CEFR-Annotated WordNet: LLM-Based Proficiency-Guided Semantic Database for Language Learning

Masato Kikuchi[♢✉], **Masatsugu Ono**[♡], **Toshioki Soga**[☆], **Tetsu Tanabe**[♣], **Tadachika Ozono**[♢]

[♢]Nagoya Institute of Technology, [♡]Kitami Institute of Technology,
[☆]Chitose Institute of Science and Technology, [♣]Hokkaido University
{kikuchi, ozono}@nitech.ac.jp, onomasa@mail.kitami-it.ac.jp,
t-soga@photon.chitose.ac.jp, ttanabe@iic.hokudai.ac.jp

Abstract

Although WordNet is a valuable resource because of its structured semantic networks and extensive vocabulary, its fine-grained sense distinctions can be challenging for second-language learners. To address this issue, we developed a version of WordNet annotated with the Common European Framework of Reference for Languages (CEFR), integrating its semantic networks with language-proficiency levels. We automated this process using a large language model to measure the semantic similarity between sense definitions in WordNet and entries in the English Vocabulary Profile Online. To validate our approach, we constructed a large-scale corpus containing both sense and CEFR-level information from the annotated WordNet and used it to develop contextual lexical classifiers. Our experiments demonstrate that models fine-tuned on this corpus perform comparably to those fine-tuned on gold-standard annotations. Furthermore, by combining this corpus with the gold-standard data, we developed a practical classifier that achieves a Macro-F1 score of 0.81. This result provides indirect evidence that the transferred labels are largely consistent with the gold-standard levels. The annotated WordNet, corpus, and classifiers are publicly available to help bridge the gap between natural language processing and language education, thereby facilitating more effective and efficient language learning.

Keywords: WordNet, CEFR Level, Language Learning, Word Sense, Corpus

1. Introduction

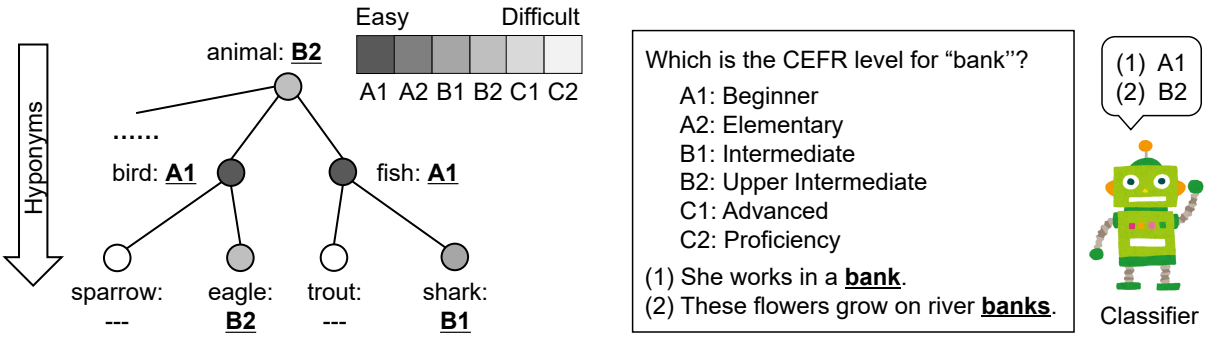
WordNet (Fellbaum, 1998) is a large-scale English lexical database that organizes approximately 155,000 words and 207,000 senses of nouns, verbs, adjectives, and adverbs into hierarchical semantic networks. It groups semantically similar words and links senses through relations such as hypernymy, hyponymy, synonymy, and antonymy. Because WordNet and its construction software are publicly available, they can be readily integrated into AI applications. Consequently, WordNet underpins a broad range of natural language processing (NLP) technologies—including machine translation (Moussallem et al., 2018), semantic analysis (Moskvoretskii et al., 2024), and natural language generation (San Vicente et al., 2014)—owing to its accessible interface and well-structured networks. These technologies also support computer-assisted language learning (CALL) by facilitating vocabulary acquisition, reading comprehension, writing assistance, automated question generation, and automated assessment.

Although leveraging semantic networks can enhance foreign-language learning (Kiritani et al., 2012), WordNet was not designed for educational purposes and presents challenges for second-language (L2) learners. Key issues are its overly fine-grained sense distinctions and the large number of senses associated with many words. This requires learners to identify the appropriate sense for a given context and proficiency level, which increases their cognitive load. While this problem is widely discussed in NLP literature (Navigli, 2006)(Lacerra et al., 2020), it has received limited attention in language education. Our goal was to develop a novel version of WordNet and leverage the resulting technologies and resources to enhance language-learning efficiency. The first step involves adapting WordNet for L2 learners and bridging the gap between NLP lexical resources and language education. Previous work (Kikuchi et al., 2024) clustered fine-grained WordNet sense definitions (glosses) using learner-oriented dictionaries. By contrast, this study integrates Common European Framework of Reference for Languages (CEFR) proficiency levels into WordNet, enabling the presentation of senses aligned with a learner's proficiency level. To build large-scale, practical resources, we employ a simple large language model (LLM)-based method for efficient and accurate semantic annotation, reducing the time, labor, and cost associated with manual annotation. This automatic approach also ensures that the adapted WordNet can be flexibly scaled.

The CEFR is an international standard for describing language proficiency across six levels, namely, A1, A2, B1, B2, C1, and C2, ranging from basic to advanced. Each level is defined by "can-

✉ Corresponding author.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3645–3661
11–16 May 2026.



(a) CEFR-Annotated WordNet.

(b) Contextual Lexical CEFR-Level Classifier.

Figure 1: Overview of the study. (a) Semantic network of hyponyms for "animal" in the CEFR-annotated WordNet. (b) Contextual CEFR-level classification for the word "bank."

do descriptors" that specify expected communicative abilities. We used an LLM to annotate WordNet senses with CEFR levels, thereby constructing a CEFR-annotated WordNet. As shown in Figure 1(a), these annotations can be used in conjunction with semantic networks to help learners acquire vocabulary while considering relationships among words and learn basic and advanced paraphrases through synonyms. The annotation pipeline involves three steps. First, we collect glosses for target words from WordNet and the English Vocabulary Profile (EVP) Online[1] (Capel, 2012), which provides CEFR levels for individual senses. Next, the LLM computes the semantic similarity between the glosses from WordNet and the EVP. Finally, we assign CEFR levels to the corresponding WordNet senses based on these similarity scores.

Because our method for annotating WordNet senses with CEFR levels is automatic, it eliminates the need for labor-intensive manual work. However, automatic labels may contain errors, and WordNet does not provide gold-standard CEFR levels. Therefore, their reliability must be verified indirectly. To address this issue, we built contextual CEFR-level classifiers that predict a sense's proficiency level from its usage, as shown in Figure 1(b). These classifiers predict the CEFR level for a word sense based on its context, not just for the word itself. We evaluate the quality of our annotations by comparing classifiers trained on our data with those trained on the EVP gold-standard levels. We also examine the effectiveness of LLMs for this task in few-shot and fine-tuning settings.

The contributions of this study can be summarized as follows:

- **CEFR-Annotated WordNet.** We developed a new resource by assigning CEFR proficiency levels to 10,644 WordNet senses corresponding to 5,645 lemmas, thereby linking the WordNet sense inventory with CEFR standards. Our annotated WordNet covers approximately 80% of all single-word senses in the EVP (8,289 out of 10,394).

- **Prompt-Only LLM Annotation.** We introduce a novel, automated method that leverages the semantic understanding of LLMs to assign CEFR levels to word senses. This is achieved by measuring semantic similarity between WordNet glosses and EVP entries. The method, implemented entirely through prompting, is simple, reproducible, and substantially less costly than manual annotation. We also provide indirect evidence that manual annotation tasks based on semantic matching can be automated with high accuracy.

- **SemCor-CEFR Corpus.** Using our annotated WordNet, we assigned CEFR levels to the word senses in SemCor 3.0[2] (Miller et al., 1993), a widely used sense-annotated corpus. This process resulted in a large-scale corpus containing more than 110,000 sense and level annotations across over 5,500 WordNet senses. Because modern NLP relies on large corpora for advanced training and analysis, this resource represents a valuable contribution to both NLP and educational technology research.

- **Contextual Lexical CEFR-Level Classifiers.** We demonstrate the validity of our CEFR-level annotations by training a classifier on our corpus that performs comparably to one trained on gold-standard EVP data. Additionally, by fine-tuning the LLM on both our annotated data and the gold-standard levels, we developed a practical classifier that achieves a Macro-F1 score of 0.81. Our analysis indicates that

[1] https://englishprofile.org/?menu=evp-online

[2] http://lcl.uniroma1.it/wsdeval/training-data

these classifiers can accurately predict CEFR levels across a broad range of contexts.

All resources developed in this study, including our WordNet, corpus, and classifier, are publicly available at https://doi.org/10.5281/zenodo.17395388.

2. Related Work

2.1. WordNets for Language Learning

As noted in the introduction, WordNet was not originally designed for educational use. To address this limitation, several learning-oriented WordNets have been developed for multiple languages (Bosch and Griesel, 2018), and their application in language learning has been studied by many researchers (Gonzalez-Dios, 2019). Some of these researchers have focused on visualizing word hierarchies and semantic relations to support learners (Sun et al., 2011; Kiritani et al., 2012; Gawde et al., 2024), while others have tailored vocabulary, glosses, and usage examples to match learners' proficiency levels (Redkar et al., 2018; Osenova and Simov, 2024). However, most of these studies have relied on manually curated resources and paid limited attention to word-sense information. By contrast, herein we introduce a novel approach that automatically annotates WordNet senses with proficiency levels. Our method can be integrated with existing techniques—such as semantic network visualization and multimodal WordNets (Marciniak, 2020)—to further enhance its utility in language-learning contexts.

2.2. Lexical Complexity Prediction

Lexical complexity prediction (LCP) (North et al., 2023; Shardlow et al., 2024) has recently attracted significant attention as a task for estimating word complexity from context. In this field, "complexity"—which is related to the CEFR levels in our study—is typically predicted as either binary (e.g., simple/complex) or on a continuous scale. Our work is closely related to SemEval-2021 Task 1 (Shardlow et al., 2021), which adopts a similar classification setting. For this task, the organizers released the CompLex 2.0 dataset[3] (Shardlow et al., 2022), in which words in context were rated by multiple annotators on a five-point Likert scale. The final scores are represented as a continuous value in $[0, 1]$, computed as the mean of these ratings. These continuous values can capture finer, context-driven differences compared with ordinal labels.

However, our approach differs from that of LCP in several key ways.

First, LCP is primarily designed as a precursor to lexical simplification (Paetzold and Specia, 2017)—which involves replacing complex words with simpler alternatives—rather than for explicitly presenting complexity information to L2 learners. Second, annotators of CompLex 2.0 were not provided with glosses; as a result, identical senses may receive different scores across contexts. Third, the dataset is limited to 9,000 nouns and excludes other parts of speech (PoS).

By contrast, our method assigns a CEFR level to each word sense in accordance with an international proficiency standard. We extend this annotation to more than 110,000 instances of nouns, verbs, adjectives, and adverbs in the large-scale SemCor corpus, making our resource more than ten times larger than CompLex 2.0. In Section 6.1, we analyze the correlation between the CompLex 2.0 complexity scores and the CEFR levels predicted by our models.

2.3. CEFR-Based Educational Technology

The CEFR is a foundational standard in educational technology and is widely applied in the automatic assessment of short sentences (Tack et al., 2017; Uchida et al., 2024), teaching materials (Pilán et al., 2016), writing skills (Kerz et al., 2021; Schmalz and Brutti, 2021), and learner proficiency (Gaillat et al., 2022). The interaction between LLMs and the CEFR has been the focus of recent studies, which have explored how well these models understand proficiency levels (Benedetto et al., 2025) and how to control the difficulty of the vocabulary and sentences they generate (Alfter, 2024; Malik et al., 2024; Barayan et al., 2025). These studies, together with the development of numerous CEFR-aligned lexical datasets, underscore the central role of the CEFR in the field.

For example, the CEFRLex project provides machine-readable lexical resources with word-level frequency counts by CEFR level (Pintard and François, 2020) for English and other languages[4] (Dürlich and François, 2018; François et al., 2014; Tack et al., 2018; François et al., 2016; Volodina et al., 2016). However, it does not assign a unique CEFR level to each sense, making it unsuitable for tasks requiring sense-specific proficiency annotations. The Sense Complexity Dataset (SeCoDa) (Strohmaier et al., 2020) provides sense-in-context CEFR annotations, but its sense labels are not aligned with those of WordNet. In addition, its small size, 1,432 words, limits its applicability within WordNet's semantic framework. Our work addresses these gaps by annotating more than

[3] https://github.com/MMU-TDMLab/CompLex

[4] Only the Dutch resource NT2Lex provides frequency information at the word-sense level.

110,000 word instances with sense-specific CEFR levels, thereby substantially expanding the available data on lexical difficulty.

Despite progress in LCP, few studies have examined the classification of vocabulary into CEFR levels based on context. Aleksandrova and Pouliot (2023) proposed ME6 Contextual, a BERT (Devlin et al., 2019)-based classifier that, like our models, is trained on a CEFR-annotated corpus to directly predict a word's level from its context. This direct prediction approach enables the model to classify words not seen during training. By contrast, Bannò et al. (2025) introduced an indirect-prediction method in which an LLM selects the appropriate EVP sense for a word in context and then maps it to a CEFR level. The performance of this approach depends on the quality of cues provided by the data source. To isolate the effect of different data sources on classification performance, we reimplemented ME6 Contextual as a baseline for our LLM-based classifiers.

3. Existing Resources

3.1. EVP Online

The EVP Online[1], developed by Cambridge University Press, provides CEFR levels for single words, phrasal verbs, phrases, and idioms. Each entry includes a PoS tag, a gloss, and both dictionary and learner examples. A key feature of the EVP is its sense-level CEFR annotation, which assigns a proficiency level to each sense. This level of granularity is beneficial for both general language education and the development of CALL systems. For this study, we used single-word entries from the American English subset of the EVP, including their CEFR levels, PoS tags, glosses, and example sentences. Unlike a previous study (Aleksandrova and Pouliot, 2023), which also included multiword expressions (MWEs), our work focuses exclusively on single words because WordNet contains very few MWEs.

3.2. SemCor Corpus

The SemCor 3.0 corpus is one of the most widely used sense-annotated corpora in NLP. It contains 226,040 sense annotations across 352 documents from the Brown Corpus. Each sense is tagged with a WordNet identifier, linking it to glosses, usage examples, and semantic relations such as hypernyms, hyponyms, synonyms, and antonyms. Moreover, its machine-readable format facilitates integration into NLP and CALL systems. However, this corpus inherits the limitations of WordNet discussed in the introduction. In particular, the absence of learner-oriented indicators, such as sense complexity or CEFR levels, limits its usefulness for educational applications. To address this issue, as described in Section 5.1, we use our CEFR-annotated WordNet to enhance the original SemCor corpus, creating a new resource annotated with both senses and CEFR levels.

4. CEFR-Annotated WordNet

To create a WordNet oriented toward L2 learners, we annotated its senses with CEFR levels by aligning them with glosses from the EVP Online. The process, illustrated in Figure 2 for the WordNet gloss g'_j of $\langle$word, PoS$\rangle = \langle$bank, noun$\rangle$, comprises three steps:

Step 1: Extraction of Gloss Sets. For each word and PoS pair, such as $\langle$bank, noun$\rangle$, we extract all corresponding glosses from both the EVP Online and WordNet. Let the set of m glosses from the EVP Online be $\{g_1, g_2, \ldots, g_m\}$, and that of n glosses from WordNet be $\{g'_1, g'_2, \ldots, g'_n\}$. In the next step, we compare the i-th gloss, g_i, from the EVP Online with the j-th gloss, g'_j, from WordNet. As shown in Figure 2, both example glosses refer to sloping land.

Step 2: Semantic Similarity Measurement. To measure the semantic similarity between g_i and g'_j, we used an LLM (GPT-4o, checkpoint `gpt-4o-2024-08-06`). The prompt is provided in Appendix 13.1. Because glosses from different resources often vary in granularity and may not align perfectly, a binary alignment judgment of same or different would be overly restrictive. Therefore, we instructed the LLM to rate similarity on a seven-point scale, where a lower score indicates higher similarity. In the example shown in Figure 2, the LLM returns a score of 1, indicating that the two glosses have identical meanings.

Step 3: CEFR-Level Annotation. If the LLM returns a score of 1 or 2—indicating that g_i and g'_j have "exactly the same" or "almost the same" meaning—we consider the glosses semantically aligned. The CEFR level associated with g_i is then transferred to g'_j. Otherwise, that is, if the output is ≥ 3, the senses are considered mismatched and no annotation is assigned. In the example, the score of 1 results in the WordNet sense being assigned the B2 level from the corresponding EVP sense. In this study, we adopt a threshold of ≤ 2 (scores 1–2) to balance accuracy and coverage. Restricting alignments to score 1 yields higher-confidence transfers but reduces coverage, whereas allowing score 3 increases coverage but tends to introduce false alignments, mainly because of partial semantic overlap and mismatched gloss granularity across

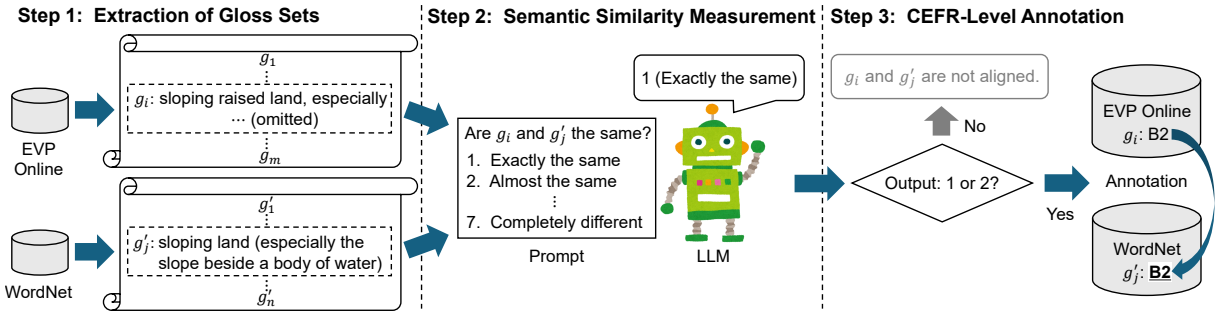


Figure 2: CEFR-level annotation process for a WordNet gloss (g'_j), illustrated with the noun "bank."

PoS	CEFR Levels						Total	Share (%)
	A1	A2	B1	B2	C1	C2		
Noun	310	626	1,021	1,426	652	853	4,888	44.46
Verb	213	263	701	948	443	595	3,163	28.77
Adjective	104	200	435	646	423	519	2,327	21.16
Adverb	40	94	127	201	92	63	617	5.61
Total	667	1,183	2,284	3,221	1,610	2,030	10,995	100.00
Share (%)	6.07	10.76	20.77	29.30	14.64	18.46	100.00	

Table 1: Distribution of senses in the CEFR-annotated WordNet by PoS and CEFR level. Share (%) indicates proportions by PoS (right) and level (bottom). Note that some senses received multiple levels because of differences in gloss granularity between the resources (see Appendix 13.2 for details).

resources. Therefore, we adopt ≤ 2 as a conservative compromise.

We applied this procedure to all gloss pairs for every ⟨word, PoS⟩ entry found in both the EVP Online and WordNet. For instance, the set of all possible gloss pairs of ⟨bank, noun⟩ is

$$\{(g,\ g') \mid g \in \{g_1,\ g_2,\ \ldots,\ g_m\},\ g' \in \{g'_1,\ g'_2,\ \ldots,\ g'_n\}\},$$

whose size is $m \times n$. This exhaustive process yielded 10,995 CEFR-level annotations for 10,644 WordNet senses across 5,645 lemmas. Table 1 lists the distribution of these annotations. Nouns constitute the largest share (4,888; 44.46%), followed by verbs (3,163; 28.77%) and adjectives (2,327; 21.16%), with adverbs comparatively scarce (617; 5.61%). The distribution across CEFR levels is concentrated in the intermediate range, with B2 (29.30%) and B1 (20.77%) together accounting for half of all annotations. The beginner levels (A1–A2) and advanced levels (C1–C2) represent 16.83% and 33.10%, respectively. Because the granularity of glosses differs between the two resources, a single WordNet sense may align with multiple EVP glosses, which can result in a sense being assigned multiple CEFR levels (see Appendix 13.2 for details). This automated procedure is generalizable and can be applied to other dictionaries or lexical databases that provide glosses. Moreover, because the pipeline relies solely on semantic similarity between annotations, it is transferable to lexical resources in other languages. However, as the process is fully automated, evaluating the accuracy of the resulting annotations is essential.

5. Experiments

To verify the accuracy of our CEFR-level annotations, we built and evaluated several contextual lexical CEFR-level classifiers (Figure 1(b)). The goal was to assess how well models trained on our automatically annotated data could predict gold-standard CEFR levels. We also trained several LLM-based classifiers to evaluate their effectiveness for this task. In this study, we did not conduct a manual spot-check of the transferred sense-level labels. Although expert validation would strengthen the reliability of the resource, rigorous CEFR-level judgment typically requires carefully designed protocols and multiple trained raters. We therefore leave such manual evaluation to future work.

5.1. Datasets and Experimental Settings

To train our classifiers, we required a corpus with CEFR-level annotations for words in context. Because the original SemCor corpus does not include this information, we created the "SemCor-CEFR

CEFR	EVP Online		SemCor-CEFR	
Level	# types	# words	# types	# words
A1	577	2,932	403	31,093
A2	1,037	4,307	680	21,065
B1	1,760	7,174	1,206	28,707
B2	2,368	8,754	1,684	23,081
C1	1,419	3,791	849	6,701
C2	1,692	4,604	992	5,647
Total	8,853	31,562	5,814	116,294

Table 2: Distribution of word types and tokens by CEFR level in the EVP Online and the SemCor-CEFR corpus.

corpus" by assigning CEFR levels to its senses using our annotated WordNet. Table 2 summarizes the word distributions in the EVP Online examples, combining dictionary and learner examples, and in the SemCor-CEFR corpus. Although our corpus contains fewer word types (# types) than the EVP examples, it includes substantially more word instances (# words) and reflects a more natural and imbalanced distribution of proficiency levels. For our experiments, we used 10% of the EVP examples as the test set. For training and validation, depending on the setting, we used (i) the remaining 90% of the EVP data, (ii) the SemCor-CEFR corpus, or (iii) a combination of both. The task for each classifier was to predict the CEFR level, formulated as a six-way classification problem, of a target word within a given sentence. We report the F1 score for each level, together with Macro-F1 and Micro-F1 scores to measure overall performance.

5.2. Classifiers

We compared the performance of a baseline model, ME6 Contextual, with several LLM-based approaches, including zero-shot, few-shot, and fine-tuned models.

ME6 Contextual. We reimplemented ME6 Contextual as a baseline. This method uses BERT-based contextual embeddings to train a support vector classifier (SVC) that predicts CEFR levels. For the hyperparameters of BERT and SVC that were not explicitly specified in the original study (Aleksandrova and Pouliot, 2023), we used their default settings. Although the model supports MWEs, we excluded them to align with the scope of WordNet, which contains very few MWEs. We trained three versions of the model: one on 90% of the EVP examples, one on the SemCor-CEFR corpus, and one on a combination of both. If the model trained on our corpus performs comparably to the model trained on the gold-standard EVP examples, this would support the accuracy of our level annotations.

As noted in Section 4, a single sense in our corpus may be associated with multiple CEFR levels. Therefore, when training on our data, we created one training example per level. These multi-level assignments arise when a single WordNet sense aligns with multiple, more fine-grained EVP senses. Rather than discarding them as noise, we split them into separate training instances—one per level—and treat them as alternative supervision signals. Because 96.84% of senses in our WordNet receive a single CEFR label (Table 6 in Appendix 13.2), we expect the overall impact of this multi-label handling to be limited. Nevertheless, since the EVP evaluation examples provide a single gold level per sentence, this strategy may introduce label ambiguity and reduce performance. Importantly, it should not provide an advantage over models trained solely on EVP data.

Zero-Shot LLM. We evaluated the LLM's inherent ability to classify CEFR levels without providing any examples. Using the prompt and parameter settings described in Appendix 13.1, we provided the model (GPT-5, checkpoint `gpt-5-2025-08-07`) with a target word and its context and asked it to output the corresponding CEFR level.

Few-Shot Prompted LLMs. We also evaluated the LLM's performance under 6-shot and 18-shot prompting, using the template provided in Appendix 13.1. This prompt provides the model with training examples to serve as clues for classifying a word sense in its context. In the 6-shot setting, we provided one training example for each of the six CEFR levels, that is, 1×6 examples. In the 18-shot setting, we used three examples per level, that is, 3×6 examples. The target words and usage examples were randomly selected from the 90% of EVP examples reserved for training. The LLM and parameter settings were identical to those used in the zero-shot experiments.

Fine-Tuned LLMs. We fine-tuned the lightweight and cost-effective GPT-4.1 mini model (checkpoint `gpt-4.1-mini-2025-04-14`) on three datasets: 90% of the EVP examples, the SemCor-CEFR corpus, and a combination of both. As in the evaluation of the ME6 Contextual baseline, the accuracy of our annotations would be supported if the model fine-tuned on our corpus achieved performance comparable to or better than that of the model trained on the gold-standard EVP examples. We also trained a model on the combined corpus to examine potential synergistic effects. Senses associated with multiple CEFR levels in our corpus were treated as separate training examples for each level. For

fine-tuning, we used a 90%/10% split for training and validation. The training data were formatted by filling the zero-shot template in Appendix 13.1 with each target word and sentence, using the corresponding CEFR level as the correct label. The default auto hyperparameters used for fine-tuning are listed in Appendix 13.1.

Fine-Tuned LLMs + Knowledge Base. For words whose CEFR level is unambiguous in the EVP Online, that is, all senses share the same level, performing a full six-level classification is computationally inefficient and increases the risk of errors. To address this issue, we developed a hybrid approach. We first constructed a knowledge base—a word-level list derived from the EVP Online that includes only words associated with a single CEFR level. For each target word, we checked this list first. If the word was present, we directly assigned its recorded level. Otherwise, we used one of the fine-tuned LLMs for classification. We applied this method to the LLMs fine-tuned on the EVP examples, the SemCor-CEFR corpus, and the combined corpus to compare differences in classification accuracy.

5.3. Results

Table 3 reports the F1 scores for each classifier. In the table, FT denotes the fine-tuned LLMs, and FT+KB refers to the fine-tuned LLMs combined with the knowledge-based approach. The training datasets used are EVP (90% of the EVP examples), SemCor-CEFR (our annotated SemCor corpus), and Mixture (a combination of both). Because the class distribution in our data is imbalanced, we use the Macro-F1 score as the primary metric for overall evaluation, as it assigns equal weight to each class and mitigates the effects of frequency imbalance.

The ME6 Contextual classifier achieved a Macro-F1 score of at least 0.5 across all training sets. However, its performance on the SemCor-CEFR corpus was 0.08 points lower than that on the EVP data. We attribute this gap to the model's vector construction method, which averages the vectors for all instances of a given word and CEFR level to produce a single vector for each word-level pair. As shown in Table 2, our SemCor-CEFR corpus has fewer unique word types than the EVP data. Consequently, despite having a higher total word frequency, it yields fewer training vectors, which likely contributed to the performance drop. Consistent with this interpretation, the classifier trained on the Mixture dataset, which included the largest number of training examples, achieved the best performance among the ME6 Contextual models.

The zero-shot LLM achieved a Macro-F1 score of 0.42, the lowest among all methods and well below that of the ME6 Contextual baseline. Its F1 scores for the C1 and C2 levels were particularly low, below 0.3, indicating that the LLM's internal knowledge alone is insufficient for classifying advanced-level senses. Providing in-context examples through few-shot prompting increased the Macro-F1 score to 0.47 in the 6-shot setting and 0.48 in the 18-shot setting. This improvement, consistent with prior findings (Enomoto et al., 2024; Smădu et al., 2024), resulted from supplementing the model's knowledge of C1 and C2 senses. Nevertheless, the performance of the few-shot models remained substantially lower than that of ME6 Contextual.

Fine-tuning proved to be a highly effective approach for developing LLM classifiers, improving the Macro-F1 score by at least 0.17 points compared with the few-shot methods. Notably, the FT model trained on the SemCor-CEFR corpus performed comparably to the model trained on the gold-standard EVP data, despite being optimized on a dataset entirely different from the test set. Moreover, an analysis of its errors (Figure 6(h) in Appendix 13.3) shows that misclassifications, particularly for C1-level senses, were often assigned to adjacent proficiency levels, which would likely minimize confusion for learners. This strong performance is likely due to the model being fine-tuned on the rich and varied usage examples in the SemCor-CEFR corpus. The model trained on the Mixture dataset achieved a Macro-F1 score of 0.73. These results provide indirect evidence supporting the quality of the CEFR annotations in our WordNet and effectiveness of combining the EVP and SemCor-CEFR corpora.

The hybrid FT+KB approach, which combines fine-tuned LLMs with a knowledge base, achieved the best overall performance. This method improved the Macro-F1 score by 0.08 to 0.13 points compared with the FT models alone, with a consistent trend across training sets. The classifier trained on the Mixture dataset achieved the highest F1 scores at all levels, exceeding 0.8 for every level except B1 and C1. This pattern suggests that a substantial portion of the test set consists of words with unambiguous CEFR levels. In such cases, the knowledge base can assign the correct level without relying on LLM inference, thereby improving both accuracy and computational efficiency.

6. Discussion

6.1. Correlation Analysis Using the CompLex 2.0 Dataset

Although the FT and FT+KB classifiers trained on EVP examples demonstrated strong performance, these results may be inflated because both the fine-tuning and test sets were drawn from the same

Classifier	Base Model	Train/Valid. Set	F1 Scores ↑							
			A1	A2	B1	B2	C1	C2	Macro	Micro
ME6 Cont.	BERT	EVP	0.77	0.61	0.54	0.53	0.51	0.59	0.59	0.57
		SemCor-CEFR	0.61	0.51	0.50	0.42	0.46	0.57	0.51	0.50
		Mixture	0.76	0.65	0.59	0.51	0.54	0.59	0.61	0.59
Zero-Shot	GPT-5	—	0.68	0.44	0.40	0.53	0.29	0.21	0.42	0.45
6-Shot		EVP	0.66	0.44	0.44	0.57	0.40	0.32	0.47	0.49
18-Shot		EVP	0.67	0.45	0.43	0.56	0.38	0.40	0.48	0.49
FT	GPT-4.1 mini	EVP	0.79	0.68	0.64	0.69	0.43	0.68	0.65	0.66
		SemCor-CEFR	0.72	0.67	0.68	0.71	0.44	0.66	0.67	0.67
		Mixture	0.81	0.76	0.73	0.75	0.61	0.73	0.73	0.73
FT+KB	GPT-4.1 mini	EVP	**0.83**	0.77	0.74	0.79	0.74	**0.81**	0.78	0.78
		SemCor-CEFR	0.75	0.72	0.76	0.81	0.75	0.77	0.76	0.76
		Mixture	**0.83**	**0.81**	**0.78**	**0.83**	**0.78**	**0.81**	**0.81**	**0.81**

Table 3: F1 scores for each classifier. **Bold** and underlined values indicate the highest and second-highest scores, respectively.

source. For real-world applications, a CEFR-level classifier must perform well across diverse domains, not only on dictionary and learner examples. However, gold-standard sense-level CEFR annotations for heterogeneous corpora are scarce. To examine generalizability, we used an indirect proxy by analyzing the correlation between the predicted CEFR levels and lexical complexity scores in the CompLex 2.0 dataset. This dataset, developed for the LCP task, spans three genres—Europarl, the Bible, and biomedical texts—and contains target words rated by multiple annotators on a continuous complexity scale from 0 to 1. We applied our classifiers to predict CEFR levels, mapped to integers 1 to 6, for 7,662 target words in the CompLex 2.0 training set. We then computed the Spearman rank correlation coefficient between the predicted levels and the dataset's complexity scores. We did not expect a high correlation because complexity scores are continuous, whereas CEFR levels are discrete.

A notable finding in Table 4 is that classifiers trained on the EVP examples, despite achieving high accuracy on the EVP test set, exhibited very low correlation with the CompLex 2.0 scores. This suggests that models fine-tuned solely on EVP data may have learned superficial, dataset-specific cues, with limited transfer to other genres. By contrast, classifiers trained on the SemCor-CEFR corpus achieved correlation coefficients above 0.5, indicating a moderate relationship between their predictions and lexical complexity. We attribute this improvement to the broad genre coverage of the SemCor corpus together with the high quality of our CEFR-level annotations. These findings indicate that classifiers fine-tuned on our corpus are better suited for application to educational materials drawn from diverse sources.

Classifier	Train/Valid. Set	Spearman ↑
ME6 Cont.	EVP	0.333
	SemCor-CEFR	0.377
	Mixture	0.362
Zero-Shot	—	0.396
6-Shot	EVP	0.494
18-Shot	EVP	0.490
FT	EVP	0.288
	SemCor-CEFR	**0.541**
	Mixture	0.529
FT+KB	EVP	0.366
	SemCor-CEFR	0.539
	Mixture	0.528

Table 4: Spearman rank correlation coefficients between predicted CEFR levels and CompLex 2.0 complexity scores. **Bold** and underlined values indicate the highest and second-highest scores, respectively.

6.2. Implications for L2 Learners

Our findings have important implications for L2 learners, who often struggle with the fine-grained sense distinctions in WordNet. By annotating WordNet senses with CEFR levels and integrating them into resources such as the SemCor-CEFR corpus, our approach aligns lexical information more closely with learner proficiency and pedagogical needs. Although the practical benefits of this approach require empirical validation through classroom-based

or longitudinal studies, it offers two main advantages. First, it allows learners to focus on senses that match their proficiency level, reducing the cognitive load associated with more advanced or nuanced meanings. Second, our high-performing classifier (Macro-F1 of 0.81) can be integrated into educational tools to quickly identify complex lexical items in a text, enabling immediate scaffolding. The model's strong performance on levels A1 through B2 is particularly beneficial for beginner and intermediate learners who are building foundational vocabulary. These advances may enable educators and independent learners to adopt more adaptive and efficient strategies for vocabulary instruction. However, further research is needed to determine whether these benefits persist across diverse learning environments and over extended periods.

7. Conclusions

In this study, we introduced an LLM-based method for annotating WordNet senses with CEFR levels and used it to construct a CEFR-annotated WordNet. This resource provides 10,995 proficiency annotations for 10,644 senses across 5,645 lemmas. Using the annotated WordNet, we also created the SemCor-CEFR corpus, a large-scale resource containing more than 110,000 sense-level CEFR annotations. To validate our approach, we trained contextual lexical CEFR-level classifiers on this corpus and found that they performed comparably to models trained on gold-standard data. Furthermore, by combining our corpus with gold-standard levels, we developed a practical classifier that achieves a Macro-F1 score of 0.81, providing indirect evidence of the utility and consistency of our CEFR annotations. Our analysis also showed that the predictions generated by our classifiers correlate with the lexical complexity scores in the CompLex 2.0 dataset, suggesting moderate alignment with lexical complexity across diverse text genres.

This work is part of the "Learner's WordNet Project," which seeks to integrate NLP methods—particularly WordNet's rich semantic network—with educational technology to support more efficient and effective L2 learning. Future work will focus on expanding CEFR-level coverage in our WordNet, evaluating its pedagogical effectiveness in real-world learning contexts and developing related applications. To support this effort, we plan to build a classifier capable of accurately assigning CEFR levels to previously unannotated word senses.

8. Acknowledgements

This work was supported in part by JSPS KAKENHI Grant Numbers JP22K02825, JP22K18006, JP25K21351, and JP24K03052. This publication/presentation/research report makes use of the English Vocabulary Profile. This resource is based on extensive research using the Cambridge Learner Corpus and forms part of the English Profile program, which aims to provide evidence about language use to support the development of improved language teaching materials. See https://englishprofile.org/ for more information.

9. Ethical Considerations

We accessed the EVP Online strictly in accordance with the EVP website's Terms of Use[5] and Cambridge University Press's text and data mining (TDM) terms[6]. Any EVP content temporarily cached to local storage for this project was deleted upon the project's completion. All released artifacts are built on WordNet and SemCor, whose licenses permit copying, modification, and redistribution, and include only derived CEFR labels mapped to WordNet sense keys. No verbatim EVP content, including entries, examples, glosses, or metadata, is included in the released resources. The EVP-derived word-level lookup list used in the FT+KB analysis (Section 5.2) was used solely for internal evaluation and is not redistributed.

For verification, we employed proprietary LLMs from OpenAI and enabled the opt-out setting to ensure that submitted data were not used for model training. For downstream applications involving personal or sensitive data, we recommend deploying open-source LLMs in a local environment to reduce the risk of unintended disclosure. The resources introduced in this work are compatible with locally hosted LLMs. The CEFR levels provided by our resources are model-based estimates and should not be used as the sole basis for high-stakes educational decisions, such as promotion, pass/fail determinations, or selective admissions.

10. Limitations

A key limitation of this work is the limited coverage of CEFR-level annotations in WordNet. Despite the automated pipeline, only 10,644 senses were annotated, representing approximately 5% of the full inventory. This limited coverage stems from the restricted availability of sense-level CEFR labels in the EVP Online, which serves as the primary source for alignment and constrains further expansion. Beyond the EVP Online, several CEFR-

[5] https://englishprofile.org/?menu=evp-terms-of-use

[6] https://www.cambridge.org/us/legal/copyright

related lexical resources are available, such as CEFRLex and SeCoDa, as well as lexical complexity datasets including CompLex 2.0. However, these resources differ in granularity and label space. Proficiency information is often provided at the word level rather than the sense level, sense inventories are not aligned with WordNet, and labels are not always expressed as CEFR levels. As a result, they cannot be directly integrated into our sense-level annotations without additional mapping. Developing principled methods to use them as auxiliary supervision remains future work. To mitigate the limited coverage, we developed lexical CEFR-level classifiers trained on the large, sense-annotated corpus, achieving a maximum Macro-F1 score of 0.81. Although these classifiers can predict levels for previously unseen senses based on contextual usage, they are currently less accurate than the primary gloss-based transfer method. Improving their accuracy, robustness, and generalization is therefore essential for reliable large-scale deployment. In parallel, it is important to assess how annotation errors affect L2 learners and establish acceptable error thresholds for educational applications. Another limitation is that the evaluation focused exclusively on single words, whereas related work on LCP, including models such as ME6 Contextual, also considers multiword expressions. Preliminary analysis suggests that most MWEs correspond to a single CEFR level, indicating that the knowledge-based component could classify them with high accuracy. However, challenges remain in identifying implicit MWEs in running text, such as in CALL systems analyzing textbooks, and in handling expressions not covered by the EVP. Addressing these issues will require more advanced and context-sensitive classification methods.

11. Bibliographical References

Desislava Aleksandrova and Vincent Pouliot. 2023. CEFR-based contextual lexical complexity classifier in English and French. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 518–527.

David Alfter. 2024. Out-of-the-box graded vocabulary lists with generative language models: Fact or fiction? In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2024)*, pages 1–19.

Stefano Bannò, Kate M. Knill, and Mark J. F. Gales. 2025. Exploiting the English Vocabulary Profile for L2 word-level vocabulary assessment with LLMs. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 632–646.

Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2025. Analysing zero-shot readability-controlled sentence simplification. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*, pages 6762–6781.

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*, 8:1–24.

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 (NAACL 2019)*, pages 4171–4186.

Taisei Enomoto, Hwichan Kim, Tosho Hirasawa, Yoshinari Nagai, Ayako Sato, Kyotaro Nakajima, and Mamoru Komachi. 2024. TMU-HIT at MLSP 2024: How well can GPT-4 tackle multilingual lexical simplification? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 590–598.

Thomas Gaillat, Andrew Simpkin, Nicolas Ballier, Bernardo Stearns, Annanda Sousa, Manon Bouyé, and Manel Zarrouk. 2022. Predicting CEFR levels in learners of English: The use of microsystem criterial features in a machine learning approach. *ReCALL*, 34(2):130–146.

Sunayana R. Gawde, Jayram Ulhas Gawas, Shrikrishna R. Parab, Shilpa Neenad Desai, and Jyoti Pawar. 2024. Konkani WordNet visualizer as a concept teaching-learning tool. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON 2024)*, pages 59–67.

Itziar Gonzalez-Dios. 2019. Textual genre based approach to use WordNet in language-for-specific-purpose classroom as dictionary. In *Proceedings of the 10th Global Wordnet Conference (GWC 2019)*, pages 222–227.

Elma Kerz, Daniel Wiechmann, Yu Qiao, Emma Tseng, and Marcus Ströbel. 2021. Automated classification of written proficiency levels on the CEFR-scale through complexity contours and

RNNs. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2021)*, pages 199–209.

Yoshie Kiritani, Naoaki Nippashi, and Yoichi Tamagaki. 2012. Effect of visualization of words relation in electronic English-Japanese dictionary. *Journal of the Science of Design*, 59(3):59–66.

Ali Malik, Stephen Mayhew, Christopher Piech, and Klinton Bicknell. 2024. From Tarzan to Tolkien: Controlling the language proficiency level of LLMs for content generation. In *Findings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, pages 15670–15693.

Jacek Marciniak. 2020. WordNet as a backbone of domain and application conceptualizations in systems with multimodal data. In *Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW 2020)*, pages 25–32.

Viktor Moskvoretskii, Ekaterina Neminova, Alina Lobanova, Alexander Panchenko, and Irina Nikishina. 2024. Taxollama: WordNet-based model for solving multiple lexical semantic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, pages 2331–2350.

Diego Moussallem, Matthias Wauer, and Axel-Cyrille Ngonga Ngomo. 2018. Machine translation using semantic web technologies: A survey. *Journal of Web Semantics*, 51:1–19.

Roberto Navigli. 2006. Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006)*, pages 105–112.

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

Gustavo H. Paetzold and Lucia Specia. 2017. A survey on lexical simplification. *Journal of Artificial Intelligence Research*, 60(1):549–593.

Ildikó Pilán, Sowmya Vajjala, and Elena Volodina. 2016. A readable read: Automatic assessment of language learning materials based on linguistic complexity. *International Journal of Computational Linguistics and Applications*, 7(1):143–159.

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 2020)*, pages 85–92.

Iñaki San Vicente, Rodrigo Agerri, and German Rigau. 2014. Simple, robust and (almost) unsupervised generation of polarity lexicons for multiple languages. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*, pages 88–97.

Veronica Juliana Schmalz and Alessio Brutti. 2021. Automatic assessment of English CEFR levels using BERT embeddings. In *Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it 2021)*, pages 1–7.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Sanja Štajner, Marcos Zampieri, and Horacio Saggion. 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 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.

Răzvan-Alexandru Smădu, David-Gabriel Ion, Dumitru-Clementin Cercel, Florin Pop, and Mihaela-Claudia Cercel. 2024. Investigating large language models for complex word identification in multilingual and multidomain setups. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*, pages 16764–16800.

Koun-Tem Sun, Yueh-Min Huang, and Ming-Chi Liu. 2011. A WordNet-based near-synonyms and similar-looking word learning system. *Educational Technology & Society*, 14(1):121–134.

Anaïs Tack, Thomas François, Sophie Roekhaut, and Cédrick Fairon. 2017. Human and automated CEFR-based grading of short answers. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2017)*, pages 169–179.

Satoru Uchida, Yuki Arase, and Tomoyuki Kajiwara. 2024. Profiling English sentences based on CEFR levels. *International Journal of Applied Linguistics*, 175(1):103–126.

12. Language Resource References

Sonja Bosch and Marissa Griesel. 2018. African WordNet: Facilitating language learning in African languages. In *Proceedings of the 9th Global Wordnet Conference (GWC 2018)*, pages 306–313.

Annette Capel. 2012. Completing the English Vocabulary Profile: C1 and C2 vocabulary. *English Profile Journal*, 3(1):1–14.

Luise Dürlich and Thomas François. 2018. EFLLex: A graded lexical resource for learners of English as a foreign language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 873–879.

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

Thomas François, Nùria Gala, Patrick Watrin, and Cédrick Fairon. 2014. FLELex: A graded lexical resource for French foreign learners. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 3766–3773.

Thomas François, Elena Volodina, Ildikó Pilán, and Anaïs Tack. 2016. SVALex: A CEFR-graded lexical resource for Swedish foreign and second language learners. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 213–219.

Masato Kikuchi, Masatsugu Ono, Toshioki Soga, Tetsu Tanabe, and Tadachika Ozono. 2024. Coarse-grained sense inventories based on semantic matching between English dictionaries. In *Proceedings of the 11th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA 2024)*, pages 1–6.

Caterina Lacerra, Michele Bevilacqua, Tommaso Pasini, and Roberto Navigli. 2020. CSI: A coarse sense inventory for 85% word sense disambiguation. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020)*, pages 8123–8130.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proceedings of the 3rd DARPA Workshop on Human Language Technology*, pages 303–308.

Petya Osenova and Kiril Simov. 2024. Towards a multimodal WordNet for language learning in Bulgarian. *Digital Presentation and Preservation of Cultural and Scientific Heritage*, 14:117–124.

Hanumant Redkar, Rajita Shukla, Sandhya Singh, Jaya Saraswati, Laxmi Kashyap, Diptesh Kanojia, Preethi Jyothi, Malhar Kulkarni, and Pushpak Bhattacharyya. 2018. Hindi Wordnet for language teaching: Experiences and lessons learnt. In *Proceedings of the 9th Global Wordnet Conference (GWC 2018)*, pages 314–323.

Matthew Shardlow, Richard Evans, and Marcos Zampieri. 2022. Predicting lexical complexity in English texts: The Complex 2.0 dataset. *Language Resources and Evaluation*, 56:1153–1194.

David Strohmaier, Sian Gooding, Shiva Taslimipoor, and Ekaterina Kochmar. 2020. SeCoDa: Sense complexity dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020)*, pages 5962–5967.

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 (BEA 2018)*, pages 137–146.

Elena Volodina, Ildikó Pilán, Lorena Llozhi, Baptiste Degryse, and Thomas François. 2016. SweLLex: Second language learners' productive vocabulary. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition (NLP4CALL 2016)*, pages 76–84.

Please assess whether the two meanings of the English word {word} are the same from a linguistic perspective.
1: {one gloss g of {word} in the EVP Online}
2: {one gloss g' of {word} in WordNet}

Please select one option from the following and answer using only the corresponding number.
1. Exactly the same meaning
2. Almost the same meaning
3. Somewhat similar meaning
4. Neither similar nor different meaning
5. Somewhat different meaning
6. Mostly different meaning
7. Completely different meaning

Figure 3: Prompt template used to measure semantic similarity between an EVP gloss (g) and a WordNet gloss (g').

The CEFR is a six-level scale, with each level corresponding to a specific English proficiency level. The levels are as follows:

A1: Beginner
A2: Elementary
B1: Intermediate
B2: Upper Intermediate
C1: Advanced
C2: Proficiency

According to the CEFR scale, which proficiency level is required to understand the sense of {word} in the following text: {sentence}

Please do not provide explanations.

Figure 4: Prompt template used for zero-shot CEFR-level classification.

13. Appendices

13.1. Parameters and Prompts

To ensure reproducibility, we report the exact checkpoint of the OpenAI model used in this study. Different checkpoints were used for gloss similarity measurements (Section 4) and classifier experiments (Section 5), as these components were conducted at different times. For each task, we used the most recent checkpoint available at the time of execution.

Semantic Similarity Measurement. To measure the semantic similarity between g_i and g'_j, we used GPT-4o (checkpoint `gpt-4o-2024-08-06`) with the prompt shown in Figure 3. The system message was set at "You are a professional linguist," and the temperature was set at 0 to ensure deterministic outputs.

Train/Valid. Set	Parameter	Value
EVP	Method	Supervised
	Seed	1900973879
	Batch size	17
	LR multiplier	2
	Epochs	1
SemCor-CEFR	Method	Supervised
	Seed	105188566
	Batch size	69
	LR multiplier	2
	Epochs	1
Mixture	Method	Supervised
	Seed	112279849
	Batch size	86
	LR multiplier	2
	Epochs	1

Table 5: Hyperparameters for the fine-tuned LLMs.

Zero-Shot and Few-Shot Classifiers. We used GPT-5 (checkpoint `gpt-5-2025-08-07`) as the base model for our classifiers. The system message was set at "You are an expert rater for the Common European Framework of Reference for Languages (CEFR)." and the parameter reasoning_effort was set at "high." Figures 4 and 5 present the prompt templates used for the zero-shot and few-shot LLM classifiers, respectively. In a preliminary experiment, we provided the LLMs with full CEFR level descriptions based on the official can-do descriptors. However, we observed no significant difference in classification performance compared to using the simplified descriptions shown in the figures. For efficiency, we therefore used the prompts with simplified descriptions in our experiments.

Fine-Tuned LLMs. As described in Section 5.2, we constructed the training data using the zero-shot template shown in Figure 4. The hyperparameters used for fine-tuning GPT-4.1 mini (checkpoint `gpt-4.1-mini-2025-04-14`) are detailed in Table 5.

13.2. Multi-Labeled Sense Keys

In WordNet, a sense key is the canonical string identifier of an individual sense, encoding the lemma, PoS, and a sense index. Because our CEFR annotation transfers levels from EVP senses to WordNet senses via gloss alignment, a single WordNet

The CEFR is a six-level scale, with each level corresponding to a specific English proficiency level. The levels are as follows:

A1: Beginner
A2: Elementary
B1: Intermediate
B2: Upper Intermediate
C1: Advanced
C2: Proficiency

According to the CEFR scale, the proficiency levels required to understand the senses of the words in the following texts are:

Word: {train_word$_1$}, Text: {train_sentence$_1$} -> CEFR: {The gold-standard level ℓ_1}
Word: {train_word$_2$}, Text: {train_sentence$_2$} -> CEFR: {The gold-standard level ℓ_2}

(...more training examples...)

Word: {test_word}, Text: {test_sentence} -> CEFR:

Please respond with only the level.

Figure 5: Prompt template used for few-shot CEFR-level classification.

# Distinct CEFR Labels	# Sense Keys	Share (%)
1	10,308	96.84
2	321	3.02
3	15	0.14

Table 6: Distribution of the number of distinct CEFR labels per WordNet sense key.

Row Level	Column Level					
	A1	A2	B1	B2	C1	C2
A1	—	33	31	13	8	2
A2	33	—	43	24	4	6
B1	31	43	—	86	12	13
B2	13	24	86	—	24	40
C1	8	4	12	24	—	27
C2	2	6	13	40	27	—

Table 7: Pairwise co-occurrence counts of CEFR levels within multi-labeled sense keys. Each cell (x, y) reports the number of sense keys whose label set contains both levels x and y; three-level cases contribute to multiple pairs.

sense key may occasionally receive multiple CEFR labels. This occurs when a WordNet gloss is sufficiently similar, with a similarity score of ≤ 2, to more than one EVP gloss and those EVP glosses carry different CEFR levels. Such multi-label assignments primarily reflect differences in gloss granularity between the two resources.

How frequent are multi-labeled sense keys? Table 6 summarizes the number of distinct CEFR labels assigned to each WordNet sense key. The vast majority of sense keys are unambiguous: 10,308 sense keys (96.84%) receive exactly one label. Multi-labeled cases are rare, with 321 sense keys (3.02%) assigned two labels and only 15 sense keys (0.14%) assigned three labels. Overall, multi-label assignments affect approximately 3.16% of all annotated sense keys. This indicates that the conservative alignment threshold produces predominantly single-level annotations while retaining coverage for borderline cases.

Which level combinations co-occur? To characterize the nature of multi-labeling, Table 7 reports pairwise co-occurrence counts of CEFR levels within the multi-labeled sense keys. Two-label cases contribute one pair, and three-label cases contribute three pairs, yielding 366 total co-occurrence pairs in Table 7. A clear pattern emerges. The most frequent co-occurrence is B1–B2 (86), followed by A2–B1 (43), B2–C2 (40), and A1–A2 (33). Overall, adjacent-level pairs (A1–A2, A2–B1, B1–B2, B2–C1, C1–C2) dominate. This pattern suggests that multi-labeling typically occurs near proficiency boundaries rather than arising from arbitrary mismatches. Less frequent and more distant pairs, such as A1–C1 and A2–C2, may reflect particularly broad WordNet glosses or EVP senses whose pedagogical sequencing differs substantially

Sense Key	Lemma	PoS	CEFR (3 Labels)	WordNet Gloss (Definition)
bad%3:00:00::	bad	Adjective	A1/A2/C1	having undesirable or negative qualities
block%2:35:02::	block	Verb	B2/C1/C2	obstruct
close%2:41:00::	close	Verb	A2/B2/C2	cease to operate or cause to cease operating
dance%1:04:00::	dance	Noun	A1/A2/B1	taking a series of rhythmical steps (and movements) in time to music
find%2:39:02::	find	Verb	A1/A2/B1	discover or determine the existence, presence, or fact of
give%2:32:02::	give	Verb	A1/A2/B1	bestow
give%2:36:00::	give	Verb	A1/A2/B1	give or supply
hard%3:00:06::	hard	Adjective	A1/B1/C1	not easy
miss%2:32:00::	miss	Verb	A2/B1/B2	fail to experience
safe%3:00:01::	safe	Adjective	A1/A2/B1	free from danger or the risk of harm
schedule%1:10:00::	schedule	Noun	A2/B1/B2	an ordered list of times at which things are planned to occur
shake%2:29:00::	shake	Verb	B1/B2/C2	move with or as if with a tremor
start%2:36:01::	start	Verb	A1/B1/B2	get off the ground
start%2:38:00::	start	Verb	A1/B1/B2	begin or set in motion
start%2:38:01::	start	Verb	A1/B1/B2	get going or set in motion

Table 8: WordNet sense keys annotated with three distinct CEFR levels. For each sense key, we report the lemma, PoS, and the WordNet gloss.

across sub-senses.

What do three-label cases look like? Table 8 lists the 15 sense keys assigned three distinct CEFR labels. These cases are dominated by highly frequent and semantically broad lemmas (e.g., *bad*, *close*, *find*, *give*, *start*), which are associated with short and general WordNet glosses. Such senses can plausibly align with EVP sub-senses introduced at different stages, for example early concrete uses versus later, more abstract or specialized uses. Importantly, the extremely small number of three-label cases suggests that wide label dispersion is exceptional; most multi-labeled sense keys involve only two nearby levels.

13.3. Confusion Matrices

Figure 6 presents the confusion matrices for each classifier. Each matrix element represents the classification probability, calculated as

$$p_{\ell,\widehat{\ell}} = \frac{n_\ell(\widehat{\ell})}{n_\ell},$$

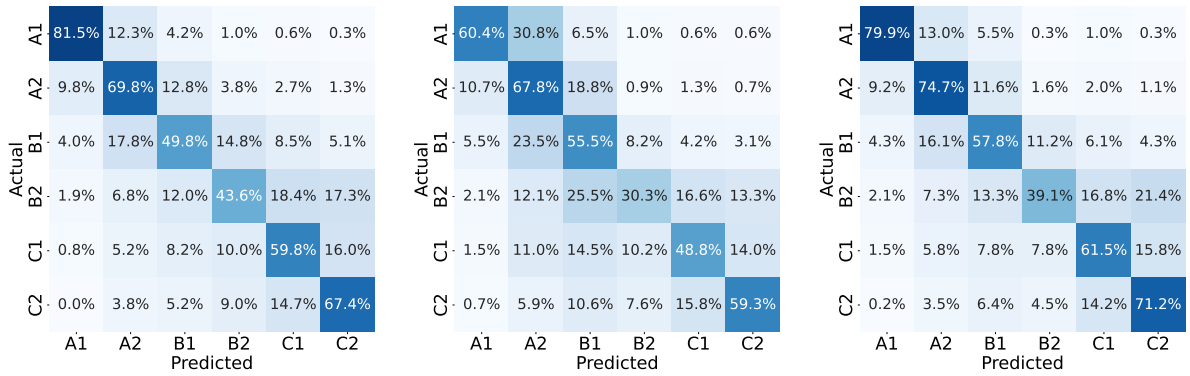
where n_ℓ denotes the number of target words with the actual CEFR level ℓ and $n_\ell(\widehat{\ell})$ is the number of those words classified as level $\widehat{\ell}$, i.e., $n_\ell = \sum_{\widehat{\ell}} n_\ell(\widehat{\ell})$. The diagonal elements represent the recall for each level; higher values along the diagonal therefore indicate greater accuracy. Because the CEFR levels are ordinal, misclassifications that fall near the diagonal, that is, those assigned to adjacent levels, are less disruptive for language learners.

The ME6 Contextual models achieve high recall at the lower levels, A1 and A2, as well as at the highest level, C2. However, as shown in Figures 6(a) and 6(b), when the model is trained on either the EVP or SemCor-CEFR corpus alone, errors at the intermediate and advanced levels, B1 to C2, are more widely distributed. By contrast, combining both resources (Figure 6(c)) reduces this dispersion, with most misclassifications occurring between adjacent levels. This result highlights the advantage of jointly leveraging both resources.

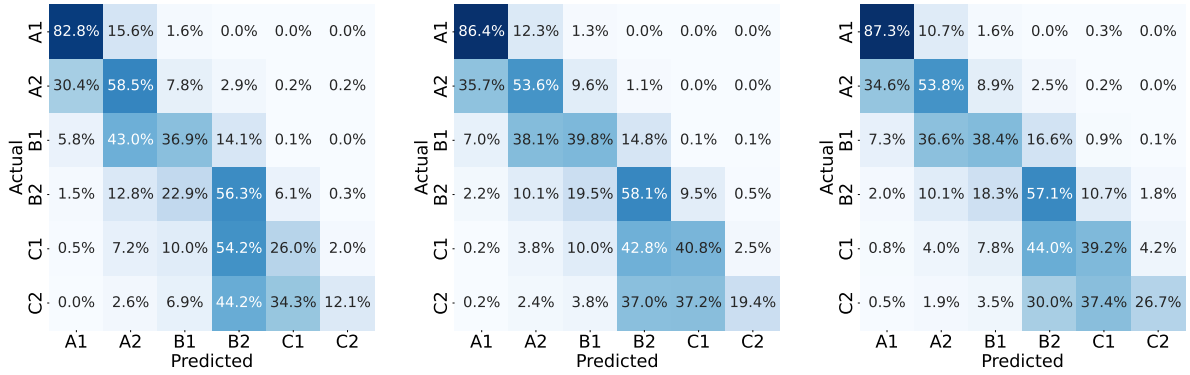
As shown in Figure 6(d), the zero-shot LLM achieves high recall for A1 (82.8%) and moderate recall for A2 (58.5%). Performance declines at B1 (36.9%) and is particularly poor at C1 and C2 (26.0% and 12.1%, respectively). The model frequently misclassifies advanced-level senses as B2, for example 54.2% of C1 and 44.2% of C2 cases, indicating a tendency to collapse more difficult senses into an intermediate level. Few-shot prompting (Figures 6(e) and 6(f)) partially alleviates this issue by improving recall at C1 and C2. However, recall at A2 declines relative to the zero-shot baseline, and performance at B1 remains similar. These results indicate that the gains from few-shot prompting are uneven across proficiency levels.

By contrast, the FT and FT+KB models substantially improve performance across all CEFR levels. When fine-tuned on the Mixture dataset (Figures 6(i) and 6(l)), the FT model achieves recall above 70% for levels A1–B2 and just below 70% for C2. The FT+KB model further improves recall, exceeding 80% for B1–B2, reaching approximately 91% for A1, remaining in the high-70% range for

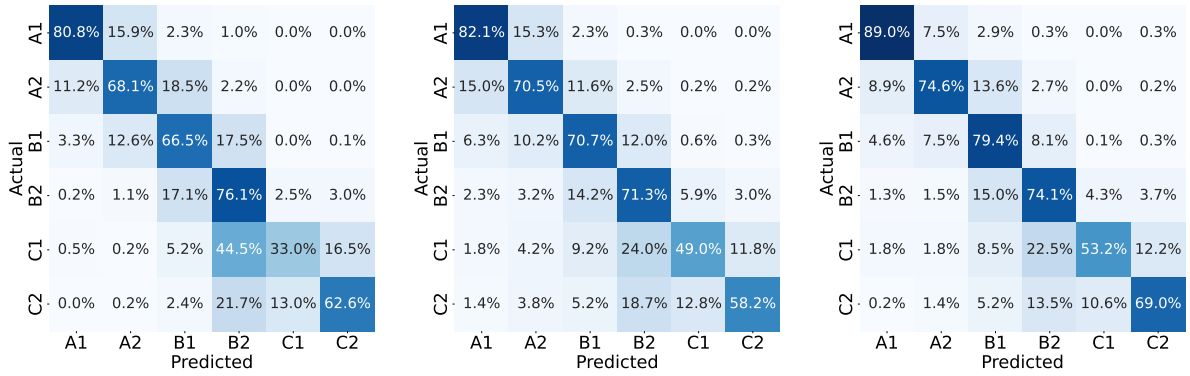
A2 and C2, and around 70% for C1. The corresponding confusion matrices show that errors are concentrated near the diagonal, meaning that most misclassifications occur between adjacent CEFR levels. This pattern reduces potential pedagogical disruption. Despite these improvements, C1 remains challenging, and C2 instances are often misclassified as B2. This pattern also appears when fine-tuning on the EVP or SemCor-CEFR corpora individually (Figures 6(g) and 6(h)), suggesting that it is not an artifact of the annotation method. Rather, it likely reflects the CEFR-level distribution in the EVP data and characteristics of the fine-tuning process. Further investigation is needed to address these residual errors.



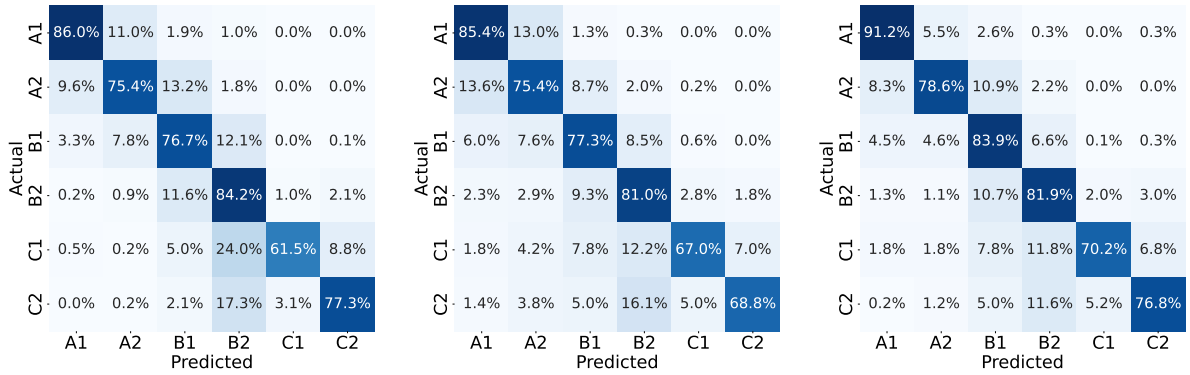
(a) ME6 Cont. using EVP. (b) ME6 Cont. using SemCor-CEFR. (c) ME6 Cont. using Mixture.



(d) Zero-Shot. (e) 6-Shot. (f) 18-Shot.



(g) FT using EVP. (h) FT using SemCor-CEFR. (i) FT using Mixture.



(j) FT+KB using EVP. (k) FT+KB using SemCor-CEFR. (l) FT+KB using Mixture.

Figure 6: Confusion matrices for each classifier.