Evaluating Collostructional Measures: On the Combination of Frequency and Log Odds Ratio in Covarying Collexeme Analysis

Kazuho Kambara[1,2], **Norihisa Takahashi**[3], **Takatsugu Kojima**[4], **Hajime Nozawa**[5]
[1] National Institute of Information and Communications Technology, [2] Ritsumeikan University,
[3]Ryukoku University, [4]Shiga University of Medical Science, [5]Kyoto University of Foreign Studies
[1,2,3,5] Kyoto, Japan, [4] Shiga, Japan
[1]kambara-k@nict.go.jp, [2]ec042070@gmail.com, [3]kojima@kojima-lab.net, [4]h_nozawa@kufs.ac.jp

Abstract

In corpus linguistics, quantifying the association strength between lexical items is standard practice, yet selecting the most appropriate association measure (AM) remains a complex challenge both theoretically and practically. This study aims to identify which AMs or their combinations best align with human subjective ratings of linguistic expressions by systematically evaluating various measures employed in collostructional analysis. Using ordered logit models fitted to 5-point Likert-scale judgements across three distinct cognitive dimensions (perceived exposure, active usage, and conceptual comprehensibility) in Japanese covarying collexemes (*N-ga V*; combinations of a noun and an intransitive verb), we evaluated univariate, additive multivariate, and interaction model configurations. The results demonstrate that a model incorporating an interaction between co-occurrence frequency ($\text{Freq}_{\{n,v\}}$) and log odds ratio (LOR) yields the overall best fit. Crucially, this interaction reflects a cognitive threshold effect: although LOR quantifies association strength independently of baseline frequency, it gains predictive utility for subjective judgements only once a sufficient level of co-occurrence frequency is reached.

Keywords: Collostructional Analysis, Covarying Collexeme Analysis, Tupleisation, Log Odds Ratio

1. Introduction

Usage-based models of language posit that linguistic knowledge is fundamentally shaped by cumulative language experience, which can be quantified through corpus-derived statistical metrics. Over the past two decades, collostructional analysis (Gries, 2019, 2023; Gries and Stefanowitsch, 2004; Stefanowitsch and Gries, 2003, 2005) has emerged as the standard quantitative methodology for measuring association strength between lexical items and grammatical slots. Empirical literature has demonstrated robust correlations between corpus-derived metrics and human cognitive processing (Gilquin and Gries, 2009; Gries et al., 2005). However, traditional single-metric approaches conflate multiple statistical variables, prompting a recent methodological shift towards tupleisation (computing multi-element tuples of independent measures) (Gries, 2019, 2024). Yet, which specific combination of disentangled corpus metrics best aligns with distinct dimensions of human cognitive judgements remains unresolved.

The present study addresses this gap by systematically evaluating how disentangled corpus metrics predict subjective human judgements across three cognitive dimensions (perceived exposure, active usage, and conceptual comprehensibility) in Japanese covarying noun-verb constructions (*N-ga V*). Using ordered logit models, we demonstrate that combining co-occurrence frequency ($\text{Freq}_{\{n,v\}}$) and log odds ratio (LOR) in an interaction specification provides the optimal account of native speaker intuitions. Crucially, this interaction reveals a cognitive threshold effect: the predictive impact of LOR (co-occurrence strength) on human judgements becomes apparent only after a necessary baseline of co-occurrence frequency ($\text{Freq}_{\{n,v\}}$) is attained.

The remainder of this paper is structured as follows. Section 2 reviews the theoretical foundations of collostructional analysis, the conflation problem, and the shift towards tupleisation. Section 3 details the stimulus selection, the psycholinguistic rating experiment, and the statistical modelling framework. Section 4 presents the empirical results and discussion across univariate, multivariate, and interaction specifications. Finally, Section 5 offers concluding remarks and broader theoretical implications.

2. Collostructional Analysis and Psycholinguistic Studies

Collostructional analysis was developed to quantify the degree of association between specific lexical items and grammatical constructions (Stefanowitsch and Gries, 2003; Gries, 2023). The term "collostructional analysis" refers to a family of three primary methods designed to capture collocational strength across varying levels of constructional schematicity (Gries, 2019, 2022):

(1) a. **Collexeme analysis**: Measures how (un)likely a given word (e.g., *sister*) occurs in a specific schematic slot (e.g., [*my X*]) (Stefanowitsch and Gries, 2003).
b. **Distinctive collexeme analysis**: Measures how (un)likely a given word (e.g., *sister*)

occurs in functionally competing schematic constructions (e.g., [*my X*] vs. [*X of mine*]) (Gries and Stefanowitsch, 2004).

c. **Covarying collexeme analysis**: Measures how (un)likely given words co-occur across two open slots within a single construction (e.g., {*sister*, *Alice*} → [*X of Y*]) (Stefanowitsch and Gries, 2005).

A central line of inquiry within usage-based linguistics examines the extent to which these corpus-derived metrics reflect human psycholinguistic reality (Bybee, 2010). Empirical studies combining corpus and psycholinguistic methods have provided converging evidence, demonstrating robust correlations between corpus-derived metrics and human cognitive processing (Gilquin and Gries, 2009; Gries et al., 2005, 2010). These findings establish that native speakers' linguistic intuitions are inherently probabilistic, reflecting fine-grained sensitivity to constructional preferences and co-occurrence patterns in input exposure.

Despite these empirical successes, traditional collostructional methods relied almost exclusively on single, composite association measures (most notably p-values derived from Fisher's exact test or Log-Likelihood Ratio (LLR)) (Gries, 2019, 2026). Gries (2019, 2024) points out that single association metrics conflate multiple distinct statistical dimensions, including raw co-occurrence frequency, marginal slot frequencies, sample size N, effect size, and directionality. Because these underlying components are merged into a single scalar value, it remains analytically opaque which specific structural property actually drives the observed statistical preference.

To achieve more accurate and granular linguistic descriptions, recent developments advocate for the systematic disentanglement (de-conflation) of these statistical variables (Gries, 2019, 2024, 2026). Gries (2019, 2024) refers to this approach as **tupleisation**: the practice of replacing a single conflated association metric with a multi-element tuple of independent, non-conflated measures (e.g., co-occurrence frequency, dispersion, sample-size-independent effect size such as LOR, and directional contingency such as ΔP). Computing these dimensions independently allows linguists to observe complex grammatical distribution patterns with precision (Gries, 2024).

While tupleisation represents a major advance in corpus-linguistic description, evaluating how disentangled measures align with psycholinguistic judgements remains a key challenge. Empirical support for the alignment between collostructional metrics and human cognitive processing was established using traditional, *conflated* measures (such as p-value-based metrics). Once these metrics are disentangled into independent tuple axes, it becomes unclear how each individual component, or which specific *combination* (tuple) of components, best predicts human subjective judgements. The present study addresses this gap by systematically testing which *combination* of disentangled corpus metrics best aligns with multi-dimensional cognitive judgements in Japanese covarying collexemes.

3. Methods

This section details the empirical methodology employed to evaluate the predictive relationship between corpus-derived metrics and human cognitive judgements. Section 3.1 describes the selection of target stimuli and the calculation of collostructional measures, Section 3.2 details the psycholinguistic rating experiment administered to native Japanese speakers, and Section 3.3 outlines the statistical framework utilising ordered logistic regression.

3.1. Target Stimuli and Corpus Analysis

To investigate the statistical alignment between corpus metrics and cognitive ratings, candidate lexical items were extracted from the Japanese Web TenTen corpus (Jakubíček et al., 2013) using Sketch Engine (Kilgarriff et al., 2004, 2014).

First, the lists of 1,000 most frequent nouns and verbs were retrieved. From these candidate lists, 10 nouns (N) and 10 verbs (V) were manually selected to encompass a wide spectrum of semantic classes. The selected nouns comprised *tema* ('effort'), *hana* ('flower'), *me* ('eye'), *jiken* ('incident'), *mēru* ('email'), *itami* ('pain'), *umi* ('sea'), *himei* ('scream'), *gimon* ('doubt'), and *namida* ('tear'). The selected verbs comprised *kakaru* ('take/cost'), *saku* ('bloom'), *sameru* ('awaken'), *okoru* ('occur'), *todoku* ('arrive'), *hashiru* ('run'), *hirogaru* ('spread'), *kikoeru* ('be heard'), *nokoru* ('remain'), and *ukabu* ('float').

All possible combinations of these items were constructed using the Japanese nominative case marker *-ga* (i.e., N-*ga* V), yielding a total of 100 unique noun-verb pairs. A covarying collexeme analysis was then conducted on these 100 combinations using the R script `coll.analysis.4.1` (Gries, 2024) to calculate statistical association measures, as explained in (2).

In (2), for a 2×2 contingency table of a target noun n and verb v (see Table 1), let a denote the co-occurrence frequency (noted as $\text{Freq}_{\{n,v\}}$), b the frequency of n with other verbs, c the frequency of v with other nouns, d the frequency of other noun-verb combinations, and $N = a + b + c + d$ the total sample size[1].

[1]Note that in covarying collexeme analysis, N represents the total number of noun-verb combination instances in the corpus for the target construction, rather than the overall word count of the corpus itself.

Table 1: Input table of covarying collexeme analysis

	n (e.g., *flower*)	$\neg n$ (i.e., the other nouns)
v (e.g., *bloom*)	a	c
$\neg v$ (i.e., the other verbs)	b	d

(2)
a. **Freq** ($\text{Freq}_{\{n,v\}}$): Observed frequency of the given noun-verb pair (i.e., co-occurrence frequency).
b. **FreqOfSlot1** (Freq_n): Observed frequency of the given noun.
c. **FreqOfSlot2** (Freq_v): Observed frequency of the given verb.
d. **Pointwise Mutual Information** (PMI): Quantifies association strength by comparing the joint probability of two words against their expected probability under independence. It tends to overstate association strength for low-frequency items.

$$\text{PMI} = \log_2 \left(\frac{a \cdot N}{(a+b)(a+c)} \right)$$

e. **Log Likelihood Ratio** (LLR): Evaluates statistical significance (G^2 test statistic) by measuring the deviation of observed frequencies (O_i) from expected baseline frequencies (E_i). Its value scales directly with overall sample size N.

$$\text{LLR} = 2 \sum_{i \in \{a,b,c,d\}} O_i \ln \left(\frac{O_i}{E_i} \right)$$

where $E_a = \frac{(a+b)(a+c)}{N}$, $E_b = \frac{(a+b)(b+d)}{N}$, $E_c = \frac{(a+c)(c+d)}{N}$, and $E_d = \frac{(b+d)(c+d)}{N}$.

f. **Log Odds Ratio** (LOR): Measures the effect size of an association by calculating the natural logarithm of the odds ratio. Independent of sample size N, it provides a stable estimate of attraction and repulsion between items.

$$\text{LOR} = \ln \left(\frac{a \cdot d}{b \cdot c} \right)$$

g. **DeltaP1To2** ($\Delta P_{n \to v}$): A directional contingency metric quantifying cue strength from noun to verb. It evaluates the probability of encountering verb v given noun n minus the probability of encountering v in the absence of n.

$$\Delta P_{n \to v} = \frac{a}{a+b} - \frac{c}{c+d}$$

h. **DeltaP2To1** ($\Delta P_{v \to n}$): A directional contingency metric quantifying cue strength from verb to noun. It evaluates the probability of encountering noun n given verb v minus the probability of encountering n in the absence of v.

$$\Delta P_{v \to n} = \frac{a}{a+c} - \frac{b}{b+d}$$

i. **KLD1To2** (i.e., $\text{KLD}_{n \to v}$): Kullback-Leibler Divergence measuring directional information divergence from noun to verb. It quantifies how much the distribution of verbs co-occurring with noun n diverges from the baseline distribution of verbs.

$$\text{KLD}_{n \to v} = P_1 \log_2 \left(\frac{P_1}{Q_1} \right) + P_2 \log_2 \left(\frac{P_2}{Q_2} \right)$$

where $P_1 = \frac{a}{a+b}$, $P_2 = \frac{b}{a+b}$, $Q_1 = \frac{a+c}{N}$, and $Q_2 = \frac{b+d}{N}$.

j. **KLD2To1** (i.e., $\text{KLD}_{v \to n}$): Kullback-Leibler Divergence measuring directional information divergence from verb to noun. It quantifies how much the distribution of nouns co-occurring with verb v diverges from the baseline distribution of nouns.

$$\text{KLD}_{v \to n} = P_1 \log_2 \left(\frac{P_1}{Q_1} \right) + P_2 \log_2 \left(\frac{P_2}{Q_2} \right)$$

where $P_1 = \frac{a}{a+c}$, $P_2 = \frac{c}{a+c}$, $Q_1 = \frac{a+b}{N}$, and $Q_2 = \frac{c+d}{N}$.

3.2. Subjective Rating Task

A total of 110 undergraduate students of Shiga University of Medical Science, all native speakers of Japanese (52 males, 58 females; mean age = 23.68 years, SD = 4.96), participated in the evaluation experiment. The experimental protocol was reviewed and approved by the Ethics Committee of Shiga University of Medical Science (Approval No. RBB-067). Participation was entirely voluntary, and all respondents were informed prior to the experiment that declining to participate or withdrawing at any stage would have no impact on their course evaluations or academic standing.

The survey was administered online via a custom-built Web interface. To familiarise participants with the task and the 5-point Likert rating scales, a practice block consisting of 10 practice trials was conducted prior to the main session as shown in (3).

(3)
a. *okome-ga neru* ('Rice sleeps')
b. *okome-ga tobu* ('Rice flies')
c. *sumaho-ga kowaru* ('A smartphone breaks')

d. *sumaho-ga hajimaru* ('A smartphone starts')
e. *oni-ga kowareru* ('A demon breaks')
f. *oni-ga naku* ('A demon cries')
g. *jugyo-ga hajimaru* ('A class starts')
h. *jugyo-ga neru* ('A class sleeps')
i. *tani-ga naku* ('An academic credit cries')
j. *tani-ga tobu* ('An academic credit flies')

Following the practice session, participants proceeded to the main experimental session. To eliminate potential presentation order and fatigue effects, the sequence of the 100 target noun-verb pairs was fully randomised for each participant. Respondents evaluated each pair across three distinct cognitive dimensions on a 5-point Likert scale (1 = lowest, 5 = highest):

(4)
a. **Exposure**: The perceived frequency of encountering (reading or hearing) the expression.
b. **Usage**: The perceived frequency of actively producing (speaking or writing) the expression.
c. **Comprehensibility**: The subjective ease of understanding the conceptual meaning of the expression.

3.3. Statistical Analysis

To evaluate how corpus-derived metrics predict human psycholinguistic judgements, **ordered logistic regression models** (ordered logit models) were fitted to the experimental data. Subjective evaluations were recorded on a 5-point Likert scale (1–5) across three distinct cognitive dimensions (Exposure, Usage, and Comprehensibility).

Because Likert-scale ratings represent ordinal categorical data where responses are rank-ordered but the interval distances between consecutive ratings cannot be assumed to be equal, standard linear regression is methodologically inappropriate. Ordered logit models resolve this issue by positing an unobserved, continuous latent variable Y_i^* that underlies the observed ordinal rating $Y_i \in \{1, 2, 3, 4, 5\}$. The latent space is mapped onto the discrete rating categories via four strictly increasing threshold cut-points ($\tau_1 < \tau_2 < \tau_3 < \tau_4$):

(5)
$$Y_i = k \quad \text{if} \quad \tau_{k-1} < Y_i^* \leq \tau_k$$
(where $\tau_0 = -\infty$ and $\tau_5 = +\infty$)

Formally, the cumulative log-odds of a subjective rating Y_i falling into or below category k ($k \in \{1, 2, 3, 4\}$) given a vector of predictor variables $\mathbf{X}_i$ is expressed as:

(6)
$$\begin{aligned}\text{logit}(P(Y_i \leq k \mid \mathbf{X}_i)) &= \ln\left(\frac{P(Y_i \leq k \mid \mathbf{X}_i)}{1 - P(Y_i \leq k \mid \mathbf{X}_i)}\right) \\ &= \tau_k - \boldsymbol{\beta}^T \mathbf{X}_i\end{aligned}$$
(where τ_k denotes the category-specific threshold intercept, and $\boldsymbol{\beta}$ is the vector of regression coefficients)

Under this parameterisation, a positive coefficient ($\hat{\beta} > 0$) indicates that an increase in the predictor variable systematically shifts the distribution towards higher subjective ratings.

All statistical analyses were executed in R (R Core Team, 2026) using the `ordinal` package (Christensen, 2026). Model comparisons across univariate, additive multivariate, and interaction specifications were conducted using Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Log-Likelihood (abbreviated as Log-Lik.) statistics.

4. Results & Discussion

This section presents the empirical findings of the study[2]. Section 4.1 provides a descriptive overview of the subjective ratings across corpus-derived association categories, while Section 4.2.1 and Section 4.2.2 detail the statistical modelling using ordered logistic regression.

4.1. Descriptive statistics

We observed how corpus-derived association categories (Attraction vs. Repulsion) distribute across the 5-point Likert ratings for each cognitive dimension[3]. As shown in Figure 1, the proportion of Attraction pairs generally increases monotonically alongside higher subjective ratings for both Exposure (encounter frequency) and Usage (use frequency). For these two dimensions, expressions assigned the highest rating (level 5) consist entirely of Attraction pairs (100%), demonstrating a tight alignment between subjective familiarity/production and corpus-derived co-occurrence statistics.

In contrast, Comprehensibility exhibits a subtle dissociation. While lower ratings (levels 1–3) are predominantly

[2] All code and data used in this study are available at: https://osf.io/pre7g.

[3] In the `coll.analysis 4.1` script (Gries, 2024), the association between Slot 1 and Slot 2 items (RELATION) was determined strictly by the algebraic sign of the LOR (log odds ratio). Item pairs yielding a positive LOR (i.e., LOR > 0) were categorised as Attraction, indicating a co-occurrence frequency higher than expected under independence. Conversely, pairs with a negative LOR (i.e., LOR < 0) were categorised as Repulsion, reflecting co-occurrence below baseline expectations. Pairings with an LOR equal to zero (i.e., LOR = 0) were designated as *chance*.

Figure 1: The proportions of attracted/repelled items in each evaluation task

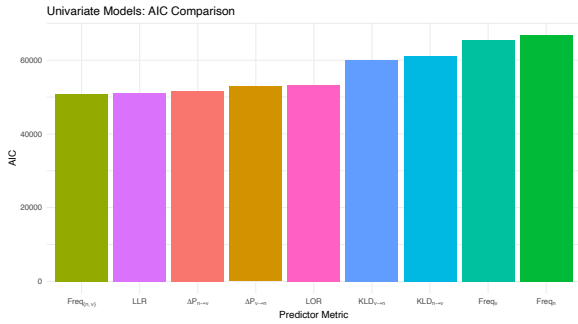

Figure 2: AIC-based comparison of each predictor

characterised by Repulsion and higher ratings lean towards Attraction, a subset of expressions given the maximum comprehensibility score (level 5) remains classified as Repulsion. Qualitative examination identifies four specific noun-verb pairs driving this divergence (i.e., *hana-ga todoku* ('flowers arrive'), *gimon-ga kikoeru* ('doubts/questions are heard'), *himei-ga okoru* ('screams arise/occur'), *me-ga todoku* ('eyes reach / keep an eye on')).

Although these combinations occur less frequently in the web corpus than expected under baseline independence (thus yielding a statistical Repulsion classification), they do not involve highly abstract metaphors or elaborate rhetorical devices. Rather, native speakers process their conceptual meanings with complete transparency, suggesting that subjective comprehensibility is governed by general semantic compositionality rather than strict statistical co-occurrence.

4.2. Statistical modelling

To evaluate how collostructional measures predict human psycholinguistic judgements, ordered logistic regression models were fitted to the rating data. We first examine the independent performance of each individual predictor using univariate models (Section 4.2.1). We then evaluate multivariate additive and interaction specifications to test whether combining disentangled metrics improves predictive capacity (Section 4.2.2). Overall, while co-occurrence frequency ($\text{Freq}_{\{n,v\}}$) serves as the single strongest univariate baseline predictor, incorporating LOR (log odds ratio) within an interaction model ($\text{Freq}_{\{n,v\}} \times \text{LOR}$) achieves the optimal explanatory power. Crucially, this pattern uncovers a cognitive threshold effect: the predictive impact of LOR (co-occurrence strength) on human judgements becomes apparent only after a requisite threshold of co-occurrence frequency ($\text{Freq}_{\{n,v\}}$) is attained.

4.2.1. Univariate ordinal logistic regression

To evaluate the independent predictive power of each corpus-derived measure, single-variable ordinal logistic regression models were fitted to the subjective ratings. Note that Pointwise Mutual Information (PMI) was excluded from regression modelling as zero co-occurrence pairs yield infinite values (i.e., $-\infty$), which precludes its application as a continuous predictor without ad hoc smoothing.

As summarised in Table 2 and visually compared in Figure 2, co-occurrence frequency ($\text{Freq}_{\{n,v\}}$) achieved the single best model fit among all individual predictors, yielding the lowest AIC (50,875.37) and BIC (50,917.39). Among the statistical association metrics, Log-Likelihood Ratio (LLR) provided the strongest univariate fit (AIC = 51,013.34), closely trailing co-occurrence frequency. Directional association metrics ($\Delta P_{n\to v}$ and $\Delta P_{v\to n}$) and LOR also exhibited substantial predictive capacity (AIC < 53,200), whereas frequency of the noun (Freq_n) and frequency of the verb (Freq_v) performed markedly worse (AIC > 65,000).

To examine potential collinearity among these predictors, Spearman rank correlations were calculated (Figure 3). While co-occurrence frequency demonstrates a strong positive correlation with LOR ($r_s = 0.93$), its correlations with LLR ($r_s = 0.50$) and directional ΔP metrics ($r_s = 0.59$ for $\Delta P_{n\to v}$) remain moderate. This structural variation necessitates multivariate modelling to ascertain whether contingency metrics contribute unique predictive information beyond co-occurrence frequency.

Interestingly, these univariate results suggest that for a relatively straightforward task such as predicting 5-point Likert-scale judgements, sophisticated co-occurrence metrics (such as those extensively debated in recent collostructional literature) may not be strictly indispensable, as co-occurrence frequency alone demonstrates remarkably robust explanatory power (cf. Bybee, 2010, 97–101). Although the specific frequency threshold required for reliable prediction remains inherently bound to the scale of the phenomenon and corpus size, these findings nevertheless highlight that even a simple, baseline metric of co-occurrence frequency can account for a substantial proportion of human psycholinguistic intuition.

Table 2: Summary of univariate ordered logit models

Predictor	Estimate	Std. Error	z-value	p-value	AIC	BIC	Log-Lik.
Freq$_{\{n,v\}}$	6.73	0.13	50.86	< 0.001	**50, 875.37**	**50, 917.39**	**−25, 432.69**
LLR	3.74	0.06	63.71	< 0.001	51, 013.34	51, 055.36	−25, 501.67
$\Delta P_{n\to v}$	1.77	0.02	99.40	< 0.001	51, 624.12	51, 666.14	−25, 807.06
$\Delta P_{v\to n}$	1.65	0.02	95.84	< 0.001	52, 838.97	52, 880.99	−26, 414.49
LOR	1.98	0.02	84.46	< 0.001	53, 181.15	53, 223.17	−26, 585.58
KLD$_{v\to n}$	1.02	0.01	73.54	< 0.001	60, 102.32	60, 144.34	−30, 046.16
KLD$_{n\to v}$	1.02	0.01	68.48	< 0.001	61, 189.85	61, 231.88	−30, 589.93
Freq$_v$	−0.49	0.01	−33.87	< 0.001	65, 567.23	65, 609.25	−32, 778.62
Freq$_n$	−0.06	0.01	−4.74	< 0.001	66, 869.52	66, 911.54	−33, 429.76

Note: All p-values are $p < 0.001$. The best-fitting model (minimum AIC/BIC) is highlighted in bold.

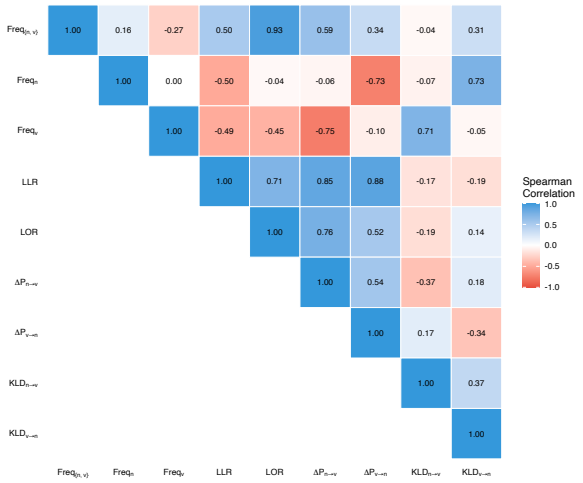

Figure 3: Correlation table among the collostructional measures

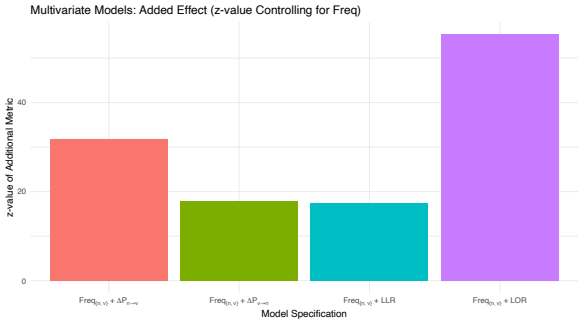

Figure 4: Comparison of added effects

4.2.2. Multivariate ordinal logistic regression

To evaluate whether statistical association metrics provide incremental predictive value beyond co-occurrence frequency, multivariate additive models (Freq$_{\{n,v\}}$ + $\{\Delta P_{\{n\to v, v\to n\}}, \text{KLD}_{\{n\to v, v\to n\}}, \text{LLR}, \text{LOR}\}$) were constructed across all cognitive dimensions. Note that predictors demonstrating markedly inferior univariate fit (AIC $> 60,000$; i.e., single-slot frequencies and KLD metrics) were omitted from multivariate testing to focus on metrics with substantial explanatory capacity. Among the evaluated combinations, incorporating any association metric alongside frequency significantly improved model fit relative to the univariate baseline (Table 3). Crucially, the combination of Freq$_{\{n,v\}}$ and LOR consistently yielded the best performance across Exposure, Usage, Comprehensibility, and the combined Overall dataset (AIC = 46, 935.1).

Figure 4 highlights the added predictive contribution (z-value) of each metric when controlling for frequency. LOR exerted the strongest additional effect ($z = 55.33$), vastly outperforming $\Delta P_{n\to v}$ ($z = 31.78$), $\Delta P_{v\to n}$ ($z = 17.77$), and LLR ($z = 17.36$). Given the strong Spearman correlation between co-occurrence frequency and LOR ($r_s = 0.93$), standard additive logic would predict substantial redundancy and negligible incremental gain. That LOR nevertheless yields the largest additive contribution suggests that its predictive capacity is not merely duplicating frequency, hinting at a more intricate, non-additive relationship between co-occurrence frequency and LOR.

To test whether incorporating the next best-performing univariate metric improves performance, a three-variable additive model combining co-occurrence frequency (Freq$_{\{n,v\}}$), LOR, and $\Delta P_{n\to v}$ was estimated. As summarised in Table 4, adding $\Delta P_{n\to v}$ yielded a statistically significant yet minor reduction in AIC across all datasets (Overall AIC = 46, 864.83, $z = 8.44$, $p < 0.001$). However, the incremental contribution of $\Delta P_{n\to v}$ ($z = 8.44$) remained substantially smaller than those of co-occurrence frequency ($z = 31.93$) and LOR ($z = 48.64$). More importantly, even this three-variable additive specification could not match the superior fit achieved by the two-variable interaction model (Freq$_{\{n,v\}} \times$ LOR; AIC = 45, 863.37), reinforcing that human linguistic intuition is driven by a non-linear threshold effect rather than the simple linear accumulation of independent statistical metrics.

Finally, pairwise interactions between predictors were systematically evaluated across candidate metric combinations to test for potential non-linear interactions (Table 5). Parallel to their dominance in additive modelling, the interaction between co-occurrence frequency and LOR (Freq$_{\{n,v\}} \times$ LOR) achieved the overall best fit by a substantial margin (AIC = 45, 863.37), decisively outper-

Table 3: Model fit statistics and parameter estimates for ordered logit models across cognitive dimensions

Dataset	Association Measure (AM)	AIC	BIC	Log-Lik.	$\text{Freq}_{\{n,v\}}$ $\hat{\beta}$ (z)		AM $\hat{\beta}$ (z)	
Overall	LOR	**46,935.1**	**46,985.5**	**−23,461.6**	4.22	(43.98)	1.11	(55.33)
	$\Delta P_{n\to v}$	49,801.8	49,852.2	−24,894.9	3.73	(29.20)	0.86	(31.78)
	LLR	50,507.9	50,558.3	−25,247.9	4.36	(22.55)	1.68	(17.36)
	$\Delta P_{v\to n}$	50,550.8	50,601.2	−25,269.4	4.69	(30.13)	0.51	(17.77)
Exposure	LOR	**14,927.2**	**14,971.1**	**−7,457.6**	4.54	(25.89)	1.03	(29.38)
	$\Delta P_{n\to v}$	15,693.9	15,737.7	−7,841.0	4.08	(17.65)	0.82	(17.27)
	$\Delta P_{v\to n}$	15,873.8	15,917.6	−7,930.9	4.64	(17.14)	0.59	(11.49)
	LLR	15,904.3	15,948.2	−7,946.2	4.72	(13.83)	1.61	(9.37)
Usage	LOR	**14,921.3**	**14,965.1**	**−7,454.7**	4.65	(25.90)	1.03	(29.31)
	$\Delta P_{n\to v}$	15,682.5	15,726.3	−7,835.3	4.16	(17.67)	0.82	(17.24)
	$\Delta P_{v\to n}$	15,867.7	15,911.5	−7,927.8	4.77	(17.09)	0.58	(11.22)
	LLR	15,898.4	15,942.2	−7,943.2	4.93	(14.26)	1.55	(9.11)
Comprehensibility	LOR	**16,837.7**	**16,881.5**	**−8,412.8**	3.64	(24.32)	1.27	(36.78)
	$\Delta P_{n\to v}$	18,215.9	18,259.7	−9,102.0	3.11	(5.27)	0.94	(20.42)
	LLR	18,496.9	18,540.7	−9,242.4	3.58	(11.21)	1.87	(11.55)
	$\Delta P_{v\to n}$	18,593.5	18,637.3	−9,290.7	4.69	(18.01)	0.39	(8.26)

Note: All p-values associated with the z-statistics are $p < 0.001$. Best fit models are highlighted in bold.

Table 4: Model fit statistics for the three-variable additive model ($\text{Freq}_{\{n,v\}}$ + LOR + $\Delta P_{n\to v}$)

Dataset	AIC	BIC	Log-Lik.	$\text{Freq}_{\{n,v\}}$ $\hat{\beta}$ (z)		LOR $\hat{\beta}$ (z)		$\Delta P_{n\to v}$ $\hat{\beta}$ (z)	
Overall	**46,864.83**	**46,923.66**	**-23,425.42**	3.64	(31.93)	1.04	(48.64)	0.25	(8.44)
Exposure	14,903.03	14,954.17	−7,444.52	3.91	(19.04)	0.96	(25.57)	0.26	(5.07)
Usage	14,896.41	14,947.55	−7,441.20	4.00	(19.04)	0.95	(25.51)	0.27	(5.14)
Comprehensibility	16,819.24	16,870.38	−8,402.62	3.13	(17.19)	1.21	(32.82)	0.23	(4.49)

Note: All p-values associated with z-statistics are $p < 0.001$.

forming every other interaction specification. Notably, a robust negative interaction coefficient was identified ($\hat{\beta} = -6.90, z = -40.94, p < 0.001$).

As depicted in Figure 5, this interaction reflects a cognitive threshold effect across frequency tiers. At low and medium frequency levels, variations in LOR have a negligible impact on subjective judgements, with ratings remaining anchored near the floor. Conversely, within the high-frequency tier, subjective ratings scale steeply with LOR. This indicates that the predictive impact of LOR on human judgements is contingent upon attaining a necessary threshold of co-occurrence frequency.

The fact that the optimal specification combines co-occurrence frequency with LOR in an interaction model provides crucial theoretical insights. In line with the core tenets of tupleisation (Gries, 2019, 2024), tupleising metrics that share redundant statistical properties yields little additional predictive power; an effective tuple must integrate distinct, mutually independent dimensions of structural information. Because LOR quantifies effect-size contingency independently of overall sample size and baseline frequency, it serves as a methodologically natural counterpart to co-occurrence frequency. Crucially, however, the interaction demonstrates that no matter how much LOR fluctuates, it carries minimal predictive utility for subjective judgements unless accompanied by a sufficient level of co-occurrence frequency. This strongly suggests that while co-occurrence frequency and LOR (i.e., co-occurrence strength) jointly shape human linguistic intuitions, co-occurrence frequency operates as a necessary prerequisite filter (indicating that association metrics only influence cognitive judgements once a required frequency threshold is reached).

Given the strong Spearman rank correlation between co-occurrence frequency ($\text{Freq}_{\{n,v\}}$) and LOR ($r_s = 0.93$), one might suspect that combining these two measures introduces statistical redundancy. In standard additive modelling, such high collinearity typically yields diminishing returns rather than substantial fit improvements. Crucially, however, the dramatic reduction in AIC achieved by the interaction specification (AIC = 45,863.37, compared to 50,875.37 for univariate frequency alone) proves that their relationship is non-linear and highly synergistic rather than redundant. Rather than duplicating explanatory variance, co-occurrence frequency and LOR fulfil distinct, complementary roles in human cognitive representation: co-occurrence frequency operates as an essential exposure gatekeeper, whereas LOR actively drives subjective judgements once this prerequisite threshold is crossed. This finding demonstrates the practical utility of tupleisation, showing that disentangling corpus metrics allows for capturing non-linear interactions between strongly correlated measures that would otherwise remain masked.

5. Conclusion

This study evaluated the predictive alignment between disentangled corpus metrics and multi-dimensional psycholinguistic judgements in Japanese covarying constructions (*N-ga V*). While co-occurrence frequency ($\text{Freq}_{\{n,v\}}$) served as the single strongest univariate pre-

Table 5: Summary of ordered logit models with interaction terms

Model / Predictors	Var 1 $\hat{\beta}$ (z)		Var 2 $\hat{\beta}$ (z)		Interaction $\hat{\beta}$ (z)		AIC	BIC	Log-Lik.
Freq$_{\{n,v\}}$ × LOR	15.53	(45.16)	−0.68	(−13.49)	−6.90	(−40.94)	**45, 863.37**	**45, 922.20**	**−22, 924.69**
LLR × LOR	6.60	(40.38)	0.55	(19.87)	−2.70	(−30.37)	46, 282.80	46, 341.63	−23, 134.40
Freq$_{\{n,v\}}$ × $\Delta P_{v\to n}$	18.30	(69.49)	−0.48	(−15.07)	−5.14	(−61.80)	46, 475.15	46, 533.98	−23, 230.57
Freq$_{\{n,v\}}$ × $\Delta P_{n\to v}$	14.09	(55.68)	−0.10	(−2.91)	−3.88	(−48.00)	47, 711.46	47, 770.29	−23, 848.73
LOR × $\Delta P_{v\to n}$	1.14	(55.40)	1.32	(27.82)	−0.14	(−4.38)	48, 545.42	48, 604.25	−24, 265.71
LOR × $\Delta P_{n\to v}$	0.95	(41.67)	1.58	(30.82)	−0.24	(−7.62)	48, 761.72	48, 820.55	−24, 373.86
LLR × $\Delta P_{v\to n}$	11.58	(41.59)	−0.54	(−12.37)	−2.87	(−35.78)	49, 067.77	49, 126.60	−24, 526.88
LLR × $\Delta P_{n\to v}$	5.13	(28.56)	0.49	(14.19)	−1.15	(−19.88)	49, 760.23	49, 819.06	−24, 873.12
Freq$_{\{n,v\}}$ × LLR	7.11	(33.41)	1.14	(12.11)	−1.27	(−38.36)	49, 774.76	49, 833.59	−24, 880.38
$\Delta P_{n\to v}$ × $\Delta P_{v\to n}$	1.67	(38.33)	0.87	(22.04)	−0.32	(−17.44)	51, 048.61	51, 107.44	−25, 517.31

Note: All predictors and interaction terms are statistically significant ($p < 0.005$). The best-fitting model is highlighted in bold.

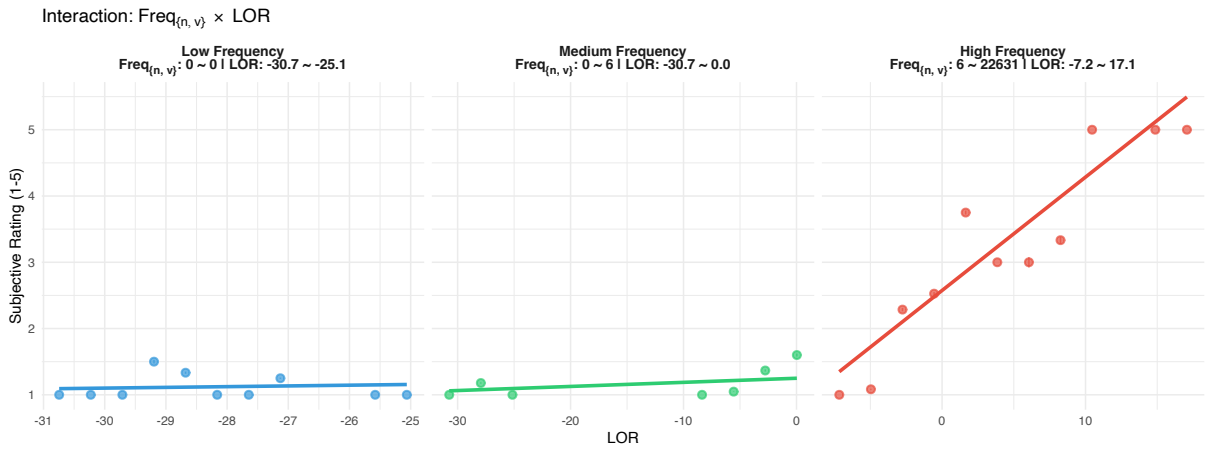

Figure 5: Interaction of co-occurrence frequency (i.e., Freq$_{\{n,v\}}$) and LOR

dictor, combining frequency with log odds ratio (LOR) in an interaction model consistently provided the optimal fit across all evaluated conditions. Crucially, the identified Freq$_{\{n,v\}}$ × LOR interaction demonstrated a robust cognitive threshold effect: the predictive impact of LOR on human judgements becomes apparent only once a requisite baseline of co-occurrence frequency is attained. These findings highlight the practical utility of tupleisation, confirming that integrating independent measures of co-occurrence frequency (Freq$_{\{n,v\}}$) and effect-size contingency (LOR) is key to capturing non-linear patterns when predicting psycholinguistic ratings.

Although the present investigation separately assessed three conceptually distinct cognitive dimensions (i.e., Exposure, Usage, and Comprehensibility), the macro-level predictive trends did not substantially diverge across tasks. This suggests that for psycholinguistic experiments operating at this level of granularity (i.e., 5-point Likert ratings), finer task distinctions may be largely redundant for broad cognitive modelling. While the selection of specific co-occurrence metrics varies depending on particular analytical objectives, our results indicate that for predicting subjective psycholinguistic judgements, combining co-occurrence frequency with LOR provides a highly effective pairing. Future research should explore tasks or experimental settings where co-occurrence frequency and LOR alone prove insufficient, testing whether combining them with predictors set aside in the present study (e.g., directional association or single-slot metrics) restores predictive power.

6. Acknowledgements

Portions of this work were supported by JSPS KAKENHI Grant Number 24K16143. Usual disclaimers apply.

7. Bibliographical References

Joan Bybee. 2010. *Language, Usage, and Cognition*. Cambridge University Press, Cambridge.

Rune H. B. Christensen. 2026. `ordinal`: *Regression Models for Ordinal Data*. R package version 2026.7-26.

Gaetanelle Gilquin and Stefan Th. Gries. 2009. Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, 5(1):1–26.

Stefan Th. Gries. 2019. 15 years of collostructions: Some long overdue additions/corrections (to/of actually all sorts of corpus-linguistics measures). *International Journal of Corpus Linguistics*, 24(3):385–412.

Stefan Th. Gries. 2022. What do (some of) our association measures measure (most)? association? *Journal of Second Language Studies*, 5(1):1–33.

Stefan Th Gries. 2023. Overhauling collostructional analysis: Towards more descriptive simplicity and more explanatory adequacy. *Cognitive Semantics*, 9(3):351–386.

Stefan Th. Gries. 2024. *Frequency, Dispersion, Association, and Keyness: Revising and Tupleizing Corpus-Linguistic Measures*. John Benjamins, Amsterdam.

Stefan Th Gries. 2026. Tupleization. In Hilary Nesi and Petar Milin, editors, *International Encyclopedia of Language and Linguistics*, 3 edition, volume 12, pages 335–339. Elsevier, Amsterdam.

Stefan Th Gries, Beate Hampe, and Doris Schönefeld. 2005. Converging evidence: bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics*, 16(4):635–676.

Stefan Th Gries, Beate Hampe, and Doris Schönefeld. 2010. Converging evidence II: More on the association of verbs and constructions. In Sally Rice and John Newman, editors, *Empirical and Experimental Methods in Cognitive/Functional Research*, pages 59–72. CSLI Stanford, CA.

Stefan Th. Gries and Anatol Stefanowitsch. 2004. Extending collostructional analysis: A corpus-based perspective on alternations. *International Journal of Corpus Linguistics*, 9(1):97–129.

R Core Team. 2026. *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Anatol Stefanowitsch and Stefan Th. Gries. 2003. Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics*, 8(2):209–243.

Anatol Stefanowitsch and Stefan Th. Gries. 2005. Covarying collexemes. *Corpus Linguistics and Linguistic Theory*, 1(1):1–43.

8. Language Resource References

Stefan Th. Gries. 2024. Coll.analysis 4.1. A script for R to compute perform collostructional analyses.

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlý, and Vít Suchomel. 2013. The TenTen corpus family. In *7th International Corpus Linguistics Conference CL 2013*, pages 125–127.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovvář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: Ten years on. *Lexicography*, 1(1):7–36.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrž, and David Tugwell. 2004. The Sketch Engine. In *Proceedings of XI EURALEX International Congress*, pages 105–116.