Conplext 1.0: A Multilingual Lexical Complexity Prediction Dataset for L2 Learning

David Alfter[1], **Jasper Degraeuwe**[2]
[1]University of Gothenburg, Sweden
[2]Ghent University, Belgium
david.alfter@gu.se, jasper.degraeuwe@ugent.be

Abstract

This paper presents `Conplext 1.0`, a multilingual dataset designed for lexical complexity prediction in the context of second language (L2) learning. The resource covers 3,901 sentence contexts for 1,000 vocabulary items across five languages (English, French, Spanish, Swedish, and Dutch), each aligned with Common European Framework of Reference (CEFR) proficiency levels. Contexts were generated using a generative large language model and subsequently filtered for pedagogical suitability. A large-scale best–worst scaling (BWS) annotation experiment is being conducted with L2 learners to derive continuous, learner-informed lexical complexity values. The resulting dataset enables the development of context-aware word difficulty models that account for variation across both languages and learning stages. In addition to its primary use in lexical complexity prediction, `Conplext` provides valuable opportunities for research in word sense disambiguation, generative model evaluation, and adaptive language learning applications. By integrating computational and educational perspectives, this work advances the study of lexical difficulty in multilingual language learning environments.

Keywords: lexical complexity, language learning, best-worst scaling

1. Introduction

Lexical complexity denotes the degree of difficulty a word poses to a reader, influenced by linguistic and contextual factors such as frequency, morphology, length, and usage. The study of lexical complexity forms the foundation of lexical simplification, a research area aimed at improving the accessibility of texts for specific audiences, including children (De Belder et al., 2010), language learners (Petersen and Ostendorf, 2007; Rets and Rogaten, 2021), and individuals with reading impairments (Devlin, 1998; Chung et al., 2013), as well as for specialized domains such as medicine (Deléger and Zweigenbaum, 2009) and law (LoPucki, 2014).

Early research addressed lexical complexity through complex word identification (Shardlow, 2013), a binary classification task that distinguishes simple words from complex words. Shardlow (2013) provided one of the first comprehensive studies of automatic lexical simplification, combining the identification of complex words with the generation of simpler alternatives. Subsequent approaches extended this paradigm using feature-based machine learning models (Paetzold and Specia, 2016b).

In parallel, the field evolved toward graded lexical complexity prediction (Gala et al., 2013, 2014), which seeks to assign continuous or discrete complexity scores reflecting educational or proficiency levels (Tack et al., 2016; Alfter et al., 2016; Alfter and Volodina, 2018; Tack et al., 2018; Pintard and François, 2020). This graded perspective closely aligns with research on (second) language acquisition and supports applications such as adaptive learning materials (Burstein et al., 2017; Alfter and Graën, 2019) and personalized vocabulary learning systems (Avdiu et al., 2019; Ehara et al., 2018; Yancey and Lepage, 2018).

However, lexical complexity is not an inherent property of words but a contextual and relative phenomenon (Alfter, 2021). The perceived difficulty of a word depends not only on its intrinsic features such as frequency, length, or morphology but also on its surrounding context and the reader's world knowledge (North et al., 2023). Even non-polysemous words may vary in complexity across contexts, as their semantic, syntactic, or domain-specific usage imposes varying cognitive demands.

Furthermore, the dimension of language learning as well as the focus on languages other than English remain comparatively underexplored. Most existing work concentrates on lexical complexity prediction for English, where the primary objective is to determine which words are complex in a given text, rather than how complex those words are for learners at different proficiency levels (Alfter, 2025). Addressing this gap requires models capable of capturing not only general lexical difficulty but also its variation across learning contexts, which in turn requires the existence of annotated resources.

This study aims to fill this gap by presenting `Conplext`[1], a multilingual (English, French, Spanish, Swedish, and Dutch) lexical **complexity** predic-

[1]The dataset is made publicly available in a GitHub repository at `https://github.com/JasperD-UGent/Conplext`.

Proceedings of the 2nd Workshop on Evaluating Text Difficulty in a Multilingual Context (DeTermIt! 2026), pages 22–32
11 May 2026.

tion dataset containing 3,901 sentence **contexts** for 1,000 vocabulary items covering five different CEFR[2] (Council of Europe, 2001, 2018) language proficiency levels. At the time of writing, a large-scale best-worst scaling annotation experiment is being performed on the dataset, with foreign/second language (L2) learners as the participants. For each in-context use of a given vocabulary item, the annotations will be converted into a numerical complexity value. These values can then be used to train context-aware word difficulty classifiers that meet the specific needs of language learners. The results of the annotation experiment will be presented in a separate follow-up study.

2. Related Research

Since the first Shared Task on Complex Word Identification (CWI) in 2016 (Paetzold and Specia, 2016a), lexical complexity research has undergone several conceptual and methodological developments. The 2016 task focused exclusively on English and framed complexity prediction as a binary classification problem, distinguishing simple from complex words. The subsequent 2018 CWI Shared Task (Yimam et al., 2018) expanded this framework to include multilingual and cross-lingual settings while maintaining the same binary distinction.[3]

A major change occurred with the 2021 SemEval Shared Task on Lexical Complexity Prediction (LCP; Shardlow et al., 2021), which introduced continuous complexity scores derived from Likert-scale annotations and provided multiple contextual instances for many words. This represented an important step toward a more fine-grained modeling of lexical difficulty, aligning the task more closely with psycholinguistic and educational perspectives on word comprehension.

Despite this progress, recent work continues to exhibit a tendency toward simplification. The 2024 MLSP Shared Task (Shardlow et al., 2024), while advancing the multilingual dimension, largely reverted to a one-to-one mapping between words and complexity labels, offering limited contextual variation. A notable exception is the LexComSpaL2 dataset, which was specifically built to train personalized word-level difficulty classifiers for L2 learners of Spanish (Degraeuwe, 2025). In a similar vein, the 2025 Shared Task on Readability-Controlled Text Simplification (Alva-Manchego et al., 2025) sought to bridge text simplification and language learning by requiring participants to simplify English paragraphs to CEFR target levels A2 or B1.

However, as highlighted in Alfter (2025), current lexical complexity prediction resources remain ill suited for language learning applications. They typically provide too few contexts per word to capture the fact that a single lexical item can exhibit multiple complexity values depending on its contextual usage. Consider the following sentences, in which the word *bear* exhibits different levels of complexity, notably due to the multi-word expression *bear market* in the second sentence:

- The brown **bear** is a wild animal.
- We currently have a **bear** market.

While the example uses a polysemous word, even non-polysemous words can exhibit different complexity values, as in the two sentences presented below. This highlights the importance of context in complexity estimation.

- The dress has **lace**.
- The antique tablecloth was edged with intricate **lace**.

Finally, although generative large language models (LLMs) may sometimes struggle to accurately interpret CEFR levels (Benedetto et al., 2025), previous studies have shown that learners tend to prefer generated example sentences over "authentic" examples selected from corpora (Degraeuwe and Goethals, 2024). Motivated by this finding, our approach employs generative LLMs to produce diverse contextual instances for the selected target words.

3. Data Compilation

3.1. Data Source

As the starting point for `Conplext`, we take the textbook-derived word lists collectively known as CEFRLex[4]. Comprising a collection of machine-readable graded lexical resources that describe the frequency distributions of words observed across the six CEFR levels[5], CEFRLex is perfectly aligned with our L2 learning target setting. The resources used in the present study are EFLLex for English (Dürlich and François, 2018), FLELex for French (Tack et al., 2016), NT2Lex for Dutch (Tack et al., 2018), SVALex for Swedish (Francois et al., 2016), and ELELex for Spanish (François and Cock, 2018). Details on the specific CEFRLex files used and the mapping of their resource-specific part-of-speech

[2] Common European Framework of Reference for Languages

[3] The 2018 task also included a continuous prediction subtask, but its labels were derived by averaging binary annotations. We therefore regard it primarily as a binary classification task.

[4] https://cental.uclouvain.be/cefrlex

[5] In practice, the last level C2 is often left out due to scarcity of data. Only French contains this level.

Language	#total	#C2	#MWEs	#nouns	#verbs	#adjectives	#adverbs
EN	15,281	0	3,852	9,244	2,230	2,630	754
ES	14,290	0	629	8,297	2,222	2,698	163
FR	14,199	490	0	7,807	2,597	3,010	603
NL	15,227	0	459	9,049	2,656	2,190	449
SV	15,686	0	1,451	9,300	1,991	2,145	367

Table 1: Statistics on CEFRLex resources prior to applying pre-processing steps. "MWE" stands for multi-word expression.

Language	NOUN					VERB				
	A1	A2	B1	B2	C1	A1	A2	B1	B2	C1
EN	1,064	1,056	1,036	1,300	1,201	348	280	368	545	463
ES	2,027	2,051	1,713	1,095	1,037	594	548	394	376	293
FR	2,290	1,534	2,133	711	870	794	511	744	216	269
NL	396	2,993	3,044	2,348	252	157	1,151	744	556	46
SV	523	1,431	2,617	2,822	1,884	178	288	595	556	374

Table 2: Statistics on CEFRLex resources after applying pre-processing steps.

(POS) tags to Universal Dependencies (UD) tags can be found in Appendix A.

3.2. Data Pre-Processing

Before selecting the vocabulary items for `Conplext`, a series of pre-processing steps was applied to the five original CEFRLex resources. First, for each entry in each of the five resources, we retrieved (1) its CEFR label (which corresponds to "the level of first occurrence", i.e. the CEFR level at which the word first occurs in the textbooks on which the resource is based) and (2) its overall frequency across all textbooks.

Secondly, as our goal is to create a dataset that provides a wide variety of different sentence contexts per target item, we decided to focus on the two part-of-speech POS categories that show the highest degree of polysemy: nouns and verbs (Raganato et al., 2017). All other POS categories were excluded. Nouns and verbs together account for around 74% of all entries included in the five CEFRLex resources used in this study (see Table 1 for the exact numbers). In future releases of `Conplext`, we plan to expand coverage by including adjectives and adverbs as well (i.e. the two remaining major content word categories).

Thirdly, to ensure a consistent approach across all five target languages, we excluded words at C2 level (only available for French) and multi-word expressions (not available for French). Finally, single-character entries that passed the above mentioned exclusion criteria (predominantly letters of the alphabet tagged as nouns) were eliminated as well, since we consider knowledge of the alphabet to be a basic prerequisite for L2 learning.

The statistics on the CEFRLex resources *prior to* applying any pre-processing steps are included in Table 1. The overview of the number of remaining candidate entries per language, POS category, and CEFR level *after* applying the pre-processing steps is presented in Table 2.

3.3. Data Sampling

On this remaining set of candidate entries, we performed a structured data sampling procedure to arrive at a final dataset that was (1) balanced in terms of frequency distribution of the vocabulary items and (2) manageable in terms of number of items to annotate in the best-worst scaling experiment to be conducted afterwards (see Section 3.7 for more details on the latter). We decided to select 1,000 target items in total, with an equal number of items for each unique language–CEFR–POS combination:

- $1,000 \div 5 = 200$ items per language
- $200 \div 5 = 40$ items per CEFR level per language
- $40 \div 2 = 20$ items per POS category (i.e. nouns and verbs) per CEFR level per language

As the data selection method, we employed random stratified sampling (n = 10), drawing from the

top 50% most frequent vocabulary items. Sampling was performed separately for each unique language–CEFR–POS combination. For instance, in the case of the 1,064 A1-level nouns for English, the 50% (i.e. 532) most frequent nouns were first ranked by overall frequency and then divided into ten distinct strata. From each stratum, two items were randomly selected to form part of the final dataset. This process was repeated for the remaining 49 unique language–CEFR–POS combinations to yield a balanced and representative set of vocabulary items. Finally, all items were manually verified: in case (1) tagging errors[6], (2) offensive vocabulary, or (3) highly specialized words occurred in the sample, these were eliminated and randomly replaced by another item from the same stratum.

3.4. Model Selection

In order to select a model for the task of context generation, we conducted a preliminary study with three current state-of-the-art models – namely OpenAI's GPT-4o, Anthropic's Claude Sonnet 4.5 and Google's Gemini 2.5 Flash – prompted through their respective APIs. We randomly selected three words per POS per language, for a total of $3 \times 2 \times 5 = 30$ words resulting in a total of $30 \times 5 = 150$ contexts. The contexts were analyzed both quantitatively (see Table 3) and qualitatively.

As the prompt included the directive to be concise yet complete (see Section 3.5 for more details), we measure the average length of the generated sentences in characters and words. Since the model was given the possibility to refuse a generation if it deemed the word too difficult for the requested level, we also count how often the models return 'Too Difficult'. Table 3 shows that Gemini had the least amount of 'Too Difficult', and also the shortest sentences. GPT-4o, on the other hand, returned 'Too Difficult' frequently, and Claude generated the longest sentences overall. The generated contexts were also qualitatively evaluated by the authors to check for potential problems such as non-target language generation, after which Gemini 2.5 Flash was chosen for the remainder of the study.

3.5. Context Generation

In order to generate the contexts, we prompt the model to generate one sentence suitable for each CEFR level, or to respond with 'Too Difficult' if no sense of the word is understandable at the requested level (according to the model's own "understanding" of what should be understandable at what level). Table 4 presents the context generation for the English A2-level noun *column*. With three different senses present in the sentences ("vertical stone post", "vertical block of words", and "regular newspaper/magazine article by same author"), the example also illustrates the importance of context in relation to the complexity of one and the same vocabulary item, as discussed in Section 2.

The system prompt, to guide the model in its generation, is set as follows:

> You are an expert language tutor AI specialized in generating concise and level-appropriate example sentences.
>
> **Your Task:**
>
> 1. Receive a target **language**, a **proficiency level** (e.g., A1, B2, C1), and a target **word**.
> 2. Generate **exactly one** complete, simple, and natural sentence in the specified language that contains the target word.
> 3. The sentence *must be suitable and fully understandable* for a language learner at the given proficiency level.
> 4. **Crucially:** If the word, in any of its common meanings, is deemed too difficult or too rare for a learner at that specific proficiency level to use or understand in a simple sentence, your **only** response must be: 'Too Difficult'.
> 5. Strictly adhere to these formatting rules:
>   - **Do not** include any translations.
>   - **Do not** include any explanations, definitions, or grammatical notes.
>   - **Do not** use quotation marks around the generated sentence.
>   - The output must be **only** the generated sentence or the phrase 'Too Difficult'.

The prompt itself ("user prompt") is set as follows:

> Generate one simple and natural sentence in [LANGUAGE] suitable for a learner at proficiency level [LEVEL], containing the word [WORD] as a [PART OF SPEECH]. If no sense of the word can be understood at the given level, reply 'Too Difficult'. Do not include translations or explanations.

[6] In the CEFRLex resource for Dutch, for example, *beelden* ('images, sculptures') is wrongfully tagged as a verb.

Language	Level	Ge D	Ge C	Ge W	Cl D	Cl C	Cl W	G D	G C	G W
Dutch	A1	4	16.00	3.00	3	28.67	5.33	5	26.00	6.00
	A2	2	29.00	5.50	2	42.50	8.25	4	35.50	6.50
	B1	1	44.40	8.00	1	54.40	9.60	2	36.25	6.75
	B2	0	53.83	9.17	0	73.17	11.33	1	49.20	8.20
	C1	0	65.67	10.67	0	103.33	15.17	1	59.20	9.80
English	A1	2	18.25	3.50	2	26.00	5.50	4	22.50	5.00
	A2	1	23.60	4.80	0	38.00	7.33	3	27.67	5.67
	B1	0	33.33	6.83	0	49.00	9.17	1	40.60	8.20
	B2	0	40.83	7.67	0	70.17	12.83	2	49.25	8.75
	C1	0	69.17	11.17	0	81.67	13.00	2	76.50	12.50
French	A1	1	21.00	4.20	2	29.00	5.50	4	24.50	5.50
	A2	0	32.00	6.00	0	45.33	8.33	1	38.40	7.40
	B1	0	33.50	6.50	0	58.33	10.83	0	40.67	7.67
	B2	0	51.00	9.00	0	77.67	13.00	0	50.17	10.17
	C1	0	80.17	12.67	0	95.83	16.50	0	55.17	9.50
Spanish	A1	2	18.75	4.00	3	27.33	5.67	5	24.00	5.00
	A2	1	30.20	6.20	1	51.60	9.20	2	36.25	6.75
	B1	1	38.20	6.60	1	59.80	10.60	1	47.40	8.80
	B2	0	50.17	7.67	1	74.20	12.00	1	49.40	8.40
	C1	0	80.00	12.50	1	93.00	15.60	1	63.80	11.60
Swedish	A1	4	21.00	4.50	3	30.33	6.33	5	15.00	4.00
	A2	2	31.75	6.00	2	49.25	9.00	5	36.00	8.00
	B1	0	36.17	7.00	0	56.17	11.00	1	34.00	6.40
	B2	0	51.67	8.83	0	68.00	11.83	1	45.60	7.20
	C1	0	63.00	9.17	0	78.50	11.50	0	54.50	9.67
Total 'Too difficult'		21			22			52		
Average length			44.15	7.67		61.58	10.63		45.74	8.32

Table 3: Quantitative results of model selection process. Ge: Gemini, Cl: Claude, G: GPT. D: Number of times 'Too Difficult' was returned, C: Length of the generated sentence in characters, W: Length of the generated sentence in words.

Requested level	**Generated sentence context**
A1	Too Difficult
A2	The old building has many tall **columns**.
B1	Please read the first **column** of the newspaper.
B2	She writes a weekly **column** for the local newspaper.
C1	She writes a weekly **column** for the local newspaper.

Table 4: Example of context generation for English A2-level noun *column*.

3.6. Data Post-Processing

After the context generation step, we obtained a provisional dataset containing a total of 5,000 sentences (i.e. five in-context uses for each of the 1,000 selected vocabulary items). To arrive at the final dataset to be used for annotation by L2 learners (Section 3.7), three post-processing steps were performed: (1) all instances labeled as 'Too Difficult' were removed, (2) all duplicates were collapsed into one single entry, and (3) generation errors[7] were fixed.

For the example presented in Table 4, this means

[7] For six instances in the English subset (nouns: *tray* at requested level C1, *failure* at C1 and *cop* at B2; verbs *promise* at A1, *associate* at B2, and *poke* at B1), the model generated two sentences instead of one. Only the first sentence was retained.

that (1) the generated context for A1 was removed and (2) the duplicate generated contexts for B2 and C1 were merged into one single dataset entry. To be able to trace back that this sentence was generated for two different levels, we assign it the new label "B2-C1". After completing the post-processing steps (performed automatically by means of a programming script), we arrived at a final dataset consisting of 3,901 instances. As the 'Too Difficult' instances contain valuable information for posterior analyses, we make available the unfiltered version of the dataset in the repository as well. Using this version of the dataset, it is possible to gain deeper insights into the "behavior" of LLMs towards CEFR levels, for example by analyzing in which particular cases the model returned 'Too Difficult'.

3.7. Data Annotation

However, for the main envisaged use of our dataset (i.e. lexical complexity prediction or LCP, see also Section 1 and 2), we still need to perform one final step: human annotation. In previous LCP studies, these annotations were usually gathered by instructing annotators to label data instances (usually a given vocabulary item presented in a sentence) on a 1 to 5 scale (with 1 being "very easy" and 5 being "very difficult"). Recent studies, however, have highlighted the benefits of using comparative judgment methods instead of rating scales (Kiritchenko and Mohammad, 2017; Alfter et al., 2021, 2022). In our study, we follow this line of research by using the technique of Best-Worst Scaling (BWS; Louviere et al., 2015), which requires participants to choose the *best* and *worst* item from a set of *n* items. Based on BWS annotations, it is possible to obtain (1) a ranking of the target items (in our case, from easiest to most difficult) and (2) a numerical score reflecting the annotated concept (in our case, the concept of lexical complexity).

Following the procedure outlined in Kiritchenko and Mohammad (2017), we converted – for nouns and verbs separately – each "language subset" of the final dataset into a set of 4-tuples (i.e. sets of four different sentence contexts, see Figure 1 for an example). To this end, all sentence contexts were – for nouns and verbs separately – randomly shuffled and assigned to different 4-tuples. The total number of 4-tuples was determined as 1.5 times the number of available sentence instances per language–POS combination, resulting in each target sentence occurring in six different 4-tuples. To achieve a balanced and representative set, 1,000 iterations[8] were performed over each language–POS subset to (1) cover as many unique sentence pairs as possible (measured by standard deviation; the lower the better) and (2) obtain a distribution as uniform as possible across all possible sentence label combinations in the 4-tuples (measured by entropy; the higher the better).

At the time of writing, we are collecting annotations from 20 L2 learners (four per language), corresponding to a projected total of 23,408 annotations (see Table 5 for the full statistics on `Conplext`). Participants are required to indicate the in-context word they find easiest (*best*) and most difficult (*worst*) for all 4-tuples in their language-specific subset. Annotations are gathered through an in-house online environment specifically built for comparative judgment experiments. The full instructions given to the L2 learners are formulated as follows (after these instructions, a tuple as shown in Figure 1 is displayed):

> **Read these instructions carefully. They will remain displayed with every instance of the task, but you do not have to read them again every time.**
>
> Below you will find a series of four vocabulary items, each of them accompanied by an example sentence illustrating the meaning of the vocabulary item. Indicate which item you find the easiest ('best') and which item you find the most difficult ('worst'). Make sure you base your judgment on **the sense the word has in the example sentence**. Go to the next item by clicking 'Next'. Your progress is saved automatically.

As soon as all annotations are collected, we will – by LCP convention – convert them into numerical scores on a continuous scale ranging from 0 to 1, using methods such as the Rescorla-Wagner model (Rescorla and Wagner, 1972) and the counting procedure (Flynn and Marley, 2014). The full results of the BWS annotation experiment will be presented and analyzed in a follow-up paper.

4. Discussion

As emphasized in previous research (Degraeuwe, 2025; Tack, 2021), the creation of relevant vocabulary learning activities for L2 learners depends to a large extent on the successful identification of difficult words. Evidently, it is impossible for L2 teachers to perform this identification process themselves if they are guiding groups of tens or hundreds of students. Rule-based methods that automatically consult computer-readable resources in which words are linked to difficulty levels provide a possible solution to this problem (Finlayson et al., 2023; Van Parys et al., 2025), but these methods

[8] The detailed results per iteration are available in the "supplementaryData" folder of the dataset repository on GitHub.

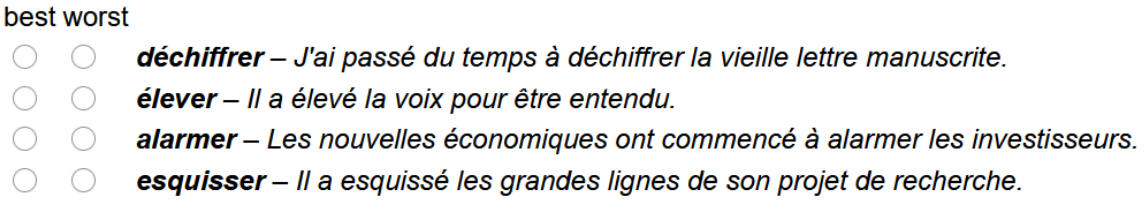

Figure 1: Example of a 4-tuple used in the BWS experiment (French subset; NOUN as POS). Translations to English are provided in Appendix B.

	#words	#sentences	#tuples	#annotations
EN	200	816	1,224	4,896
ES	200	762	1,143	4,572
FR	200	718	1,077	4,308
NL	200	795	1,193	4,772
SV	200	810	1,215	4,860
Total	1,000	3,901	5,852	23,408

Table 5: Statistics on Conplext: number of vocabulary items, filtered sentences, 4-tuples, and (projected) number of annotations.

come with one major disadvantage, in that they can only assign a difficulty label to words included in the resources. To overcome this limitation, more advanced systems using machine learning techniques can be designed: these systems learn to generalize over the training data and can, theoretically, classify any type of textual input into a given set of difficulty levels.

It is for this specific purpose that the annotations to be obtained from the BWS experiment (Section 3.7) will be particularly useful: they provide the necessary, relevant, and labeled training data to develop context-aware word difficulty classifiers that are specifically tailored to L2 learners. In turn, these classifiers can be integrated into proper educational applications, such as computer-assisted language learning environments or intelligent language tutoring systems. Furthermore, the classifiers can be used to customize lexical simplification pipelines according to the specific needs of L2 learners.

While the development of context-aware word difficulty classifiers constitutes its main envisaged use, we believe that the Conplext dataset can also serve other purposes. First, as we generated multiple sentences for each vocabulary item in order to factor in the importance of polysemy and context (Section 3.5), the dataset constitutes a new valuable resource for word sense disambiguation (WSD) studies targeting L2 learners. In a future study, we plan to enrich Conplext by adding sense labels to the sentences, which will render the dataset even more relevant for WSD purposes.

Second, as we also release the unfiltered version of the dataset as a part of the repository (Section 3.6), Conplext can be used to gain deeper insights into the capabilities of generative artificial intelligence models to "think" in terms of CEFR levels. Possible types of analyses that can shed light on this matter include (1) evaluating for which vocabulary items the model returned 'Too Difficult' and (2) studying which differences (if any) can be identified across the generated contexts of the five CEFR levels (e.g., in terms of lexis, semantics, and word frequency).

5. Conclusion

We present Conplext, a new type of resource specifically aimed at lexical complexity prediction for language learning purposes. Our resource bridges a gap in current research by (1) providing words in multiple contexts to potentially allow for the learning of multiple complexity values depending on the context, (2) covering five languages, and (3) being manually annotated by language learners of the respective languages. This design allows for a more nuanced representation of lexical difficulty across diverse linguistic settings.

At the time of writing, the large-scale best–worst scaling annotation process with L2 learners is ongoing. The resulting data will provide empirically grounded complexity values that reflect learner perceptions across languages and proficiency levels, enabling a more reliable evaluation of lexical difficulty models once complete. This ongoing effort also ensures that the dataset will scale in both size and representativeness over time.

This dataset establishes a foundation for developing context-aware lexical difficulty models and evaluating multilingual complexity prediction in educational natural language processing. By combining large language model output with empirically grounded learner data, Conplext contributes a reproducible and extensible benchmark for advancing research in lexical complexity and second language acquisition. It further supports future work on adaptive learning systems that rely on fine-grained difficulty estimation.

Limitations

Our study relies on a specific model (Gemini 2.5 Flash) to generate contexts. This may introduce model-specific stylistic choices and biases.

As pointed out in the paper, models may have limited understanding of the CEFR scale. However, as we also conduct a large-scale annotation with language learners, we believe this risk is mitigated, as the final dataset will draw its difficulty assignments from the human-annotated data.

Although our resource aims at providing more contexts per word, the number of contexts is still limited, mainly due to the large-scale annotation effort required. We plan on extending the number of contexts in subsequent work.

Despite presenting a multilingual perspective, we acknowledge the Eurocentric nature of the languages involved.

Finally, generating contexts using large language models and the follow-up large-scale human annotation are costly both in terms of time and money. These factors might hinder the extension of our approach to other languages. It is also for this reason that we have not experimented with different prompting strategies, or executing the same prompt multiple times to test for stability of the generated responses.

Ethical Considerations

The contexts were automatically generated. This fact is explicitly stated to prevent any misunderstanding or inappropriate comparison with authentic, human-produced material.

While models nowadays are carefully adjusted to avoid producing or perpetuating biases and prejudices, we cannot exclude that such biases or prejudices might still be present in the final resource. Any such cases identified during evaluation or use will be promptly reviewed and, where appropriate, removed from the resource.

For the annotation process, no personal or identifying information is collected. All responses are anonymous.

6. Bibliographical References

David Alfter. 2021. *Exploring natural language processing for single-word and multi-word lexical complexity from a second language learner perspective*. Ph.D. thesis, University of Gothenburg, Sweden.

David Alfter. 2025. The need for truly graded lexical complexity prediction. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 326–333, Vienna, Austria. Association for Computational Linguistics.

David Alfter, Yuri Bizzoni, Anders Agebjörn, Elena Volodina, and Ildikó Pilán. 2016. From Distributions to Labels: A Lexical Proficiency Analysis using Learner Corpora. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC, Umeå, 16th November 2016*, 130, pages 1–7. Linköping University Electronic Press.

David Alfter, Rémi Cardon, and Thomas François. 2022. A dictionary-based study of word sense difficulty. In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*, pages 17–24.

David Alfter and Johannes Graën. 2019. Interconnecting lexical resources and word alignment: How do learners get on with particle verbs? In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 321–326.

David Alfter, Therese Lindström Tiedemann, and Elena Volodina. 2021. Crowdsourcing relative rankings of multi-word expressions: Experts versus non-experts. In *Northern European Journal of Language Technology, Volume 7*.

David Alfter and Elena Volodina. 2018. Towards Single Word Lexical Complexity Prediction. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 79–88.

Fernando Alva-Manchego, Regina Stodden, Joseph Marvin Imperial, Abdullah Barayan, Kai North, and Harish Tayyar Madabushi. 2025. Findings of the TSAR 2025 shared task on readability-controlled text simplification. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility, and Readability (TSAR 2025)*, Suzhou, China. Association for Computational Linguistics.

Drilon Avdiu, Vanessa Bui, Klára Ptacinová Klimci, et al. 2019. Predicting learner knowledge of individual words using machine learning. In *Proceedings of the 8th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2019), September 30, Turku Finland*, 164, pages 1–9. Linköping University Electronic Press.

Luca Benedetto, Gabrielle Gaudeau, Andrew Caines, and Paula Buttery. 2025. Assessing how

accurately large language models encode and apply the common European framework of reference for languages. *Computers and Education: Artificial Intelligence*, 8:100353.

Jill Burstein, Nitin Madnani, John Sabatini, Dan McCaffrey, Kietha Biggers, and Kelsey Dreier. 2017. Generating Language Activities in Real-Time for English Learners using Language Muse. In *Proceedings of the Fourth (2017) ACM Conference on Learning@ Scale*, pages 213–215. ACM.

Jin-Woo Chung, Hye-Jin Min, Joonyeob Kim, and Jong C Park. 2013. Enhancing readability of web documents by text augmentation for deaf people. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, pages 1–10.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Press Syndicate of the University of Cambridge.

Council of Europe. 2018. Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors. Accessed 09.03.2019 from www.coe.int/lang-cefr.

Jan De Belder, Koen Deschacht, and Marie-Francine Moens. 2010. Lexical simplification. In *Proceedings of ITEC2010 : 1st International Conference on Interdisciplinary Research on Technology, Education and Communication*.

Jasper Degraeuwe. 2025. You Shall Know a Word's Difficulty by the Family It Keeps: Word Family Features in Personalised Word Difficulty Classifiers for L2 Spanish. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 312–325, Vienna, Austria. Association for Computational Linguistics.

Jasper Degraeuwe and Patrick Goethals. 2024. Leading by example: The use of generative artificial intelligence to create pedagogically suitable example sentences. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 33–48, Rennes, France. LiU Electronic Press.

Louise Deléger and Pierre Zweigenbaum. 2009. Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora (BUCC)*, pages 2–10.

Siobhan Devlin. 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic databases*.

Yo Ehara, Issei Sato, Hidekazu Oiwa, and Hiroshi Nakagawa. 2018. Mining Words in the Minds of Second Language Learners for Learner-specific Word Difficulty. *Journal of Information Processing*, 26:267–275.

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.

T.N. Flynn and A.A.J. Marley. 2014. Best-worst scaling: theory and methods. In Stephane Hess and Andrew Daly, editors, *Handbook of Choice Modelling*. Edward Elgar Publishing.

Núria Gala, Thomas François, Delphine Bernhard, and Cédrick Fairon. 2014. Un modèle pour prédire la complexité lexicale et graduer les mots. In *TALN 2014*, pages 91–102.

Núria Gala, Thomas François, and Cédrick Fairon. 2013. Towards a French lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. *E-lexicography in the 21st century: thinking outside the paper., Tallin, Estonia*.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.

Lynn M LoPucki. 2014. System and method for enhancing comprehension and readability of legal text. US Patent 8,794,972.

Jordan J Louviere, Terry N Flynn, and Anthony Alfred John Marley. 2015. *Best-worst scaling: Theory, methods and applications*. Cambridge University Press.

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

Gustavo Paetzold and Lucia Specia. 2016a. Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569.

Gustavo Paetzold and Lucia Specia. 2016b. Sv000gg at semeval-2016 task 11: Heavy gauge complex word identification with system voting. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 969–974.

Sarah E Petersen and Mari Ostendorf. 2007. Text simplification for language learners: a corpus analysis. In *Workshop on Speech and Language Technology in Education*.

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.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word Sense Disambiguation: A Unified Evaluation Framework and Empirical Comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain. Association for Computational Linguistics.

Robert A. Rescorla and Allan R. Wagner. 1972. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A.H. Black and W.F. Prokasy, editors, *Classical Conditioning II*, pages 64–99. Appleton-Century-Crofts, New York.

Irina Rets and Jekaterina Rogaten. 2021. To simplify or not? Facilitating English L2 users' comprehension and processing of open educational resources in English using text simplification. *Journal of Computer Assisted Learning*, 37(3):705–717.

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

Matthew Shardlow, Fernando Alva-Manchego, Riza Theresa Batista-Navarro, Stefan Bott, Saul Calderon-Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, and Anna Huelsing. 2024. The BEA 2024 Shared Task on the Multilingual Lexical Simplification Pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589.

Matthew Shardlow, Richard Evans, Gustavo 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.

Anaïs Tack, Thomas François, Piet Desmet, and Cédrick Fairon. 2018. NT2Lex: A CEFR-Graded Lexical Resource for Dutch as a Foreign Language Linked to Open Dutch WordNet. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 137–146.

Anaïs Tack, Thomas François, Anne-Laure Ligozat, and Cédrick Fairon. 2016. Evaluating Lexical Simplification and Vocabulary Knowledge for Learners of French: Possibilities of Using the FLELex Resource. In *LREC*.

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.

Amaury Van Parys, Vanessa De Wilde, Lieve Macken, and Maribel Montero Perez. 2025. LexPro: A plurilingual lexical profiling tool to assist teachers and researchers in analysing vocabulary of L2 input. *Language Teaching Research*, page 13621688251352259.

Kevin Yancey and Yves Lepage. 2018. Korean L2 Vocabulary Prediction: Can a Large Annotated Corpus be Used to Train Better Models for Predicting Unknown Words? In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

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, United States. Association for Computational Linguistics.

7. Language Resource References

Dürlich, Luise and François, Thomas. 2018. *EFLLex: A graded lexical resource for learners of English as a foreign language*.

Francois, Thomas and Volodina, Elena and Pilán, Ildikó and Tack, Anaïs. 2016. *SVALex*. ISLRN 854-377-992-687-3.

Thomas François and Barbara De Cock. 2018. *ELELex: a CEFR-graded lexical resource for Spanish as a foreign language.* PID http://hdl.handle.net/2078.1/204347.

Anaïs Tack and Thomas Francois and Anne-Laure Ligozat and Cédrick Fairon. 2016. *FLELex*. ISLRN 742-240-876-017-1.

Tack, Anaïs and François, Thomas and Desmet, Piet and Fairon, Cédrick. 2018. *NT2Lex: A CEFR-Graded Lexical Resource for Dutch as a Foreign Language Linked to Open Dutch WordNet*. Association for Computational Linguistics.

A. Additional Information on CEFRLex Data Used

The additional information on the CEFRLex data used in the study can be found in Table 6.

Language	CEFRLex file used	Mapping CEFRLex tags to UD tags
EN	`EFLLex_NLP4J.tsv`	`{"NN": "NOUN", "VB": "VERB", "JJ": "ADJ", "RB": "ADV"}`
ES	`ELELex_Freeling.tsv`	`{"NCM": "NOUN", "NCF": "NOUN", "NCC": "NOUN", "VM": "VERB", "AQ0": "ADJ", "RG": "ADV"}`
FR	`FLELex_TreeTagger.tsv`	`{"NOM": "NOUN", "VER": "VERB", "ADJ": "ADJ", "ADV": "ADV"}`
NL	`NT2Lex_Frog-CGN.tsv`	`{"N(soort)": "NOUN", "WW()": "VERB", "ADJ()": "ADJ", "BW()": "ADV"}`
SV	`SVALex_Korp.tsv`	`{"NN_NEU": "NOUN", "NN_UTR": "NOUN", "VB": "VERB", "JJ": "ADJ", "AB": "ADV"}`

Table 6: Additional details on the CEFRLex data used in the study: specific file and dictionary mapping CEFRLex tags to Universal Dependencies (UD) tags.

B. English Translations

The translation of the BWS tuple presented as an illustration in Figure 1 can be found in Table 7.

best	worst	
○	○	**decipher** - I spent time deciphering the old handwritten letter
○	○	**raise** - He raised his voice to be heard
○	○	**alarm** - The economic news began to alarm investors
○	○	**outline** - He outlines the main points of his research project

Table 7: Translation of the example in Figure 1.