Graduate School of Science and Technology Master's Thesis Abstract

Laboratory name (Supervisor)	Natural Language Processing (WATANABE TARO (Professor))		
Student ID	2411339	Submission date	2026 / 1 / 19
Name	RIERA MACHIN MARIA ANGELICA		
Thesis title	Using the CEFR for guiding LLMs in Lexical Complexity Prediction		

Abstract

Lexical complexity prediction is a key task for language teaching, as word difficulty varies across learners and proficiency levels. While prior work has largely focused on generating CEFR-aligned content, less attention has been paid to whether lexical complexity judgments themselves align with CEFR standards. This work focuses on whether large language models (LLMs) can reliably predict the CEFR (Common European Framework of Reference for Languages) level of words, using a zero-shot and few-shot prompt setting. We use three independent CEFR-annotated lexical datasets as a reference for evaluation. Results show that performance can vary substantially depending on prompt design: simple zero-shot prompts and prompts with level-specific examples produce the most stable predictions, while prompts that include CEFR descriptions tend to bias models toward higher levels. Across datasets, which vary in how they classify certain words, models seem to align more strongly with one specific dataset, suggesting that LLM judgments reflect underlying patterns or biases in specific lexical resources depending on what data is more readily available during their training. These findings highlight both the potential and limitations of LLMs for CEFR-based lexical assessment.