

COREWO4500: A New Corpus-based Word List for Japanese High School Students —Using AI for Word List Development—

Shin'ichiro ISHIKAWA

Kobe University
Tsurukabuto 1-2-1, Nada-ku, Kobe, Japan
iskwshin@gmail.com

Abstract

This paper introduces the development of the COREWO4500, a newly released word list for Japanese high school students. The COREWO4500 is based on texts from Japan's Common Test for University Admissions (including its predecessor tests) covering approximately 40 years (between 1987 and 2025). The list includes around 4,500 headwords, accompanied by Japanese translations, short sample phrases, and CEFR levels. Sophistication of entry items and generation of translations and sample phrases were supported by AI. The CORE4500 would help students aiming to enlarge their vocabulary knowledge, instructors teaching vocabulary to students, and material writers and test developers who need to conduct vocabulary control.

Keywords: corpus, word list, Japanese high school students, AI

1. Introduction

Corpus-based word lists offer a reliable guide for vocabulary learning. Though a variety of word lists have been available, there are not many lists that meet the needs of Japanese high school students aiming to enter university, featuring a reliable selection of words, Japanese translations, and example phrases as an aid to vocabulary memorisation.

To address this shortfall, the author created a 310,000-token corpus from approximately 40-years' English exam data from Japan's Common Test for University Admissions (including its predecessor tests), extracted 4,500 words from the corpus, and released the Core English Words 4500 for Japanese High School Students (COREWO4500), which is now publicly available from the URL (language.sakura.ne.jp/s/doc/voc/corewo4500.xlsx).

Unlike the author's past word list developments, this project utilised AI for the purposes of entry check, translation generation, and sample phrase generation.

This paper introduces corpus construction, word extraction, sophistication of entries, generation of translations and sample phrases, and comparison with the existing vocabulary lists.

2. Literature

Corpus compilation and word list development often come hand in hand. According to Youngblood and Folse (2017), corpus-based word lists can be classified into general lists, academic lists, disciplinary lists, and formulaic lists.

Past releases of large-scale representative corpora have often led to publications of full-scale word lists, which represent the whole vocabulary of a target language. These include Francis, Kučera, and Mackie (1982) based on the Brown Corpus, Leech, Rayson, and Wilson (2001) based on BNC1994, Davies and Gardner (2010) based on COCA, and Brezina and Gablasova (2023) based on BNC2014.

Recent corpus-based word lists offer rich information, not only listing entry items. Figure 1 shows an entry of "night" (noun) from Brezina and Gablasova (2023).



Figure 1: Entry of "night" in BNC2014 Word List

The entry above shows (1) the rank in the list (#217), (2) inclusion in the New General Service List (NGSL), (3) the per-million-word frequency (419/1M), (4) the deviation of proportions (DP) value (0.28: The value of 0 means completely even dispersion, and the value of 1.0 means completely random and uneven dispersion), (5) the genre with the highest frequency (WS: Written to be spoken texts), (6) the genre with the lowest frequency (O: Official documents), (7) the dispersion across eight genres (From the left: O, A (Academic writing), N (newspapers), M (Magazines), F (Fictions), E (E-language), WS, and S (Informal speeches), and (8) high-frequency collocating words (out, before, with, for, when, etc.).

In addition to these large general-purpose word lists, corpus linguists have also created several academic word lists. Academic Word List (AWL) (Coxhead, 2000) is a representative example. She created a corpus of academic texts from 28 subfields of the four major fields of art, business, law, and science (3.5M words) and extracted 570 word families, which occur more than 100 times in all the four major fields and in more than 15 subfields, and they are not included in the first 2,000 words in the original General Service List (West, 1953). By taking both frequency and range into consideration, AWL excludes both non-academic vocabulary and genre-specific academic vocabulary. As a similar project, Gardner and Davies (2013) collected 120-million-words written academic texts from the COCA and extracted 2,000 words (lemmas) that occur significantly more in academic texts than in non-academic texts, which are included in their New Academic Vocabulary List (NAVL).

Some researchers have also been interested in compiling disciplinary or ESP (English for Specific Purposes) word lists. For instance, Vuković-Stamatović (2024) compiled a corpus of 200 physics research articles (1.3 M words) and created a physics ESP word list including 929 words, which

the author intends to effectively support students who have difficulties in reading the research papers. Chang (2025) constructed a 760,000-word corpus of mechanical engineering articles and extracted 398 words (289 word families). Chang’s list is reported to cover the vocabulary occurring in mechanical engineering more efficiently than AWL.

Also, pedagogical word lists for school students have been compiled. Some focus on disciplinary lists for students. Kruawong and Phoocharoensil (2025) created a corpus from English science textbooks for lower secondary Thai students (2M words) and extracted 480 science words, which include vocabulary in the fields of physical science, biological science, and earth/space science. Arndt (2022) created a corpus of digital science resources for middle school students (2.7 M words) and developed a digital science word list. Others focus on more general lists. Sunthornklam (2018) analysed ten sets of Grade 12 English Ordinary National Education Test (O-Net) administered between 2005 and 2017 in Thailand, and based on the word selection criterion adopted in Coxhead (2000), extracted 419 word families. Most of them overlap with the words included in GSL and AWL, but around 1.2% of the entries of the O-Net List were found to be unique.

Although a variety of base data can be chosen according to the purpose of the word list creation, its method is rather simple and lacks diversity: creating a corpus, counting the frequency, and choosing the words that meet some criteria, which often concern frequency (how often a word occurs in total) and range (how widely a word occurs across different types of texts). In the Routledge Handbook, Dang (2019) mentions factors such as the decision of a lexical unit (lemmas vs word families), corpus construction, and word selection criteria (subjective or objective) as variables in word list creation. Dang also suggests that future word list creation needs to pay more attention to mid-frequency vocabulary, spoken vocabulary, subject-specific vocabulary, and multi-word units.

3. Development of COREWO4500

3.1 Aim and Scope

Japan’s Course of Study suggests how many English words high school students should learn, but there are no guidelines indicating which words should be prioritised for learning, and to what level of vocabulary they should have mastered by the time they graduate. Thus, this project aims to compile a word list showing a range of general English vocabulary that high school students who aim to enter colleges and universities need to learn before graduating.

3.2 Data

When generating a word list for students, teachers and researchers often use English school textbooks as data. Though this seems to be a reasonable choice, different textbooks include considerably different types of vocabulary, and whether they represent the vocabulary that all students really need to learn remains unclear. Therefore, this project uses data from the Common Test for University Admissions (CTUA).

In Japan, around 450,000 to 480,000 students take the CTUA every year. As the total number of third-graders at high schools is around 900,000 to 1,000,000, approximately half of them take the test.

A history of the common tests for students who aim to enter colleges and universities in Japan dates back to 1979, when the Common First-Stage Exam (FSE) was introduced. Its English tests focused heavily on language knowledge rather than communicative skills, and it included reading comprehension questions based on shorter passages.

In 1990, the National Centre Test for University Admissions (NCT) began. Its English test still focused on language knowledge, but it included longer passages for reading. In 2006, it came to include a listening section in order to catch up with the change in the trend in English language teaching in Japan and the world.

Then, in 2021, the current CTUS started. Its English test pays more attention to measuring students’ communicative English skills by putting equal emphasis on reading and listening, and knowledge-based questions that characterised the first and the second generation tests were eliminated. In the reading section, it now requires test takers to read texts and charts/graphs/posters in combination and process information based on real settings and situations.

Years	Content	N
1987-1989	FSE: [M/S] x 3 yrs	6
1990-2005	NCT: [M/S] x 16 yrs	32
2006	NCT: [L-sample, MR/ML/SR/SL]	5
2007-2020	NCT: [MR/ML/SR/SL] x 14 yrs	56
2021-2024	CTUS: [MR/ML/SR/SL] x 4 yrs	16
2025	CTUS: [MR/ML]	2
Total	---	117

Table 1: Collected English exam data

This project collected 117 kinds of English tests from these three types of tests. In Table 1, M and S stand for the main test and the supplementary test (resit), and R and L stand for the reading test and listening test (not incl. transcripts), respectively.

3.3 Method

3.3.1 Deletion of Japanese texts

As each test data includes Japanese instructions, they were all deleted by replacing `[\^a-zA-Z0-9.\-'\r\n]` (i.e., any character excluding alphabets, numbers, periods, hyphens, single quotes, spaces, and new lines) with null.

3.3.2 Lemmatising

Although some of the existing word lists define words as word families (e.g., “play” includes “player” and “replay” etc.), it seems inappropriate for Japanese students without enough knowledge of affixes and derivations. Thus, COREWO4500 has chosen to define words as lemmas. Lemmatisation was conducted with TagAnt V2.1.1 (Anthony, 2024), which uses `en_core_web_sm-3.7.0` (a small English language pipeline for the spaCy NLP library).

Before	After
Susan and Tim were waiting for their friend Akira at the airport. Tim shouted There he is...	Susan and Tim be wait for their friend Akira at the airport . Tim shout there he be...

Table 2: Example of lemmatisation

Automatic lemmatisation is convenient, but the result is not perfect, of course. Some derivations, especially those in the words appearing at the head of the sentences, were found to remain unlemmatised. In addition, lemmatisation created several non-words such as “ll” from “you’ll,” “ca” and “n’t” from “can’t,” and “wo” from “won’t,” for instance.

3.3.3 Frequency and Range Count

Then, lemmatised files were processed with the Word list function of AntConc V4.4.2 (Anthony, 2026). The generated list was saved as a CSV file.

	A	B	C	D	E	F
1	Type	Rank	Freq	Range	NormFreq	NormRange
2	the	1	16372	117	52284.16	1
3	be	2	13019	117	41576.32	1
4	to	3	9511	117	30373.48	1
5	a	4	8743	117	27920.87	1
6	of	5	6416	117	20489.57	1
7	in	6	5492	117	17538.76	1
8	and	7	5436	116	17359.93	0.991
9	i	8	5422	109	17315.22	0.932

Figure 2: A part of the CSV file

This tentative list included 8,681 entries (lemmas, all in lower case), and the total frequency was 313,135.

3.3.4 Exclusion of Low-ranged words

As mentioned above, Coxhead (2000) established the criteria of a minimum frequency and a minimum range for choosing the words to be included in the list. However, the size of the exam corpus is much smaller and setting the threshold of the minimum frequency seems to exclude potentially important words. Therefore, this project set only the minimum range criterion and excluded the words that appeared only in one of the 117 tests. This rule decreased the total number of entries from 8,681 to 5,101.

3.3.5 Frequency adjustment

The frequencies of “ca,” “wo,” and “n” were added to the frequencies of “can,” “will,” and “not,” respectively. Also, the estimated frequency of “a” used as a code (e.g., “A:, B:, C:...”) was subtracted from the frequency of “a” as an article. Also, the frequency of plural nouns that were not automatically lemmatised was added to that of singular nouns (e.g., the frequency of “Americans” was added to that of “American”). The headword “cans” was changed to “can (n).”

3.3.6 Detection of questionable entries

The list automatically created from the exam corpus still includes many inappropriate entries. Some are word fragments. The early-stage common tests included many accent quizzes, where a test taker was requested to choose a stressed syllable from “con - di - tion - al,” “dem - o -

crat - ic, and “pro - fes - sion - al,” for instance. Due to this, many word fragments are included in the list. Others include person names (e.g., Yuko), city names (e.g., Akiharabara), and abbreviations (e.g., GDP), which students may not need to learn with priority.

As a consistent and exhaustive detection of such questionable entries is highly challenging for humans, this project used AI (Google Gemini 3.6 Flash) as support. The prompt given to AI and the received feedback are shown in Table 3.

[Author] I am now making an English word list for students. Please identify (1) non-word fragments, (2) person names, (3) place names, (4) abbreviations, (5) foreign words, and (6) other inappropriate entries from the list below...

[Gemini] Here is the breakdown of the entries from your word list, organised by the six requested categories.

1. Non-Word Fragments & Structural Noise---These are incomplete word roots, suffixes, standalone letters, or formatting errors: Single-letter / Sub-word fragments: ac, ad...

2. Person Names---Proper names should be capitalised or evaluated if they belong on a general vocabulary list: Given Names & Surnames: aaron, aki...

3. Place Names & Geographical Terms---Proper nouns referring to countries, cities, or landmarks: Countries / Regions: africa, america, Americas... Cities, Landmarks & Physical Geography: akiharabara, alaska...

4. Abbreviations & Contractions---Non-standard single words or contractions that may need standard expansion or capitalisation: Abbreviations: co, dr, dvd, etc, fri, gnp... Contractions / Written Conversational Forms: ca (as in can’t fragment), im (I’m), ll, ve, wo (won’t fragment)

5. Foreign Words & Loanwords---Words directly borrowed from non-English languages (some are recognised loanwords in English, but worth noting): Japanese Terms: anime, janken, kimono... Other Languages / Cultural Terms: caf (café), flamenco, hula, karate, kayak, swahili

6. Other Inappropriate or Questionable Entries---Interjections, informal dialogue noise, plurals mixed into base forms, typos, and proper adjectives: Interjections & Conversational Noise: ah, bye, gosh, hello, hey... Proper Adjectives / Nationalities (Require Capitalisation): african, africans, american... Plural Inconsistencies (List generally uses singular base forms): adventures, affairs, ages, bags...

Table 3: Prompt for a check of the entries and received feedback

The feedback above helped the project to clarify its word selection criteria. Considering the specific needs of Japanese high school students in L2 English vocabulary learning, the project has established the policies about entry inclusion.

To be excluded	To be included
(1) fragments	(1) country names and their adjective forms (e.g., Japan, Japanese)
(2) person names	(2) abbreviations and acronyms (e.g., UN)
(3) city names (e.g., Los Angeles)	(3) non-Japanese foreign words (e.g., kayak)
(4) Japanese words (e.g., kimono)	(4) planet names (e.g., Jupiter)
(5) truncated forms (e.g., Fri for Friday, Jan for January)	(5) interjections
	(6) numbers
	(7) days of the week and month names

Table 4: COREWO4500's word inclusion policy

As regards (5) and (6) in the included items, two kinds of codes, Intj (interjections) (e.g., year, hi) and Num (numbers: e.g, one, two), were given.

3.3.7 Capitalisation

As all the entries appear in lower case, several words were capitalised manually (e.g., america→America, wednesday→Wednesday). Then, four kinds of group word codes—Abb (abbreviations: e.g., GDP, TV, Mr.), Day etc. (e.g., Sunday, January, Jupiter), Country etc. (Italy, Italian), and Religion (e.g., Christmas, Buddhist)—were added to capitalised words.

3.3.8 Multiple checks of the entries by humans and AI

After each of the manual corrections of the entries, the revised list was sent to the AI for a further check. However, the AI often mentioned the entries already deleted and the problems already corrected, which suggests that the AI may check the stored data rather than the newly input data.

[Author] Please double-check the new list and identify the inappropriate entries. The list goes as follows:...

[Gemini] Here is the double-check of your revised list. While you successfully removed many of the personal proper names, several issues remain. Below is the complete audit of inappropriate or problematic entries left in your list, categorised by type....

[Author] You may have wrongly checked the previous list. Please recheck the new list shown in this prompt.

[Gemini] My apologies for any confusion in the previous check! I have thoroughly re-examined the exact text of the list provided in this prompt. Here is the precise breakdown of entries that are non-words, typos, proper nouns, abbreviations, or structural artefacts...

Table 5: Prompt for a double-check of the entries and received feedback

3.3.9 Adding FR values

After the process of multiple checks of the entries based on the communication with the AI, the number of entries has decreased from 5,101 to 4,505.

As suggested by the literature, ranking of the entries solely based on their frequency might not be appropriate, especially in the word list for students. This project, therefore, added the FR value to each entry.

The FR value, which can be obtained by multiplying normalised frequency (NFreq) by normalised range (NRange). Take two words as samples.

	Freq	Range	NFreq	NRange	FR
but	1175	99	3752.4	0.85	3174.5
by	1089	108	3477.7	0.92	3209.9

Table 6: Calculation of the FR values

As seen in the table, “but” occurs more often, while “by” occurs more widely across different texts. Though it is highly difficult to decide subjectively which should be prioritised in vocabulary learning, the FR value shows that, in this case, “by” should be regarded as more important for students, who need to prioritise acquiring the words that are used frequently and at the same time widely. The FR values, which are simple and clear in definition, offer a consistent guide to decide the ranking.

3.3.10 Translation

Although translations are a good support for non-native L2 English learners, adding translation to each entry is a time-consuming task. Thus, this project converted an Excel file to a Google document and automatically added translations by its unique function: “=googletranslate(word, en, ja).” This function gives only one typical definition to each entry, even when it is polysemous. Google Translate is based on the data for machine translation, and it sometimes tends to offer somewhat technical definitions, but they still could help learners to grasp the meanings of words.

3.3.11 Sample phrases

As the literature suggests, it would be more beneficial for learners to learn new words in context (Krashen, 1982). This project, therefore, has decided to prepare a short phrase including a target word for each entry. As this is also a highly time-consuming task, however, it utilised AI.

[Author] For the list of English words below, please create one short English phrase of no more than three words—suitable for university entrance exams—and one Japanese translation for it. Omit any unnecessary explanations and ensure that your output follows the format below exactly, in the order listed.

[Format example] vital choice (重要な選択)

[List of words] the, be, to...

Table 7 Prompt for phrase generation

As processing all 4,500 words at one time was rejected, they were divided into nine sets of 500 words, and each was sent to AI sequentially.

3.3.12 Addition of the word-level information from the other lists

Mainly for teachers, COREWO4500 also offers (1) three difficulty levels based on the (original) GSL and AWL (1K, 2K, and AWL: head-word-based lists were obtained from www.lexutor.ca/freq/lists_download/), (2) the frequency-based rank in the New GSL v1.2 (Browne & Culligan,

2013; obtained from www.newgeneral-service-list.com/new-general-service-list), and (3) CEFR-bands based on *Cambridge Assessment Word lists: Pre A1 Starters, A1 Movers, and A2 Flyers (For Exams from 2018)* (CUPA, n.d.) and *Cambridge Assessment Word lists: B1 Preliminary and B1 Preliminary for Schools* (CUPA, 2025: obtained from www.cambridgeenglish.org/).

4. Conclusion

The figure below shows a part of the COREWO4500.

Rank	Lemma	Group	Note	GSL	CSL	Ph	CEFL	Freq	Range	NFreq	NRange	FR	Definition	Phrase
2001	subway						B1	14	8	44.788	0.068	3.62下鉄	by subway (地下鉄で)	
2002	stomach							12	9	38.322	0.077	3.50胃腸	stomach C (ロウキン)	
2003	container							15	7	47.663	0.06	2.9容器	shipping container (B)	
2004	elephant		2K				-A1	21	5	67.264	0.043	2.9象	african elephant (アフリカ)	
2005	emotional						B1	15	7	47.663	0.06	2.9感情的	emotional support (E)	
2006	hamburger							18	6	57.483	0.053	2.9ハンバーガー	eat a hamburger (ハ)	
2007	predict							15	7	47.663	0.06	2.9予測する	predict the future (F)	
2008	Spain	Region					B1	18	6	57.483	0.053	2.9スペイン	tour around Spain (コ)	
2009	tip		2K				1669 B1	15	7	47.663	0.06	2.9ヒント	helpful tip (役立つ)	
2010	agreement						TB1	13	8	43.538	0.068	2.9合意	reach an agreement (I)	
2011	athlete						2559 B1	13	8	43.538	0.068	2.9アスリート	professional athlete (I)	
2012	backpack						A1	17	6	54.29	0.053	2.9バックパック	travel backpack (旅行)	
2013	CD	Abb					-A1	13	8	43.538	0.068	2.9CD	listen to a CD (CDを)	
2014	correctly							13	8	43.538	0.068	2.9正しく	answer correctly (正)	
2015	crash		2K				1867 B1	13	8	43.538	0.068	2.9クラッシュ	car crash (自動車事故)	
2016	cycle						1384 A2	13	8	43.538	0.068	2.9サイクル	bike cycle (バイクサイ	
2017	gold		2K				1453 A2	13	8	43.538	0.068	2.9金	gold medal (金メダル)	
2018	lane							17	6	54.29	0.053	2.9レーン	traffic lane (車線)	
2019	origin		2K				1989	13	8	43.538	0.068	2.9起源	country of origin (原産	
2020	trail		2K				1180 B1	17	6	54.29	0.053	2.9トレール	rock a trail (トレールを	

Figure 3: COREWO4500 (Rank 2001 to 2014)

The table below illustrates information offered in the list.

Header	Content
Rank	Rank based on the FR value
Lemma	Entries
Group	Word groups (e.g., Abb, Region, Day/Month/Others)
Note	British spelling and other notes
GSL	Levels based on GSL and AWL (1K, 2K, and AWL)
NGSL Rank	Rank based on the Standard frequency index (SFI) in NGSL
CEFR	CEFL levels (-A1, A2, B1) based on the Cambridge Assessment Word lists
Freq	Raw frequency in the CTUS corpus (313,135 words in total)
Range	Raw range values in the CTUS corpus including 117 texts
NFreq	Normalised (Per-million-word) frequency
NRange	Normalised range
FR	FR value obtained by multiplying NFreq with NRange
Definition	Japanese translations obtained from Google Translate
Phrase	2-3-word phrases and Japanese translations automatically generated by AI

Table 8 Structure of the COREW4500

Comparison with existing lists has shown that COREWO4500 includes many entries that have not been covered in them.

vitamin, container, hamburger, Spain, correctly, lane, climber, colorful, discourage, fifth, flyer, highway, shortage, assignment, sunlight, whom, advertising, bacteria, trash, hearing, Indian, lifelong, Russia, semester, audio, interestingly, yell, clip, multi, civilization, collector, housing, laser, mammal, Spanish, sustainable, ancestor, enthusiastic, foolish, huh, nap, amusement, Australian, kingdom, medal, New Zealand, satellite, accordingly, actively, angrily, chew, closing, detailed, digest, disagreement, enjoyment, fascinating, grocery, wisdom, workplace, decoration, Denmark, flying, intercultural, Korea, layout, oral, underlined, aquarium, auction, belonging, breathing, commuter, deadline, feedback, forty, logical, making, means, melody, planning, seasonal, sixty, unfamiliar, warming, allergy, bilingual, dye, orientation, timing, African, beneficial, bowling, Brazilian, camper, commute, downtown, historically, hunter, influential, misunderstanding, physically, pole

Table 9: Entries that have not been covered in other major word lists (Between Ranks 2001 and 2500)

Systematically learning these words step-by-step would help high school students in Japan develop their L2 English communication skills, not only prepare for the college entrance exam.

Learning more than 4,000 words may seem to be a highly challenging task for students. However, the number of words that they encounter for the first time is not necessarily large. The tables below show breakdowns of the 4,500 words according to the three criteria: type of group words, GSL/AWL-levels, and CEFR-levels.

Type	Entry Samples
Abb (12)	AM, CD, Dr., DVD, etc.
Day etc. (19)	April, August, December, February, Friday
Intj (14)	ah, bye, gosh, hello, hey
Num (28)	eighteen, eighteenth, eighth, eighty, eleven
Country etc. (73)	Africa, African, America, American, Asia
Religion (5)	Buddha, Buddhist, Christian, Christmas, God
Others (4354)	a, abandon, ability, able, about

Table 10: Breakdown of the words in terms of word groups

Type	Entry Samples
GSL 1K (934)	able, about, above, accept, accord
GSL 2K (753)	abroad, absence, absent, absolute, absolutely
AWL (368)	abandon, academy, access, accompany, accurate

Others (2450)	a, ability, absorb, absurd, academic
---------------	--------------------------------------

Table 11: Breakdown of the words in terms of GSL/AWL levels

Type	Entry Samples
Pre A1/ A1 (721)	a, about, above, ache, add
A2 (377)	across, act, actor, actually, adventure
B1(1353)	ability, abroad, absent, absolutely, accent
Others (2054)	abandon, absence, absolute, absorb, absurd

Table 12: Breakdown of the words in terms of CEFR levels

If the students think they have already finished learning GSL words or A1/A2-level words, which are usually learned at junior high schools, the number of words that high school students need to learn newly decreases to around 2800–3400 words. Excluding a set of group words, it further decreases. Using a filter button on the Excel sheet, students can easily extract the words that meet a particular criterion they have chosen (e.g., listing the words beyond B1 levels only).

Rank	Lemma	Glos	Note	GSL	NGSL Rank	CEFL	Freq	Rang	NFre
3445	異議(異)				1965	7	7	22.355	
2878	異議(異)				1758	5	5	15.968	
4017	色で書く(色)				2398	2	2	6.387	
1960	色で書く(色)					17	9	38.322	
4018	色で書く(色)					2	2	6.387	
1961	色で書く(色)				1751	12	9	38.322	
3569	色で書く(色)					3	3	9.581	
3396	色で書く(色)					4	3	12.774	
4020	色で書く(色)					2	2	6.387	
3570	色で書く(色)					3	3	9.581	
3177	色で書く(色)					4	4	12.774	
2448	色で書く(色)				2456	8	6	25.548	
3572	色で書く(色)					3	3	9.581	
179	色で書く(色)				522	202	80	645.09	

Figure 4: Extracting entries except for GSL 1K-level words by using Excel's filter function

In addition to students, teachers and material writers would also be able to utilise COREWO4500 to judge whether they need to control the vocabulary in a text for students. The author hopes that COREWO4500 is used widely by those interested in vocabulary learning and teaching in Japan.

Bibliographical References

- Arndt, R. (2022). A specialized vocabulary list from an original corpus of digital science resources for middle school learners. *Journal of English for Academic Purposes*, 60, Article: 101187, 1–11. [10.1016/j.jeap.2022.101187](https://doi.org/10.1016/j.jeap.2022.101187)
- Brezina, V., & Gablasova, D. (2023). *A frequency dictionary of British English: Core vocabulary and exercises for learners*. Routledge.
- Browne, C., & Culligan, B. (2013). The new general service list. Retrieved from <https://www.newgeneralservicelist.com>.
- Chang, L. (2025). A corpus-based mechanical engineering academic word list. *International Journal of TESOL Studies*, 5(3), 126–142. <https://www.tesolunion.org/archives-info/181>

- Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), 213–238. <https://doi.org/10.2307/3587951>
- CUPA (Cambridge University Press and Assessment) (n.d.). *Cambridge Assessment Word lists: Pre A1 Starters, A1 Movers, and A2 Flyers (For Exams from 2018)*. <https://www.cambridgeenglish.org/Images/wordlists-pre-a1-starters-a1-movers-and-a2-flyers.pdf>
- CUPA (Cambridge University Press and Assessment) (2025). *Cambridge Assessment Word lists: B1 Preliminary and B1 Preliminary for Schools*. <https://www.cambridgeenglish.org/Images/506887-b1-preliminary-vocabulary-list.pdf>
- Dang, T. N. Y. (2019). Corpus-based word lists in second language vocabulary research, learning, and teaching. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 288–303). Routledge.
- Davies, M., & Gardner, D. (2010). A frequency dictionary of contemporary American English: Word sketches, collocates and thematic lists. Routledge.
- Francis, N. W., Kučera, H., & Mackie, A. W. (1982). *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin.
- Gardner, D., & Davies, M. (2013). A new academic vocabulary list. *Applied Linguistics*, 35(3), 305–327. <https://doi.org/10.1093/applin/amt015>
- Kruawong, T., & Phoocharoensil, S. (2025). Establishing English science textbook word lists (ESTWLs) for Thai EFL secondary students: A corpus-based approach. *PASAA*, 71(1), 386–418. [10.58837/CHULA.PASAA.71.13](https://doi.org/10.58837/CHULA.PASAA.71.13)
- Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: Based on the British National Corpus*. Longman.
- Sunthornklam, T. (2018). A corpus-based word list of grade-12 English O-NET examinations. [Master's thesis, Thammasat University Language Institute]. https://ethesisarchive.library.tu.ac.th/thesis/2018/TU_2018_6021042368_10695_10358.pdf
- Vuković-Stamatović, M. (2024). Creating and validating a corpus-based English academic word list for physics. *Revista Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics*, 38(1), 80–106. <https://doi.org/10.1075/resla.22041.vuk>
- West, M. (1953). General service list of English words: With semantic frequencies and a supplementary word-list for the writing of popular science and technology. Longman.
- Youngblood, A. M., & Folse, K. S. (2017). Survey of corpus-based vocabulary lists for TESOL classes. *MEXTESOL Journal*, 41(1), 1–14. <https://www.mextesol.net/journal/public/files/f71f44170615a7efa63468aa46aaf885.pdf>