

「ソフトウェア紹介」

Estimating the CEFR-J Level of English Reading Passages: Development and Accuracy of CVLA3

Satoru UCHIDA and Masashi NEGISHI

1. Introduction

Assessing the difficulty level of English texts is essential for effective, personalized education. Numerous applications such as Bax's (2012) Text Inspector and Mizumoto's (2022) New Word Level Checker have been developed, demonstrating the high demand for these tools. This paper reports on the CEFR-based Vocabulary Level Analyzer, Version 3 (CVLA3; <https://cvla.langedu.jp/>), designed to estimate the CEFR-J level of reading texts.

The previous version, CVLA2 (Uchida and Negishi, 2018), has been used in various studies (Azemoto & Uchida, 2022; Jodoi, 2023; Miura, 2021; Sato & Yamada, 2020). Feedback from these studies highlights the need for more stable assessment results, improved processing speeds, file-based processing, and the option for locally hosted versions. To address these needs, a new version was developed, with several enhancements. This study outlines the updates in CVLA3, followed by a report on the accuracy validation and comparative experiments with CVLA2.

2. Updates in CVLA3

2.1 Backend Update

In CVLA2, the TreeTagger is employed for backend processing, utilizing the treetaggerwrapper library in Python for part-of-speech (POS) analysis. Recently, spaCy, a native Python library, has been widely adopted, offering not only POS tagging, but also dependency parsing and named entity recognition with proven high performance and accuracy (cf. Altinok, 2021; Vasiliev, 2020). Considering potential future developments, such as local application deployment, CVLA3 has transitioned to an entirely Python-based backend. POS tagging and syntactic analysis leverage spaCy 3.7.2, with the `en_core_web_sm` dictionary. Additionally, textstat 0.7.4 was used to calculate the Automated Readability Index (ARI), which may result in differences from CVLA2's calculations. This update enables more accurate POS tagging and supports the integration of the new metrics introduced in subsequent sections.

2.2 Update to Training Data

The data used in CVLA2 were early CEFR-aligned materials published before 2013. Given the

increased adoption and refined understanding of the CEFR in recent years, CVLA3 has shifted to the use of EFL textbooks published between 2014 and 2020 for statistical training. To ensure clear representation, textbooks spanning multiple levels, such as A1-A2 or A2-B1, were excluded. Instead, 539 texts specifically classified as A1, A2, B1, B2, or C1 were selected (a sufficient number of C2-level texts were unavailable). These were then randomly split, with 431 texts (80%) designated for training and 108 texts (20%) for evaluation testing. This update reflects a more current interpretation of CEFR levels and includes C1-level texts, an addition from the CVLA2 that only covers levels A1–B2.

2.3 Update to Metrics

In CVLA2, four metrics were used: *AvrDiff* (average difficulty of content words), *BperA* (ratio of B-level to A-level content words), *ARI* (Automated Readability Index), and *VperSent* (average number of verbs per sentence) (for details, see Uchida and Negishi, 2018). The first two metrics represent the lexical complexity, whereas the latter two reflect the sentence and text complexities.

In CVLA3, *AvrDiff* was calculated by adding C1 (470 words) and C2 (381 words) words from the English Vocabulary Profile wordlist. Previously, C-level words were highlighted in red in the output, but they were not included in the calculation, which may have been confusing to users. Since the EVP C-level list is limited, its inclusion is not expected to have a significant impact on the calculation results (but improves interpretability). In addition, CVLA3 expanded the set of metrics to eight, adding *CVV1*, *AvrFreqRank*, *POSTypes*, and *LenNP*, allowing for a more detailed analysis of English texts and potentially enhancing the accuracy of level estimation.

CVV1 is defined as “the number of verb tokens divided by the square root of twice the number of verbs” (Spring & Johnson, 2022) and has been validated as an effective measure for evaluating English writing. Essentially, this metric represents lexical diversity, particularly in verb use, an area not covered by CVLA2. Note that be-verbs were not included in this calculation. **AvrFreqRank** represents the average rank of words based on their frequency in the Corpus of Contemporary American English (COCA). Items ranked above 10,000 were uniformly calculated as 10,000 to prevent outliers. Additionally, the three most infrequent words were excluded from the overall calculations to compute the average. This approach minimizes the impact of low-frequency words, particularly when the passages are short. Unlike *AvrDiff* and *BperA*, which focus exclusively on content words, *AvrFreqRank* includes all the words, allowing for a comprehensive lexical-level analysis. Furthermore, it assigns unique values to each word based on rank rather than broad-level categories (*AvrDiff* calculates levels as 1 for A1, 2 for A2, and so on). Thus, *CVV1* and *AvrFreqRank* offer more detailed assessments of lexical complexity.

POSTypes is used to calculate the average number of distinct POS tags per sentence. More complex and longer sentences tend to include a wider range of tags, making a higher POS-type value indicative of greater grammatical complexity. Whereas *VperSent* focuses solely on verb counts, *POSTypes* accounts for all parts of speech. **LenNP** represents the average length of the noun phrases calculated using spaCy POS

tagging and dependency parsing. It measures the lengths of noun phrases that serve primarily as subjects or objects in a sentence. Longer noun phrases are presumed to increase the sentence difficulty, suggesting a higher level of complexity as LenNP increases. Together, POSTypes and LenNP provide new perspectives on sentence complexity beyond those offered by the CVLA2 metrics.

2.4 Update to Evaluation Method

Considering the updates described above, CVLA3 assesses CEFR-J levels by utilizing new metrics in the updated corpus. Table 1 presents the average values for each metric across CEFR levels in the new textbook corpus, highlighting the linear trend in which each metric increased with higher levels. This linear relationship allows the construction of simple regression equations for each metric, providing a clear and interpretable framework. Therefore, users can easily identify the metrics that most strongly indicate higher or lower difficulty levels.

Table 1 Average values of each metric by CEFR level

CEFR	AvrDiff	BperA	CVV1	AvrFreqRank	ARI	VperSent	POSTypes	LenNP
A1	1.28	0.06	1.93	367.99	4.10	1.51	7.16	2.94
A2	1.44	0.12	2.95	445.92	6.22	2.05	8.14	3.36
B1	1.57	0.18	3.90	514.55	7.82	2.66	8.73	3.64
B2	1.74	0.26	4.67	613.05	9.19	2.95	9.04	3.99
C1	1.91	0.36	5.58	739.30	10.79	3.28	9.36	4.51

Based on the results in this table, we developed regression equations using the level assignments of A1 = 1, A2 = 2, B1 = 3, B2 = 4, and C1 = 5, following the approach used in CVLA2. To prevent outliers from skewing the results, an upper limit of 7 was applied to these equations.

$$CVV1_CEFR = \min(CVV1 \times 1.1059 - 1.208, 7)$$

$$AvrDiff_CEFR = \min(AvrDiff \times 6.417 - 7.184, 7)$$

$$BperA_CEFR = \min(BperA \times 13.146 + 0.428, 7)$$

$$AvrFreqRank_CEFR = \min(AvrFreqRank \times 0.004 - 0.608, 7)$$

$$POSTypes_CEFR = \min(POSTypes \times 1.768 - 12.006, 7)$$

$$VperSent_CEFR = \min(VperSent \times 2.203 - 2.486, 7)$$

$$ARI_CEFR = \min(ARI \times 0.607 - 1.632, 7)$$

$$LenNP_CEFR = \min(LenNP \times 2.629 - 6.697, 7)$$

To ensure stability, the final value was calculated by excluding the minimum and maximum values from the regression results and averaging the six middle values. Notably, a value of zero is not necessarily excluded as the lowest value owing to the nature of the regression equation. For example, when BperA is zero, it yields a value of 0.428; therefore, if there are values lower than this value, they cannot be excluded from the calculation.

The conversion to CEFR-J levels followed the method outlined by Uchida and Negishi (2018), as shown

in Table 2. Figure 1 presents the sample analysis results, with gray-shaded metrics indicating those that were not used in the calculation.

Table 2 Mapping to CEFR-J levels

Range	CEFR-J	Range	CEFR-J
$x < 0.5$	preA1	$2.5 \leq x < 3$	B1.1
$0.5 \leq x < 0.84$	A1.1	$3 \leq x < 3.5$	B1.2
$0.84 \leq x < 1.17$	A1.2	$3.5 \leq x < 4$	B2.1
$1.17 \leq x < 1.5$	A1.3	$4 \leq x < 4.5$	B2.2
$1.5 \leq x < 2$	A2.1	$4.5 \leq x < 5.5$	C1
$2 \leq x < 2.5$	A2.2	$x \geq 5.5$	C2

CVLA: CEFR-based Vocabulary Level Analyzer (ver. 3.0)









Input Text

Writing is the act of recording language on a visual medium using a set of symbols. The symbols must be known to others, so that the text may be read. A text may also use other visual systems, such as illustrations and decorations. These are not called writing, but may help the message work. Usually, all educated people in a country use the same writing system to record the same language. To be able to read and write is to be literate.

Legend:

A1: A1 Level Word, A2: A2 Level Word, B1: B1 Level Word, B2: B2 Level Word, NA content words: NA Content Word, NA others: NA Other Word

Estimated CEFR-J Level: B1.1

Index	AvrDiff	BperA	CVV1	AvrFreqRank	ARI	VperSent	POSTypes	LenNP
A1	1.28	0.06	1.93	367.99	4.10	1.51	7.16	2.94
A2	1.44	0.12	2.95	445.92	6.22	2.05	8.14	3.36
B1	1.57	0.18	3.90	514.55	7.82	2.66	8.73	3.64
B2	1.74	0.26	4.67	613.05	9.19	2.95	9.04	3.99
C1	1.91	0.36	5.58	739.30	10.79	3.28	9.36	4.51
Input Text	1.86	0.21	2.16	486.51	6.00	3.33	7.50	4.20
Score (0-7)								
CEFR-J Level	C1	B1.2	A1.3	A2.1	A2.2	C1	A1.3	B2.2

#Cells highlighted in gray are not used for level assessment.

Figure 1 Analysis results of sample text using CVLA3

For the sample text, the CEFR scores for each metric were AvrDiff = 4.73, BperA = 3.13, CVV1 = 1.18, AvrFreqRank = 1.58, ARI = 2.01, VperSent = 4.86, POSTypes = 1.26, and LenNP = 4.34. Excluding the

minimum (CVV1 = 1.18) and maximum (VperSent = 4.86) values, the average of the six remaining values was 2.84. According to Table 2, this score corresponds to a CEFR of B1.1.

2.5 Addition of File Mode

To facilitate the processing of large volumes of files, CVLA3 includes a file mode that supports batch processing. Users can upload up to 30 text files with a maximum size of 10 KB per file. The results are output as a summary table, which can be downloaded in the CSV format, enabling efficient analysis of extensive datasets. Figure 2 shows a sample screen of the results generated in file mode.

CVLA: CEFR-based Vocabulary Level Analyzer (ver. 3.0)

File Analysis Results

Filename	AvrDiff	BperA	CVV1	AvrFreqRank	ARI	VperSent	POSTypes	LenNP	Predicted Level	CEFR Score	Unused Features
shikou_1A.txt	1.84	0.31	2.56	650.61	7.80	2.10	8.40	4.39	B1.2	3.25	CVV1, LenNP
shikou_1B.txt	1.43	0.12	2.45	344.92	7.40	1.21	7.21	2.87	A1.3	1.33	VperSent, ARI
shikou_2A.txt	1.55	0.09	2.65	745.64	4.90	1.47	6.76	3.41	A2.1	1.73	POSTypes, AvrDiff
shikou_2B.txt	1.44	0.09	3.40	351.09	9.50	2.50	8.67	2.54	A2.2	2.25	LenNP, ARI

Download CSV

Back

Figure 2 Example of results in the file mode

3. Accuracy Validation

This section reports the accuracy of the CVLA3. Although CVLA3 was designed to estimate CEFR-J levels, no corpus with pre-assigned CEFR-J levels currently exists. Therefore, we conducted validation using texts labeled with standard CEFR levels, converting the levels as follows for consistency: preA1, A1.1, A1.2, and A1.3 were converted to A1; A2.1 and A2.2 to A2; B1.1 and B1.2 to B1; and B2.1, B2.2, to B2. In a previous study, the CVLA2 achieved an accuracy of approximately 53% on the CEFR scale (Uchida and Negishi, 2021).

The evaluation dataset used for the validation consisted of 108 English texts from an updated CEFR-aligned corpus. Table 3 shows the accuracy results of CVLA3, with rows representing the actual text levels and columns representing CVLA3's predicted levels. Of the 108 texts, CVLA3 correctly identified 71 cases (highlighted in dark blue), resulting in an accuracy of 65.74%. When accounting for adjacent levels (highlighted in light blue), the accuracy increased to 107 matches, indicating a high stability of 99.07%.

Table 3 Level estimation results of CVLA3 on the evaluation dataset

	A1	A2	B1	B2	C1	C2	total
A1	16	2					18
A2	2	17	3				22
B1		8	14	4			26
B2			5	14	2		21
C1			1	9	10	1	21
total	18	27	23	27	12	1	108

4. Comparison with CVLA2

Table 4 presents the validation results for CVLA2, using the same evaluation dataset. CVLA2 correctly identified 65 of 108 cases, yielding an accuracy rate of 60.19%. Although slightly lower than CVLA3's accuracy, this result reaffirms that CVLA2 still offers a practical level of accuracy for practical applications.

Table 4 Level estimation results of CVLA2 on the evaluation dataset

	A1	A2	B1	B2	C1	C2	total
A1	16	2					18
A2	6	15	1				22
B1		9	12	4	1		26
B2			4	13	4		21
C1			1	7	9	4	21
total	22	26	18	24	14	4	108

Table 5 presents a cross-tabulation of the results based on CEFR-J levels using the same dataset. The match rate at the CEFR level (six categories, highlighted in light blue) was 77 out of 108 (71.30 %). For CEFR-J levels (12 categories, highlighted in dark blue), the match rate was 55 of 108 (50.93 %). Although the judgment results may vary depending on the CVLA version, the validation results and increased number of metrics suggest that CVLA3 is likely to provide higher accuracy and greater stability.

Table 5 Comparison of CEFR-J level estimation results between CVLA2 (row) and CVLA3 (column)

	preA1	A1.1	A1.2	A1.3	A2.1	A2.2	B1.1	B1.2	B2.1	B2.2	C1	C2	total
preA1	4	2	1										7
A1.1	2		1	1									4
A1.2				1	1								2
A1.3			1	2	6								9
A2.1				3	7	3	1						14
A2.2					2	6	4						12
B1.1						1	5	1					7
B1.2						1		9	1				11
B2.1								3	9	4			16
B2.2									3	4	1		8
C1											6	8	14
C2												3	4
total	6	2	3	7	16	11	10	13	13	14	12	1	108

5. Conclusion and Future Directions

CVLA3 has achieved substantial enhancements through updates to its backend, corpus foundation, metrics, evaluation methods, and the addition of a file mode, resulting in a faster and more stable web application. With an accuracy rate of 65.74% in predicting CEFR levels (6-level classification) using the evaluation dataset, it serves as a valuable tool for assessing text difficulty.

While a listening mode was not implemented in this revision because of challenges such as insufficient data and the need for audio-based metrics, such as Words Per Minute, for accurate assessment. However, incorporation of this feature should be considered in future studies. Additionally, we plan to develop a local version that enables users to analyze sensitive data without transmitting them online.

References

- Altinok, D. (2021). *Mastering spaCy: An end-to-end practical guide to implementing NLP applications using the Python ecosystem*. Packt Publishing Ltd.
- Azemoto, R., & Uchida, S. (2022). The importance of criterial features for CEFR-based textbooks: A case study using CVLA. *Studies in Languages and Cultures*, 48, 35–47. <https://doi.org/10.15017/4773109> [畔元里沙子・内田諭. (2022). 「CEFR レベル別英語教科書における基準特性の重要度 : CVLA の指標を用いて」『言語文化論究』48, 35–47.]
- Bax, S. (2012). Text Inspector: Online text analysis tool. Available at: <https://textinspector.com/>.
- Jodoi, K. (2023). The correlations between parliamentary debate participation, communication competence, communication apprehension, argumentativeness, and willingness to communicate in a Japanese context. *Argumentation*, 37(1), 91–118. <https://doi.org/10.1007/s10503-022-09591-5>
- Miura, A. (2021). Identifying the CEFR-J Levels of the Reading Texts Introduced in a Course for Current English 1 (Reading). *Journal of Multilingual Pedagogy and Practice*, 1, 1–15. <https://doi.org/10.14992/00020483>
- Mizumoto, A. (2022). An overview of New Word Level Checker. *Proceedings of the Methodology Study Group*, 12, 1–24. <https://doi.org/10.69194/methodologysig.12> [水本篤. (2022). 「New Word Level Checker の概要」『メソドロジー研究部会報告論集』12, 1–24.]
- Sato, T., & Yamada, Y. (2020). Comparison of the text difficulty in new university entrance examinations. *Bulletin of the Chubu English Language Education Society*, 49, 149–156. https://doi.org/10.20713/celes.49.0_149 [佐藤選・山田裕也. (2020). 「新大学入試におけるリーディング文章の難易度比較」『中部地区英語教育学会紀要』49, 146-156.]
- Spring, R., & Johnson, M. (2022). The possibility of improving automated calculation of measures of lexical richness for EFL writing: A comparison of the LCA, NLTK and SpaCy tools. *System*, 106, 102770. <https://doi.org/10.1016/j.system.2022.102770>
- Uchida, S., & Negishi, M. (2018). Assigning CEFR-J levels to English texts based on textual features. In Y. Tono and H. Isahara (eds.) *Proceedings of the 4th Asia Pacific Corpus Linguistics Conference (APCLC 2018)*, pp. 463-467.
- Uchida, S., & Negishi, M. (2021). Estimating the CEFR levels of English reading materials: Evaluation of CVLA. *Journal of Corpus-Based Lexicology Studies*, 3, 1-14. [内田諭・根岸雅史(2021)「英語読解教材の CEFR レベルの推定 : CVLA の妥当性評価」*Journal of Corpus-based Lexicology Studies*, 3, 1-14.]
- Vasiliev, Y. (2020). *Natural language processing with Python and spaCy: A practical introduction*. No Starch Press.