OPEN ACCESS

Vocabulary Learning and Instruction
ISSN 2981-9954
Volume 14, Number 2 (2025)
https://doi.org/10.29140/vli.v14n2.103597

Castledown

A Commentary on Uchida (2025) and Kang and Nichol (2025)

Jeff Stewart
Tokyo University of Science, Japan
jeffjrstewart@gmail.com

Abstract

It has been my pleasure to read and review the studies presented at the 2024 JALT Vocabulary SIG Symposium. In this short paper I will overview two of those studies, conducted by Uchida and Kang & Nichol, and share my thoughts on them.

Uchida, S. (2025). Generative AI and CEFR Levels: Evaluating the Accuracy of Text Generation with ChatGPT-4o Through Textual Features

Recent developments in large language models (LLMs) such as ChatGPT have prompted a surge of research into their potential applications within L2 vocabulary acquisition and the fields of computer-assisted language learning and second language acquisition more broadly. As Uchida's study demonstrates, the use of ChatGPT (OpenAI, 2025) for generating texts categorized by CEFR (Common European Framework of Reference for Languages) (Council of Europe, 2001) level represents an emerging direction in this area. The broader trend suggests that LLM-based artificial intelligence is not a passing fad but a transformative technology that will continue to shape language education and assessment practices (whether we like it or not).

Data Availability Statement: All relevant data are within this paper.

Research on uses of ChatGPT and similar models in SLA has focused primarily on three domains: test item writing, automated scoring and assessment, and materials development. Several testing organizations have already begun adopting these tools, particularly where large-scale content generation can improve efficiency. Although the text produced by LLMs can be generic and predictable, these characteristics are arguably beneficial (or at least unproblematic) when writing for readers of an L2. Frankly, the temptation for publishers and test makers to reduce labor costs by relying on AI-generated readings will likely lead to their widespread use very soon. However, questions remain regarding the actual quality and pedagogical appropriateness of AI-generated materials. Given this, it is necessary to evaluate their ability to level texts for learners appropriately, as Uchida does in this first study.

Most current studies employ "generalist" LLMs such as OpenAI's ChatGPT, systems trained on a vast array of data and capable of performing a wide range of tasks, yet rarely excelling at any specific one. While AI-generated materials and automated scoring outputs often appear accurate and coherent at first glance, closer examination frequently reveals more subtle flaws. In the future, general artificial intelligence may be obtained and AI may begin to not only match but, perhaps, even exceed humans in many intellectual endeavors. But as of this writing, publicly available LLMs are often less akin to geniuses who can do work for human specialists better than the specialists can do it themselves and more analogous to having hundreds of not-so-brilliant people at our command who can do our menial tasks for us. However, it can only do these tasks if we explain them in sufficient detail.

In practice, this means that effective use of LLMs currently depends on careful and explicit prompt engineering. Users must provide highly detailed instructions to guide the model's output toward the desired results. Yet doing so presupposes that educators and test developers themselves have clearly defined and operationalizable goals. In language education, and particularly in materials or assessments aligned to the CEFR, such clarity is often lacking. Many aspects of CEFR-level differentiation still rely heavily on intuitive judgments and qualitative descriptors rather than measurable linguistic parameters.

Uchida's study focused on using ChatGPT to generate texts across different CEFR levels and then evaluating those texts with an objective, data-driven approach. To do this, he employed the CEFR-Based Vocabulary Level Analyzer (CVLA), developed by Uchida and Negishi (2018). This tool is particularly notable because it provides a quantitative framework for assessing CEFR alignment, a welcome contrast to the often intuitive or impressionistic methods used in language materials development.

The CVLA was built on a large corpus of textbook materials that had been explicitly written to match various CEFR levels. These materials were analyzed using a range of lexical complexity formulas designed to capture different aspects of linguistic difficulty. Among the metrics used by the CVLA are:

- **ARI (Automated Readability Index):** A traditional readability formula that combines word and letter length with words per sentence.

- **VperSent (Verbs per Sentence):** The number of verbs appearing per sentence. This metric reflects grammatical complexity, as more complex verb forms (e.g., "was being written" versus "wrote") tend to appear in higher-level texts.
- **AvrDiff (Average Difficulty):** A measure of word difficulty based on the CEFR-J Wordlist (Tono, 2019). Importantly, this metric relies on CEFR-level categorizations rather than raw or log-transformed frequency data, meaning it does not include C-level distinctions.
- **BperA (B-to-A ratio):** The ratio of CEFR B-level words to A-level words. A higher ratio indicates greater lexical difficulty.

Although these formulas are relatively simple, they have the advantage of being transparent and easily adjustable, which makes them especially practical for materials writers who need to fine-tune texts to match specific difficulty bands.

Using these metrics, Uchida and colleagues analyzed the corpus of CEFR-graded textbook materials and then derived a regression formula that could be used to "work backward", that is, to predict or generate text that fits a desired CEFR level. This approach allowed for a systematic comparison between human-written materials and AI-generated texts, shedding light on how well ChatGPT can mimic the linguistic characteristics of level-appropriate educational content.

Uchida found that for "zero-shot" prompts with no example text, the vast majority of A1 and A2 targeted texts created by ChatGPT were judged as "PreA1" by CVLA. B1 texts showed great variability and were judged as ranging from A1.1 to B1.2 proficiency. B2-targeted texts and above were nearly all rated at the most difficult C2 level. Although providing example texts improved appropriateness somewhat (particularly for A2-targeted texts), the general pattern was the same, and Uchida concluded that "the A1 and A2 levels tend to be overly simplified, B1 exhibits some variability, and B2, C1, and C2 levels are excessively complex." Uchida's research therefore revealed very middling success in using LLMs as generators of CEFR-aligned reading materials, which may underscore not just technical limitations of current LLMs for such tasks but also deeper conceptual challenges inherent in CEFR-based text writing more generally. As I will argue, LLMs' effectiveness for this purpose may depend on a deeper theoretical and empirical understanding of how linguistic complexity aligns with CEFR descriptors to begin with.

The most straightforward solution to lackluster LLM output is more detailed prompt engineering. The AI needs to be told what to do in great detail so that it knows exactly what we want. But for it to get that, we need to know what we want. However, a lot of the materials researchers and educators make for EFL and ESL learners are still based largely on intuition and qualitative descriptions of targets and goals. This is often true of materials and tests aimed at the CEFR, which may itself be poorly understood even by human materials writers. Uchida's results may therefore illustrate not just limitations with current LLMs, but also potential challenges with writing to the CEFR. Table 1 below collapses Uchida's results into simpler A-B-C categories, greatly improving accuracy rates for the highest and lowest classifications.

Table 1 *Zero-Shot Texts, Original Table, Reduced to Standard CEFR Levels only and Reduced to A-C Categories only*

Category interpretation		"Beginner"	"Intermediate"	"Advanced"
Classification rate		100%	15%	100%
Targeted text level		A	B	C
CVLA Classification	**A**	*60*	**21**	
	B		*9*	
	C		**30**	*60*
	total	30	60	60

Based on Uchida's results, ChatGPT appears to separate English text into roughly three difficulty modes:

1. "Beginner" (simplified English);
2. A much more variable and haphazard "Intermediate" level; and
3. "Advanced" (college level English).

The intermediate category is much less clearly defined and does not appear to adhere to a consistent set of rules or guidelines. However, do human materials writers do all that better?

To illustrate this possibility, below is a similar analysis of readings taken from the EFL Textbook Business Plus 3 (Helliwell, 2015), which was written prior to the advent

Table 2 *CVLA Reading Mode Analyses for "Asia Online" Text Readings from the Textbook Business Plus 3 ("A2-B1"), Cambridge Press*

	"A2-B1"
PreA1	
A1.1	
A1.2	
A2.1	
A2.2	
B1.1	
B1.2	1
B2.1	
B2.2	1
C1	3
C2	1
Total	6

of ChatGPT. The textbook was rated at A2-B1 CEFR levels by its publisher. However, if the reading exercises from its units are examined with the CVLA, one text rated at the B1.2 level by CVLA on Reading mode, one B2.2 level, and 4 texts at the C1 level or higher. The similar patterns of miscategorization suggest that difficulty leveling texts to CEFR guidelines may extend to human materials writers.

In regard to the topics AI chose for given CEFR levels, if we re-sort the topics in Uchida's Table 8 we see the following pattern by CEFR level:

Table 3 *ChatGPT-Generated Topics by CEFR Level, Adapted from Uchida*

Topic	A1	A2	B1	B2	C1	C2	total
daily_life	28	28					56
family	**2**						2
hobbies		**2**					2
travel			**17**				17
health			**12**				12
culture			1	1			2
environment				**13**	6		19
sustainability				6	**9**	1	16
climate_change				2	**7**	4	13
technology				4	4	3	11
SNS					3		3
AI					1	**19**	20
others				4		3	7
total	30	30	30	30	30	30	180

As Uchida notes, for levels A1 and A2: the topics are themes related to daily life, such as family and hobbies. For B1, there is a focus on health and travel. For B2, themes such environment and sustainability commonly re-occur. For C1 we see issues such as sustainability and climate change are common, and for C2 AI is prevalent. Uchida concluded that A1 and A2 levels appeared to be overly simplified, B1 showed variability, and B2, C1 and C2 levels were excessively complex.

However, are the topics per level inappropriate given the CEFR level descriptors? Refer to the descriptors below. Judged qualitatively, the main separator between levels appears to be between B1-B2 for reading as we move from "standard English but on familiar topics" to "standard text with more abstraction (high school level)". After that, complexity of content appears to become a bigger issue than the topics themselves.

Table 4 *CEFR Descriptor Levels and Suggested Topics*

CEFR level		Descriptor	Suggested topic
Proficient User	C2	Can understand with ease virtually everything heard or read.	NA
	C1	Can understand a wide range of demanding, longer texts, and recognise implicit meaning.	College level or higher
Independent User	B2	Can understand main ideas of complex text on both concrete and abstract topics, including technical discussions in field of specialisation.	Standard text with more abstraction (high school level)
	B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, etc.	Standard English but on familiar topics
Basic User	A2	Can understand sentences and common expressions of immediate personal relevance (e.g. very basic personal info, shopping, local area, work).	Basic Q's expanded to directions, basic personal info about family, local area
	A1	Can understand and familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type.	Most basic "classroom English" questions

Source: coe.int/en/web/common-european-framework-reference-languages

In summary, we tend to scrutinize AI and trust human judgement as sacrosanct, but this perspective may be flawed. The CVLA is based on CEFR judgements of text difficulty by textbook makers, but it is worth reconsidering how accurate publishers' judgements are in the first place. To properly level generated texts, we may need more quantitatively definable text features as part of prompts. My own experience discussing these issues with materials makers is that as of this writing, educational company employees that use AI for materials often are not even aware that ChatGPT knows what the CEFR is, and instead use more detailed prompts unique to those companies. Researchers that work for such private companies can be the most driven to improve AI-generated materials, but also are often the least likely to share how they do so with the public. While this secrecy may give individual materials makers and publishers a competitive edge in the marketplace, it can prevent the establishment of universal standards in text leveling. More clarity on how materials makers who use AI specify their prompts would be useful in the name of open science, because proprietary research is unreplicable and hard to build on. Clarifying to AI how to level for CEFR could also help clarify the process for ourselves.

More research on quantifiable metrics of CEFR level is needed in order to avoid taking textbook materials at face value. The CVLA could benefit by *convergent validity* studies on empirical difficulty of texts for learners at different CEFR levels. This could be done by testing learners on their comprehension of texts, and then comparing how the CVLA's classifications of text difficulty measure up to the empirical metrics of text difficulty made by testing real learners on the same materials.

Kang & Nichol (2025): Effect of Word-Focused Activity on Reading and Learning of L2 Vocabulary: A Self-Paced Reading Study

Kang and Nichol's study, titled "Effect of word-focused activity on reading and learning of L2 vocabulary: A self-paced reading study" investigated how targeted attention to unfamiliar words during reading might influence vocabulary acquisition and reading behavior. The research was grounded in Swain's noticing hypothesis, which posits that learners are more likely to remember linguistic items when their attention is explicitly drawn to them. In other words, noticing unknown words during reading can facilitate deeper processing and improve retention.

The study had two main aims:

1. To determine whether engaging in a word-focused activity would affect reading time during a subsequent self-paced reading task; and
2. To examine whether the level of engagement in that activity contributed to better vocabulary learning outcomes.

The participants were 95 Korean college students, whose proficiency levels were estimated to range from high-intermediate to advanced. Their placement was based on Korean CSAT scores classified as "above third grade." According to approximate equivalencies, students at this level correspond to Grade 1 (top 4%, roughly TOEIC 800–990) or Grade 2 (top 11%, roughly TOEIC 700–899), which would place them around the CEFR B2 level. The below table summarizes the study's methodology:

Table 5 *Flow Chart of Kang and Nichol's Methodology with Descriptions*

Test/task	Details
Practice reading	Same for control and treatment
First reading task	16 real words within the English reading were replaced with non-words, in order to ensure the tested words had not been previously learned. Two versions with non-words were swapped around.
Fill-in-the-blank activity	Based on words from story. Two versions- non-words (treatment) or real words elsewhere in story (control).
Second reading task + self-paced reading	Read story again. This time reading speed checked to see if different per group.
Vocabulary test	Test non-words for both groups. Scored for exact meaning and "semantic category".

Below is a sample of the reading task. I have bolded the tested pseudwords for emphasis.

> This is the story of Hugo. His life started in a poor **melnawg** in the south of Italy. There were only a few buildings and one of them was

> a **glabe** for poor boys who had no money and nowhere to live. It was run by a very angry master and an old woman. They took in infants from a short time after their **nagid** with the idea that each of them would be an unpaid **zamper** for years. This was not a secret in the community. Everybody knew about it.

For the word-focused activity, learners were asked to fill in the blanks on the tested words, as can be seen in Figures 2A and 2B in Kang and Nichol's paper. The treatment group received a version where the non-words were tested, whereas the control group were tested on the real/original words.

The result of the experiments was that there was no real differences in reading speed / processing across conditions. When the final vocabulary test was scored conventionally, there was no significant difference between the groups ($p = 0.22$), but according to the authors the results "approached significance" for the semantic category ($p = 0.06$). Despite this still-insignificant result, the authors concluded that "Results showed that the engagement with the activity involving target non-words improved recall of the non-word's meaning".

The results were largely insignificant despite multiple significance tests. The finding that the word-focused activity improves recall is questionable: the fact that the test only had 16 items likely means the reliability was low. I recommend reporting internal consistency with a metric such as Cronbach Alpha, as the test's standard error of measurement is likely wide. Although there are many significance tests in this study, only one even approaches $p = 0.05$. The authors focus on this result and downplay the other more flatly insignificant results. Frankly, this introduces a risk of subconscious cherrypicking of the results.

While there is room for disagreement regarding the results, it's clear effect sizes are at best very low. In light of this it is worthwhile to conduct a post-mortem on the study to determine why this experiment was less than fully successful. I can think of two non-exclusive possibilities: 1) The noticing effect is real, but small and difficult to detect, but also perhaps 2) The noticing effect is context-dependent, leading to differing results between experiments. Perhaps our experiments often do not recreate relevant context well because noticing may come down to salience and interest on the part of individual learners.

To give an example of how the noticing effect could have been less salient in this study, I will refer to a possibly analogous case from my own anecdotal experience. I often read my young daughter bedtime stories somewhat above her own speaking level with words unknown to her. Below is an example of such a story, take from the Richmal Crompton book *William in Trouble* (1927). Generally, she tolerates unknown words if she feels she can guess their meaning from context. Examples of such words have been placed in italics in the passage below. However, she inquires about the meanings of the more salient unknown words that seem more likely to signal things and concepts still entirely unknown to her, indicated below in bold font.

> William assumed the *stern air* of a leader of men. "We've gotter go to the **Vicarage**?" he said, "and get back Robert's cup. It's on a *bracket* in

> the **Vicar's** study, 'cause his son took it (though it b'longs to Robert), and it's gotter go on to the silver-table in our drawing-room, where Robert's other cup is an' where it b'longs." The Outlaws cheered *lustily*. The explanation *conveyed* little to them, but they understood that an *exploit* of a more or less unlawful nature was in progress, and they swung joyously up the hill to the **Vicarage**. They peered through the gate. A gardener came up and threatened them with a hose. "Run off, ye *saucy* little 'ounds," he said.
>
> From "William In Trouble" by Richmal Crompton (1927, p.196)

To refer back to the present study, how salient were the non-words in this study to the learners? As the authors note, about 25% of participants were cut from the study for signs they did not take it seriously under the criteria that they 1) attained a score of less than 70% on the True/False reading comprehension test, 2) gave "non-sensical" responses to fill-in-the-blank activity (including English answers instead of pseudowords), or were "barely reading each paragraph" (<0.1 seconds). Despite these filters, the remaining students still might not have taken noticing or learning the non-words very seriously.

Learners knew they would encounter "unknown words". However, given the non-words used in the study, they may have guessed the unknown words were not real due to a lack of "plausibility", and words could have been disregarded simply because learning them was not strictly essential to comprehension of the passage. In regard to plausibility, compare the pseudowords used in this study to Meara's "plausible non-words" (Meara, 2012), which include words with real prefixes and suffixes, such as "Abrogative" and "Benevolate".

Table 6 *Non-words in Story A and Real Words, Adapted from Kang and Nichol*

Real words	Non-words in Story A
Prisoner	Redaster
village	Melnawg
ring	Salp
baby	Semz
son	Mer
mother	flugit
food	Norg
bowel	Nuse
clothes	blaunts

Arguably, it is easier to guess that many of the non-words used in this study (e.g., "semz", "melnawg") are not real English words as they do not conform to typical

English orthographic patterns, which could reduce learners' motivation to take further notice of them despite their novelty. A further complication is that the general meaning of these non-words can often be inferred from the passage by context, which could reduce any sense of urgency to learn their meanings and master them. For example, the non-word "melnawg" appears in the reading in the context of the sentence "His life started in a poor **melnawg** in the south of Italy". In such cases a learner may assume "melnawg" functions as a synonym of the general concept of "area" and may not feel the need to memorize the term's precise meaning further, particularly if they suspect the word is not real and will not be of further use to them after the end of the study anyway.

Although it cannot be stated conclusively that these issues affected the results of the experiment, it seems plausible that some unknown words in a reading and their corresponding meanings will have more salience and be viewed with more importance by learners than others, and that such considerations could affect experimental designs and the results of such studies. Often, learners interest (or lack thereof) in our experiments remains an unexamined variable that can affect results.

Conclusion

Overall, these were both strong studies with good methodologies. The best studies are the ones that inspire and provoke questions, and both of them accomplish that. In regard to Kang and Nichol's study, I will conclude by noting that salience of material for learners and their interest in tasks are important considerations for research that are often overlooked when evaluating the effects on teaching interventions. Perhaps it is time to start making explicit controls for these variables a standard component of such research. Salience is an important consideration for research. Indicators of perceived task relevance and motivation could be useful "pretests" for these considerations. We could then use learner interest as a moderating variable when assessing the effect size of given teaching and learning methods.

Uchida's study makes it clear how important prompting remains when using AI to create materials for L2 learners. Unfortunately, many detailed prompts used by publishers and test makers remain proprietary and secret. More clarity on prompts used in the creation of EFL and ESL materials would be useful in the name of open science, as proprietary research is unreplicable and hard to build on. In addition to helping establish standards for how to make CEFR graded materials with LLMs, the establishment of universal prompts to clarify to AI how to level to the CEFR could also help clarify the process for ourselves. More research on quantifiable metrics of CEFR levels and the factors that underly lexical and textual diversity metrics are needed.

As Uchida and Negishi's CVLA tool makes use of CEFR-levelled textbook material in order to determine how to level nominal texts, it is also worth noting that research should not take textbook materials at face value, as evaluation of what texts match a given CEFR level can vary from publisher to publisher or even unit to unit within a single textbook. The CVLA could benefit by convergent validity studies on empirical difficulty of texts for learners at different CEFR levels. This could be accomplished by testing learners at given proficiency levels on texts, and determine features of text fully comprehended (or still not comprehended) by learners at given CEFR levels.

References

Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Council of Europe. (n.d.). *Common European Framework of Reference for Languages (CEFR)*. Retrieved November 1, 2025, from https://www.coe.int/en/web/common-european-framework-reference-languages

Crompton, R. (1927). *William in trouble.* Newnes.

Helliwell, M. (2015). *Business Plus Level 3 Student's Book* (Vol. 3). Cambridge University Press.

Kang, H., & Nicol, J. (2025). Effect of word-focused activity on reading and learning of L2 vocabulary: A self-paced reading study. *Vocabulary Learning and Instruction*, *14*(1), 2070. https://doi.org/10.29140/vli.v14n1.2070

Meara, P. (2012). Imaginary words. *The encyclopedia of applied linguistics*.

OpenAI. (2025, October 31). *ChatGPT* [Large language model]. https://chat.openai.com/

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

Uchida, S. (2025). Generative AI and CEFR levels: Evaluating the accuracy of text generation with ChatGPT-4o through textual features. *Vocabulary Learning and Instruction*, *14*(1), 2078. https://doi.org/10.29140/vli.v14n1.2078

Uchida, S., & Negishi, M. (2018, September). Assigning CEFR-J levels to English texts based on textual features. In *Proceedings of Asia Pacific Corpus Linguistics Conference* (Vol. 4, pp. 463–467).