Improving CEFR Level Prediction for Multiword Expressions Using Large Language Models

Manako Yamada, Satoru Uchida, Yasutake Ishii
School of Interdisciplinary Science and Innovation, Kyushu University,
Faculty of Languages and Cultures, Kyushu University,
Faculty of Social Innovation, Seijo University
yamada.manako.979@s.kyushu-u.ac.jp, uchida@flc.kyushu-u.ac.jp, ishii@seijo.ac.jp

Abstract

Multiword expressions (MWEs) pose challenges for CEFR-level prediction because their difficulty cannot always be determined from the difficulty of their component words alone. This study investigated how prompt design and few-shot example selection affect the accuracy of phrase-level CEFR prediction using large language models (LLMs). Using ChatGPT 5.5 and Claude Opus 4.7, we compared 12 prompting conditions that combined English phrases, Japanese translations, example sentences, few-shot examples, and instructions explicitly directing the models to consider MWE characteristics such as semantic opacity, idiomaticity, formulaicity, and pragmatic functions. The results showed that directing the models to consider these characteristics improved prediction accuracy, whereas the effect of adding example sentences was limited. We then conducted a second round of predictions using items with large prediction errors in the first round as new few-shot examples. This error-based selection further improved accuracy, with Claude Opus 4.7 achieving the highest mean accuracy of 60.4%. These findings suggest that LLMs have the potential to support phrase-level CEFR prediction and that prediction accuracy can be improved by explicitly directing models to consider MWE-specific characteristics and by selecting few-shot examples based on initial prediction errors.

Keywords: multiword expressions, CEFR-level prediction, few-shot prompting

1. Background of This Research

The Common European Framework of Reference for Languages (CEFR) provides a common basis for language learning, teaching, and assessment and describes language proficiency across six levels, ranging from A1 to C2 (Council of Europe, 2001, 2020). Within the CEFR, lexical competence is regarded as a major component of language proficiency. For each proficiency level, descriptors are provided for both the breadth of vocabulary that learners can use (vocabulary range) and their ability to use vocabulary appropriately in context (vocabulary control) (Council of Europe, 2020, pp. 131–133). Vocabulary knowledge is also closely associated with overall language proficiency, and vocabulary size has been shown to serve as a useful indicator corresponding to CEFR levels (Milton & Alexiou, 2009). Accordingly, appropriately assigning CEFR levels to individual lexical items is important for assessing learners' lexical development on a common proficiency scale and for understanding its relationship with overall language proficiency.

Furthermore, information on CEFR-based lexical difficulty can be applied in various educational and assessment contexts, including the selection of learning materials appropriate to learners' proficiency levels, the sequencing of vocabulary instruction, syllabus design, and test development. Given that the CEFR is intended to provide a common basis for language education and assessment, mapping individual lexical items onto CEFR levels provides an important foundation for applying the framework to educational practice. In actual implementation, several CEFR-aligned lexical resources have been developed. However, existing research and resources have primarily focused on single words, while the coverage of multiword expressions (MWEs) remains limited. Consequently, the assessment of the difficulty of phrasal vocabulary items has received comparatively less attention.

The difficulty of MWEs cannot be determined solely from the difficulty of their component words. MWEs exhibit a range of characteristics, including non-compositionality and varying degrees of fixedness and variability, and there is no consensus regarding their definition or scope (Myles & Cordier, 2017). This diversity also makes it difficult to consistently extract and identify MWEs from corpora (Constant et al., 2017). In addition, individual MWEs occur much less frequently than their component words, making it difficult to obtain enough usage examples. For these reasons, it is challenging to comprehensively assign CEFR levels to a wide range of MWEs based solely on their frequency of occurrence or the number of learners using them in proficiency-leveled learner corpora.

In contrast, large language models (LLMs) have the potential to process phrasal items by considering not only the meanings of individual words but also their combinations, grammatical structures, contexts of use, and the functions of expressions. Previous research has shown that LLMs can be effective in estimating lexical difficulty while considering contextual information and semantic ambiguity (Bannò et al., 2025). Therefore, using LLMs to predict the CEFR levels of MWEs may provide a practical approach that complements conventional methods.

Against this background, the present study aims to investigate the potential of LLMs for predicting the CEFR levels of MWEs and to evaluate their prediction performance.

2. Related Work

2.1 CEFR-Based Lexical Profiling

Research on CEFR-based lexical profiling has sought to map lexical items onto CEFR levels using various sources of evidence, including learner corpora, teaching materials, and expert judgments. The outcomes of the research have been made available through a range of lexical resources, including the English Vocabulary Profile (EVP)

(Cambridge University Press, 2012), the Global Scale of English (GSE) Teacher Toolkit (Vocabulary) (Pearson, 2016), hereafter referred to as the GSE Vocabulary, the EFLLex (Dürlich & François, 2018), and the CEFR-J Wordlist (Tono, 2020).

One of the most widely recognized CEFR-aligned lexical resources is the EVP. The EVP assigns CEFR levels to vocabulary items (individual meanings and uses rather than word forms) that English learners are expected to understand and use at different proficiency levels. These levels are assigned based on multiple sources of evidence, including the Cambridge Learner Corpus, examination vocabulary lists, teaching materials, and native-speaker corpora (Capel, 2010, 2012). The EVP also includes some MWEs, such as phrasal verbs and idioms.

Another major CEFR-aligned lexical resource is the GSE Vocabulary. The GSE is based on the CEFR and describes English proficiency on a more fine-grained scale ranging from 10 to 90. It is intended to support materials development, the specification of learning objectives, and syllabus design (Benigno & de Jong, 2017). The GSE Vocabulary was developed using evidence from large-scale corpora together with judgments from teachers.

The CEFRLex project has developed lexical resources for multiple languages by analyzing CEFR-graded learner-oriented materials. These resources provide frequency distributions of words and MWEs at each CEFR level (Graën et al., 2020). The EFLLex is the English component of the CEFRLex project.

The CEFR-J Wordlist was developed as a CEFR-aligned lexical resource tailored to the context of English language education in Japan (Tono, 2019). CEFR-J is an adaptation of the CEFR that subdivides its proficiency levels to better reflect the proficiency of English learners in Japan, and the CEFR-J Wordlist was developed as a lexical resource to support its use. The wordlist was developed based on a corpus of English textbooks used in primary and secondary schools in China, Taiwan, and South Korea, classified by CEFR level. Although the CEFR-J framework itself consists of 12 levels ranging from pre-A1 to C2, lexical items in the CEFR-J Wordlist are classified into four levels: A1, A2, B1, and B2.

These resources have contributed to the development of CEFR-based lexical profiling. However, the criteria used to assign CEFR levels vary across resources. The EVP uses learner corpora, examination materials, teaching materials, and expert judgments. The GSE uses large-scale corpus data and teacher judgments to position vocabulary and learning objectives on a common scale, while the EFLLex is based on learner-oriented materials classified by CEFR level. Therefore, the same word or expression may be assigned to different CEFR levels across resources. Moreover, even for single words, lexical difficulty may vary depending on factors such as polysemy and the context in which a word is used.

This issue becomes even more complex in the case of MWEs. MWEs encompass a wide variety of expressions with different linguistic properties, including phrasal verbs, idioms, collocations, discourse markers, and formulaic expressions. These expressions vary considerably in terms of semantic compositionality, lexical fixedness, syntactic variability, and pragmatic function, making the criteria for assessing their difficulty more complex. Furthermore, as discussed in Section 1, identifying MWEs and consistently extracting them from proficiency-leveled learner corpora in order to determine their CEFR levels is not straightforward. For example, Negishi et al. (2012) found that the difficulty ranges of EVP phrasal verbs overlapped considerably across adjacent CEFR levels in a test administered to approximately 1,600 Japanese learners, indicating that assigned levels do not always reflect actual learner difficulty.

Consequently, no fully established method exists for mapping semantically, syntactically, and pragmatically diverse MWEs to CEFR levels in a large-scale and consistent manner based on just one of the following criteria: corpus frequency, occurrence in instructional materials, expert judgment, or learner tests.

2.2 LLM-Based Lexical Difficulty Prediction

Research on the automatic estimation of lexical difficulty has evolved from traditional machine learning methods—which utilize word frequency and morphological features—to supervised learning using contextual language models such as BERT, and further to zero-shot and few-shot estimation using general-purpose LLMs. In previous research, tasks that classified or predicted lexical complexity as a continuous value—such as Complex Word Identification (CWI) and Lexical Complexity Prediction (LCP)—were predominant. These tasks focus on lexical complexity and do not predict CEFR levels directly. In recent years, research on lexical difficulty estimation targeting CEFR levels has also been advancing.

As a study directly targeting CEFR levels, Aleksandrova and Pouliot (2023) predicted the CEFR levels of words and multiword lexical items in English and French by combining BERT-based contextualized representations with supervised classifiers. Their study highlights the importance of assessing lexical difficulty by taking into account context-dependent meanings and uses rather than relying solely on word forms. However, such supervised approaches require CEFR-labeled training data to be prepared in advance.

More recently, researchers have explored the use of general-purpose LLMs for lexical difficulty prediction without task-specific fine-tuning. Bannò et al. (2025) proposed an approach in which an LLM selects the sense listed in the EVP that best corresponds to a target word in context, and the CEFR level assigned to that sense is then used as the predicted level. This approach was effective in handling polysemy and lexical ambiguity; however, the LLM does not directly predict the CEFR level but instead selects among senses already included in the EVP. Riera Machin et al. (2026), in contrast, investigated whether LLMs could directly predict the CEFR levels of individual words. They compared multiple prompting conditions combining zero-shot and few-shot prompting with information such as CEFR descriptors and example words representing different proficiency levels. Their results showed that prediction performance was highly sensitive to prompt design and that providing additional information did not necessarily lead to improved accuracy.

These studies demonstrate that LLMs have some capacity for selecting contextually appropriate word senses and

directly predicting CEFR levels. However, previous research has focused primarily on individual words, and even when MWEs have been included, the extent to which MWE-specific characteristics affect prediction performance has not been sufficiently investigated. MWEs vary considerably in terms of semantic transparency, idiomaticity, fixedness, syntactic variability, and discourse and pragmatic functions, and their overall difficulty often cannot be determined solely from the difficulty of their component words. Furthermore, the individual effects of different prompt components, such as example sentences, translations, CEFR-related information, and few-shot examples, on prediction accuracy have not yet been fully identified.

To address these gaps, the present study focuses on MWEs and examines how different prompt components—English phrases and their Japanese translations, example sentences and their translations, few-shot examples, and gap-awareness instructions—affect the accuracy of CEFR-level prediction.

2.3 Our Previous Research

In our previous study, we investigated LLM-based CEFR-level prediction using 360 English phrases (Uchida et al., 2026). The target phrases were short, decontextualized expressions that were not embedded in sentences or discourse (see Section 4). Therefore, the LLMs were required to predict their CEFR levels based on characteristics of the expressions themselves, such as their semantic properties, combinations of component words, idiomaticity, fixedness, and pragmatic functions.

In the previous study, ChatGPT 5.4 and Claude Opus 4.6 were employed as representative general-purpose LLMs with strong performance at the time of the experiment. By comparing models from different model families, the study also examined whether the observed patterns were consistent across LLMs rather than being specific to a particular model.

The experiment compared multiple prompting conditions in which the information provided to the models was systematically varied. The baseline condition presented only the English phrase. Additional conditions provided a Japanese translation of the phrase, few-shot examples illustrating the criteria for CEFR-level judgments, or an instruction directing the models' attention to semantic and pragmatic discrepancies between the English phrase and its Japanese translation, hereafter referred to as "the semantic/pragmatic gap." These conditions were designed to examine the extent to which Japanese translations and few-shot examples improved CEFR-level prediction and whether explicitly directing the models' attention to potential semantic and pragmatic gaps relevant to Japanese learners was effective.

In the previous study, Claude Opus 4.6 achieved an accuracy of 39.1% compared with the gold standard (see Section 4) when only the English phrase was provided. Adding Japanese translations and few-shot examples resulted in higher prediction accuracy than the phrase-only condition, with the combination of the English phrase, its Japanese translation, and few-shot examples achieving the highest accuracy of 49.1%. These results suggest that, whereas phrases presented in isolation may be ambiguous in meaning and context of use, providing Japanese translations and concrete CEFR-level examples may help LLMs apply more consistent criteria when making CEFR-level judgments. A similar improvement was observed for ChatGPT 5.4 in conditions that included few-shot examples, indicating that Japanese translations and few-shot examples were beneficial across both models.

In contrast, adding a gap-awareness instruction did not improve prediction accuracy and instead resulted in a slight decrease. The instruction used in the previous study was as follows: "You are a CEFR expert. ... Be aware of the gap between the English phrase and its Japanese translation. The literal translation of individual Japanese words and the translation of the phrase as a whole may not always match, and this mismatch can be a source of difficulty. Focus on the difficulty of the English expression for a learner. For EACH phrase, determine its CEFR level (A1, A2, B1, B2, C1, or C2)." This result suggests that simply instructing LLMs to attend to discrepancies between an English phrase and its Japanese translation may not have provided sufficiently specific guidance regarding which characteristics should be considered when assessing difficulty. The difficulty of a phrase may depend not only on semantic transparency—that is, the extent to which its meaning can be inferred from its component words—but also on factors such as idiomatic meaning, conventionalized use as a fixed or formulaic expression, and pragmatic functions in interaction, including indirectness and discourse-marking functions. The gap-awareness instruction used in the previous study did not explicitly specify these dimensions, which may explain why it did not lead to improved prediction accuracy.

Meanwhile, as a method for improving the prompt itself, we also conducted additional experiments in which we reselected items that showed large errors relative to the gold standard in the initial estimation as few-shot examples. These items were used as examples intended to calibrate the models' judgments by explicitly presenting phrases that the LLMs had tended to underestimate or overestimate. The aim was to examine whether such examples could help the models adjust their criteria for CEFR-level prediction. Whereas the highest accuracy in the initial experiment was 49.1%, reselecting the few-shot examples increased the accuracy to 58.9% with Claude Opus 4.6. This substantial improvement indicates that the selection of few-shot examples can have a considerable impact on the accuracy of CEFR-level prediction.

Although the refinement of the prompting conditions improved overall prediction accuracy, a closer analysis of the misclassification patterns revealed that substantial challenges remained for higher-level items. Even under the condition that achieved the highest overall accuracy, only 1 of the 13 C1 items was correctly predicted, corresponding to an accuracy of 7.7%, which was considerably lower than the accuracy observed for items at the A1–B2 levels. This finding indicates that LLM prediction performance declined substantially for higher-level phrases. Some higher-level phrases consist of relatively basic component words but convey more advanced idiomatic meanings or pragmatic functions as whole expressions. One possible explanation for these misclassifications is that phrases presented in isolation may not provide sufficient information for the models to identify their intended meanings and contexts of use.

The previous study also did not examine the effects of providing example sentences and their Japanese translations. Because phrases presented without context may be ambiguous in meaning and use, example sentences may help clarify their meanings, contexts of use, and pragmatic functions. This may lead to more accurate CEFR-level predictions.

To address the issues identified in the previous study, this study uses the same dataset and extends the previous research by employing ChatGPT 5.5 and Claude Opus 4.7. In the newly conducted experiments, we updated the models and adopted the method for selecting few-shot examples that proved effective in the previous study. Furthermore, we verify the impact on the accuracy of CEFR level estimation for MWEs of "refined gap-awareness instructions"—which more explicitly account for semantic opacity, idiomaticity, formulaicity, fixed collocations, discourse and pragmatic functions, and the difficulty of acquisition for Japanese EFL learners—as well as the presentation of example sentences and their Japanese translations.

3. Research Goals and Research Questions

Against this background, the present study aims to systematically investigate which prompt components contribute to improving the accuracy of CEFR-level prediction for MWEs using LLMs. In particular, the study focuses on the effects of instructions directing the models to consider the semantic/pragmatic gap between English phrases and their Japanese translations, as well as the provision of example sentences and their Japanese translations. Accordingly, the following research questions were formulated:

RQ1: Do refined instructions directing LLMs to consider semantic/pragmatic gaps between English phrases and their Japanese translations increase accuracy?

RQ2: Does providing example sentences and their translations improve predictions?

4. Dataset

We used Ishii (2026) as the dataset. This is a phrase-based vocabulary study book designed to support beginner- to intermediate-level Japanese EFL learners in acquiring high-frequency, pedagogically relevant formulaic phrases. The study book consists of 50 high-frequency lexical verb and noun headwords (e.g., *get*, *go*, *know*, *say*, *think*; *day*, *person/people*, *thing*, *time*, *way*). The headwords were selected from among the top frequency words in three large corpora, i.e., the Spoken BNC 2014, COCA (TV/MOVIES, SPOKEN, and FICTION sections), and enTenTen21. Words with the lowest summed weighted ranks were selected to prioritize items that are both frequent and communicatively useful. Several headwords appear two or three times, totaling 72 units.

Each headword entry is accompanied by five pedagogically relevant formulaic phrases, yielding 360 phrases in total. Example phrases given under the headword *go* include *go home*, *go shopping*, *go bad*, *go up* ('increase'), *go on talking*, *go to the bathroom*, *to go* ('for takeout'), *I have to go.* (used to end a conversation), and *go ahead*. The full set of phrases includes collocations (e.g., *make a choice*, *in person*), phrasal verbs (e.g., *find out*, *take off*),

Figure 1: Sample entry from Ishii (2026)

communicative expressions (e.g., *Could you say that again?*, *Nice talking to you.*, *No way!*, *You might want to ...*), and other (semi-)fixed phrases (e.g., *as far as I know*, *I've heard a lot about ...*, *see ... with my own eyes*). Each phrase is accompanied by a Japanese translation.

The formulaic phrases associated with each headword were chosen through a combined approach drawing on (1) corpus frequency, (2) inclusion in learner dictionaries as sense units or listed phrases, and (3) coverage in CEFR-related resources (e.g., the EVP and the GSE Vocabulary). Expressions were further screened for their communicative relevance to Japanese learners, especially in everyday conversational contexts. Kawamura (2024) was referred to for this purpose, as it lists around 500 functional expressions that are considered relevant for Japanese EFL learners.

Each expression was manually assigned a CEFR level on a six-band scale (A1–C2) by the author based on several factors: the difficulty of the lexical and grammatical components (the EVP, the GSE Vocabulary, the CEFR-J Wordlist, the English Grammar Profile (Cambridge University Press, 2015), and the CEFR-J Grammar Profile (Ishii, 2025)), semantic transparency, pragmatic complexity and communicative relevance to Japanese learners (Kawamura, 2024), the degree of cross-linguistic mismatch with Japanese, and levels of the phrases indicated in CEFR-related reference materials when available. For example, *a young child* and *for a long time* are rated at A2, whereas *make sense* and *You know what?* are rated at B2. The goal was to provide learners and instructors with a principled indication of learning necessity and difficulty that reflects both linguistic and pedagogical considerations. No phrase was assigned to C2. The CEFR levels assigned in this work were used as the gold standard in the present study.

Each phrase is also illustrated with two short spoken-style example sentences, accompanied by their Japanese translations. The example sentences were created to be short and simple so that learners can easily visualize the contexts in which the phrase is used.

Figure 1 is a sample page of the entry *want*. This verb is classified at A1 in the EVP, while the phrases given under this entry (i.e., *Do you want a coffee?*, *if you want*, *I want you to know ...*, *I just wanted to ...*, *You might want to ...*) are rated at higher levels. The reasons for this discrepancy

include certain communicative functions served by these phrases, the use of a higher-level word (*might* (POSSIBLY TRUE) is judged to be at B1 in the EVP), and a complicated grammatical structure (*want someone to do*; cf. *do you want me to...?* rated at A2+ in the GSE Vocabulary).

5. Experimental Setup

The present study followed the experimental design of our previous study (see Section 2.3) while using the latest models available from the same model families at the time of the experiment. Whereas the previous study used ChatGPT 5.4 and Claude Opus 4.6, the present study employed their respective successor models, ChatGPT 5.5 and Claude Opus 4.7. To ensure comparability with the previous study, the basic experimental design was maintained as closely as possible, except for the models used and the experimental modifications described below.

The dataset consisted of 360 English phrases from Ishii (2026), together with their Japanese translations, assigned CEFR levels, which were treated as the gold standard, two English example sentences for each phrase, and Japanese translations of each example sentence (see Section 4).

Each LLM was instructed to assign a CEFR level to each target phrase and to return only the predicted level, without any reasoning process or explanation. The experiments were conducted via OpenRouter (https:// openrouter.ai) using the default parameter settings. Each estimation was performed independently, ensuring that the input data and output results from the previous prediction were not carried over to subsequent predictions. For each prompt condition, predictions were conducted five times, and the accuracy was averaged across the five runs. In this study, 12 prompt conditions were defined by progressively varying the input of information and instructions (Table 1). The baseline condition presented only the English phrase, to which the Japanese translation, an instruction to consider the semantic/pragmatic gap, few-shot examples, example sentences, and Japanese translations of the example sentences were progressively added.

Cond. #	Prompt condition	Cond. #	Prompt condition
1	Phrase only	7	Phrase + JP + Gap2
2	Phrase + JP	8	Phrase + JP + Gap2 + FS
3	Phrase + JP + Gap1	9	Phrase + JP + Ex1 + Ex1-JP
4	Phrase + FS	10	Phrase + JP + Ex1 + Ex1-JP + Ex2 + Ex2-JP
5	Phrase + JP + FS	11	Phrase + JP + Ex1 + Ex1-JP + Gap2 + FS
6	Phrase + JP + Gap1 + FS	12	Phrase + JP + Ex1 + Ex1-JP + Ex2 + Ex2-JP + Gap2 + FS

Table 1: Prompt conditions used in the experiments

Note.
JP: Japanese translation of the phrase
Gap1: Instruction to consider the semantic/pragmatic gap
Gap2: Refined semantic/pragmatic gap awareness instruction
FS: Few-shot examples
Ex1: Example sentence 1
Ex1-JP: Japanese translation of example sentence 1
Ex2: Example sentence 2
Ex2-JP: Japanese translation of example sentence 2

For the Gap1 instruction, see Section 2.3. The Gap2 instruction, a refined version of the instruction used in the previous study, was as follows: "You are a CEFR expert. ... For EACH phrase, determine its CEFR level (A1, A2, B1, B2, C1, or C2). When estimating the CEFR level of each phrase, pay attention to the difficulty of understanding and acquiring the phrase for English learners whose first language is Japanese. For example, if the semantic transparency of the phrase is low, if the communicative function performed by the phrase is advanced, if the meaning of the phrase cannot be fully captured by a simple literal Japanese translation, or if the phrase is highly idiomatic and formulaic, assume that the phrase is likely to correspond to a higher CEFR level."

The experiment was conducted in two rounds. In the first round, a fixed set of few-shot examples covering CEFR levels from A1 to C1 was used to clarify the criteria for CEFR-level judgments. Specifically, two items were selected from each CEFR level, resulting in a total of 10 items (Table 2), and these were presented in the prompt as labeled reference examples. These examples served as anchors for the models' CEFR-level judgments and were excluded from the accuracy evaluation. The remaining 350 items were used for evaluation.

Phrase	Manually assigned level
get up	A1
take a picture	A1
make a phone call	A2
Take an umbrella with you.	A2
I mean it.	B1
local people	B1
come along	B2
There is no such thing as ...	B2
call a meeting	C1
pull the car over	C1

Table 2: Few-shot examples used in the first round

In the second-round predictions, based on the results of the most accurate first-round prediction (Condition 11, run 1 on Claude Opus 4.7), 12 items showing large discrepancies between the gold-standard and predicted CEFR levels were selected as error-based few-shot examples. The 12 items (Table 3) were presented as few-shot examples together with their correct CEFR levels. These items were excluded from the evaluation, and the remaining 348 items were evaluated. For the six conditions containing few-shot examples (Conditions 4, 5, 6, 8, 11, and 12), predictions were conducted again five times for each model.

We evaluated accuracy based on exact matches with the gold-standard CEFR levels. Specifically, a prediction was considered correct when the CEFR level predicted by the LLM matched the manually assigned CEFR level. Accuracy was then calculated for each model and prompt condition.

Phrase	Manually assigned level	Estimated level
go on talking	B2	A2
I see.	B1	A1
What do you mean?	B1	A1
can tell ...	C1	B1
Leave me alone.	B2	A2
see what happens	C1	B1
It's been a long time.	B2	A2
local time	C1	B1
people person	A2	B2
in place	C1	B1
Good question.	B2	A2
next door	B2	A2

Table 3: Few-shot examples used in the second round

6. Results

6.1 First-Round Predictions

Table 4 presents the results of first-round predictions. The highest mean accuracy across the five runs was 52.5%, achieved by Claude Opus 4.7 under Condition 11 (Phrase + JP + Ex1 + Ex1-JP + Gap2 + FS). Among the individual runs, the highest accuracy was 55.4%, also achieved by Claude Opus 4.7 under Condition 11.

Cond. #	Prompt condition	Opus-4.7 Mean (*SD*)	GPT-5.5 Mean (*SD*)	Best single-run accuracy
1	Phrase only	43.9% (1.9)	45.7% (1.5)	47.1% (gpt-5.5, run 2)
2	Phrase + JP	45.2% (1.9)	45.8% (1.9)	47.4% (gpt-5.5, run 2)
3	Phrase + JP + Gap1	43.4% (1.3)	46.1% (1.3)	48.3% (gpt-5.5, run 3)
4	Phrase + FS	45.1% (1.7)	44.4% (1.3)	46.9% (opus-4.7, run 2)
5	Phrase + JP + FS	43.8% (2.0)	46.5% (1.1)	48.0% (gpt-5.5, run 4)
6	Phrase + JP + Gap1 + FS	45.8% (1.4)	48.1% (1.6)	50.3% (gpt-5.5, run 3)
7	Phrase + JP + Gap2	47.7% (1.4)	47.8% (0.9)	49.1% (opus-4.7, run 3)
8	Phrase + JP + Gap2 + FS	49.8% (2.2)	50.5% (0.9)	52.6% (opus-4.7, run 1)
9	Phrase + JP + Ex1 + Ex1-JP	46.7% (1.7)	46.5% (0.9)	48.9% (opus-4.7, run 1)
10	Phrase + JP + Ex1 + Ex1-JP + Ex2 + Ex2-JP	48.9% (1.9)	47.8% (1.6)	50.6% (opus-4.7, run 2)
11	Phrase + JP + Ex1 + Ex1-JP + Gap2 + FS	52.5% (2.2)	50.3% (1.4)	55.4% (opus-4.7, run 1)
12	Phrase + JP + Ex1 + Ex1-JP + Ex2 + Ex2-JP + Gap2 + FS	51.1% (1.6)	50.4% (1.3)	53.4% (opus-4.7, run 2)

Table 4: Accuracy of first-round predictions

Note.
n = 350 per run

Comparing conditions with and without Gap2, the mean accuracy of Claude Opus 4.7 increased slightly from 45.2% under Condition 2 (Phrase + JP) to 47.7% under Condition 7 (Phrase + JP + Gap2). Similarly, the mean accuracy increased from 43.8% under Condition 5 (Phrase + JP + FS) to 49.8% under Condition 8 (Phrase + JP + Gap2 + FS), indicating higher accuracy when the Gap2 instruction was added. For the conditions with example sentences, the mean accuracy was 52.5% under Condition 11. When a second example sentence and its Japanese translation were added in Condition 12, the mean accuracy decreased slightly to 51.1%.

6.2 Second-Round Predictions

Table 5 presents the results of the second-round predictions using the error-based few-shot examples. For Claude Opus 4.7, Condition 12 (Phrase + JP + Ex1 + Ex1-JP + Ex2 + Ex2-JP + Gap2 + FS) achieved a mean accuracy of 60.4% (SD = 0.7) across the five runs, which was the highest mean accuracy among all conditions. Among the individual runs, the highest accuracy was 61.5%, achieved by Claude Opus 4.7 under Condition 12.

Cond. #	Prompt condition	Opus-4.7 Mean (*SD*)	GPT-5.5 Mean (*SD*)	Best single-run accuracy
4	Phrase + FS	54.8% (1.9)	54.4% (1.6)	57.8% (opus-4.7, run 2)
5	Phrase + JP + FS	55.9% (0.6)	55.9% (1.9)	58.6% (gpt-5.5, run 4)
6	Phrase + JP + Gap1 + FS	55.7% (0.9)	55.6% (0.7)	57.2% (opus-4.7, run 3)
8	Phrase + JP + Gap2 + FS	58.8% (1.7)	56.3% (1.2)	60.3% (opus-4.7, run 5)
11	Phrase + JP + Ex1 + Ex1-JP + Gap2 + FS	58.9% (0.8)	54.3% (0.9)	60.1% (opus-4.7, run 2)
12	Phrase + JP + Ex1 + Ex1-JP + Ex2 + Ex2-JP + Gap2 + FS	60.4% (0.7)	52.4% (0.8)	61.5% (opus-4.7, run 3)

Table 5: Accuracy of second-round predictions

Note.
n = 348 per run

For Claude Opus 4.7, the conditions combining Gap2 and example sentences with few-shot examples (Conditions 8, 11, and 12) all achieved high mean accuracies.

Compared with the first-round predictions, Claude Opus 4.7 showed substantial increases in mean accuracy across all conditions examined in the second round. In particular, under Condition 12, the mean accuracy increased from 51.1% in the first round to 60.4% in the second round, representing an improvement of 9.3 percentage points. For GPT-5.5, notable improvements were observed under Conditions 4, 5, 6, and 8, whereas Conditions 11 and 12 did not show improvements comparable to those observed for Claude Opus 4.7.

7. Discussion and Conclusion

The present study investigated the effects of prompts that explicitly specified evaluation criteria, the provision of example sentences, and the selection of few-shot examples on the accuracy of CEFR-level prediction for English phrases using LLMs. The results showed that refined instructions explicitly directing the models to consider the semantic/pragmatic gap contributed to improved prediction accuracy, whereas increasing the number of example sentences did not necessarily lead to consistent

improvements in accuracy (compare Conditions 9 with 10, and 11 with 12). In addition, using items with large prediction errors in the initial predictions as few-shot examples was shown to have the potential to further improve prediction accuracy.

7.1 Effect of Refined Instructions

Regarding RQ1 (Do refined instructions directing LLMs to consider semantic/pragmatic gaps between English phrases and their Japanese translations increase accuracy?), the conditions including Gap2 in the first-round predictions achieved higher accuracy than the corresponding conditions with Gap1 or those without Gap2. This pattern suggests that the refined instructions contributed to improving CEFR-level prediction accuracy by explicitly directing the models to consider the semantic/pragmatic gap.

In idiomatic and set expressions, even when all constituent words are basic words, knowledge regarding semantic opacity, idiomaticity, pragmatic functions, and contextual usage is often required. Research on second language acquisition has also pointed out that the acquisition of multiword expressions presents challenges distinct from those associated with single-word vocabulary knowledge (Myles & Cordier, 2017). The above results suggest a possible approach for enabling LLMs to better capture these challenges, which differ from those associated with individual word knowledge.

When only the target phrases were presented, the LLMs may have relied primarily on the frequency of the component words and their general lexical difficulty when making CEFR-level judgments. As a result, higher-level phrases composed of relatively basic words may have been underestimated. In contrast, the refined instructions used in the present study explicitly directed the models to consider semantic transparency, idiomaticity, formulaicity, pragmatic functions, and the difficulty of acquiring the expressions for Japanese learners of English. As a result, the models may have been more likely to incorporate characteristics of the phrases, rather than relying solely on the surface-level difficulty of their component words, which may have contributed to the improvement in prediction accuracy.

7.2 Effect of Example Sentences

Regarding RQ2 (Does providing example sentences and their translations improve predictions?), no consistent improvement in prediction accuracy was observed in the conditions where example sentences and their Japanese translations were provided. In the first-round predictions, the mean accuracy under Condition 11 with Claude Opus 4.7, which included one example sentence, was 52.5%, exceeding the 51.1% obtained under Condition 12, which included two example sentences. However, because Conditions 11 and 12 both combine multiple prompt components, the independent effect of providing example sentences cannot be determined from this comparison alone. These results suggest that, at least under the conditions examined in the present study, simply increasing the number of example sentences did not lead to improved accuracy in CEFR-level prediction.

One possible reason why increasing the number of example sentences did not improve the accuracy is that the example sentences and their Japanese translations used in the present study may not have provided sufficient additional information beyond what was already available from the phrases and their Japanese translations. Even if the example sentences illustrated the basic meanings or typical contexts of use of the phrases, if such information could be easily inferred from the other input provided to the LLMs, the examples may not have provided additional cues for CEFR-level judgments and therefore may not have contributed to improved prediction accuracy.

A second possible reason is that the example sentences may have highlighted the simplicity of the sentences rather than the inherent difficulty of the target phrases. Even when a phrase performs an advanced pragmatic function, if it is embedded in a short example sentence consisting primarily of basic words, the LLMs may perceive the sentence as relatively easy and consequently underestimate the CEFR level of the target phrase.

Taken together, these findings suggest that the effectiveness of providing example sentences may depend not simply on whether examples are provided, but on the extent to which the example sentences and their Japanese translations provide information that is not already available from the other inputs.

7.3 Selection of Few-Shot Examples

One of the key findings of the present study is the effectiveness of using items with large prediction errors in the initial predictions as few-shot examples. In the second-round predictions, presenting items that had shown large prediction errors in the first-round predictions as new few-shot examples helped adjust the models' criteria for CEFR-level judgments and improved prediction accuracy. Furthermore, conditions combining these error-based few-shot examples with Gap2 and example sentences consistently achieved high prediction accuracy in the second-round predictions. This suggests that combining error-based few-shot examples with information directing the models' attention to the semantic and pragmatic characteristics of the phrases may have contributed to improved prediction accuracy.

These findings suggest that few-shot examples may serve not only to demonstrate the task format but also to adjust the models' criteria for CEFR-level judgments. Many of the items with large prediction errors were expressions that had been manually assigned relatively high CEFR levels despite being composed of basic words, or expressions whose semantic and pragmatic characteristics were difficult to infer from their surface forms. By presenting these items together with their correct CEFR labels, the LLMs may have been encouraged to take into account not only the difficulty of the individual component words but also phrase-level characteristics such as idiomaticity and pragmatic function.

The selection of few-shot examples has already been shown to affect the performance of in-context learning. Gupta et al. (2023) demonstrated that model performance can be improved by selecting a set of few-shot examples that collectively cover important features of the input and a broad range of reasoning patterns. However, to the best of our knowledge, no previous study has examined the reuse of items with large errors in initial predictions as few-shot examples for CEFR-based phrase-level lexical difficulty

prediction. Therefore, the findings of the present study can be regarded as preliminary evidence for the effectiveness of error-based example selection in CEFR-level prediction of phrases using LLMs.

7.4 Potential and Limitations of LLM-Based CEFR-Level Prediction

The results of the present study demonstrate that LLMs have the potential to serve as a promising tool for supporting CEFR-based assessment of phrase-level lexical difficulty. With appropriate prompt design and few-shot examples, LLMs were able to estimate CEFR levels by considering, to some extent, not only word-level difficulty but also characteristics specific to MWEs, such as semantic opacity, idiomaticity, formulaicity, and pragmatic functions.

At the same time, several limitations were identified. First, even under the most accurate condition, the exact agreement rate remained at approximately 60%, and higher-level phrases continued to be underestimated. Second, predictions varied across runs even under the same conditions. Although this variation was not large, the consistency of LLM judgments remains an issue. Third, the gold standard used in the present study has not yet been sufficiently validated against actual comprehension and production data from Japanese learners of English.

Therefore, at present, LLMs should not be regarded as a complete replacement for human assessment. Rather, they may be more appropriately used for large-scale preliminary assessment or as a tool to support human evaluation.

7.5 Future Research

Future research should systematically examine the quality of example sentences. Although the effect of providing example sentences was limited in this study, this does not necessarily mean that example sentences are ineffective. Designing example sentences that more clearly illustrate semantic opacity, idiomaticity, and pragmatic functions may improve prediction accuracy.

For few-shot examples, systematic selection criteria should also be developed by considering not only the magnitude of prediction errors but also factors such as the distribution of CEFR levels, expression types, semantic transparency, and pragmatic features. It is also important to evaluate the stability of LLM outputs by conducting multiple predictions under the same conditions and to compare the performance of multiple models. In addition, the validity of the manually assigned CEFR levels should be further examined by comparing them with comprehension and production data from Japanese learners of English and by assessing inter-rater reliability.

Overall, this study demonstrates the potential of LLMs to support CEFR level prediction for MWEs, while also showing that their performance depends substantially on prompt design and few-shot example selection. Systematic improvements in these areas may further enhance the reliability of LLM-based phrase-level lexical difficulty assessment.

8. Acknowledgements

This work was partly supported by JSPS KAKENHI Grant Numbers JP22H00677 / JP24K04167.

9. Bibliographical References

Aleksandrova, D., & Pouliot, V. (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) (pp. 518–527). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.43

Bannò, S., Knill, K. M., & Gales, M. J. F. (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) (pp. 632–646). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.bea-1.45

Benigno, V., & de Jong, J. (2017). *The Global Scale of English: Learning objectives for English language proficiency*. Pearson.

Capel, A. (2010). A1–B2 Vocabulary: Insights and issues arising from the English Profile Wordlists. *English Profile Studies*.

Capel, A. (2012). Completing the English Vocabulary Profile: C1 and C2 vocabulary. English Profile Studies.

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., & Todirascu, A. (2017). Multiword expression processing: A survey. *Computational Linguistics*, *43*(4), 837–892.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press.

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment: Companion volume*. Council of Europe Publishing.

Dürlich, L., & François, T. (2018). EFLLex: A graded lexical resource for learners of English as a foreign language. In Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018) (pp. 873–879). European Language Resources Association.

Graën, J., Alfter, D., & Schneider, G. (2020). Using Multilingual resources to evaluate CEFRLex for learner applications. In Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020) (pp. 346–355). European Language Resources Association.

Gupta, S., Gardner, M., & Singh, S. (2023). Coverage-based example selection for in-context learning. In Findings of the Association for Computational Linguistics: EMNLP 2023 (pp. 13924–13950). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.930.

Ishii, Y. (2026). *Kiite patto hanasu shunkan eisakubun training: Kikusaku [Kikusaku: Instant English composition training through listening and speaking]*. Alc Press.

Kawamura, A. (Ed.) (2024). *Gurobaru shakai no eigo komyunikeshon handobukku: Hatsuwakoui poraitonesu hyougen jiten tsuki [The English communication handbook for a global society: With a dictionary of speech acts and politeness expressions]*. Sanseido.

Milton, J., & Alexiou, T. (2009). Vocabulary size and the Common European Framework of Reference for Languages. In B. Richards, M. Daller, D. D. Malvern, P. Meara, J. Milton, & J. Treffers-Daller (Eds.),

Vocabulary studies in first and second language acquisition (pp. 194–211). Palgrave Macmillan.

Myles, F., & Cordier, C. (2017). Formulaic sequence (FS) cannot be an umbrella term in SLA: Focusing on psycholinguistic FSs and their identification. *Studies in Second Language Acquisition*, *39*(1), 3–28.

Negishi, M., Tono, Y., & Fujita, Y. (2012). A validation study of the CEFR levels of phrasal verbs in the English Vocabulary Profile. *English Profile Journal, 3*, Article e3.

Riera Machin, M. A., Nohejl, A., & Watanabe, T. (2026). Using the CEFR for guiding LLMs in lexical complexity prediction. In Proceedings of the 32nd Annual Meeting of the Association for Natural Language Processing.

Tono, Y. (2019). Coming full circle—From CEFR to CEFR-J and back. *CEFR Journal—Research and Practice*, *1*, 5–17.

Uchida, S., Yamada, M., & Ishii, Y. (2026, May). When words look easy but usage isn't: Functional complexity and the persistent difficulty of CEFR-level prediction by LLMs [Paper presentation]. International Conference on Applied Language Sciences (ALS-Methoken 2026), Hong Kong.

10. Language Resource References

Cambridge University Press. (2012). *English Vocabulary Profile*. English Profile. https://englishprofile.org/?menu=english-vocabulary-profile

Cambridge University Press. (2015). *English Grammar Profile*. English Profile. https://englishprofile.org/?menu=english-grammar-profile

Ishii, Y. (2025). *CEFR-J Grammar Profile 2025*. CEFR-J Project. https://www.cefr-j.org/PDF/sympo2020/CEFRJGP_JPTEXTBOOK2025.pdf

Pearson. (2016). *GSE Teacher Toolkit*. Global Scale of English. https://www.english.com/gse/teacher-toolkit/user/vocabulary

Tono, Y. (2020). *CEFR-J Wordlist (Version 1.6)*. CEFR-J Project. https://www.cefr-j.org/download.html#cefrj_wordlist