framework of Reference for Languages [CEFR]).

By comparing a base classifier to a "word family-enriched" one, we highlighted the notable added value of feeding information on word families – and of adding linguistic features in general – to word difficulty classifiers. With the top-performing model obtaining an MCC value of 0.45, an F1 score of 0.62, and an RMSE score of 1.05, our classifier shows great potential to be included in real-life ICALL scenarios, for instance as a "difficult word detector" in a personalised reading assistant.

In future studies, we aim to test other machine learning architectures (e.g., using LLMs) and compare them against the BiLSTM classifier presented in this study. Other directions for future research include (1) studying the effect of replacing the static fastText embeddings as an input feature by contextualised word embeddings, (2) analysing the addition of more linguistic features next to word family information, and (3) collecting – in a GDPR-compliant fashion – more information about the participants in order to expand the "learner profile" input feature. Possible additional types of participant information include personal interests (e.g., hobbies), reading behaviour (in L1 or L2), and mastery of other languages than Dutch as L1 and Spanish as L2.

Limitations

A first important observation to be made concerns the real-life applicability of the classifier. Despite the promising results, it could be argued that the predicted values of the model – even for the best-performing classifier – are not yet close enough to the expected values for the model to be integrated *as is* in real-life settings. As shown by the (R)MSE scores[11] of 1.11 and 1.05 for *combi*, there is still a considerable difference between what the classifier should predict and what it actually predicts. Especially in a pedagogical setting it is crucial to obtain relatively high accuracy and precision rates (which is why the F1 and accuracy metrics were included in the analysis), because it should be avoided at all cost that learners lose precious time over errors in their learning materials. For example, if the classifier were used to identify vocabulary items that are known passively but not actively by a given learner (i.e. label 3) and ended up selecting words which are known (very) well by the learner, the exercise would lose most of its pedagogical value.

As obtaining promising but not (yet) pedagogically usable results is a recurrent finding in research on automated word difficulty prediction – Pintard and François (2020), for example, report a top accuracy score of 0.54 for their French CEFR classifier –, one might be left to wonder if the concept of word difficulty (especially for individual learners) is too sophisticated for machine learning classifiers to fully capture. In fact, despite the existence of clear patterns (e.g., high-frequency words tend to be easier than low-frequency words and long words tend to be more difficult than short words), there is also a wide range of factors that may affect perceived word difficulty but that are much harder to model using computational techniques (e.g., a word's degree of abstractness/concreteness, a learner's world knowledge, or a learner's ability to deduce meaning based on morphological knowledge or contextual clues). Yet, using (generative) LLMs as classifiers and/or expanding the number of features related to the learner (see also Section 5) are two research avenues which have the potential of providing the domain of automated word difficulty prediction with a new *élan* and leading to a considerable increase in performance. Additionally, the integration of features indicating how specialised words are for a given domain – for instance using "keyness" (Gabrielatos, 2018) or "termhood" (Rigouts Terryn et al., 2021) metrics – could also be a direction worth pursuing.

[11] These metrics penalise predictions that are far from the true label more severely than near-correct predictions (e.g., a predicted value of 1 while the true label is 5).

Secondly, in the current setup, every learner who wants to get personalised predictions from the classifier first needs to annotate all 200 sentences in the LexComSpaL2 corpus, as this information is used to build the "learner profile" input feature. As suggested by Degraeuwe and Goethals (2024), to facilitate the implementation in a real-life setting, an item analysis could be performed on the dataset to identify the most "informative sentences" and have new learners annotate this "concentrated" set of sentences instead. Another limitation of the dataset is that, currently, only annotations from L1 Dutch speakers are included. To assess the true personalisation potential of the classifier, the LexComSpaL2 dataset would need to be expanded with annotations coming from L2 Spanish who do not have Dutch as their L1.

Finally, regarding the analyses conducted, the present paper did not provide an in-depth evaluation of the personalisation potential of the classifier. In future research, we aim to isolate this aspect of the model, for example by performing a comparative analysis of the results per learner in order to identify potential differences and look for factors that might explain these differences (e.g., by studying if they correlate with the learners' proficiency level). Additionally, the study did not address if and how the sentence context impacts the perceived difficulty of a word (perspective of the learner) and how this relates to the predicted difficulty of that word (perspective of the computer). Finally, it should be noted that we did not apply any word sense disambiguation (WSD) method to the data. As a result, homonymic and polysemous words (e.g., *banco* as a bench and as a financial institution) were not considered as two separate tokens or lemmas.

Acknowledgments

The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, FWO and the Flemish Government – department EWI. Additionally, a sincere word of gratitude goes to Janiça Hackenbuchner (for acting as a soundboard and proofreading), to the people who attended the LATILL workshop at the University of Tübingen (for giving me the inspiration that led to this paper), and to the anonymous reviewers (for their valuable feedback and suggestions).

References

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

Laurie Bauer and I.S.P. Nation. 1993. Word Families. *International Journal of Lexicography*, 6(4):253–279.

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

Bram Bulté, Alex Housen, and Gabriele Pallotti. 2025. Complexity and Difficulty in Second Language Acquisition: A Theoretical and Methodological Overview. *Language Learning*, 75(2):533–574.

José Cañete. 2019. Spanish Word Embeddings.

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

Mark Davies and Kathy Hayward Davies. 2018. *A frequency dictionary of Spanish: Core vocabulary for learners*, 2 edition. Routledge frequency dictionaries. Routledge, London ; New York.

Dirk De Hertog and Anaïs Tack. 2018. Deep Learning Architecture for Complex Word Identification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 328–334, New Orleans, Louisiana. Association for Computational Linguistics.

Vanessa De Wilde, Marc Brysbaert, and June Eyckmans. 2022. Formal versus informal L2 learning: How do individual differences and word-related variables influence french and English L2 vocabulary learning in Dutch-speaking children? *Studies in Second Language Acquisition*, 44(1):87–111.

Tamar Degani and Miri Goldberg. 2019. How Individual Differences Affect Learning of Translation-Ambiguous Vocabulary. *Language Learning*, 69(3):600–651.

Jasper Degraeuwe and Patrick Goethals. 2024. LexComSpaL2: A lexical complexity corpus for Spanish as a foreign language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10432–10447, Torino, Italia. ELRA and ICCL.

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.

Zoltán Dörnyei. 2014. *The Psychology of the Language Learner*. Routledge.

Holly Else. 2023. Abstracts written by ChatGPT fool scientists. *Nature*, 613(7944):423–423.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

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.

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.

Robert Godwin-Jones. 2019. In a World of SMART Technology, Why Learn Another Language? *Educational Technology & Society*, 22(2):4–13.

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, Gent, Belgium. Éditions Universitaires Européennes.

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.

Bernd Kortmann and Benedikt Szmrecsanyi, editors. 2012. *Linguistic Complexity: Second Language Acquisition, Indigenization, Contact*. Mouton de Gruyter.

William H. Kruskal and W. Allen Wallis. 1952. Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, 47(260):583–621.

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.

Karen Lichtman and Bill VanPatten. 2021. Was Krashen right? Forty years later. *Foreign Language Annals*, 54(2):283–305.

Katherine I. Martin and Nick C. Ellis. 2012. The roles of phonological short-term memory and working memory in L2 grammar and vocabulary learning. *Studies in Second Language Acquisition*, 34(3):379–413.

B.W. Matthews. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451.

Detmar Meurers, Kordula De Kuthy, Florian Nuxoll, Björn Rudzewitz, and Ramon Ziai. 2019. Scaling Up Intervention Studies to Investigate Real-Life Foreign Language Learning in School. *Annual Review of Applied Linguistics*, 39:161–188.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space.

I.S.P. Nation. 2016. *Making and Using Word Lists for Language Learning and Testing*. John Benjamins Publishing Company, Amsterdam.

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

I.S.P. Nation and Stuart Webb. 2011. *Researching and analyzing vocabulary*, 1 edition. Heinle, Cengage Learning, Boston, MA.

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

Roland R. Nyns. 1989. Is intelligent computer-assisted language learning possible? *System*, 17(1):35–47.

Jenny A. Ortiz-Zambrano, César H. Espín-Riofrío, and Arturo Montejo-Ráez. 2025. Deep Encodings vs. Linguistic Features in Lexical Complexity Prediction. *Neural Computing and Applications*, 37(3):1171–1187.

Gustavo H. Paetzold and Lucia Specia. 2017. A Survey on Lexical Simplification. *Journal of Artificial Intelligence Research*, 60:549–593.

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.

Alice Pintard and Thomas François. 2020. Combining Expert Knowledge with Frequency Information to Infer CEFR Levels for Words. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 85–92, Marseille, France. European Language Resources Association.

Ayla Rigouts Terryn, Véronique Hoste, and Els Lefever. 2021. HAMLET: Hybrid Adaptable Machine Learning approach to Extract Terminology. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 27(2):254–293.

Simón Ruiz, Patrick Rebuschat, and Detmar Meurers. 2023. Supporting Individualized Practice through Intelligent CALL. In *Practice and Automatization in Second Language Research*, 1 edition, pages 119–143. Routledge, New York.

Norbert Schmitt. 2010a. 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.

Matthew Shardlow, Kai North, and Marcos Zampieri. 2024. Multilingual resources for lexical complexity prediction: A review. In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, pages 51–59, Torino, Italia. ELRA and ICCL.

Peter Skehan. 1991. Individual Differences in Second Language Learning. *Studies in Second Language Acquisition*, 13(2):275–298.

Thomas J. Smith, David A. Walker, and Cornelius M. McKenna. 2021. A coefficient of discrimination for use with nominal and ordinal regression models. *Journal of Applied Statistics*, 48(16):3208–3219.

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. 2021. The lemma dilemma: How should words be operationalized in research and pedagogy? *Studies in Second Language Acquisition*, 43(5):941–949.

D. A. Wilkins. 1972. *Linguistics in language teaching*. Edward Arnold, London.

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.

Haomin Zhang and Jiexin Lin. 2021. Morphological knowledge in second language reading comprehension: examining mediation through vocabulary knowledge and lexical inference. *Educational Psychology*, 41(5):563–581.

A Full Visualisation Classifier Architecture

The full visualisation of the architecture of the BiLSTM classifier is presented in Figure 3.

B Overview of Observations per Cross-Validation Fold

The overview of the observations per cross-validation fold (overall and per LCP label) is provided in Table 7.

C Class Weights for Cross-Validation

The class weights used by the BiLSTM classifier are presented in Table 6.

Fold	1	2	3	4	5
1	0.41	0.9812	1.2587	2.2369	3.333
2	0.4137	0.9859	1.2493	2.1925	3.2078
3	0.4126	0.9924	1.2446	2.1808	3.2616
4	0.4068	0.9899	1.2553	2.2354	3.4804
5	0.4094	0.9847	1.2505	2.2173	3.4355
6	0.4129	0.9877	1.2444	2.2119	3.2232
7	0.4116	0.9984	1.2529	2.1998	3.161
8	0.411	0.9974	1.2524	2.1979	3.2137
9	0.4059	0.9845	1.2613	2.2673	3.4856
10	0.405	0.9801	1.267	2.2647	3.5718

Table 6: Class weights of BiLSTM word difficulty classifier per cross-validation fold.

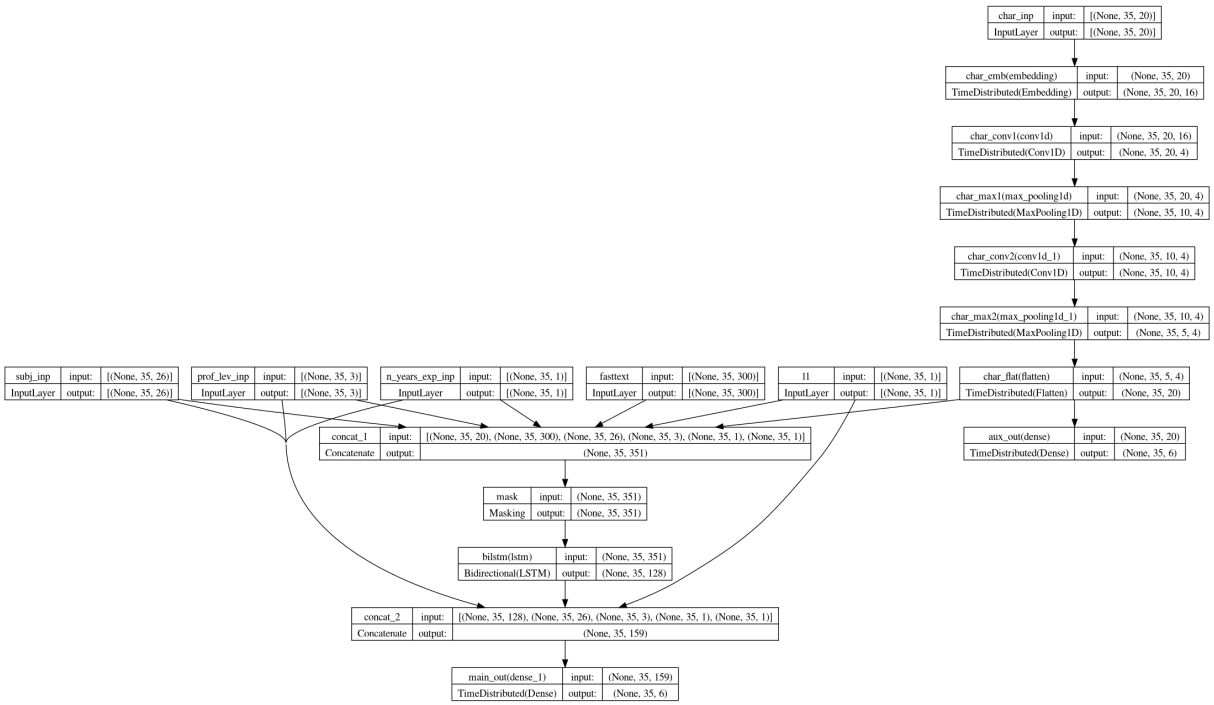

Figure 3: Full representation of BiLSTM word difficulty classifier.

Fold	#observations (target words | annotations)			#observations per label				
	TR	VA	TE	1	2	3	4	5
1	1,796 | 46,696	216 | 5,616	228 | 5,928	TR: 22,781 VA: 2,747 TE: 2,889	TR: 9,518 VA: 1,146 TE: 1,123	TR: 7,420 VA: 924 TE: 948	TR: 4,175 VA: 516 TE: 556	TR: 2,802 VA: 283 TE: 412
2	1,797 | 46,722	227 | 5,902	216 | 5,616	TR: 22,589 VA: 3,081 TE: 2,747	TR: 9,478 VA: 1,163 TE: 1,146	TR: 7,480 VA: 888 TE: 924	TR: 4,262 VA: 469 TE: 516	TR: 2,913 VA: 301 TE: 283
3	1,787 | 46,462	226 | 5,876	227 | 5,902	TR: 22,522 VA: 2,814 TE: 3,081	TR: 9,364 VA: 1,260 TE: 1,163	TR: 7,466 VA: 938 TE: 888	TR: 4,261 VA: 517 TE: 469	TR: 2,849 VA: 347 TE: 301
4	1,775 | 46,150	226 | 5,876	226 | 5,876	TR: 22,692 VA: 2,911 TE: 2,814	TR: 9,324 VA: 1,203 TE: 1,260	TR: 7,353 VA: 1,001 TE: 938	TR: 4,129 VA: 601 TE: 517	TR: 2,652 VA: 498 TE: 347
5	1,799 | 46,774	202 | 5,252	239 | 6,214	TR: 22,851 VA: 2,655 TE: 2,911	TR: 9,500 VA: 1,084 TE: 1,203	TR: 7,481 VA: 810 TE: 1,001	TR: 4,219 VA: 427 TE: 601	TR: 2,723 VA: 276 TE: 498
6	1,818 | 47,268	220 | 5,720	202 | 5,252	TR: 22,893 VA: 2,869 TE: 2,655	TR: 9,571 VA: 1,132 TE: 1,084	TR: 7,597 VA: 885 TE: 810	TR: 4,274 VA: 546 TE: 427	TR: 2,933 VA: 288 TE: 276
7	1,803 | 46,878	217 | 5,642	220 | 5,720	TR: 22,776 VA: 2,772 TE: 2,869	TR: 9,391 VA: 1,264 TE: 1,132	TR: 7,483 VA: 924 TE: 885	TR: 4,262 VA: 439 TE: 546	TR: 2,966 VA: 243 TE: 288
8	1,804 | 46,904	219 | 5,694	217 | 5,642	TR: 22,822 VA: 2,823 TE: 2,772	TR: 9,405 VA: 1,118 TE: 1,264	TR: 7,490 VA: 878 TE: 924	TR: 4,268 VA: 540 TE: 439	TR: 2,919 VA: 335 TE: 243
9	1,775 | 46,150	246 | 6,396	219 | 5,694	TR: 22,738 VA: 2,856 TE: 2,823	TR: 9,375 VA: 1,294 TE: 1,118	TR: 7,318 VA: 1,096 TE: 878	TR: 4,071 VA: 636 TE: 540	TR: 2,648 VA: 514 TE: 335
10	1,766 | 45,916	228 | 5,928	246 | 6,396	TR: 22,672 VA: 2,889 TE: 2,856	TR: 9,370 VA: 1,123 TE: 1,294	TR: 7,248 VA: 948 TE: 1,096	TR: 4,055 VA: 556 TE: 636	TR: 2,571 VA: 412 TE: 514

Table 7: Overview of observations per cross-validation fold for training ("TR"), validation ("VA"), and test ("TE") sets. The training set always contains 160 sentences, the validation and test sets always contain 20. The sets always contain an equal number of sentences per domain (economics, health, law, and migration).