Recovering Registers from Leveled Wordlists

Yo Ehara
Tokyo Gakugei University
4-1-1 Nukuikita-machi, Koganei-shi, Tokyo, Japan.
ehara@u-gakugei.ac.jp

Abstract

For vocabulary learning in language acquisition, it is desirable for learners to acquire words that they are likely to need in the language environments they will encounter. Such language environments are referred to as "registers" in general corpora, which are typically designed to include diverse registers. However, the proportion of registers included, that is, which registers are included and to what extent, is determined by the circumstances under which each general corpus was compiled and is not necessarily optimized for language learning. To bridge this gap, various leveled wordlists have been created in language education using linguistic resources other than word frequency, such as expert judgment and learner responses. However, it has not been quantitatively clear what gap in register proportions in general corpora these leveled wordlists were designed to fill. This study proposes a method that, given a leveled wordlist and a general corpus, estimates the register ratio that best aligns the frequency ordering of words across registers with the leveled wordlist. This makes it easier for learners and educators to interpret which wordlists are appropriate for particular learning goals. Our method is formulated as a linear programming problem and yields a globally optimal solution. Unlike neural networks, it is less susceptible to variation due to initial values or approximation and is therefore easier to interpret. We evaluated the proposed method on two languages, English and Japanese, through a range of experiments. We further show that it can also be used to evaluate vocabulary lists created for specific contexts, such as those generated by Large Language Models like ChatGPT.

Keywords: register analysis, lexical difficulty, linear programming, general corpora

1. Introduction

Language education has long emphasized that vocabulary learning should be aligned with learners' expected language use rather than treated as the memorization of decontextualized word lists. In applied linguistics, vocabulary goals are commonly defined with respect to learners' needs and purposes of language use (Richards, 2015), and high-frequency vocabulary is motivated pedagogically because it provides the most useful initial coverage for learners (Nation, 2022a). After this core vocabulary, subsequent learning is often directed toward more specialized areas depending on learners' aims, such as academic or occupational domains (Nation, 2022b). This perspective requires some account of *register*, that is, varieties of language associated with situational characteristics and recurrent communicative settings (Conrad, 2015). For NLP readers, a convenient intuition is that registers correspond to systematically different usage environments, for example, conversation, fiction, news, or academic prose, rather than to topic alone.

General corpus projects such as the British National Corpus (BNC) have therefore been highly influential in pedagogy because they provide large-scale frequency evidence across written and spoken English (Aston and Burnard, 1998). However, the BNC was designed as a broad sample of British English rather than as a corpus specifically constructed around the language-use conditions of second-language learners. This design choice matters for vocabulary ordering. In Nation's BNC-

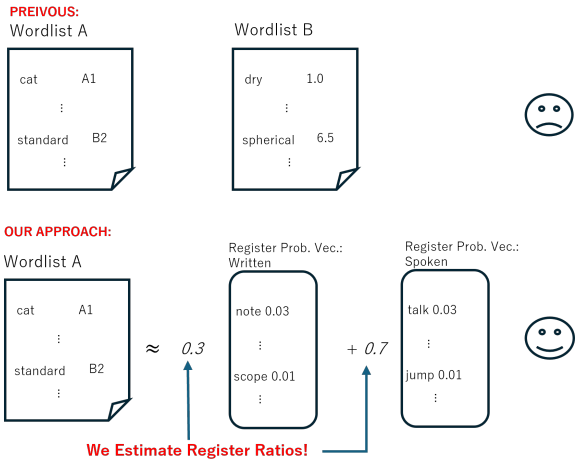

Figure 1: Overview of this study. Previously, different metrics were used for each educational wordlist, and one had to consult manuals written by the wordlist creators to understand which learners they targeted. Consequently, learners were unable to determine which wordlists they should use to progress in their studies. Our approach estimates the register ratio in a general corpus whose resulting word-frequency ordering best matches the difficulty ordering in a given wordlist. This makes it easier for learners and educators to interpret which wordlists should be used for learning.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1143–1154
11–16 May 2026.

based analysis, the corpus itself is mostly written rather than spoken, which pushes several common spoken words downward in the frequency bands while elevating words that reflect the adult and institutional character of the written corpus (Nation, 2006). Nation therefore notes that separate written and spoken corpora may yield more appropriate lexical orderings (Nation, 2006). Consequently, directly deriving lexical profiles or difficulty signals from aggregate corpus counts can misrepresent the linguistic environment presupposed by a learning wordlist, especially when the underlying corpus was not designed for language learners and large written collections dominate the frequency statistics. The same concern also applies to first-language education, where general corpora need not reflect the age-specific mixture of spoken, written, and instructional language encountered by children.

Mixture estimation is attractive for pedagogical analysis because it produces an interpretable vector of register weights instead of opaque embeddings. An important motivation for our study is that existing lexical difficulty resources encode difficulties on incomparable scales. Syllabi aligned with staged proficiency scales reflect the judgments of experienced teachers, whereas lexical complexity-shared tasks and vocabulary size assessments capture learners' perceptions through continuous scores, majority votes, or graded responses (Shardlow et al., 2021; Paetzold and Specia, 2016; Yimam et al., 2018; Palacios et al., 2018). Our framework only requires pairwise ordering information such that one lexical item is known to be easier than another and is, therefore, agnostic to the absolute calibration of each resource. This scale invariance enables the ingestion of teacher-authored syllabi alongside learner-generated annotations without re-normalizing the underlying scales or assuming linear relationships between them.

Figure 1 shows an overview of this research. With conventional methods, it is difficult to understand the assumed linguistic environment even when given a learning wordlist. Furthermore, words are described using different levels of measurement in each wordlist. This study proposes a method for estimating the register ratio underlying a wordlist. Unlike conventional methods, this study only requires the input wordlist to be level-classified; it can be classified using real numbers or by stages.

Our contributions are fourfold:

- We propose a scale-agnostic method for estimating register ratios when given a created wordlist and a general corpus.
- We provide an extensive empirical study contrasting expert-curated and learner-derived resources across coarse and fine-grained registers in general English and Japanese corpora, and we demonstrate that the same pipeline can audit vocabulary lists generated by large language models. The release includes open-source scripts for caching register counts and integrating item response theory estimates derived from a vocabulary size test [1].
- Being scale-agnostic, our method can also be used for evaluating wordlists created by prompting large language models (LLMs), such as ChatGPT, with prompts designed for specific situations, such as studying abroad, by estimating their registers.
- Because the proposed method is a linear programming problem, it remains tractable even for large problems and can be applied to a wide range of wordlists and general corpora. Furthermore, the quality of the obtained optimal solution can be interpreted independently of the wordlist or general corpus. We plan to present the register ratios calculated in this study as a dataset.

2. Related Work

Register variation has been studied extensively in corpus linguistics, with early quantitative analyzes highlighting systematic differences between written and spoken communication (Biber, 1988). Subsequent work explored how register-sensitive lexicons can support pedagogical sequencing, for instance through graded vocabularies in CEFR-aligned curricula (Tono et al., 2013). Concurrently, computational models have begun to infer register mixtures by optimizing over topic proportions or by balancing objective functions that reward register-specific distributions (Petrenz and Webber, 2010).

Regarding word frequencies in general corpora, most studies calculate the word frequency of a corpus by simply summing the frequencies of the registers included while maintaining their respective proportions. For example, the word frequency in the BNC is calculated by summing the word frequencies for each register in the BNC according to their respective proportions within the corpus. However, the proportion of registers in the BNC depends on factors like the corpus collection process and is not necessarily optimized for language learning. Of course, this often does not match the actual language environment of English learners. To bridge this gap, various studies have created wordlists.

CEFR-J aggregates teacher judgments to extend the Common European Framework of Refer-

[1] The resources will be available at https://rebrand.ly/lrec2026 or https://researchmap.jp/yo_ehara?lang=en

ence (CEFR) for Japanese learners, whereas CompLex summarizes crowd-sourced pairwise comparisons elicited at SemEval 2021 (Shardlow et al., 2021). Earlier shared tasks such as CWI 2016 (Paetzold and Specia, 2016) and CWI 2018 (Yimam et al., 2018) contributed frequency-annotated corpora with absolute difficulty scores derived from learner annotations. Item response theory has recently been applied to vocabulary assessment to separate item difficulty from participant ability (Chen and Li, 2011). We extend this line by fitting one- and two-parameter logistic models to multiple-choice responses, thereby generating continuous constraints that enrich the optimization problem.

Optimization-based approaches to mixture estimation often rely on convex objectives with slack penalties. Soft-margin support vector machines balance hinge loss with slack, and more recent constrained NLP models integrate expectation constraints with ℓ_1 relaxations (Ganchev et al., 2010). Our linear program inherits this intuition but focuses on interpretable register weights α_r. The proposed minimax variant parallels robustness techniques that bound the largest residual (Ben-Tal et al., 2009), which we adopt to reveal disagreements between lexical resources. We implement the optimization using Gurobi (Gurobi Optimization, LLC, 2023) and provide caching mechanisms that guarantee reproducibility while reducing the overhead of scanning entire corpora.

This study interprets LLM-generated vocabulary lists by reconstructing their registers. While many recent studies involve generating vocabulary lists with LLMs (Alfter, 2024; Nikolova-Stoupak et al., 2024a,b; Schaaff et al., 2023; Chow et al., 2024; Durward and Thomson, 2024; Kelious et al., 2024; Bannò et al., 2025; Degraeuwe and Goethals, 2024), to the authors' knowledge, this study is the first to create lists tailored to specific language learner contexts and to interpret them by reconstructing their registers.

3. Proposed Methods

Our pipeline transforms lexical difficulty resources into inequalities that constrain register mixtures. We denote the set of registers by R and the set of lexemes observed in the corpus and covered by a difficulty resource by V. Register counts are computed from the BNC in both a coarse setting with $R = \{\text{spoken}, \text{written}\}$ and a fine-grained setting with eight registers, such as *written:fiction* or *spoken:convrsn*. Let $c_r(w)$ be the count of lexeme w in register r, and let $C_r = \sum_w c_r(w)$ be the total number of tokens for register r. We approximate the conditional probabilities using normalized counts; that is, $p(w \mid r) = c_r(w)/C_r$. The mixture weights α_r satisfy $\alpha_r \geq 0$ and $\sum_r \alpha_r = 1$. The expected probability of observing w in the mixture is $q(w) = \sum_r \alpha_r p(w \mid r)$.

3.1. Constraint Generation

Resources that yield ordinal difficulty levels, such as CEFR-J, are converted into ordered pairs (w_i, w_j), with w_i easier than w_j. For datasets where difficulty is represented by real numbers (e.g., CompLex), we convert this into ordering constraints by sampling pairs. We require $q(w_i) + s_{ij} \geq q(w_j)$, where $s_{ij} \geq 0$ is called a "slack variable". A slack variable of 0 indicates that the constraint is not violated. That is, the current α_r ensures that the word frequency comparison (i, j) follows the wordlist constraint: i is easier than j. Conversely, if $s_{ij} > 0$, the constraint is violated; a larger value indicates a more significant violation. Therefore, the linear inequality is

$$\sum_{r \in R} \alpha_r p(w_i \mid r) + s_{ij} \geq \sum_{r \in R} \alpha_r p(w_j \mid r). \quad (1)$$

When the resource supplies numeric scores, we symmetrize the constraints to encourage a greater probability of easier words. Ordinal resources remain asymmetric.

3.2. Optimization Objectives

Recall that a larger s_{ij} indicates a stronger violation of the ordering constraint that i is easier than j. We introduce the auxiliary variable t and solve the following optimization problem:

$$\begin{aligned} \min_{\{\alpha_r\},\{s_{ij}\},t} \quad & t \\ \text{s.t.} \quad & \sum_{r \in R} \alpha_r p(w_i \mid r) + s_{ij} \\ & \qquad \geq \sum_{r \in R} \alpha_r p(w_j \mid r), \quad \forall (i,j), \\ & s_{ij} \leq t, \quad \forall (i,j), \\ & \alpha_r \geq 0, \quad \forall r \in R, \\ & \sum_{r \in R} \alpha_r = 1, \\ & s_{ij} \geq 0, \quad \forall (i,j). \end{aligned} \quad (2)$$

The objective t bounds the largest violation and stabilizes the mixture. This remains effective even when the number of constraints in the optimization problem grows to two hundred thousand. The solution highlights the specific pairs that attain the maximum slack, enabling a targeted error analysis.

Throughout this paper, we use Gurobi to solve this optimization problem on a machine with 256 GB of memory. However, as this is a linear programming problem, we can also use non-commercial solvers such as the "pulp" library.

4. Experiments

We now describe the data, experimental settings, and results for coarse and fine-grained register mixtures. All experiments use random sampling with constraint budgets between one thousand and five hundred thousand, depending on the resource. Unless noted otherwise, we report the minimax objective because it prevented the degenerate solutions observed with the sum objective. The meanings of budget and minimax objective are explained in Section 4.2.

4.1. Datasets

First, we used the British National Corpus, which offers balanced coverage of written and spoken English and includes register annotations down to specific genres (Consortium, 2007). For coarse experiments, we aggregate registers into written and spoken categories and compute probabilistic lexeme profiles from the token counts. Fine-grained experiments retain eight BNC registers: *spoken:convrsn*, *spoken:othersp*, *written:acprose*, *written:fiction*, *written:news*, *written:nonac*, *written:otherpub*, and *written:unpub*. To test cross-corpus robustness, we replicate the full pipeline on the Open American National Corpus (OANC) (Ide, 2008), whose spoken component comprises *face-to-face* and *telephone* conversations and whose written component spans *fiction*, *journal*, *letters*, *non-fiction*, *technical*, and *travel_guides*. This dual setting allows us to assess whether corpus-specific register inventories materially alter the inferred mixtures.

Lexical difficulty resources cover diverse elicitation protocols. CEFR-J assigns CEFR levels to entries collected for Japanese learners of English and includes part-of-speech-specific variants (Tono et al., 2013). CompLex aggregates pairwise comparisons from crowd workers, producing continuous scores that approximate the probability of one lexeme being more complex than another (Shardlow et al., 2021). The CWI 2016 shared task provides sentence-level annotations transformed into unigram scores by aggregating annotator votes (Paetzold and Specia, 2016). CWI 2018 expands the coverage with Wikipedia, News, and WikiNews partitions annotated by learners (Yimam et al., 2018). Finally, the EVKD vocabulary size test includes responses from one hundred participants on one hundred multiple-choice items; we infer item parameters through item response theory to derive continuous difficulty scores.

These resources differ markedly in provenance. CEFR-J (Tono et al., 2013) is curated by expert teachers who align items with curricular expectations, while CompLex (Shardlow et al., 2021), CWI16 (Paetzold and Specia, 2016), CWI18 (Yimam et al., 2018), and EVKD (Ehara, 2018) capture judgments or performances of language learners. Consequently, their scales are not directly comparable: CEFR-J uses discrete CEFR bands, CompLex outputs probabilities in $[0, 1]$, CWI resources average integer difficulty votes, and EVKD produces logits derived from item response theory. The proposed optimization treats them uniformly by converting each signal into ordered pairs, making it possible to contrast teacher-oriented syllabi with learner-centric evidence in a single mixture estimation problem.

4.2. Experimental Settings

For each resource, we intersected the vocabulary with the BNC lexicon. Because the number of possible pairwise constraints can be very large, we randomly sample a fixed number of constraints in each experiment; we refer to this number as the budget. Unless noted otherwise, we use the minimax objective introduced in Eq. (2), which minimizes the largest slack over all sampled constraints. The random sampler draws constraints until the budget is exhausted or the unique pairs are depleted. For CEFR-J, we experimented with budgets of one thousand, fifty thousand, one hundred thousand, two hundred thousand, and five hundred thousand to assess saturation effects. CompLex and CWI resources use budgets up to two hundred thousand. The vocabulary size test constraints naturally cap at a few thousand pairs owing to the small lexicon.

We computed register probabilities from lemma counts without smoothing, reflecting the pedagogical motivation to analyze lexemes attested at least once in the corpus. Shared-vocabulary experiments, particularly with CWI16, require each lexeme to appear in every register; consequently, the fine-grained mixture collapses into a single register when the resource is concentrated on academic prose. The optimization problem was instantiated with uniform weights such that each constraint contributed equally to the slack calculation.

We also explored the effect of varying the minimum frequency threshold from one to five tokens. Increasing the threshold amplifies the influence of high-coverage registers because low-frequency spoken lexemes are filtered out, which slightly increases the slack of the CEFR-J constraints. However, the recovered mixtures remained within two percentage points of the values reported in Tables 1 and 2, reinforcing the conclusion that scale heterogeneity rather than frequency sparsity drives the observed differences. Additional experiments switching back to the sum objective confirmed that the degeneracy associated with the objective in (2) persisted even after frequency filtering, further motivating the adoption of the minimax formulation.

4.3. Fine-Grained Mixtures

Fine-grained estimates provide rich insights into how resources prioritize registers beyond the written-versus-spoken dichotomy. Table 1 reports the nonzero mixtures for a constraint budget of fifty thousand. CEFR-J and CompLex both favor fiction, yet CEFR-J leaves nearly twenty percent of the mass on spoken registers, whereas CompLex concentrates on written genres. CWI18 Wikipedia distributes weights across academic prose, fiction, and other publications, whereas CWI18 News highlights spoken conversations, followed by a mixture of written news-related registers. CWI18 WikiNews amplifies written fiction almost exclusively, underscoring the sensational style of the annotated articles. Finally, the vocabulary size test indicated that the items were predominantly drawn from academic prose and news, reflecting its role as an academic placement instrument.

Two-parameter logistic estimates from the vocabulary test produce mixtures similar to the one-parameter model, but highlight different slack patterns. Discrimination-driven constraints emphasize items whose responses sharply separate high- and low-ability participants, causing the optimizer to stress the contrast between academic prose and news. Difficulty-driven constraints focus on rare lexemes, such as *marsupial*, which are absent from spoken registers, reinforcing the written emphasis observed in Table 1.

4.4. Coarse Register Mixtures

Table 2 visualizes the minimax solutions after sampling two hundred thousand constraints (seed 99). Color intensity tracks mixture weights and highlights that CEFR-J retains a sizeable spoken component, whereas learner-driven resources concentrate on written registers. The News partition of CWI18 collapses to a purely spoken solution, whereas the IRT-derived estimates place all mass on written evidence.

Tracking the mixtures across budgets preserves this picture: CEFR-J moves from spoken dominance at one thousand constraints to the written-heavy mixture shown in Table 2, CompLex and CWI18 Wikipedia remain close to a 0.2/0.8 split, and WikiNews stabilizes at approximately fifteen percent spoken mass. These trajectories confirm that the snapshot was representative rather than a single random draw.

The minimax objective maintains objective values lower than 10^{-4} for CompLex and CWI resources, indicating that only a small fraction of the constraints reach the bound. In contrast, CEFR-J retains an objective value of approximately 0.10 even after two hundred thousand constraints, highlighting the substantial deviation between the syllabus and BNC frequencies. Examination of the slack ranking files shows that rare academic lexemes, such as *onomatopoeia* and *negligible*, contribute the largest violations because they are frequent in written academic texts, yet are assigned higher difficulty levels in CEFR-J.

4.5. Cross-Corpus Comparison

Table 3 extends the color-coded analysis to the Open American National Corpus (OANC) using one hundred thousand constraints. The teacher-oriented CEFR-J resource favors telephone speech alongside fiction, CompLex places comparable weight on travel guides and technical writing, and learner annotations from CWI18 Wikipedia reinforce fiction and journalistic prose.

4.6. Sensitivity Analysis

To understand the robustness of the recovered mixtures, we varied two dimensions: the sampling mode and the random seed. Switching to prefix sampling accentuated the dominance of the easiest portions of each resource, producing more extreme mixtures (e.g., CEFR-J allocating 82% weight to spoken registers for the coarse setting) but preserving the relative ranking of resources. Repeating the random sampling with seed 101 changed individual mixture components by at most 0.02, indicating that the estimation is not overly sensitive to stochastic variation in constraint selection. We also observed that the minimax objective consistently yielded lower maximum slack than the sum objective even when both found similar mixtures, which suggests that the minimax solution is preferable for auditing constraint violations because it highlights a smaller set of high-impact contradictions.

4.7. Slack Diagnostics

Table 4 summarizes the largest slack observed for representative runs. CEFR-J saturates the 0.10 bound with pairs such as *zoo* $\succ$ *yearn* in the BNC and *zoo* $\succ$ *widen* in the OANC, underscoring the tension between syllabus expectations and conversational evidence. CompLex produces violations an order of magnitude smaller, typically contrasting *zone* with rare proper nouns. Learner-generated constraints from CWI18 align most closely with corpus frequencies, keeping the maximum slack below 4×10^{-4}.

4.8. Qualitative Case Studies

We manually inspected the top twenty constraints with non-zero slack for each resource to interpret the disagreements highlighted in Table 4. CEFR-J violations repeatedly juxtapose concrete spoken

Resource	Non-zero Register Weights (50k pairs)
CEFR-J	spoken:othersp 0.194, spoken:convrsn 0.024, written:fiction 0.609, written:news 0.136, written:unpub 0.037
CompLex	spoken:convrsn 0.170, written:acprose 0.202, written:fiction 0.628
CWI18 Wikipedia	spoken:othersp 0.102, written:acprose 0.382, written:fiction 0.336, written:otherpub 0.181
CWI18 News	spoken:convrsn 0.521, written:acprose 0.185, written:fiction 0.053, written:nonac 0.083, written:otherpub 0.159
CWI18 WikiNews	spoken:convrsn 0.112, written:fiction 0.888
VST 1PL	written:acprose 0.797, written:news 0.203
CWI16 (shared)	written:acprose 1.000

Table 1: Fine-grained register mixtures for random sampling with fifty thousand constraints and the minimax objective. Values rounded to three decimals.

Resource	Spoken	Written
CEFR-J	0.24	0.76
CompLex	0.19	0.81
CWI18 Wikipedia	0.19	0.81
CWI18 News	1.00	0.00
CWI18 WikiNews	0.15	0.85
CWI16	0.00	1.00
VST 1PL	0.00	1.00
VST 2PL Difficulty	0.00	1.00
VST 2PL Discrimination	0.00	1.00

Table 2: Coarse BNC mixtures estimated with the minimax objective and two hundred thousand random constraints (seed 99). Darker values indicate higher mixture weight.

nouns such as *zoo* with abstract verbs like *yearn* or *widen*. The syllabus promotes the verbs to higher proficiency bands even though conversational registers rely on them heavily, forcing the optimizer to increase spoken weight to satisfy the constraint.

CompLex exhibits the opposite tendency: crowd workers often marked geographically flavored nouns (e.g., *zither*, *Zimbabwe*) as easier than the polysemous *zone*. This reverses the corpus evidence, which treats *zone* as the more prevalent item, and hints that annotators were comfortable with proper nouns drawn from global media.

Learner annotations from CWI18 Wikipedia cluster around scientific adjectives such as *zoological* being judged easier than action-oriented nouns like *vandalism*. These decisions push the mixture toward written informational registers even when the contrasted noun remains common in spoken narratives. Finally, the EVKD-derived constraints surface archaic vocabulary such as *zephyr* and *yonder*; the minimax solution responds by increasing weight on fiction and nonacademic prose, where such items sporadically appear.

These qualitative observations provide practical guidance for resource designers. When slack repeatedly points to the same semantic domain (e.g., technical nouns or political entities), revisiting the difficulty assignments for those domains could produce resources that align more faithfully with corpus evidence. Conversely, when slack exposes rare lexemes that remain absent from all registers, corpus enrichment rather than resource revision may be necessary.

5. Japanese Register Reconstruction

To examine the generality of this method beyond English, we constructed a Japanese pipeline based on the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Omura and Asahara, 2018). The official frequency tables distribute each lemma across thirteen registers (for example, *fiction*, *telephone*, and *opinion magazine*). We adopted the Kyoiku-Kihon-Goi (Basic Education Vocabulary) lists distributed by the National Institute for Japanese Language and Linguistics (NIN-

Resource	Face-to-face	Fiction	Journal	Letters	Non-fiction	Technical	Telephone	Travel guides
CEFR-J	0.00	0.35	0.00	0.00	0.00	0.13	0.45	0.07
CompLex	0.20	0.21	0.00	0.00	0.00	0.34	0.00	0.25
CWI18 Wikipedia	0.04	0.46	0.32	0.03	0.00	0.02	0.00	0.13

Table 3: OANC mixtures estimated with the minimax objective, one hundred thousand random constraints, and seed 99. The color intensity reflects the magnitude of the value.

Corpus	Wordlist	Word Pair	Max Slack
BNC	CEFR-J	*zoo* $\succ$ *yearn*	0.100
BNC	CompLex	*zone* $\succ$ *zither*	1.6×10^{-3}
BNC	CWI18 Wikipedia	*zoological* $\succ$ *vandalism*	3.7×10^{-4}
OANC	CEFR-J	*zoo* $\succ$ *widen*	0.100
OANC	CompLex	*zone* $\succ$ *vertebrates*	1.0×10^{-3}
OANC	CWI18 Wikipedia	*zoological* $\succ$ *volunteered*	2.4×10^{-4}

Table 4: Largest slack values for selected runs with the minimax objective. Smaller values indicate closer agreement between the difficulty resource and the corpus registers.

JAL)[2], which assign staged levels such as *Shin-Sakamoto A1* through *ShinSakamoto C4*.

We mapped these labels to an ordinal scale (A1 easiest, C4 hardest) and converted them into pairwise constraints, as in the English experiments. While NINJAL distributes many Japanese educational wordlists, we chose (Sakamoto, 1958) and (Sakamoto, 1984) because they are older datasets constructed long before the BCCWJ. This clearly indicates that their wordlists were created independently of the BCCWJ.

Furthermore, while the 1958 dataset was created through a detailed process involving voting by multiple school teachers, a revised edition was published in 1984 without the author providing a specific explanation for the changes. Clarifying what changes occurred during this period is useful for Japanese vocabulary education.

Table 5 shows the results. First, although the BCCWJ naturally includes spoken registers, neither dataset contained any spoken registers at all. This reveals a significant bias toward written registers. Furthermore, while (Sakamoto, 1958) reflects multiple types of written registers, (Sakamoto, 1984) consolidates them into a single written register, indicating a loss of diversity.

6. Evaluating LLM-created Wordlists

One of the greatest advantages of the proposed method is that, unlike conventional approaches, it can accept any word difficulty metric as input as long as the order is known. Therefore, as a completely new approach, we analyze wordlists generated by an LLM from specific situations and then estimate their registers.

[2] https://www.ninjal.ac.jp/english/

Specifically, under the same setting of "Japanese university students studying computer science," we imagined scenarios in which users would actually input prompts based on their post-graduation career paths. We then input the following two prompts in Japanese to generate wordlists:

- (JP): Create a list of 100 English words for Japanese computer science undergraduate students who have been accepted into Japanese graduate schools and intend to read English papers for international conference submissions as part of their graduate research. Organize the words into five levels, from the simplest to the most difficult.
- (US): Create a list of English words for Japanese computer science students who will study at a graduate school in the US after graduation and need to communicate orally with native speakers. The list should consist of 100 words across five levels, from the simplest to the most difficult.

Table 6 drills down into vocabulary lists generated by LLMs. GPT-5 Pro allocates most mass to conversations; however, the Japanese-oriented list preserves a small academic-prose footprint, whereas the US-focused list shifts written evidence toward general nonacademic prose. Claude Sonnet converges to an entirely conversational mixture for both regions. The table exemplifies how the proposed pipeline can audit automatically constructed study materials without any bespoke adjustments.

7. Discussion

The experiments demonstrate that lexical difficulty resources encode distinct views of register expo-

Register	Description	Sakamoto	ShinSakamoto
OP	Opinion magazines	0.113	1.000
OL	Local newspapers	0.166	0.000
OM	General magazines	0.169	0.000
OY	Youth magazines	0.304	0.000
PN	National newspapers	0.247	0.000

Table 5: Fine-grained BCCWJ mixtures (objective=max, 50k constraints, seed 99) for the Sakamoto and ShinSakamoto basic vocabulary lists. Values are register contribution α_r rounded to three decimals.

Resource	spoken:convrsn	written:acprose	written:nonac
GPT5Pro (JP)	0.922	0.078	0.000
GPT5Pro (US)	0.937	0.000	0.063
Claude (JP)	1.000	0.000	0.000
Claude (US)	1.000	0.000	0.000

Table 6: Fine-grained BNC mixtures (objective=max, 50k constraints, seed 99) for vocabulary lists produced by large language models. Remaining registers obtain zero weight.

sure. Expert-crafted CEFR-J levels emphasize spoken interaction at low-constraint budgets; however, increasing the number of sampled pairs pushes the mixture toward written registers. This behavior indicates that the underlying syllabus covers both conversational and academic vocabulary, yet short samples over-represent the conversational tiers. Therefore, educators who rely on small subsets of CEFR-J may underestimate the written exposure required for advanced learners.

The LLM-derived lists in Table 6 indicate that the same diagnostics can be extended to automatically generated study materials. Differences between the GPT and Claude vocabularies become visible as shifts in written mass from academic to nonacademic prose or as a complete collapse to conversation, signaling where an instructor might need to rebalance synthetic resources before classroom deployment.

The scale-invariant nature of the optimization supports juxtaposing teacher and learner perspectives without redesigning the inference procedure. Teacher-produced CEFR-J levels and learner-produced CompLex or CWI annotations enter the linear program through the same pairwise inequalities, making their disagreements transparent in the recovered mixtures. When fine-grained mixtures attribute most of the mass to written fiction for CEFR-J, but to academic prose for CompLex, the difference can be interpreted as a genuine divergence in pedagogical priorities rather than an artifact of incompatible scales.

Item response theory further moderates the influence of specific examinee cohorts. Because EVKD item parameters are conditioned on latent ability scores (Chen and Li, 2011; Palacios et al., 2018), the resulting constraints remain stable even if test takers are unusually proficient or struggling. This robustness is visible in Table 1: the vocabulary size test aligns with academic prose and news regardless of other resources, suggesting that the recovered mixture is not dominated by the abilities of the sampled participants. Therefore, this approach is suitable for incorporating classroom-level assessments provided that the assessments can be calibrated using IRT or a similar model.

Learner-derived resources, such as CompLex and CWI18 Wikipedia, present consistent written-dominant mixtures, underscoring the fact that learners perceive difficulty in terms of informational density rather than conversational spontaneity. The News partition is an exception; its mixture collapses into the spoken form, revealing that the annotated items stem from speech-heavy contexts. This observation cautions against pooling the three CWI18 partitions when deriving pedagogical recommendations because the News portion would bias the mixture toward oral registers.

It is instructive to contrast our framework with topic models. Latent Dirichlet Allocation and its derivatives are designed for the largely unsupervised discovery of latent themes when the semantic structure of a corpus is unknown; the inferred topics are often inspected post hoc to understand the corpus properties. In our setting, the register taxonomy of the BNC is known and carefully curated. The goal is not to discover new latent structures, but to express the variation encoded in lexical difficulty resources in terms of predefined registers. The learning signal is provided by external difficulty judgments rather than by word co-occurrence alone, so the recovered mixture directly addresses the pedagogical question of “Which register mixture best explains this difficulty ordering?” rather

than uncovering the hidden topics.

To ensure that the zero-frequency artifacts did not overly bias the mixtures, we introduced an optional Lidstone smoothing parameter that perturbs the register likelihoods using a small constant. Sweeping the parameter over $\lambda \in \{0.0, 0.1, 1.0\}$ changed the coarse BNC mixtures by at most two percentage points (for example, CEFR-J shifted from 0.535/0.465 to 0.547/0.453 for spoken/written), and learner-driven resources, such as CompLex and CWI18, effectively remained unchanged.

Furthermore, while this study adopts an approach that minimizes the maximum value of slack variables, we have also experimented with an approach that minimizes the total sum of slack variables. However, the total-sum minimization approach was heavily influenced by the sample size of constraints from wordlists, leading to unstable results. In practice, for most wordlists, enumerating all pair constraints is practically impossible. For example, a 5000-word wordlist would yield approximately 12 million pair constraints because $5000 \times 4999/2$ amounts to this value. Moreover, the sum-of-slack objective continued to collapse into degenerate solutions irrespective of smoothing. These observations suggest that the qualitative conclusions reported here stem from disagreements in the evidence of difficulty rather than from accidental zeros in the corpus frequencies.

This study also connects to our previous work in educational data mining and AI in education on learner-oriented lexical modeling and difficulty interpretation. In earlier work, we investigated learner-specific lexical knowledge estimation from vocabulary tests and learner data, including learner-specific word difficulty estimation, learner-oriented readability assessment, and the prediction of learners' knowledge of atypical meanings of words from tests targeting their typical meanings (Ehara et al., 2012; Ehara, 2021, 2022a). We also proposed a method for selecting reading texts for incidental vocabulary learning by estimating the distribution of acquired vocabulary for individual learners (Ehara, 2022b). More recently, we explored how difficulty-related structure can be interpreted in embedding spaces through inverse embedding and geometric modeling (Ehara, 2025b,a). This study complements these studies by shifting the focus from learner-specific lexical knowledge estimation, learner-specific text selection, and embedding-based difficulty interpretation to the reconstruction of register ratios implicitly assumed by pedagogical wordlists. Together, these studies support a broader research program on making lexical difficulty representations more interpretable for educational use.

8. Conclusions

We proposed a methodology for estimating register mixtures from lexical difficulty constraints and applied it to a broad collection of learner-oriented resources. We evaluated our method using general corpora of two languages: English and Japanese.

In addition to curated datasets, we showed that the pipeline audits LLM vocabularies and exposes register biases implicit in automatically produced study lists, providing practitioners with a quick method for vetting synthetic materials prior to distribution.

For future work, the methodology can support adaptive reading platforms that tailor register exposures to learner goals. By plugging into domain-specific corpora and learner-produced difficulty judgments from the target population, stakeholders can rapidly assess whether materials reflect the registers that learners expect to encounter.

9. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 22K12287 and by JST, PRESTO Grant Number JPMJPR2363. We are deeply grateful to the anonymous reviewers for their constructive feedback.

10. Limitations

Our study relies on frequency estimates from general corpora. Although the BNC is balanced, its register taxonomy may not reflect contemporary digital communication. In this study, register-conditioned word probabilities are estimated using raw frequencies in the corpus, although smoothing can also be applied. However, because our method targets only the ordering of difficulty, applying Lidstone smoothing changed the estimated register ratio by only about 0.01.

The number of sampled constraints is typically set to 50,000. However, since our method only considers the maximum frequency difference of constraint violations, we tested up to 500,000 constraints and found no significant impact on the recovered registers.

Another limitation is that we treat lexical items as independent, ignoring morphological families and multiword expressions that often drive readability in pedagogical materials. Extending the framework to operate on lemmas linked through morphological or semantic relations would better capture how teachers and learners generalize difficulty assessments. Moreover, while item response theory mitigates cohort effects for the EVKD resource, we did not calibrate the other learner-produced datasets, so their mixtures may still reflect the demographic profile of

the annotators. Future work should explore anchoring strategies that align multiple learner cohorts before constructing the constraint sets.

11. Ethical Considerations

Since we used previously collected publicly available datasets only, we believe that this study does not require any particular ethical considerations.

12. Bibliographical References

David Alfter. 2024. Out-of-the-box graded vocabulary lists with generative language models: Fact or fiction? In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 1–19, Rennes, France. LiU Electronic Press.

Guy Aston and Lou Burnard. 1998. *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh University Press, Edinburgh.

Stefano Bannò, Kate M. Knill, and Mark J. F. Gales. 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)*, pages 632–646, Vienna, Austria. Association for Computational Linguistics.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. 2009. *Robust Optimization*. Princeton University Press, Princeton.

Douglas Biber. 1988. *Variation Across Speech and Writing*. Cambridge University Press, Cambridge.

Chung-Hua Chen and Shu-Yu Li. 2011. Applying item response theory to vocabulary assessment. *Language Testing*, 28(4):681–703.

Siew Yeng Chow, Chang-Uk Shin, and Francis Bond. 2024. This word mean what: Constructing a Singlish dictionary with ChatGPT. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI) @ LREC-COLING 2024*, pages 41–50, Torino, Italia. ELRA and ICCL.

Susan Conrad. 2015. Register variation. In Douglas Biber and Randi Reppen, editors, *The Cambridge Handbook of English Corpus Linguistics*, pages 309–329. Cambridge University Press, Cambridge.

Jasper Degraeuwe and Patrick Goethals. 2024. Leading by example: The use of generative artificial intelligence to create pedagogically suitable example sentences. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 33–48, Rennes, France. LiU Electronic Press.

Matthew Durward and Christopher Thomson. 2024. Evaluating vocabulary usage in LLMs. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 266–282, Mexico City, Mexico. Association for Computational Linguistics.

Yo Ehara. 2018. Building an english vocabulary knowledge dataset of japanese english-as-a-second-language learners using crowdsourcing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Yo Ehara. 2021. Lurat: a lightweight unsupervised automatic readability assessment toolkit for second language learners. In *Proceedings of 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 806–814.

Yo Ehara. 2022a. No meaning left unlearned: Predicting learners' knowledge of atypical meanings of words from vocabulary tests for their typical meanings. In *Proceedings of the 15th International Conference on Educational Data Mining (EDM, short paper)*, pages 492–499, Durham, United Kingdom. International Educational Data Mining Society.

Yo Ehara. 2022b. Selecting reading texts suitable for incidental vocabulary learning by considering the estimated distribution of acquired vocabulary. In *Proceedings of the 15th International Conference on Educational Data Mining (EDM, poster)*, pages 767–772, Durham, United Kingdom. International Educational Data Mining Society.

Yo Ehara. 2025a. Educational cone model in embedding vector spaces. In *Proceedings of ICCE 2025: The 33rd International Conference on Computers in Education (short paper)*.

Yo Ehara. 2025b. Generating diverse difficulty examples in embedding vector spaces via inverse embedding. In *Proceedings of Artificial Intelligence in Education (AIED) Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium, Blue Sky, and WideAIED*, pages 68–76, Cham. Springer Nature Switzerland.

Yo Ehara, Issei Sato, Hidekazu Oiwa, and Hiroshi Nakagawa. 2012. Mining words in the minds

of second language learners: Learner-specific word difficulty. In *Proceedings of COLING 2012*, pages 799–814, Mumbai, India. The COLING 2012 Organizing Committee.

Kuzman Ganchev, Joao Graca, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11:2001–2049.

Gurobi Optimization, LLC. 2023. *Gurobi Optimizer Reference Manual*. Gurobi Optimization, LLC.

Nancy Ide. 2008. The american national corpus: Then, now, and tomorrow. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages, Summerville, MA. Cascadilla Proceedings Project*, volume 127.

Abdelhak Kelious, Mathieu Constant, and Christophe Coeur. 2024. Complex word identification: A comparative study between ChatGPT and a dedicated model for this task. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3645–3653, Torino, Italia. ELRA and ICCL.

I. S. P. Nation. 2006. How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, 63(1):59–82.

I. S. P. Nation. 2022a. The goals of vocabulary learning. In *Learning Vocabulary in Another Language*, 2 edition. Cambridge University Press, Cambridge.

I. S. P. Nation. 2022b. Specialised uses of vocabulary. In *Learning Vocabulary in Another Language*, 2 edition. Cambridge University Press, Cambridge.

Iglika Nikolova-Stoupak, Serge Bibauw, Amandine Dumont, Françoise Stas, Patrick Watrin, and Thomas François. 2024a. Generating contexts for ESP vocabulary exercises with LLMs. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 153–175, Rennes, France. LiU Electronic Press.

Iglika Nikolova-Stoupak, Serge Bibauw, Amandine Dumont, Françoise Stas, Patrick Watrin, and Thomas François. 2024b. LLM-generated contexts to practice specialised vocabulary: Corpus presentation and comparison. In *Actes de la 31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position*, pages 472–498, Toulouse, France. ATALA and AFPC.

Mai Omura and Masayuki Asahara. 2018. Ud-japanese bccwj: Universal dependencies annotation for the balanced corpus of contemporary written japanese. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 117–125.

Gustavo Paetzold and Lucia Specia. 2016. Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California. Association for Computational Linguistics.

Laura Palacios, Koji Yamamoto, and Yuki Sato. 2018. Evkd vocabulary size test response set. Educational Vocabulary Knowledge Diagnostic Project Dataset. Version 1.0.

Philipp Petrenz and Bonnie Webber. 2010. Genre inference and the lack of distributional evidence. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 2186–2193, Valletta, Malta. European Language Resources Association.

Jack C. Richards. 2015. Vocabulary. In *Key Issues in Language Teaching*. Cambridge University Press, Cambridge.

Ichiro Sakamoto. 1958. *Kyoiku Kihon Goi (Educational Basic Wordlists)*. Maki-Shoten.

Ichiro Sakamoto. 1984. *Shin Kyoiku Kihon Goi (New Educational Basic Wordlists)*. Gakugei-Tosho.

Kristina Schaaff, Tim Schlippe, and Lorenz Mindner. 2023. Classification of human- and AI-generated texts for English, French, German, and Spanish. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 1–10, Online. Association for Computational Linguistics.

Matthew Shardlow, Marcos Zampieri, Mihael Pasov, and Cassandre Boulc. 2021. Semeval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online. Association for Computational Linguistics.

Yukinori Tono, Satoko Kawaguchi, and Masashi Negishi. 2013. Developing a CEFR-based word list for Japanese learners. In *Research and Practice in Assessing Academic Writing*, pages 55–76. Cambridge University Press.

Seid Muhie Yimam, Chris Biemann, Gustavo Paetzold, and Lucia Specia. 2018. Multilingual and cross-lingual complex word identification. In *Proceedings of the 2018 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 401–411, New Orleans, Louisiana. Association for Computational Linguistics.

13. Language Resource References

British National Corpus Consortium. 2007. *British National Corpus, Version 3 (BNC XML Edition)*. Oxford University Press. PID http://hdl.handle.net/20.500.12024/2554.