Verb-Specific Timing of Semantic Expansion Across CEFR Levels: A Layer-wise von Mises–Fisher Analysis of Learner English

Yukio Tono, Christopher R. Cooper
Tokyo University of Foreign Studies, Waseda University
y.tono@tufs.ac.jp, cooper@waseda.jp

Abstract

High-frequency verbs such as "make", "take" and "like" broaden their range of use as second language (L2) proficiency develops, yet whether this semantic expansion unfolds uniformly across verbs has rarely been quantified. We analyse learner writing from the JEFLL Corpus (Japanese EFL learners) at three CEFR bands (A1, A2, B1+) and extract contextual representations of three target verbs from all thirteen layers of BERT, including the embedding layer. Semantic dispersion is measured as the concentration parameter κ of a von Mises–Fisher distribution fitted for each verb, level and layer, with essay-level cluster-bootstrap confidence intervals. Concentration decreases with proficiency for all three verbs, but the timing differs: "take" and "make" diffuse early (A1→A2), whereas "like" diffuses late (A2→B1+). This asymmetry holds at every one of the twelve contextual layers. Because raw κ is inflated by the anisotropy of contextual embeddings, we additionally report a scale-free log-ratio, adapting the alignment-free variance measure of Nagata et al. (2023); it confirms the timing effect and resolves an embedding-layer artefact, relocating the dispersion peak of "make" to the upper layers. Semantic expansion in L2 verb use is thus verb-specific in its developmental timing.

Keywords: learner corpora, contextual embeddings, semantic dispersion, von Mises–Fisher, CEFR

1. Introduction

High-frequency verbs are among the earliest items a second language (L2) learner acquires, yet mastering their full range of use is a protracted process. Beginners rely on a small set of core, concrete senses — "take a bus", "make a cake", "like apples" — and only gradually extend these verbs into the collocational and constructional patterns that characterise proficient use. Capturing this gradual broadening has traditionally relied on counts of frequency and collocation. Contextual word representations open a complementary route: they let us measure, directly and in a graded way, how varied a learner's use of a verb has become.

By a contextual representation we mean the following. Earlier distributional models such as word2vec or GloVe assign a single vector to each word type, for example, one fixed point for "take", no matter how it is used. Contextual models such as BERT (Devlin et al., 2019) instead assign every individual occurrence its own vector, computed from the whole sentence, so that "take a bus" and "take a photograph" receive different representations. BERT is a bidirectional encoder: each token's vector is conditioned on both its left and its right context, so the vector reflects how the verb is actually used in that sentence. This is exactly the information we need in order to gauge the spread of a verb's uses. Although decoder-only large language models such as ChatGPT have since become the dominant paradigm, they are trained to predict the next word from left context alone; encoder models of the BERT family, which read the whole sentence at once, remain the standard — and for word-sense and word-in-context analysis often the strongest — tool when the goal is to represent a token's meaning in context rather than to generate text. We therefore adopt BERT for this representation-focused study.

A further property of BERT is that it constructs its representation through a stack of layers. A useful simplification, supported by probing studies (Ethayarajh, 2019), is that the lower layers stay close to surface, lexical information, encoding which word form is present and its typical collocates, while the higher layers encode increasingly context-dependent meaning. Reading a verb's dispersion layer by layer therefore lets us ask not only whether a verb's use broadens with proficiency but where in this lexical-to-contextual hierarchy the broadening resides, separating expansion that is essentially lexical from expansion that is genuinely contextual.

We address three questions. (RQ1) Does the semantic dispersion of high-frequency verbs increase with CEFR level? (RQ2) Is the timing of that expansion uniform across verbs, or verb-specific? (RQ3) At which layers is the effect located, and does it survive the known geometric biases of contextual representations? We contribute a layer-wise von Mises–Fisher framework with essay-level bootstrap confidence intervals. The resulting analysis shows that the timing of semantic expansion is verb-specific and robust across all contextual layers. An anisotropy-aware log-ratio analysis then connects our concentration measure to the variance-based corpus-comparison method of Nagata et al. (2023).

2. Background

2.1. Semantic Expansion in Learner Corpora

The Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001) describes L2 development as a progression through can-do levels. Learner corpora annotated for CEFR make it possible to track how a linguistic feature changes along that progression. Vocabulary development is commonly framed in terms of both breadth (how many words a learner knows) and depth (how richly each is known). For high-frequency verbs, depth includes the widening of a verb's constructional and idiomatic range, reflecting the shift from a few core senses to the many patterned uses of proficient writing.

We operationalise this widening as the dispersion of a verb's contextual representations within a CEFR band. In the sense introduced in Section 1, each occurrence of a verb has its own vector (768 dimensions in BERT for a single instance of a given verb in a sentence); the more varied a verb's uses at a given level, the more these occurrence vectors point in different directions, i.e. the more they spread out. Comparing that spread across A1, A2 and B1+

thus turns an abstract claim about "richer verb use" into a quantity we can estimate and test. We focus on verbs rather than nouns because their extensibility is carried by constructions, and on high-frequency verbs because they furnish enough occurrences at every level for stable estimation.

2.2. The Geometry of Contextual Embeddings

Measuring spread requires care, because of how these vectors are arranged in space. It is convenient first to normalise every vector to unit length; each representation then becomes a point on the surface of a sphere (the unit hypersphere), where the direction a point faces encodes its meaning and the angle between two points measures how differently the verb is used. Ideally, a verb's many uses would spread out over this surface. In practice they do not: contextual vectors crowd into a narrow cone, occupying a small region in which almost all points face a similar direction. This tendency is known as anisotropy, or the representation-degeneration problem (Gao et al., 2019; Ethayarajh, 2019; Mu and Viswanath, 2018; Timkey and van Schijndel, 2021).

As illustrated in Figure 1, anisotropy has a direct consequence for us. Any statistic that summarises the concentration of a set of vectors will be distorted upward, because the vectors already share a common direction before we look at their meaning: everything appears more tightly clustered than it really is. The distortion is strongest at the embedding layer, where the model has not yet pulled representations apart by context. Since our entire method rests on comparing how spread out a verb's uses are, we cannot take a raw concentration value at face value; we must control for this shared component, which is the strategy set out in the next section.

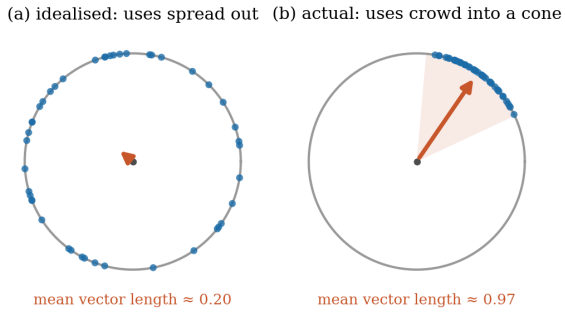

Figure 1: Why raw concentration is inflated (a two-dimensional schematic, not from the corpus data). (a) if a verb's uses spread over the sphere, the mean vector is short; (b) real contextual vectors crowd into a cone (anisotropy), so the mean vector is long, and concentration looks high, regardless of meaning.

2.3. Concentration as von Mises–Fisher κ

To quantify spread on the sphere we need a probability distribution defined on the sphere itself. The von Mises–Fisher (vMF) distribution is the natural counterpart of the familiar bell-shaped (Gaussian) distribution, wrapped onto the spherical surface. It has two parameters: a mean direction, μ, which is the "average" direction the vectors face — the verb's typical use — and a concentration parameter, κ (kappa), which behaves like the inverse of variance. A large κ means the occurrences huddle tightly around the mean direction, i.e. a fixed, narrow use; a small κ means they fan out, i.e. a varied, diffuse use. On this reading, a fall in κ from A1 to B1+ therefore means that the verb's uses have spread apart as proficiency rises. Because κ cannot be solved for exactly in high dimensions, we estimate it with the standard closed-form approximation of Banerjee et al. (2005).

As noted in Section 2.2, raw κ is inflated by anisotropy, so we do not compare κ values by simple subtraction. A closely related idea comes from Nagata et al. (2023). They detect semantic differences between two corpora from the norm of a word's mean vector, which is a variance-based measure of how much of the word's meaning space is covered. This requires no word or corpus alignment, and the authors show it to be robust to differences in corpus size (demonstrated, among others, on native versus non-native English). Their measure and our κ are geometrically related, as both derive from the length of the mean vector. Adapting this idea, we compare levels on a ratio (logarithmic) scale rather than by subtraction. This is because the shared anisotropic component enters both κ estimates in the same way, so taking the ratio cancels it and leaves only the genuine difference in spread. We report this scale-free log-ratio alongside raw κ as a robustness check. The same comparison, a log-ratio of vMF concentration between two corpora, was recently found by Kishino et al. (2025) to be the strongest form-based indicator of whether a word's meaning has broadened or narrowed, which further supports its use here.

3. Data

We use the JEFLL Corpus of writing by Japanese EFL learners (Tono, 2007), in which essays carry CEFR labels. Due to the small samples for B2, we merge B1 and B2 into a single B1+ band, giving three levels A1, A2 and B1+. We focus on three that are frequent, polysemous, and span a lexical-to-constructional continuum of extensibility: *like*, *make* and *take*. Sentences containing a target verb were extracted from JEFLL with CQL queries in Sketch Engine (Kilgarriff et al., 2014), and the target verb was located within each sentence using spaCy (Honnibal et al., 2020). Each sentence keeps its essay identifier, so that dependence among sentences from the same essay can be handled at estimation time. Table 1 reports the resulting counts.

Verb	A1	A2	B1+	Total
take	277	930	888	2,095
make	330	1,146	939	2,415
like	1,324	2,309	1,245	4,878

Table 1: Sentence counts per verb and CEFR level (B1/B2 merged into B1+).

4. Method

4.1. All-layer Verb Vector Extraction

We use bert-base-uncased with output_hidden_states enabled, yielding thirteen representations per token (the embedding layer L0 plus twelve transformer layers). For each sentence the target verb is located by character offset; when the verb is split into sub-words the sub-word vectors are averaged. Every vector is L2-normalised so that it lies on the unit sphere, the domain on which the vMF distribution is defined.

4.2. Layer-wise Probing

To anchor the concentration analysis to a principled layer, we probe each layer independently. For probing, 200 occurrences were randomly sampled from each of the three CEFR bands for each verb, yielding a balanced dataset of 600 instances per verb. Each verb was probed separately. At each layer, a logistic-regression classifier was trained to predict CEFR level from the target verb representation. Cross-validation used StratifiedGroupKFold with essay ID as the grouping variable, ensuring that sentences from the same essay never occurred in both training and test sets while maintaining approximately balanced CEFR distributions across folds. Because the three classes were equally represented, chance-level accuracy was .33.

Importantly, layer-wise probing should not be interpreted as identifying a layer that encodes more advanced semantic uses of the verb. Rather, it indicates how readily CEFR-relevant information can be decoded from the representation at each layer. Such information may reflect not only verb sense, but also collocation, syntactic construction, surrounding vocabulary, grammatical complexity, topic, and other contextual properties. The peak layer should therefore be understood as the layer at which CEFR-relevant contextual information associated with the target verb is most accessible to a simple linear classifier.

Probing each verb separately locates a distinct peak layer, which we take as that verb's anchor for the concentration analysis: *take* at L6, *make* at L11, and *like* jointly at L2 and L11 (identical accuracies). As Section 5.4 shows, the central findings do not in any case depend on this choice.

4.3. Concentration Estimation

For every verb, level and layer we fit a vMF distribution to the corresponding set of unit vectors and record κ (Banerjee et al., 2005). Higher κ indicates a more concentrated, more fixed usage; lower κ indicates a more dispersed, more varied usage.

4.4. Bootstrap Confidence Intervals

Point estimates alone cannot tell us whether an observed change in κ is more than sampling noise. We therefore attach 95% confidence intervals by an essay-level cluster bootstrap (B = 2,000): essays are resampled with replacement and all sentences of a resampled essay are taken together, which respects the non-independence of sentences within an essay and yields appropriately conservative intervals. Two adjacent levels are treated as reliably different when their intervals do not overlap.

4.5. Anisotropy-aware Comparison

Because raw κ is inflated by anisotropy (Section 2.2), we complement it with the log-ratio ln(κ_i / κ_j) between levels. The ratio is invariant to any multiplicative component shared by the two κ estimates and therefore expresses the difference on a relative rather than an absolute scale. This reduces sensitivity to common scaling effects associated with anisotropy, although it does not eliminate anisotropy itself. This is in the spirit of the variance-based comparison of Nagata et al. (2023). As an independent check we also compute the mean pairwise cosine distance within each set of vectors, a dispersion measure that increases as vectors spread apart.

4.6. Collocation Analysis

To triangulate the findings, a more traditional corpus-based analysis was undertaken. Collocates of the verbs were identified by running spaCy's dependency parser with the en_core_web_sm model on all of the sentences containing *like, make*, and *take*. Words with the following relationships to the verb were classed as collocates: noun, coordinated noun, proper noun, pronoun, gerund, to-infinitive, predicate adjective, and passive subject. Where no collocate was identified, the sentence was coded as "none", except in the case of clausal and phrasal relationships, where only the relationship was recorded. The percentage of the usage of collocates from each category within each CEFR level for each verb was recorded. Then, specific collocates were extracted and ranked by their increase and decrease in usage across CEFR levels. Examples were manually examined to account for parsing errors. These usage patterns are described in Section 6.

As *take, make* and *like* have differing functions, a further dimension of the collocates was captured by measuring their lexical diversity. HD-D (McCarthy & Jarvis, 2010) was calculated for verb-noun collocations using only the noun collocate in each sentence. Proper nouns were not included. Verb-noun collocations were selected, as they accounted for the most frequent category across all three verbs. The number of samples was determined by the CEFR level with the lowest number of sentences featuring verb-noun collocations. The Python package LexicalRichness (Shen, 2023) was used for the analysis.

5. Results

5.1. Layer-wise Probing

The logistic regression results for *take, make* and *like* across layers are shown in Figure 2. Each verb is classified above the three-class chance level of .33, but with clearly different strengths and peak layers.

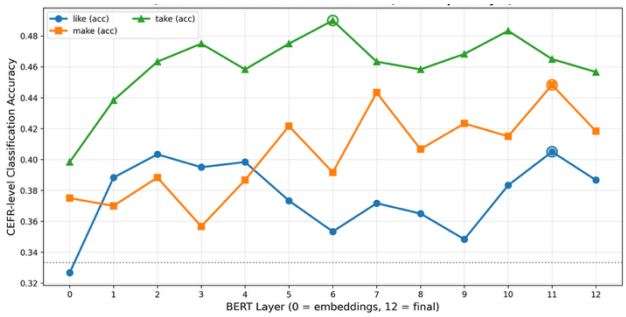

Figure 2: Layer-wise CEFR-classification accuracy for each verb (circles mark peak layers; dotted line = chance).

The verb *take* is the most separable, peaking at L6 with an accuracy of .49 (+.16 over chance), indicating that CEFR level is most accurately predicted from the representation of *take* at this layer. *Make* is next, peaking at L11 (.45, +.11), while *like* is the weakest, though still above chance, with identical peaks at L2 and L11 (.40, +.07). The peak layers therefore differ markedly in depth: lower for *like*, middle for *take*, and upper for *make*. We return to the interpretation of this pattern in Section 5.5.

The verb-specific nature of these layer profiles becomes clearer when the three verbs are pooled in a single probing analysis. When tokens of *take*, *make*, and *like* are combined, the resulting profile is flatter and peaks at L2 with an accuracy of .42 (Figure 3).

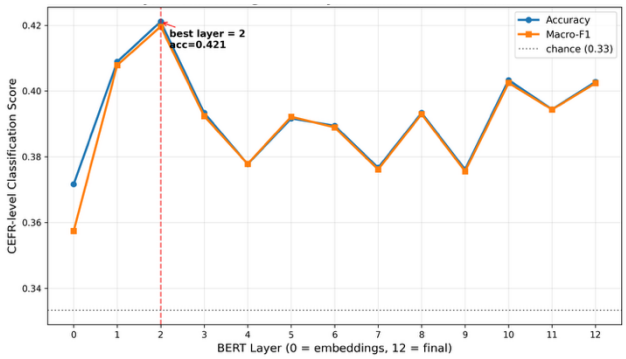

Figure 3: Layer-wise probing: all three verbs combined

This suggests that there is no single BERT layer at which CEFR-related information is optimally represented for all verbs. Rather, the accessibility of such information appears to depend on the lexical and contextual properties of the individual verb. The middle-layer peak for *take*, the upper-layer peak for *make*, and the lower-layer peak for *like* are therefore partly obscured when the verbs are analysed together, because their different layer-wise profiles average one another out.

5.2. Concentration Decreases with Proficiency

Across all three verbs, κ at the anchor layer falls from A1 to B1+, and in every case the A1 and B1+ 95% CIs do not overlap (Table 2, Figure 4): *take* 1626→1368, *make* 1430→1305, *like* 2333→2137. Higher-proficiency writing thus uses these verbs in a measurably more dispersed way.

Verb	Level	κ	95% CI
take	A1	1626	[1545, 1742]
take	A2	1362	[1326, 1406]
take	B1+	1368	[1322, 1423]
make	A1	1430	[1384, 1493]
make	A2	1322	[1300, 1348]
make	B1+	1305	[1279, 1337]
like	A1	2333	[2268, 2406]
like	A2	2334	[2287, 2385]
like	B1+	2137	[2060, 2229]

Table 2: vMF κ with 95% essay-level bootstrap CIs (B = 2,000) at each verb's anchor layer (*take* L6, *make/like* L11).

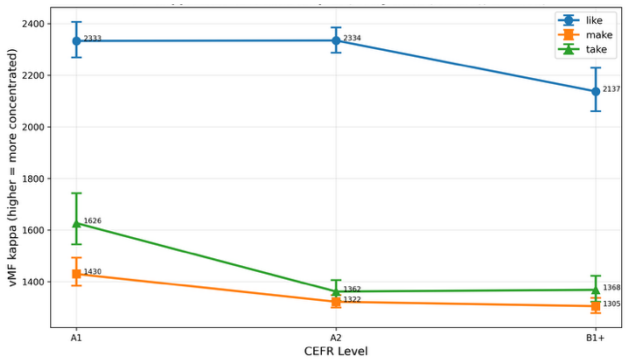

Figure 4: vMF κ by CEFR level with 95% cluster-bootstrap CIs.

5.3. The Timing of Expansion is Verb-Specific

The overall decline conceals a difference in when it happens. For *take*, the drop is concentrated in A1→A2 (1626→1362) while A2→B1+ is flat. The verb *make* behaves the same way: a substantial A1→A2 drop (1430→1322) followed by a lesser A2→B1+ change in the same direction. The verb *like* is the mirror image of the other two: A1→A2 is essentially flat (2333→2334) and a notable decline appears only at A2→B1+ (2334→2137). The verbs *take* and *make* therefore expand early, while the verb *like* expands late.

5.4. The Asymmetry is Robust Across Layers

This timing pattern is not an artefact of the anchor layer. Comparing the size of the early (A1→A2) and late (A2→B1+) κ drops at each of the twelve contextual layers, *take* and *make* show a larger early drop at all 12/12 layers, and *like* shows a larger late drop at all 12/12 layers. The verb-specific timing thus holds regardless of which layer is chosen for reporting, and does not depend on the probing peaks of Section 5.1.

5.5. Anisotropy and the Layer Profile of Expansion

The A1–B1+ gap in raw κ peaks at the embedding layer L0 for all three verbs (*take* 1026, *make* 142, *like* 4296). L0 is also the most anisotropic layer, with mean pairwise cosine distance lowest there (*like* .08, *take* .14, *make* .22) and rising to .25–.42 in contextual layers. Therefore, a large raw gap at L0 partly reflects anisotropy, not a genuine semantic difference (Figure 5a). The log-ratio controls for this effect and separates three genuine profiles (Figure 5b). For *like* the gap stays L0-dominant even after normalisation (.63), which suggests that part of the CEFR difference may arise from input-level properties such as surface-form and positional variation rather than contextual semantics. For *make* the peak moves up to L11 (.09), marking a context-dependent profile. The verb *take* remains front-loaded but flatter, with a mid-layer peak, i.e. a mixed profile. We thus describe *like* as input-lexical, *make* as context-dependent, and *take* as mixed. Reassuringly, these profiles line up with the probing peaks of Section 5.1: the layer at which each verb's single-token representation best separates CEFR levels is low for *like* (L2), middle for *take* (L6) and upper for *make* (L11), which is an independent corroboration of the verb-specific profiles identified above.

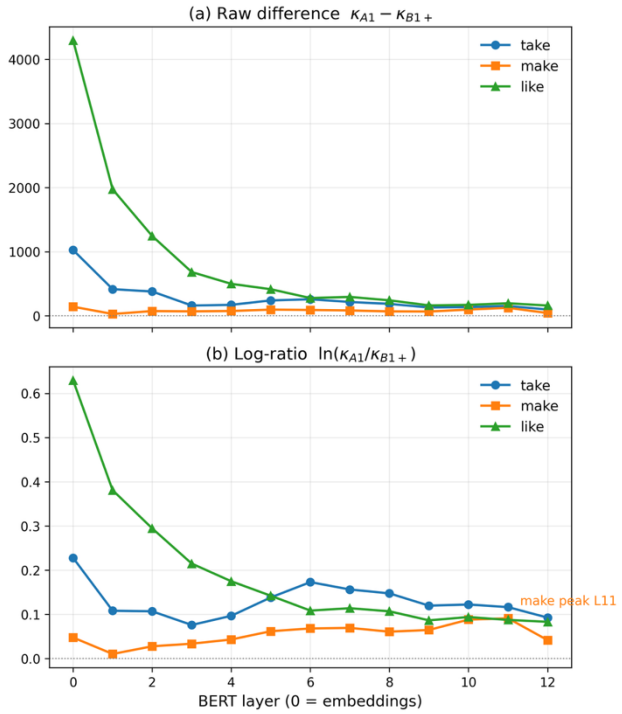

Figure 5: A1–B1+ dispersion gap by layer. (a) raw κ difference peaks at L0 for all verbs; (b) the scale-free log-ratio separates three profiles and moves make's peak to L11.

5.6. Comparison with Mean Pairwise Cosine Difference

As a geometric consistency check, we compared vMF κ with mean pairwise cosine distance across layers and CEFR levels. The two measures are strongly negatively correlated within each verb (r = −.96 for *take*, −.98 for *make*, and −.92 for *like*; Figure 6): more tightly clustered contextual representations show larger κ values and smaller average cosine distances, whereas more dispersed representations show lower κ values and larger cosine distances. This close relationship is expected rather than independent evidence, because both measures are mathematically related to the length of the mean resultant vector. Figure 6 therefore confirms that the estimated κ values behave consistently with the underlying geometry of the contextual representations. Independent linguistic support for interpreting this dispersion as diversification of verb use is provided by the collocational analysis in Section 6.

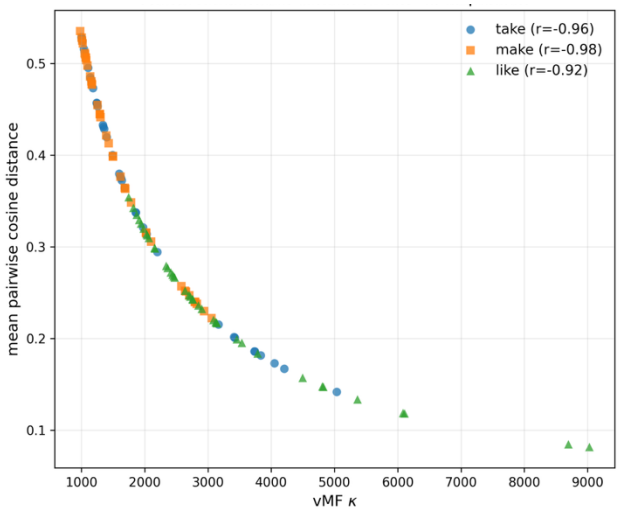

Figure 6: κ against mean pairwise cosine distance across all layers and levels; the two dispersion measures are tightly and negatively related.

6. Collocation Analysis

The most frequent collocation category across all three verbs was nouns. They accounted for collocations in the following percentages of sentences depending on the CEFR level: *take*: 70–72%, *make*: 42–46%, *like*: 52–54%. The lexical diversity results of these verb-noun collocations are shown in Table 3. The pattern for *take* and *like* is similar, as there is little or no difference between A1 and A2, then an increase between A2 and B1+. This suggests that at B1+, learners increase the range of nouns they can use with these two verbs. For *make*, the highest lexical diversity score is at the A1 level, decreasing to the A2 level, then increasing slightly to the B1+ level.

Verb	A1	A2	B1+	n
take	.46	.45	.49	198
make	.67	.61	.64	137
like	.26	.26	.29	521

Table 3: Lexical diversity (HD-D) of noun collocates in verb-noun collocations.

Qualitative examination of the collocates showed that the development across CEFR levels was not only lexical, but also constructional. The collocates of *take* that increased in usage the most, particularly from A1 to A2 level, but also from A1 to B1+ level were *take part* (2.5%→5.4%→6.4%), *take time* (0.7%→3.3%→3.3%), and *take place* (0%→1.7%→1.7%). These collocations, used in sentences such as "Our school festival took place on September" (A2) are more abstract and phraseological in nature than the most frequent collocations at A1 level. These included *money* (13.4%→5.2%→3.8%) and *bankbook* (6.9%→3.5%→0.3%) in sentences such as "I will take out money and bank book first" (A1). These two collocations were clearly influenced by the writing prompt, which asked learners to describe what they would take with them in the event of an earthquake. The earthquake topic was the most dominant across all CEFR levels for *take* (67.1%→63.3%→59%), but the starkest difference in usage within a topic was *breakfast* (3.2%→9.6%→11.2%). Although some of these collocations were related to food

items or meals (e.g. “I always take a breakfast on weekend”, A1), many of them were related to time, such as “But it takes a lot of time” (A2) and “It takes about 10-12 minutes to eat” (B1+).

For *make*, the noun collocations were less revealing, as the most frequent collocations were related to concrete concepts. The biggest increases from lower to higher levels were *house* (0.6%→3.2%→2.6%) and *movie* (1.8%→5.8%→3.7%). These were largely from essays on the *festival* topic about making ghost houses and movies, particularly by A2 and B1+ learners. This may partly explain the lower lexical diversity for *make* at these levels compared with A1. The biggest decrease in usage was for the collocate *rice* (1.5%→0.5%→0.3%), which also does not represent a change in usage. A different perspective was revealed when collocations with a clausal relationship were examined. These accounted for a substantial proportion of the collocations for *make* across A1, A2 and B1+ (17.1%→26%→24.3%). There was an increase in usage between A1 and A2, before a plateau at B1+. Upon manual examination, these collocations were largely causative constructions such as “milk makes me strong” (A1), “Baking them well made me so happy” (A2), and “Bread has a lot of butter in it, and that makes me feel bad after eating bread.” (B1+). Although these constructions are used from the A1 level, the percentage increase in usage provides evidence of progression across CEFR levels.

There was a similar pattern in the *like* collocations. The most frequently increasing and decreasing noun collocates across CEFR levels were concrete nouns. The biggest increase was *bread* (10.7%→12.5%→13.1%) and the biggest decrease was *milk* (4.7%→4.7%→2.3%). To-infinitive collocates were also fairly well represented for *like* and usage increased across CEFR levels (3.7%→8%→10.9%). The most frequent collocate was “to eat” (0.6%→2.6%→3.3%), used in examples such as “I don't like to eat rice” (A1), “I like to eat bread in the morning” (A2), and “Really, I would like to eat rice, because my family like to eat rice.” (B1+).

7. Discussion

A natural question is why the timing should differ across verbs. One possibility is that there are differences in the diversity of contexts each verb affords. The verbs *make* and *take* are highly polysemous: they head light-verb (e.g. *take charge*), causative (e.g. *make me tired*), phrasal verb (e.g. *make up*) and idiomatic (e.g. *take care of*) constructions, and this constructional range exposes them to varied contexts early, broadening their representations by the A2 level. The verb *like*, by contrast, has a narrower range as a main verb. It forms neither light-verb nor phrasal verb constructions, so its diversification depends on the range of nouns it takes, which broadens only at B1+ (Table 3). This account is broadly consistent with the layer profiles reported in Section 5.5. For *like*, the proficiency-related difference is strongest toward the input and lower layers, whereas *make* shows a stronger upper-layer profile. This pattern is compatible with the possibility that the development of *like* depends more heavily on relatively lexical properties of its use, while the diversification of *make* involves more context-dependent constructional information.

These profiles also resolve an apparent tension in our results. The layer-wise probe predicts proficiency only weakly from a single occurrence of a given verb. *Like* is the weakest of the three verbs, and yet it is plainly used in more varied ways at B1+. The tension is only apparent, because the two analyses answer different questions. The probe asks whether one verb token, on its own, reveals the writer’s level. On the other hand, the concentration measure κ asks whether the full set of a verb’s uses has spread out. A verb’s range of use can widen without any single instance becoming a reliable cue to the writer’s level, which is precisely what we find for *like*.

The development across CEFR levels varying by verb was supported in part by the qualitative evidence discussed in Section 6. For *take*, the use of abstract nouns followed the same early increase trajectory as the κ results. The same was true for *make* with causative constructions. The increase in to-infinitive collocates for *like* seems counterintuitive, as the largest increase was also between A1 and A2 level. However, the lexical diversity results align with the κ results, with an increase in diversity only evident between A2 and B1+ level. Although not fully acquired at the A1 level, to-infinitives might not be as difficult to master as other constructions and are one example of an earlier acquisition of one collocation type. On the other hand, the increase in noun diversity is a more gradual development that aligns with κ and may be more reflective of the semantic development of *like*. Lexical diversity for *take* also only increased at the B1+ level. This contrasts with the early rise in κ, suggesting that the developmental path for *take* represents information beyond individual word usage, and is likely influenced by the shift towards abstract, phraseological collocations (e.g. place/part/time) that is not captured by lexical diversity.

Methodologically, the verb profiles only come into focus once the embedding layer is handled with care. The raw concentration measure is inflated there by anisotropy — the tendency of contextual vectors to crowd into a narrow region of the space rather than spreading across it — which exaggerates the differences seen at that layer. Reporting the scale-free log-ratio alongside raw κ removes this distortion and, as Section 5.5 showed, sharpens rather than blurs the contrast between verbs.

The verb-specific timing itself, with earlier broadening for the more polysemous verbs, is reminiscent of proposed regularities in historical semantic change. Hamilton et al. (2016) report a law of innovation, whereby more polysemous words change faster independently of frequency, alongside a law of conformity, whereby more frequent words change more slowly (see also Winter et al., 2014, on cognitive factors motivating meaning change). If a similar dynamic operated during acquisition, then the greater polysemy of *make* and *take* could plausibly drive the earlier broadening we observe, with the diversification of *like* lagging behind. We offer this as an interpretive direction rather than a demonstrated mechanism.

Two cautions delimit the inference. First, polysemy and frequency are confounded in our three-verb sample: *like* is not only the least constructionally diverse but also by far the most frequent (Table 1), and the law of conformity would on its own predict slower change for a high-frequency verb. Disentangling the two would require a design that varies polysemy while holding frequency constant, or vice versa, which is a clear next step. Second, historical change is a community-level process over time, whereas the trajectory studied here is developmental and

learner-internal. Whether the same statistical regularities transfer across these scales is itself an open question. Testing for a law-of-innovation-like effect in L2 verb acquisition, with frequency controlled for and a larger, more varied verb set, is a promising avenue. A complementary direction is to ask which individual usage instances drive each verb's expansion; instance-level optimal-transport measures such as Sense Usage Shift (Kishino et al., 2025) could quantify this, refining the aggregate picture that κ provides.

For language assessment, dispersion trajectories of this kind could complement frequency- and can-do-based vocabulary profiles, as they indicate not only which verbs learners use, but how flexibly they use them and when that flexibility emerges.

8. Limitations

Three limitations qualify these results. First, although every verb is classified above chance, the probing accuracies are modest (peak .40–.49); single-token representations therefore encode CEFR level only weakly, and the anchor layers should be read as where that weak signal is strongest rather than as sharp boundaries. Second, we use a single model (bert-base-uncased) with three verbs, and the JEFLL writing prompts may introduce topic effects; broader model and verb coverage is needed. Third, although the log-ratio controls for a shared anisotropic component, it does not remove anisotropy entirely; stronger corrections such as all-but-the-top postprocessing (Mu and Viswanath, 2018) are left for future work.

9. Conclusion

Semantic expansion of high-frequency verbs in L2 writing is not uniform but verb-specific in its timing: *take* and *make* expand early, *like* expands late. The pattern is supported by essay-level bootstrap confidence intervals, robust across all twelve contextual layers. An anisotropy-aware log-ratio both confirms the effect and reveals three distinct layer profiles: input-lexical, context-dependent and mixed. Interpretable, statistically grounded analysis of contextual embeddings can thus offer a fine-grained view of L2 lexical development.

10. Data Availability Statement

The data and code used in the current study are available in the semantic_expansion_verbs_vmf GitHub repository: https://github.com/cooperchris17/semantic_expansion_verbs_vmf

References

Banerjee, A., Dhillon, I. S., Ghosh, J., and Sra, S. (2005). Clustering on the unit hypersphere using von Mises–Fisher distributions. *Journal of Machine Learning Research*, 6:1345–1382.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge University Press, Cambridge.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT,* pages 4171–4186.

Ethayarajh, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of EMNLP-IJCNLP,* pages 55–65.

Gao, J., He, D., Tan, X., Qin, T., Wang, L., and Liu, T.-Y. (2019). Representation degeneration problem in training natural language generation models. In *Proceedings of ICLR*.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the ACL* (Volume 1: Long Papers), pages 1489–1501.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python*.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, *1*(1), 7–36.

Kishino, R., Yamagiwa, H., Nagata, R., Yokoi, S., and Shimodaira, H. (2025). Quantifying lexical semantic shift via unbalanced optimal transport. In *Proceedings of the 63rd Annual Meeting of the ACL* (Volume 1: Long Papers), pages 15913–15933.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2), 381–392.

Mu, J. and Viswanath, P. (2018). All-but-the-top: Simple and effective postprocessing for word representations. In *Proceedings of ICLR*.

Nagata, R., Takamura, H., Otani, N., and Kawasaki, Y. (2023). Variance matters: Detecting semantic differences without corpus/word alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15609–15622.

Shen, L. (2023). *LexicalRichness* (Version 0.5.1) [Computer software]. https://pypi.org/project/lexicalrichness/

Timkey, W. and van Schijndel, M. (2021). All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of EMNL*P, pages 4527–4546.

Tono, Y. (2007). *Nihonjin chūkōsei ichiman-nin no eigo kōpasu: JEFLL Corpus* [*An English corpus of 10,000 Japanese junior and senior high school students: The JEFLL Corpus*]. Tokyo: Shogakukan.

Winter, B., Thompson, G., and Urban, M. (2014). Cognitive factors motivating the evolution of word meanings: Evidence from corpora, behavioral data and encyclopedic network structure. In *Proceedings of EVOLANG*, pages 353–360.