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, Muroran Institute of Technology, Chitose Institute of Science and Technology, Hokkaido University {kikuchi, ozono}@nitech.ac.jp, onomasa@muroran-it.ac.jp, t-soga@photon.chitose.ac.jp, ttanabe@iic.hokudai.ac.jp

Abstract

Although WordNet is a valuable resource owing to its structured semantic networks and extensive vocabulary, its fine-grained sense distinctions can be challenging for second-language learners. To address this, we developed a 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 method, we constructed a large-scale corpus containing both sense and CEFR-level information from our annotated WordNet and used it to develop contextual lexical classifiers. Our experiments demonstrate that models fine-tuned on our corpus perform comparably to those trained on gold-standard annotations. Furthermore, by combining our corpus with the gold-standard data, we developed a practical classifier that achieves a Macro-F1 score of 0.81, indicating the high accuracy of our annotations. Our 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

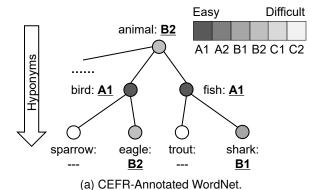
1. Introduction

WordNet (Fellbaum, 1998) is a large-scale English lexical database that organizes approximately 155,000 words and 207,000 senses for nouns, verbs, adjectives, and adverbs into hierarchical semantic networks. It groups semantically similar words and links senses through rich relations such as hypernymy, hyponymy, synonymy, and antonymy. Because WordNet and its construction software are publicly available, they can be easily integrated into Al applications. Consequently, Word-Net 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 (Vicente et al., 2014)—owing to its accessible interface and well-structured networks. These technologies also support computerassisted language learning (CALL) by aiding in 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. A key issue is its overly finegrained sense distinctions and the sheer number of senses for 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-related 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, bridging the gap between NLP lexical resources and language education. A previous work (Kikuchi et al., 2024) clustered fine-grained WordNet sense definitions (glosses) from learneroriented dictionaries. This study adopts a different approach by integrating Common European Framework of Reference for Languages (CEFR) proficiency levels into WordNet, enabling the presentation of senses that align with a learner's proficiency. To build large-scale, practical resources, we employ a simple, large language model (LLM)based method for efficient and accurate semantic annotation that reduces the time, labor, and cost associated with manual annotation. This automatic approach also ensures that our WordNet can be scaled flexibly.

The CEFR is an international standard for describing language proficiency across six levels (A1, A2, B1, B2, C1, C2), from basic to advanced. Each level is defined by "can-do descriptors" that specify expected communication abilities. We used an LLM to annotate WordNet senses with these CEFR lev-



Which is the CEFR level for "bank"?

A1: Beginner

A2: Elementary
B1: Intermediate
B2: Upper Intermediate

C1: Advanced C2: Proficiency

(1) She works in a **bank**.

(2) These flowers grow on river banks

(1) A1 (2) B2 Classifier

(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."

els, thereby constructing a CEFR-annotated Word-Net. 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, as well as to learn basic and advanced paraphrases through synonyms. The annotation pipeline involves three steps. First, we gather glosses for target words from WordNet and the English Vocabulary Profile (EVP) Online¹ (Capel, 2012), which provides CEFR levels for individual senses. Next, an 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.

As our method for annotating WordNet senses with CEFR levels is automatic, it eliminates the need for labor-intensive manual work. However, because automatic labels can contain errors and WordNet lacks gold-standard CEFR levels, their reliability must be verified indirectly. To address this, 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 assess the quality of our annotations by comparing classifiers trained on our data with those trained on the EVP goldstandard levels. We also examine the effectiveness of the LLMs for this task in few-shot and fine-tuning settinas.

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, effectively linking the WordNet sense inventory with CEFR standards. Our annotated WordNet covers approx-

- 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 the semantic similarity between WordNet glosses and EVP entries. The method, implemented entirely through prompting, is simple, reproducible, and significantly less costly than manual annotation. We also indirectly demonstrate 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 3.0² (Miller et al., 1993), a widely used sense-annotated corpus. This resulted in a large-scale corpus with over 110,000 sense and level annotations across more than 5,500 WordNet senses. As modern NLP relies on large corpora for advanced training and analysis, our resource is a valuable contribution to 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 that 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 MacroF1 score of 0.81. Our analysis suggests that
 these classifiers can accurately predict CEFR
 levels in a broad range of contexts.

All resources developed in this study, including our WordNet, corpus, and classifier, are pub-

imately 80% of all single-word senses in the EVP (8,289 out of 10,394).

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

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

licly 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 numerous studies have explored their application in language learning (Gonzalez-Dios, 2019). Some of this work has focused on visualizing word hierarchies and semantic relations to aid learners (Sun et al., 2011; Kiritani et al., 2012; Gawde et al., 2024). 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 efforts rely on manually curated resources and pay limited attention to word-sense information. By contrast, 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 languagelearning contexts.

2.2. Lexical Complexity Prediction

Lexical complexity prediction (LCP) (North et al., 2023; Shardlow et al., 2024) has recently gained significant attention for estimating word complexity from context. In this field, "complexity"which aligns with 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 shares a similar classification setting. For this task, the organizers released the CompLex 2.0 dataset³ (Shardlow et al., 2022), wherein 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, contextdriven 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)—replacing complex words with simpler ones—rather than for explicitly presenting complexity information to L2 learners. Second, annotators for CompLex 2.0 were not provided with glosses; therefore, identical senses could receive

different scores in different contexts. Third, the dataset is limited to 9,000 nouns, excluding other parts of speech (PoS).

By contrast, our method assigns a CEFR level to each word sense, adhering to an international proficiency standard. We extend this annotation to over 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, 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). Recent research has also focused on the interaction between LLMs and the CEFR, exploring 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 efforts, along with the development of numerous CEFR-aligned lexical datasets, underscore the CEFR's central role in the field.

For example, the CEFRLex project provides machine-readable lexical resources with word and sense frequency counts by CEFR level (Pintard and François, 2020) for English and other languages (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) does provide sense-in-context CEFR annotations, but its sense labels are not aligned with those of WordNet, and its small size (1,432 words) limits its applicability within WordNet's semantic framework. Our work addresses these gaps by annotating over 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 focused on classifying vocabulary into CEFR levels based on context. Aleksandrova and Pouliot (2023) proposed ME6 Contextual, a bidirectional encoder representations from transformers (BERT) (Devlin et al., 2019)-based classifier that, similar to our models, is trained on a CEFR-annotated corpus to directly predict a word's level from its con-

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

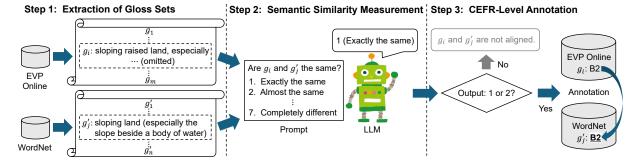


Figure 2: CEFR-level annotation process for a WordNet gloss (g'_i) , illustrated with the noun "bank."

text. This direct-prediction approach allows it to classify words not seen during training. By contrast, Bannò et al. (2025) introduced an indirect-prediction method wherein an LLM selects the appropriate EVP sense for a word in context before mapping it to a CEFR level. The performance of this approach depends on the quality of cues from 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 Online¹, 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 the 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) that also included multiword expressions (MWEs), our work focuses exclusively on single words, as WordNet contains very few MWEs.

3.2. SemCor Corpus

The SemCor 3.0 corpus is the most widely used sense-annotated corpus in NLP. It contains 226,040 sense annotations across 352 documents from the Brown Corpus. Each sense is tagged with a Word-Net identifier, which links 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 mentioned in the

introduction. Notably, the lack of learner-oriented indicators, such as sense complexity or CEFR levels, limits its utility for educational applications. To address this, 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 $\langle \text{bank}, \text{noun} \rangle$, 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 focus on comparing 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 of the 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-40-2024-08-06). The prompt is shown in Appendix 13.1. Because glosses from different resources often vary in granularity and may not align perfectly, a binary alignment (same/different) judgment would be too restrictive. Therefore, we instructed the LLM to rate the 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 the two glosses have identical meaning.

Step 3: CEFR-Level Annotation. If the LLM returns a score of 1 or 2—indicating that g_i and

| PoS | | | CEFR Levels Tota | | | | | Share (%) | |
|-----------|------|-------|------------------|-------|-------|-------|--------|-------------|--|
| F03 | A1 | A2 | B1 | B2 | C1 | C2 | Total | Silale (70) | |
| 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 owing to differences in gloss granularity between resources.

 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 (i.e., output ≥ 3), the senses are considered mismatched, and no annotation is made. In our example, the score of 1 results in the WordNet sense being assigned the B2 level from the corresponding EVP sense.

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, \dots, g_m\},\$$

 $g' \in \{g'_1, g'_2, \dots, g'_n\}\},\$

This exhaustive prowhose size is $m \times n$. cess 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%) accounting for half of all annotations. The beginner (A1-A2) and advanced (C1-C2) levels represent 16.83% and 33.10%, respectively. Because the granularity of glosses differs between the two resources, a single WordNet sense can align with multiple EVP glosses, sometimes resulting in a sense being assigned multiple CEFR levels. This automated procedure is generalizable and could be applied to other dictionaries or lexical databases that provide glosses. However, because it is fully automated, evaluating the accuracy of the resulting annotations is essential.

5. Experiment

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 various LLM-based classifiers to evaluate their effectiveness for this task.

5.1. Datasets and Experimental Settings

To train our classifiers, we needed a corpus with CEFR-level annotations for words in context. As the original SemCor corpus lacks this information, we created the "SemCor-CEFR corpus" by assigning CEFR levels to its senses using our annotated WordNet. Table 2 summarizes the word distributions in the EVP Online examples (dictionary and learner examples combined) and our SemCor-CEFR corpus. While our corpus has fewer word types (#types) than the EVP examples, it contains substantially more word instances (#words) and reflects a more natural, imbalanced distribution of proficiency levels. For our experiments, we used 10% of the EVP examples as the test set, whereas the remaining 90% and the SemCor-CEFR corpus were used for training and validation. The task for each classifier was to predict the CEFR level (a six-way classification) of a target word within a given sentence. We report the F1 score for each level, along with Macro-F1 and Micro-F1 scores for overall performance.

5.2. Classifiers

We compared the performance of a baseline model, ME6 Contextual, with several LLM-based approaches: 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

| | CEFR | EVP Online | | SemCo | or-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.

their default settings. Although this model supports MWEs, we exclude MWEs to align with the scope of WordNet, which contains almost none. We trained three versions of this model: one on 90% of the EVP examples, one on our SemCor-CEFR corpus, and one on a mixture of both. As noted in Section 4, a single sense in our corpus may have multiple levels; therefore, we created one training example per level when training on our data. If the model trained on our corpus performs comparably to the one trained on the gold-standard EVP examples, it would support the accuracy of our level annotations.

Zero-Shot LLM. We evaluated an LLM's inherent ability to classify CEFR levels without any examples. Using the prompt and parameter settings 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 Trained LLMs. We also evaluated the LLM's performance with 6- 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. For the 6-shot setting, we provided one training example for each of the six CEFR levels (i.e., 1×6 examples), whereas for the 18-shot setting, we used three examples per level (i.e., 3×6 examples). The target words and their usage examples were randomly selected from the 90% of EVP examples reserved for training. The LLM and parameter settings were the same as those used for 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 different datasets: 90% of the EVP examples, our SemCor-CEFR corpus, and a mixture of both. Similar to

the ME6 Contextual baseline evaluation, the high accuracy of our annotations would be confirmed if the model fine-tuned on our corpus achieved comparable or better performance than that trained on the gold-standard EVP examples. We also trained a model on the mixed corpus to test for any synergistic effects. Senses 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 populating the zeroshot template from Appendix 13.1 with each target word and sentence, using the corresponding CEFR level as the correct answer. 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 (i.e., all senses share the same level), running a full six-level classification is computationally inefficient and increases the risk of classification errors. To address this, we developed a hybrid approach. We first built a knowledge base—a wordlevel list from the EVP Online containing only words with a single CEFR level. For each target word, we first checked this list. 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, our SemCor-CEFR corpus, and the mixed corpus to observe the differences in classification accuracy.

5.3. Results

Table 3 reports the F1 scores for each classifier. In this 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 was imbalanced, we used the Macro-F1 score as the primary metric for evaluating overall performance, as it assigns equal weight to each class and thus 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, resulting in a single vector per word-level pair. As shown in Table 2, our SemCor-CEFR corpus has fewer unique word types than the EVP data. Con-

| Classifier | Classifier Base Model | Train/Valid. | F1 scores ↑ | | | | | | | |
|------------|-----------------------|--------------|-------------|------|------|------|-------------|------|-------|-------|
| Ciassillei | | Set | A1 | A2 | B1 | B2 | C1 | C2 | Macro | Micro |
| | | EVP | 0.77 | 0.61 | 0.54 | 0.53 | 0.51 | 0.59 | 0.59 | 0.57 |
| ME6 Cont. | BERT | 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 | | _ | 0.68 | 0.44 | 0.40 | 0.53 | 0.29 | 0.21 | 0.42 | 0.45 |
| 6-Shot | GPT-5 | 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 |
| | | EVP | 0.79 | 0.68 | 0.64 | 0.69 | 0.43 | 0.68 | 0.65 | 0.66 |
| FT | GPT-4.1 mini | 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 | <u>0.75</u> | 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 <u>underlined</u> values indicate the highest and second-highest scores, respectively.

sequently, 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 contained the most training examples, achieved the best performance among the ME6 Contextual models.

The zero-shot LLM achieved a Macro-F1 of 0.42, the lowest score 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 via few-shot prompting raised the Macro-F1 score to 0.47 (6-shot) and 0.48 (18-shot). This improvement, which aligns with prior findings (Enomoto et al., 2024; Smădu et al., 2024), stemmed from supplementing the model's knowledge of C1 and C2 senses. Nevertheless, the performance of the few-shot models remained significantly 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 over the few-shot methods. Notably, the FT model trained on our SemCor-CEFR corpus performed comparably to that trained on the gold-standard EVP data, despite being optimized on an entirely different dataset from the test set. Moreover, an analysis of its errors (Figure 6(h) in Appendix 13.2) shows that misclassifications, especially for C1-level senses, were often assigned to adjacent proficiency levels, which would minimize confusion for learners. This strong performance is likely because the LLM could be trained directly on the rich and varied usage examples in the SemCor-CEFR corpus. The model trained on the Mixture dataset achieved a MacroF1 of 0.73. These results confirm both the high accuracy of the CEFR annotations in our WordNet and the effectiveness of combining the EVP and SemCor-CEFR corpora.

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

6. Discussion

6.1. Correlation Analysis Using the CompLex 2.0 Dataset

While our FT and FT+KB classifiers trained on EVP examples demonstrated strong performance, these results may be inflated, as both the fine-tuning and test sets were drawn from the same source. For real-world applications, a CEFR-level classifier must be effective across various domains, not just dictionary and learner examples. However, gold-standard, sense-level CEFR annotations for diverse corpora are scarce. To address this, we evaluated our classifiers' generalizability by examining the correlation between their predicted CEFR levels and the lexical complexity scores in the CompLex 2.0 dataset. This dataset, used for the LCP task, spans three distinct genres—Europarl, the

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

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

Bible, and biomedical texts— and contains target words rated by multiple annotators on a continuous complexity scale from 0 to 1. We used our classifiers to predict CEFR levels (mapped as integers 1–6) for 7,662 target words in the CompLex 2.0 training set and then computed the Spearman rank correlation coefficient between our predictions and the dataset's complexity scores. We did not expect a very high correlation, as complexity scores are continuous while CEFR levels are discrete.

A notable finding from the results presented in Table 4 is that classifiers trained on the EVP examples, despite their high accuracy on the EVP test set, showed significantly 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 and therefore exhibit poor generalizability to other text genres. By contrast, classifiers trained on our SemCor-CEFR corpus achieved correlation coefficients exceeding 0.5, indicating a moderate relationship between their predictions and lexical complexity. We attribute this significant improvement to the broad genre coverage of the SemCor corpus combined with the high accuracy of our CEFR-level annotations. Consequently, classifiers fine-tuned on our corpus are better suited for use with educational materials from diverse sources.

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 method helps align lexical information with learner proficiency and pedagogical needs. Although the practical benefits of this approach require empirical validation through classroom or longitudinal studies, it offers two key 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 superior performance on levels A1 through B2 is particularly beneficial for beginner and intermediate learners who are building their foundational vocabulary. These advancements can allow educators and autonomous learners to adopt more adaptive and efficient strategies for vocabulary instruction. However, further research is required to confirm 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 Word-Net. This new resource provides 10,995 proficiency annotations for 10,644 senses across 5,645 lemmas. Using this annotated WordNet, we also created the SemCor-CEFR corpus, a large-scale resource containing over 110,000 sense-level CEFR annotations. To validate our approach, we trained contextual lexical CEFR-level classifiers on our corpus and found that they performed comparably to those trained on gold-standard data. Moreover, by combining our corpus with the gold-standard levels, we developed a practical classifier that achieves a Macro-F1 score of 0.81, confirming the high accuracy and utility of our CEFR annotations. Our analysis also showed that the predictions from our classifiers correlate with the lexical complexity scores in the CompLex 2.0 dataset, suggesting that they can generalize effectively across diverse text genres.

This work is part of the "Learner's WordNet Project," which aims 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 the CEFR-level coverage in our WordNet, evaluating its pedagogical effectiveness in real-world learning scenarios, and developing related applications. To support this expansion, we plan to build a classifier that can accurately assign CEFR levels to previously unannotated word senses.

8. Acknowledgements

This work was supported in part by JSPS KAK-ENHI Grant Numbers JP22K02825, JP22K18006, JP25K21351, and JP24K03052. This publication/presentation/research report has made use of the English Vocabulary Profile. This resource is based on extensive research using the Cambridge Learner Corpus and is part of the English Profile programme, which aims to provide evidence about language use that helps to produce better language teaching materials. See https://englishprofile.org/for more information.

9. Ethical Considerations

We accessed the EVP Online strictly under the EVP website's Terms of Use⁴ and Cambridge University Press's text and data mining (TDM) terms⁵. Any EVP content temporarily cached to local storage for this project was deleted upon the project's completion. All released artifacts are built exclusively from WordNet and SemCor, whose licenses permit copying, modification, and redistribution. No EVP data (including entries, examples, glosses, or metadata) are included in any of the released resources.

For verification, we employed proprietary LLMs from OpenAI and enabled the opt-out setting to ensure that data were not used for model training. However, in any downstream use cases that involve 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 such locally hosted LLMs. The CEFR levels provided by our resources are model-based estimates and must not be used as the sole criterion for high-stakes educational decisions, such as promotion, pass/fail judgments, or selective admissions.

10. Limitations

A key limitation of our work is the need to expand the coverage of CEFR-level annotations in Word-Net. Despite our novel automated pipeline, we annotated only 10,644 senses, which represents approximately 5% of WordNet's total. This low coverage is a direct result of the limited availability of sense-level CEFR labels in the EVP Online, a foundational resource for many dictionaries that constrains further expansion. To mitigate this, we developed lexical CEFR-level classifiers trained on our large, sense-annotated corpus, which achieved a maximum Macro-F1 score of 0.81. While these

classifiers can predict levels for previously unseen senses from their usage context, they are currently less accurate than our primary gloss-based method. Consequently, improving their accuracy, capability, and robustness is essential for reliable, large-scale applications. In parallel, it is crucial to quantify the impact of annotation errors on L2 learners and establish acceptable error thresholds for educational use. Another limitation is that our evaluation targeted only single words, whereas related work in tasks such as LCP and models like ME6 Contextual also includes MWEs. Our preliminary analysis indicates that most MWEs map to a single CEFR level, suggesting that our knowledge-based approach could classify them with high accuracy. However, significant challenges remain, particularly in identifying implicit MWEs within texts (such as in CALL systems analyzing textbooks) or handling MWEs that are not present in the EVP. Addressing these limitations will require the development of more powerful and context-aware classifiers.

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 zeroshot 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.

⁴https://englishprofile.org/?menu= evp-terms-of-use

⁵https://www.cambridge.org/us/legal/
copyright

- 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.
- 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.
- 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.

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 Con*ference (GWC 2018), pages 314–323.
- Matthew Shardlow, Richard Evans, and Marcos Zampieri. 2022. Predicting lexical complexity in English texts: The Complex 2.0 dataset.

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').

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.

13. Appendices

13.1. Parameters and Prompts

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 content was set to "You are a professional linguist," and the temperature was set to 0 to ensure deterministic outputs.

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.

Zero-Shot and Few-Shot Classifiers. We used GPT-5 (checkpoint gpt-5-2025-08-07) as the base model for our classifiers, setting the system content to "You are an expert rater for the Common European Framework of Reference for Languages (CEFR)." and the parameter reasoning effort to "high." Figures 4 and 5 show the prompt templates used for the zero-shot and few-shot LLM classifiers, respectively. In a preliminary experiment, we provided the LLMs with full descriptions of the CEFR levels 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. Therefore, for efficiency, we opted for 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. 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 correspond to the recall for each level; thus, higher values along the diagonal

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

A1: BeginnerA2: ElementaryB1: 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 \ell_1\} Word: \{train\_word_2\}, Text: \{train\_sentence_2\} -> CEFR: \{The gold\_standard level \ell_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.

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

Table 5: Hyperparameters for the fine-tuned LLMs.

indicate greater accuracy. Because the CEFR levels are ordinal, misclassifications that fall closer to the diagonal (i.e., to an adjacent level) are less disruptive for language learners.

The ME6 Contextual models achieve high recall for the lower levels (A1 and A2) and 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–C2) are

more broadly distributed. By contrast, combining both resources (Figure 6(c)) narrows the distribution of these errors, with most misclassifications occurring between adjacent levels. This result underscores the benefit of leveraging both resources jointly.

As shown in Figure 6(d), the zero-shot LLM achieves high recall for A1 (82.8%) and moderate recall for A2 (58.5%), but its performance declines for B1 (36.9%) and is notably poor for C1 and C2 (26.0% and 12.1%, respectively). The model tends to misclassify advanced-level senses as B2 (e.g., 54.2% of C1 and 44.2% of C2), indicating a tendency to collapse more challenging senses into an intermediate level. Few-shot prompting (Figures 6(e) and 6(f)) partially mitigates this issue, improving recall for C1 and C2. However, recall for A2 declines compared to that with the zero-shot baseline, and B1 performance remains comparable, indicating 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 attains recall above 70% for levels A1–B2 and just under 70% for C2. The FT+KB model improves recall further, surpassing 80% for B1–B2, reaching approximately 91% for A1, remaining in the high-70% range for A2 and C2, and around 70% for C1. Moreover, the confusion matrices for these models show that errors are concentrated near the diagonal, indicating that they typically involve adjacent CEFR levels,

which minimizes pedagogical disruption. Despite these gains, C1 remains a challenging level, and C2 instances are often misclassified as B2. This pattern persists even when fine-tuning on the individual EVP or SemCor-CEFR corpora (Figures 6(g) and 6(h)), suggesting the issue is not an artifact of our annotation method. Instead, it likely stems from the CEFR-level distribution in the EVP data and the properties of the fine-tuning process. Addressing these residual errors will require further investigation.

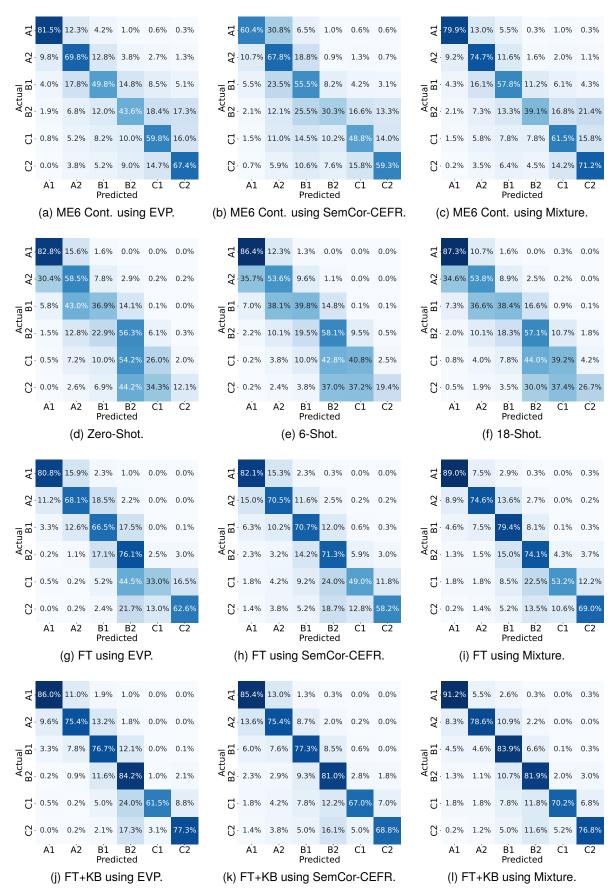


Figure 6: Confusion matrices for each classifier.