

Article

Fatih Ünal Bozdağ*

Modeling distributional stability in learner English passives: The Linguistic Clusters Framework

<https://doi.org/10.1515/cllt-2025-0141>

Received December 10, 2025; accepted May 28, 2026; published online June 10, 2026

Abstract: Frequency-based learner corpus approaches struggle to separate systematic grammatical patterning from artifacts of prompts or register. This study introduces the Linguistic Clusters (LC) Framework, a multi-metric procedure for identifying stable distributional patterns in learner passives. The framework implements a three-layer architecture integrating instance-level filters, schema-level acceptance criteria, and validation procedures. Applied to over 57,000 valid passive constructions from the International Corpus of Learner English, the LC Framework identified eight validated schemas from argumentative essays across diverse first-language backgrounds. Permutation testing confirmed non-random TAM $\times$ complement structure and demonstrated that schema acceptance is structurally rather than lexically driven. Sensitivity analysis showed that the majority of schemas remain stable across threshold variations, while ablation analysis attributed most false-positive reduction to instance-level filters. The validated schemas cluster around modal passives with spatial prepositional complements, revealing a frequency–formulaicity dissociation: modal-be passives account for a small fraction of corpus constructions yet comprise the majority of validated schemas. These findings support usage-based accounts that treat L2 constructional development as driven by the convergence of multiple distributional cues, not by raw token frequency alone.

Keywords: construction grammar; learner corpus; passive voice; formulaic language; multi-metric validation

*Corresponding author: **Fatih Ünal Bozdağ**, English Translation and Interpretation, Faculty of Humanities and Social Sciences, Osmaniye Korkut Ata University, Osmaniye, Türkiye, E-mail: fatihbozdag@osmaniye.edu.tr. <https://orcid.org/0000-0002-9959-4704>

 Open Access. © 2026 the author(s), published by De Gruyter.  This work is licensed under the Creative Commons Attribution 4.0 International License.

1 Introduction

Learner corpus research has established frequency-based approaches as foundational for analyzing L2 grammatical patterns (Gilquin and Paquot 2008; Granger 2009). However, frequency counts alone cannot distinguish systematic patterning from artifacts of corpus composition. High-frequency forms may reflect clustering induced by writing prompts (i.e., task instructions that direct learners to write about specific topics) or register-specific recycling rather than genuine distributional regularities in learner grammars (Larsson et al. 2021). When writing prompts concentrate lexical material around specific themes, and when task familiarity varies across learner populations, raw frequency provides an unreliable index of formulaic status.

Usage-based accounts of grammar emphasize that conventional form-meaning pairings are shaped by experience with multiple distributional dimensions, including frequency, contingency, and contextual dispersion (Bybee 2010; Ellis 2006; Wulff and Ellis 2018). Formulaic sequences become entrenched through repeated use in consistent contexts, exhibiting both strong internal cohesion and predictable external environments (Wray and Perkins 2000). Identifying such sequences in corpus data requires methods sensitive to these multiple dimensions, distinguishing patterns that are genuinely prefabricated from combinations that merely recur due to topical or grammatical necessity.

Quantifying the degree to which recurrent sequences constitute genuinely formulaic patterns remains a central methodological challenge. Forsyth and Grabowski (2015) demonstrate that productivity-based indices can distinguish formulaic from non-formulaic registers, while Maziarz et al. (2023) show that lexicalization criteria for multiword expressions vary across languages, with compositionality judgments playing a central role. The effectiveness of such metrics depends on the length of the target unit (collocations vs. multi-word expressions), contiguity of components (adjacent vs. non-adjacent items), and the typological properties of the language under investigation.

Standard collocation statistics are insufficient for this task. For instance, point-wise mutual information inflates rare co-occurrences because low marginals exaggerate surprise, while chi-square and related tests can overstate effects under sparsity (Church and Hanks 1990; Dunning 1993). Association measures such as Delta-P (Gries 2013) capture contingency but not contextual predictability. Conversely, entropy-based measures (Degaetano-Ortlieb and Teich 2019; Shannon 1948) assess predictability but not association strength. No single metric captures the multidimensional nature of formulaicity, and applying multiple metrics without principled integration risks either excessive false positives or excessive false negatives.

A grammatical domain that exhibits this kind of combinatorial complexity is the English passive, offering a revealing test case for distributional organization in learner writing. In argumentative prose, which constitutes the primary register examined in this study, passives support stance-taking, agent suppression, and an appearance of objectivity central to the academic register (Biber and Conrad 2009). The construction exhibits systematic variation along two dimensions. Tense-aspect-modality (TAM) categories determine auxiliary composition, while complement structure introduces further variation through agent phrases, prepositional complements, and infinitival extensions. This combinatorial space creates numerous potential sites for formulaic patterning, as specific TAM $\times$ complement configurations may be stored and deployed as prefabricated units for recurrent rhetorical purposes.

This study introduces the Linguistic Clusters (LC) Framework, a multi-metric procedure for detecting stable distributional patterns in learner passive production. The framework implements a three-layer filtering architecture that operationalizes formulaicity through converging evidence. Instance-level prefilters assess transitional probability (Aslin and Newport 2012; Saffran et al. 1996), directional contingency (Ellis 2006), and boundary entropy (Mansfield 2021). Schema-level criteria evaluate association strength and lexical productivity through dual pathways. Validation procedures confirm robustness through permutation testing and sensitivity analysis. Section 3 details each component.

Because the framework draws on concepts from several traditions, the following terms are used consistently throughout this paper. A *formulaic instance* refers to an individual passive construction token that passes the instance-level filters (Layer 1). Once a TAM $\times$ complement configuration type passes schema-level acceptance criteria (Layer 2), it is defined as a *schema*. Within this layer, two distinct types of schemas are recognized: a *lexically-cohesive schema* exhibits a strong association between a specific TAM and complement, functioning as a fixed collocation, whereas a *productive schema* accommodates diverse verb lemmas within a consistent TAM $\times$ complement frame, functioning as a schematic construction with an open verb slot. Finally, the term *pattern* serves as a general cover term for any recurring TAM $\times$ complement configuration, whether or not it has passed validation.

The framework captures two theoretically distinct types of distributional stability. Lexically-cohesive patterns exhibit strong association between specific TAM and complement categories, functioning as fixed collocations. Productive templates exhibit consistent TAM $\times$ complement frames that accommodate diverse verbs, functioning as schematic constructions with open slots. By implementing complementary acceptance criteria for each type, the LC Framework recognizes both as valid manifestations of formulaic organization. Section 2 develops the theoretical

rationale for this distinction within Construction Grammar and usage-based linguistics.

Whether these two types of stability emerge depends in part on the register examined. This study focuses specifically on argumentative essay writing, a register where passives serve stance-marking and epistemic hedging functions (Biber and Conrad 2009; Gilquin and Paquot 2008). Whether the observed patterns generalize to other registers remains an empirical question requiring cross-register replication.

Drawing on learner writing from the International Corpus of Learner English (ICLE; Granger et al. 2020), the study addresses three research questions:

RQ1: How can stable patterns in learner passive production be modeled through multi-metric integration?

RQ2: Which TAM $\times$ complement configurations exhibit statistically robust stability when association strength, lexical productivity, and validation criteria converge?

RQ3: What do these distributions reveal about the organization of learner passive usage in argumentative writing, and what types of formulaic patterning emerge?

The remainder of the paper is organized as follows. Section 2 reviews the theoretical and methodological literature motivating the LC Framework. Section 3 details the three-layer filtering architecture, schema-level acceptance criteria, and statistical validation procedures. Section 4 presents the results organized around the three research questions. Section 5 discusses the findings in relation to Construction Grammar, usage-based accounts, and methodological contributions. Section 6 concludes with a summary, limitations, and future directions.

2 Theoretical and methodological background

2.1 Construction Grammar and usage from a distributional perspective

Construction Grammar treats grammatical knowledge as an inventory of form-function pairings ranging from fully lexical to highly abstract (Goldberg 1995, 2006; Hilpert 2014), with evidence that some pairings are stored holistically (Conklin and Schmitt 2012; Wray 2002). For corpus analysis, this framework requires metrics that assess not only how frequently forms recur, but how reliably they pair with specific functions and complements across texts and contexts (Bybee 2010; Ellis 2006). Classic statements emphasize that schematic and item-specific knowledge coexist within a single descriptive space, with frames, collocations, and fully abstract constructions

belonging to an inventory sensitive to distributional experience in discourse (Ellis and Cadierno 2009; Ellis and Ferreira-Junior 2009; Goldberg 2006; Suttle and Goldberg 2011).

For learner data, this framing matters because repeated strings can arise from prompt content or topic recycling without constituting configurations that generalize across writers (Gries 2024; Larsson et al. 2021). A distributional perspective therefore asks whether a candidate pairing is not only frequent, but also reliably configured in context, cohesive when measured against background expectations, and visible across documents written by different authors.

Within this framework, systematic patterning manifests along a specificity-schematicity continuum (Goldberg 2006). At one pole, lexically-cohesive sequences exhibit strong association between specific lexical items (e.g., *kick the bucket*), approaching the traditional conception of formulaic sequences as prefabricated chunks (Wray 2002). At the other pole, productive templates exhibit systematic lexical diversity within stable structural frames (e.g., [SUBJ] *be V-en* [PP-directional]), where multiple verbs instantiate schematic slots. Both types can achieve systematic organization through frequency and functional specialization (Bybee 2010), though they differ in the degree to which form is lexically fixed versus productively variable.

2.2 Limitations of frequency-based approaches in learner corpus analysis

As noted above, learner corpora present distinctive statistical challenges because writing prompts, proficiency variation, and individual writers' habits can generate surface patterns that do not reflect generalizable configurations. Under these conditions, standard association statistics exhibit predictable limitations, including rare-event inflation, sparse-table instability, and susceptibility to topical clustering (Church and Hanks 1990; Dunning 1993; Gries 2008; Lijffijt et al. 2016).

Recent methodological work addresses these challenges by complementing raw counts with measures of pairing reliability, interpretable cohesion scales, continuation predictability, and cross-document breadth (Bentz and Alikaniotis 2016; Bouma 2009; Degaetano-Ortlieb and Teich 2018, 2019; Mansfield 2021; Rastelli 2020, 2022; Sun and Wang 2021). Two observations sharpen the analytical target. First, many recurrent strings in learner essays are local to a topic or prompt and do not travel across documents. Second, some configurations are modest in frequency yet highly selective in their complement choices, making them predictable in context and reproducible across texts.

2.3 Learner english passives in argumentative writing

Descriptions of expert academic prose consistently link passive voice constructions to the management of agency, information structure, and the foregrounding of concepts over agents (Biber and Conrad 2009). When writers background agents, foreground processes or results, and present claims as shared knowledge or as outcomes of prior work, they often do so through complement-sensitive passive frames. Comparisons of learner and expert writing report that learners significantly underuse passives and rely on a limited set of frequent collocations and multiword sequences, underusing the broader phraseological repertoire characteristic of expert academic prose (Granger 2009).

The analytic unit in this literature is therefore not the abstract string *be + V-en*, but recurrent configurations that bundle auxiliary morphology, lexical heads, and complement types into recognizable discourse moves. Typical examples in argumentative prose include *has been shown to* followed by a non-finite clause, *is widely considered* followed by a noun phrase, *is expected to* followed by a verb phrase, and *was born in* followed by a prepositional phrase indicating location or time. These frames differ in how tightly their components co-occur and in how predictable their subsequent elements are. Some exhibit tight pairing and low boundary uncertainty across many documents, whereas others are more dispersed across diverse contexts and cluster within specific topics.

Viewed from this perspective, work on learner passives suggests a distributional contrast between expert and learner production. Expert writing distributes passive use across a broader range of functions – foregrounding concepts, managing agency, and structuring dense informational prose (Biber and Conrad 2009) – whereas learner writing significantly underuses passives and relies on a limited phraseological repertoire, underusing the multiword sequences characteristic of expert academic register (Granger 2009).

2.4 Corpus methods for constructional patterns in learner data

Construction-oriented corpus methods began with single-measure association and have moved toward triangulated evidence. Collostructional analysis estimates attraction or repulsion between lexemes and constructional slots and remains a standard tool when the slot is well defined and counts are adequate for exact tests (Gries and Stefanowitsch 2004; Stefanowitsch and Gries 2003; Stefanowitsch and Gries 2005). Strengths include clear interpretation and a direct focus on slot behavior. At the same time, the limitations in learner settings are well documented, notably

sensitivity to low counts, limited treatment of document dispersion unless it is modeled explicitly, and susceptibility to topical skew.

Recent discussions recommend reporting effect sizes, handling dispersion directly, and complementing slot-based tests with frequency-robust, directional estimates that travel better across corpora with different sizes and topic mixes, including log-odds frameworks with appropriate corrections and comparative designs that stabilize inference across subcorpora (Gilquin 2025; Gries 2024; Gries and Wulff 2025).

Information-theoretic measures add a second lens that addresses known weaknesses of the classic toolkit. Normalized pointwise mutual information (NPMI) puts association on a bounded, interpretable scale and reduces the rare-event inflation typical of learner data (Bouma 2009). Entropy profiles distinguish cohesive, phrase-like subsequent elements from loose co-occurrence by quantifying boundary uncertainty in local context (Bentz and Alikaniotis 2016; Mansfield 2021). Relative entropy has proved stable for characterizing register and diachronic or proficiency differences because it summarizes distributional change rather than absolute totals, which helps separate stylistic shifts from volume effects in large corpora (Degaetano-Ortlieb and Teich 2018, 2019; Sun and Wang 2021).

Work that imports directional dependence into second-language contexts shows how asymmetries in conditional probabilities separate chunk-like behavior from broader variation as distributions flatten with growth, again cautioning against frequency as the sole arbiter of patterning (Rastelli 2020, 2022).

The present study builds on and extends this body of work in three ways. First, it integrates multiple complementary metrics (transitional probability, directional contingency, boundary entropy, association strength, and slot entropy) into a single filtering architecture, addressing calls for multi-metric validation in corpus-based construction research (Gries 2024). Second, it introduces a dual-lane acceptance system that captures both lexically-cohesive and productive patterns, operationalizing the specificity-schematicity continuum within a single framework. Third, it applies permutation-based validation to distinguish genuine distributional stability from corpus artifacts, providing a replicable template for construction mining in learner data.

3 Methodology

3.1 Corpus and data preparation

The analysis uses the International Corpus of Learner English version 3 (ICLE; Granger et al. 2020), specifically 9,529 texts (94 % argumentative essays, with literary

essays in some subcorpora) from advanced L2 English learners (CEFR B2-C1) with diverse L1 backgrounds. Texts range from 500 to 1,000 words, with a total corpus size of approximately 5.8 million tokens. The texts were written under varied task conditions – approximately 40 % timed, 38 % under examination conditions – with some writers permitted access to reference tools, representing typical university coursework and assessment contexts.

ICLE metadata identifies 364 literary essays (3.8 %) and 200 texts classified as “other” (2.1 %), with the remaining 8,965 texts (94.1 %) classified as argumentative. All authentic examples attested across the eight validated schemas originate from argumentative texts. The literary and other texts were retained in the pipeline to preserve the full corpus, but the validated schemas are driven entirely by argumentative production.

Passive voice constructions were identified using ASC-Tagger v2 (Sung and Kyle 2024), a specialized tool for automatic identification of argument structure constructions in English texts. The tagger employs a spaCy transformer model fine-tuned on gold-standard annotated data (ASC Treebank v2) to classify clauses into nine ASC types, including passives (defined as Subj-auxV_past.participle constructions with theme-undergoes-action semantics). Unlike bottom-up approaches that rely solely on dependency representations, the model leverages sentence-level contextual information to disambiguate structurally similar but semantically distinct constructions, achieving F1 scores of 0.912–0.928 across L1 and L2 text types.

For each identified passive, dependency parsing was performed using spaCy (en_core_web_trf, v3.7; Honnibal et al. 2020) to extract the complete construction including auxiliary chain, grammatical subject, and post-verbal complements. Each detected passive construction was analyzed for its TAM auxiliary configuration and complement structure.

Basic filtering removed constructions with invalid TAM categories (e.g., progressive perfect: **has been being* VBN), extraposed *it*-passives (*it + be + VBN + that-S*), and constructions where the past participle could not be reliably distinguished from adjectival predicates.

Recent psycholinguistic work shows that speakers track multiple distributional dimensions of multiword expressions, including frequency, directional association strength, and contextual breadth (Zhong et al. 2025). Building on this work, we explicitly distinguish frequency, directional contingency ($\Delta P_{\text{backward}}$), and distributional predictability (H_r) as complementary dimensions of distributional structure. The notation is kept consistent across methods and results. While the first two dimensions parallel those studied by Zhong et al., our third dimension (boundary entropy) captures local continuation predictability rather than cross-context dispersion.

3.2 Instance-level prefiltering (Layer 1)

Layer 1 applies three sequential filters to assess surface formulaicity at the construction-token level. Each formulaic instance must pass all three thresholds to advance to schema-level evaluation. These instance-level metrics serve as coarse gates, and final acceptance is determined at the schema level through converging metrics that provide more robust evidence of formulaic status.

3.2.1 Average transition probability (ATP)

ATP measures bidirectional transitional probability between words in the passive construction's auxiliary chain. For each adjacent word pair in the chain, forward and backward Delta-P values are computed and averaged. The pair-level ATP is the average of forward and backward Delta-P (Eq. (1)). The instance-level ATP is the mean of these pair-level averages across all adjacent pairs in the construction (Eq. (2)). Schema-level ATP is then the mean across all instances assigned to that schema (Eq. (3)).

Transitional probability as a statistical learning mechanism was introduced by Saffran et al. (1996) and has since been applied across language and cognition research (Aslin and Newport 2012). In L1/L2 comparison, higher transitional probability values characterize formulaic sequences used preferentially by native speakers (Appel and Trofimovich 2017).

$$ATP_{\text{pair}} = \frac{\Delta P_{\text{forward}} + \Delta P_{\text{backward}}}{2} \quad (1)$$

$$ATP_{\text{instance}} = \text{mean}(ATP_{\text{pair}}) \text{ across adjacent pairs} \quad (2)$$

$$ATP_{\text{schema}} = \text{mean}(ATP_{\text{instance}}) \text{ across instances} \quad (3)$$

Forward Delta-P ($\Delta P_{\text{forward}} = P(w_2|w_1) - P(w_2)$) measures how much the preceding word increases the probability of the following word beyond its baseline rate. Backward Delta-P ($\Delta P_{\text{backward}} = P(w_1|w_2) - P(w_1)$) captures the reverse. Both are computed against the marginal baseline $P(w)$ and thus index conditional-probability lift — the directional form of Delta-P used here, distinct from the strict contingency $\Delta P = P(O|C) - P(O|\neg C)$ (Ellis 2006). Averaging both directions ensures that ATP reflects mutual predictability rather than unidirectional association. The threshold $ATP \geq 0.10$ requires that the mean bidirectional Delta-P across word pairs reach at least 0.10, indicating that adjacent words in the auxiliary chain reliably predict each other beyond chance. Sensitivity analysis confirmed this threshold is non-critical, as schema identification remains stable across ATP values ranging from 0.05 to 0.15.

ATP is interpreted only as a coarse gate, and final acceptance depends on converging schema-level metrics.

3.2.2 Delta-P backward ($\Delta P_{\text{backward}}$)

$\Delta P_{\text{backward}}$ measures backward contingency, a directional association metric that captures the asymmetric predictive relationship between two co-occurring elements (Ellis 2006). Unlike symmetric measures such as mutual information, directional contingency distinguishes the predictive strength of element A for element B from the predictive strength of B for A. In the present context, $\Delta P_{\text{backward}}$ at Layer 1 assesses how much the following word (typically the participle) increases the probability of the preceding auxiliary above its baseline rate, computed on the final transition in the auxiliary chain, as expressed in Eq. (4).

$$\Delta P_{\text{backward}} = P(w_1 | w_2) - P(w_1) \quad (4)$$

Positive $\Delta P_{\text{backward}}$ values indicate that the participle is a reliable cue to the preceding auxiliary, meaning that encountering the participle increases the probability of the auxiliary above its baseline rate. The threshold $\Delta P_{\text{backward}} \geq 0.10$ requires that the participle increase the auxiliary's probability by at least 10 percentage points relative to its baseline rate. Because ΔP can be sensitive to sparse counts, it is not interpreted in isolation but works in conjunction with ATP and H_r to identify candidates with multiple formulaic signatures.

3.2.3 Right-boundary entropy (H_r)

H_r measures the Shannon entropy of the immediate right-context (the word following the passive construction), computed via Eq. (5).

$$H_r = -\sum P(w|\text{pattern}) \log_2 P(w|\text{pattern}) \quad (5)$$

The sum ranges over all word types w appearing in right-boundary position for the given TAM $\times$ complement pattern, and $P(w|\text{pattern})$ is estimated as the relative frequency of w among right-boundary tokens for the pattern. No smoothing is applied. Low entropy indicates that the pattern tends to occur in a restricted set of right contexts, consistent with formulaic status; high entropy indicates diverse, unpredictable subsequent elements. The threshold $H_r \leq 2.8$ bits selects constructions with relatively predictable right contexts, corresponding to an effective vocabulary of approximately 7 equally likely continuations ($2^{2.8} \approx 7$). Entropy estimates are noisy for patterns with few tokens; this role is restricted to filtering candidates with grossly unpredictable contexts.

Constructions passing all three Layer 1 thresholds ($ATP \geq 0.10$, $\Delta P_{\text{backward}} \geq 0.10$, $H_r \leq 2.8$) are classified as formulaic instances. These thresholds were determined empirically through preliminary analysis of the passive construction data. $ATP \geq 0.10$ retains pairs whose bidirectional transitional probability exceeds chance association in the corpus. $\Delta P_{\text{backward}} \geq 0.10$ ensures that the participle provides a meaningful cue to the preceding auxiliary beyond baseline rates. $H_r \leq 2.8$ excludes constructions whose right-boundary contexts are too variable to suggest predictable continuation. The robustness of these thresholds is evaluated through sensitivity analysis (Section 3.4.3), which tests schema stability across systematic threshold variations. Formulaic instances aggregate into unique TAM $\times$ complement patterns for schema-level evaluation.

To illustrate, the schema *modal-be+pp_up_NP* is attested in constructions such as “the decision should be left up to the industry rather than to the government” and “euthanasia should be allowed up to now.” These instances pass Layer 1 because the TAM $\times$ complement combination (*modal-be+pp_up_NP*) exhibits high bidirectional transitional probability ($ATP = 0.31$), reliable backward contingency ($\Delta P_{\text{backward}} = 0.14$), and constrained right-boundary entropy. These converging surface signatures mark the sequence as formulaic at the instance level.

3.3 Schema-level acceptance (Layer 2)

Layer 2 evaluates aggregated patterns through a dual-lane architecture that admits schemas through two complementary pathways. The NPMI lane targets fixed collocations and the Hslot lane targets productive templates.

3.3.1 NPMI lane: fixed collocations

The NPMI (Normalized Pointwise Mutual Information) lane identifies patterns with strong TAM-complement association, defined in Eq. (6).

$$NPMI = \frac{\left[\log_2 \left(\frac{P(\text{TAM}, \text{COMP})}{P(\text{TAM}) \times P(\text{COMP})} \right) \right]}{[-\log_2 P(\text{TAM}, \text{COMP})]} \quad (6)$$

NPMI normalizes PMI to the $[-1, 1]$ range, where $NPMI = 1$ indicates perfect co-occurrence, $NPMI = 0$ indicates independence, and $NPMI = -1$ indicates perfect complementary distribution. The threshold $NPMI \geq 0.05$ requires positive association, filtering patterns where TAM and complement co-occur less than expected under independence. This threshold is deliberately conservative because the accompanying G^2 criterion provides additional filtering.

Association strength is supported by the log-likelihood ratio statistic G^2 (Dunning 1993), shown in Eq. (7).

$$G^2 = 2 \times \sum O \times \ln\left(\frac{O}{E}\right) \quad (7)$$

where O and E represent observed and expected frequencies under independence, the sum is over all four cells of the 2×2 contingency table, and $\ln$ denotes the natural logarithm. The threshold $G^2 \geq 3.84$ corresponds to the chi-square critical value at $p < 0.05$ ($df = 1$). However, given the small counts typical of many patterns in learner corpus data, asymptotic G^2 approximations may be imprecise. G^2 is therefore interpreted as an approximate indicator of association strength rather than a standalone inferential test. The key inferential claim – that accepted schemas reflect non-random structure – is validated through permutation testing (Section 3.4.2), which provides non-parametric assessment without distributional assumptions. A pattern is accepted in the NPMI lane only if it satisfies both $\text{NPMI} \geq 0.05$ and $G^2 \geq 3.84$.

For instance, the schema `modal-be+pp_over_NP` qualifies through the NPMI lane. This schema arises from the combination of modal auxiliaries with over-phrases in passive constructions (e.g., “People of the same nation may be scattered all over the world,” “entire encyclopedias can be transmitted over the Internet”). These constructions exhibit high normalized co-occurrence ($\text{NPMI} = 0.199$) supported by significant association ($G^2 = 37.58$, $p < 0.001$).

3.3.2 Hslot lane: productive templates

The Hslot lane identifies patterns with high lexical productivity, measured through Shannon entropy of the verb slot, captured in Eq. (8).

$$H_{\text{slot}} = -\sum P(\text{verb}|\text{TAM}, \text{COMP}) \log_2 P(\text{verb}|\text{TAM}, \text{COMP}) \quad (8)$$

where `verb` ranges over all verb lemmas appearing in the $\text{TAM} \times \text{complement}$ frame. High Hslot indicates that the pattern accommodates diverse verbs, signaling a productive template. The threshold $\text{Hslot} \geq 1.5$ bits corresponds to approximately 2.8 equiprobable verb types ($2^{1.5} \approx 2.83$), ensuring meaningful lexical diversity beyond chance variation.

High Hslot values can arise from three sources. First, genuine productivity occurs when the $\text{TAM} \times \text{complement}$ frame systematically accommodates diverse verbs serving similar discourse functions. Second, accidental heterogeneity occurs when verbs co-occur without systematic constraint. Third, functional polysemy occurs when different sub-functions within a schema license different verbs. The requirement for positive NPMI (≥ 0) and statistical consistency ($G^2 \geq 1.0$) in the Hslot lane helps distinguish genuine productivity from accidental heterogeneity. Patterns

with negative NPMI indicate anti-association (grammatical necessity) rather than formulaic stability, while low G^2 values suggest random co-occurrence. Schemas passing the Hslot lane thus exhibit systematic lexical diversity rather than arbitrary verb selection. A pattern is accepted in the Hslot lane only if it simultaneously satisfies $\text{Hslot} \geq 1.5$, $G^2 \geq 1.0$, and $\text{NPMI} \geq 0$.

For example, the schema *perf-be+pp_within_NP* qualifies through the Hslot lane. The frame accommodates diverse verb lemmas (e.g., “it has been proved within the field of marketing,” “genetic engineering has been developed within several decades,” “This has been demonstrated within the past few years”). This yields high slot entropy ($\text{Hslot} = 2.32$), indicating genuine productivity rather than reliance on a single lexical item.

3.3.3 Dual-lane logic

Patterns qualifying through either lane are accepted (OR logic). This dual-lane architecture reflects two theoretically distinct types of systematic patterning that emerge in learner production. The NPMI lane targets lexically-cohesive collocations, that is, patterns where specific TAM-complement combinations exhibit strong association ($\text{NPMI} \geq 0.05$) and conventional statistical significance ($G^2 \geq 3.84$). These represent the traditional conception of formulaic sequences as recurrent pairings with relatively fixed lexical content (Wray 2002; Wray and Perkins 2000). The Hslot lane targets productive templates, that is, patterns where TAM-complement frames exhibit systematic lexical diversity ($\text{Hslot} \geq 1.5$) within stable structural configurations. These represent schematic constructions where multiple verbs systematically instantiate the same TAM-complement frame, signaling productive rather than lexically-fixed formulaic organization.

This distinction operationalizes the specificity-schematicity continuum discussed in Section 2.1. The OR logic ensures that both types of systematic organization receive recognition, with lexically-cohesive sequences qualifying through strong association and productive templates qualifying through systematic slot-filling patterns. Patterns qualifying through both lanes exhibit strong association and lexical productivity simultaneously, representing an intermediate position on the continuum.

Table 1: Schema acceptance under asymmetric versus symmetric G^2 thresholds.

Configuration	Schemas accepted	Patterns lost
Asymmetric (NPMI: $G^2 \geq 3.84$; Hslot: $G^2 \geq 1.0$)	8	–
Symmetric (both: $G^2 \geq 3.84$)	4	4

Under symmetric G^2 thresholds (≥ 3.84 for both lanes), four Hslot-lane schemas would be lost: past-be+pp_out_NP ($G^2 = 2.78$), perf-be+pp_within_NP ($G^2 = 2.78$), modal-be+pp_around_NP ($G^2 = 1.28$), and modal-be+pp_outside_NP ($G^2 = 1.28$). All four have $n \leq 5$, illustrating the sample-size constraint that motivates the relaxed threshold for the Hslot lane.

3.4 Statistical validation

3.4.1 Asymmetric G^2 thresholds: justification

The dual-lane architecture employs asymmetric G^2 thresholds. The NPMI lane requires $G^2 \geq 3.84$ (corresponding to $p < 0.05$, $df = 1$), while the Hslot lane requires $G^2 \geq 1.0$. This asymmetry reflects the theoretical distinction between fixed collocations and productive templates, and addresses a fundamental statistical constraint. G^2 is proportional to sample size, creating systematic bias against low-frequency patterns regardless of their genuine constructional status.

For patterns with $n \leq 6$, achieving $G^2 \geq 3.84$ is mathematically constrained even under perfect association. The G^2 statistic increases linearly with sample size for a given association strength, such that low-frequency patterns cannot reach conventional significance thresholds regardless of their NPMI or Hslot values. This creates a systematic Type II error problem, as genuine low-frequency productive templates are rejected not due to weak association but due to insufficient statistical power. Table 1 presents the threshold comparison, showing which schemas would be accepted under asymmetric versus symmetric G^2 thresholds.

The relaxed G^2 threshold (≥ 1.0) for the Hslot lane accommodates this sample size constraint while maintaining filtering rigor. $G^2 = 1.0$ still represents a meaningful threshold, corresponding to observed-expected discrepancy larger than expected under independence and filtering patterns with minimal or negative association. Combined with the Hslot ≥ 1.5 criterion (ensuring genuine lexical productivity) and the NPMI ≥ 0 floor (excluding negatively-associated patterns representing grammatical necessity), the threshold $G^2 \geq 1.0$ effectively distinguishes productive templates from random co-occurrence.

Different types of formulaicity – lexically-fixed collocations versus productive templates – require different statistical signatures. Requiring uniform statistical thresholds across these distinct constructional types imposes a one-size-fits-all criterion that systematically excludes theoretically important patterns. The dual-lane

architecture with asymmetric G^2 thresholds provides a principled solution that recognizes two valid pathways to formulaic status while maintaining statistical accountability for each pathway.

3.4.2 Permutation testing: dual validation design

The validation design employs two complementary permutation tests with distinct null hypotheses. TAM-shuffling tests non-randomness of TAM $\times$ complement association by randomizing TAM labels across constructions while preserving complement structure. Verb-shuffling tests lexical invariance (structural determination) by randomizing verb lemmas within schemas while preserving TAM $\times$ complement frames. Together, these tests distinguish structurally determined patterns from lexically driven artifacts.

Permutation testing validates that identified schemas represent genuine distributional regularities rather than statistical artifacts. Two complementary permutation designs test different hypotheses about schema organization and provide convergent evidence for the methodological validity of the dual-lane acceptance system.

The TAM-shuffling design tests whether observed TAM $\times$ complement patterns exhibit systematic co-occurrence beyond chance. For each of 1,000 permutation iterations, TAM categories are randomly reassigned across the formulaic instances while preserving the complement distribution. This breaks the association between specific TAM categories and specific complement types while maintaining the marginal distributions of both variables. After shuffling, the full dual-lane acceptance procedure (NPMI and Hslot calculations, G^2 testing, OR logic) is re-applied to identify which patterns would be accepted under randomized TAM assignments. Jaccard similarity is computed between the observed schema set and each permuted schema set, measuring the degree of overlap. The permutation p -value is computed as the proportion of permutations yielding Jaccard similarity greater than or equal to the observed Jaccard value. For TAM-shuffling, we treat the schema set as non-random when the permutation p -value associated with Jaccard is < 0.05 . Jaccard values below 0.30 indicate that permuted runs typically recover only a small fraction of the observed schemas, providing an interpretable effect-size benchmark.

The p -value is computed as a one-sided test, $P(\text{Jaccard}_{\text{perm}} \geq \text{Jaccard}_{\text{observed}})$, testing whether permuted runs produce schema sets as similar to the observed set as would be expected under the null hypothesis. The interpretive benchmark of Jaccard < 0.30 is an empirical threshold indicating low overlap between permuted and observed schema sets. This threshold is not a fixed statistical criterion but reflects the specific corpus and method; its adequacy is supported by the observed mean Jaccard of 0.244 (SD = 0.111), which falls well below 0.30.

The verb-shuffling design tests whether schema acceptance depends on specific lexical items or operates at the structural TAM $\times$ complement level. For TAM-shuffling, we compute p as the proportion of permutations with Jaccard $\geq$ observed. For verb-shuffling, high Jaccard is expected under the null. We therefore treat the null as supported when 97.7 % of permutations yield Jaccard values greater than or equal to the observed value. If schema acceptance were driven by specific verb $\times$ TAM $\times$ complement triples (lexically-fixed collocations (e.g., modal-be+pp_up_NP instantiated as should be left up to)), verb-shuffling would disrupt this and yield different schema sets. If schema acceptance operates at the structural TAM $\times$ complement level (productive templates), verb-shuffling should recover similar schemas because the metrics target pattern-level rather than verb-level associations.

The complementary nature of these designs addresses a key theoretical question in construction research, namely whether formulaic patterns in learner language reflect item-specific memorization (verb $\times$ TAM $\times$ complement triples stored as chunks) or schematic abstraction (TAM $\times$ complement frames with variable verb slots). If both permutation tests validate the identified schemas – TAM-shuffling confirms non-randomness, verb-shuffling confirms structural rather than lexical determination – this provides strong evidence that the LC Framework correctly identifies validated schemas (both fixed collocations and productive templates) rather than spurious lexical associations.

3.4.3 Sensitivity analysis

Sensitivity analysis assesses the robustness of identified schemas across threshold variations. For each of the five critical thresholds (ATP, $\Delta P_{\text{backward}}$, H_r , NPMI, Hslot), values are systematically varied around their default settings while holding other thresholds constant, using a one-factor-at-a-time (OFAT) design. Table 2 presents the parameter grid.

Table 2: Sensitivity analysis parameter grid.

Parameter	Values tested
ATP	0.05, 0.08, 0.10 (default), 0.12, 0.15
$\Delta P_{\text{backward}}$	0.05, 0.08, 0.10 (default), 0.12, 0.15
H_r	2.0, 2.5, 2.8 (default), 3.0, 3.5
NPMI	0.02, 0.03, 0.05 (default), 0.08, 0.10
Hslot	1.0, 1.25, 1.50 (default), 1.75, 2.0

OFAT design yields 5 parameters $\times$ 5 values = 25 threshold configurations.

For each parameter combination, the complete three-layer pipeline is re-run and schema acceptance is recorded. Schemas are classified according to their stability across the 25 configurations. Globally stable schemas appear in more than 80 % of configurations (≥ 21 of 25), representing robust patterns insensitive to threshold choice. Parameter-sensitive schemas appear in 50–80 % of configurations (13–20 of 25), indicating moderate robustness with some threshold dependency. Parameter-dependent schemas appear in fewer than 50 % of configurations (< 13 of 25), indicating critical reliance on specific threshold settings.

3.4.4 Ablation analysis

Ablation analysis quantifies the contribution of instance-level prefiltering (Layer 1) to false positive control. Without Layer 1 filters, all valid constructions would be aggregated into TAM $\times$ complement patterns and evaluated directly at Layer 2. To assess Layer 1's contribution, the pipeline is run in ablated configuration with ATP, $\Delta P_{\text{backward}}$, and H_r filters disabled, aggregating all valid constructions into TAM $\times$ complement patterns and applying Layer 2 dual-lane acceptance criteria. The number of patterns passing Layer 2 without Layer 1 prefiltering is compared to the number passing with the full pipeline, quantifying the false positive reduction attributable to Layer 1.

4 Results

4.1 Multi-metric modeling of stable patterns (RQ1)

Following the extraction and filtering pipeline described in Section 3, the LC Framework processed 9,529 texts from ICLE, yielding 81,138 passive constructions. After basic filtering removed invalid TAM categories and problematic constructions (extraposed it-passives, ambiguous adjectival predicates), 57,090 valid passive constructions remained (70.4 % retention), representing the input to the three-layer filtering architecture.

Starting from 57,090 passive constructions that passed basic validity filters, the three-layer architecture identified 8 accepted schemas. At Layer 1 (instance-level prefiltering), constructions passed through three sequential gates assessing bidirectional transitional probability ($ATP \geq 0.10$), backward contingency ($\Delta P_{\text{backward}} \geq 0.10$), and right-boundary entropy ($H_r \leq 2.8$). Of the 57,090 valid constructions, 183 (0.32 %) passed all three Layer 1 thresholds and were classified as formulaic instances. The H_r filter proved most restrictive, eliminating 53.8 % of

Table 3: Complete filtering cascade from constructions to schemas.

Stage	Count	Retention
Initial passive constructions	81,138	–
After basic filtering (valid TAM)	57,090	70.4 %
Layer 1: formulaic instances	183	0.32 %
Unique TAM × COMP patterns	85	–
Layer 2: accepted schemas	8	9.4 %

Layer 1 retention is the percentage of valid constructions (183 of 57,090).

constructions that had passed the ATP filter, followed by ATP (36.3 % reduction from input) and $\Delta P_{\text{backward}}$ (8.0 % additional reduction).

These 183 formulaic instances aggregated into 85 unique TAM × complement patterns for schema-level evaluation. At Layer 2 (schema-level acceptance), dual-lane evaluation identified 8 patterns meeting either NPMI lane criteria ($\text{NPMI} \geq 0.05$ AND $G^2 \geq 3.84$) or Hslot lane criteria ($\text{Hslot} \geq 1.5$ AND $G^2 \geq 1.0$ AND $\text{NPMI} \geq 0$). Table 3 summarizes the complete filtering cascade, clarifying the relationship between Layer 1 and Layer 2 outputs.

Within the passive domain, the TAM distribution shows clear preferences. Present-*be* passives account for 48.2 % of valid constructions (27,516 of 57,090), making them the dominant passive type in argumentative writing. Past-*be* contributes 26.7 % (15,243), perfect-*be* contributes 8.7 % (4,967), progressive-*be* contributes 4.0 % (2,284), and modal-*be* contributes 12.4 % (7,080).

Despite modal-*be*'s relatively low corpus frequency (12.4 %), it accounts for 6 of 8 validated schemas (75 % of the schema inventory), demonstrating a striking disassociation between corpus frequency and formulaic stability. This pattern – where the least frequent TAM category generates the most validated schemas – suggests that

Table 4: Validated schema inventory with distributional metrics.

Schema	<i>n</i>	NPMI	Hslot	G^2	Lane	Stability
modal-be+pp_up_NP	20	0.183	1.02	55.28	NPMI	100 %
modal-be+pp_above_NP	7	0.058	2.52	6.94	Both	96 %
modal-be+pp_towards_NP	6	0.115	2.25	4.69	Both	92 %
perf-be+pp_within_NP	5	0.588	2.32	2.78	Hslot	92 %
modal-be+pp_over_NP	16	0.199	3.75	37.58	Both	88 %
past-be+pp_out_NP	5	0.551	1.92	2.78	Hslot	84 %
modal-be+pp_around_NP	4	0.037	1.50	1.28	Hslot	76 % ^a
modal-be+pp_outside_NP	4	0.085	1.50	1.28	Hslot	72 % ^a

n, number of formulaic instances; NPMI, normalized pointwise mutual information, Hslot = Shannon entropy of verb slot (bits). G^2 = log-likelihood ratio. Lane = qualifying pathway (NPMI, Hslot, or both). Stability = percentage of 25 sensitivity analysis configurations where schema appears. ^aParameter-sensitive (50–80 % stability).

formulaic organization in learner passive production operates independently of raw frequency, supporting the theoretical position that systematic patterning emerges through constraint and predictability rather than through repetition alone.

4.2 TAM × complement configurations with robust stability (RQ2)

Table 4 presents the eight validated schemas with their distributional metrics, qualifying lane, and stability classification. All eight schemas involve spatial or directional prepositional complements, and six involve modal-*be* auxiliaries.

The schema inventory exhibits clear TAM clustering. Six of eight schemas (75 %) involve modal-*be* auxiliaries paired with directional or spatial prepositions (*up*, *over*, *above*, *towards*, *around*, *outside*). One schema involves perfect-*be* (*perf-be+pp_within_NP*) and one involves past-*be* (*past-be+pp_out_NP*). No present-*be* or progressive-*be* schemas reached acceptance, despite present-*be*'s dominance in the corpus (48.2 % of valid constructions).

The lane distribution reveals the complementary roles of association and productivity. One schema qualifies exclusively through the NPMI lane (*modal-be+pp_up_NP*), exhibiting strong association but lower lexical productivity (Hslot = 1.02). Three schemas qualify through both lanes (*modal-be+pp_above_NP*, *modal-be+pp_towards_NP*, *modal-be+pp_over_NP*), indicating strong association and high productivity simultaneously. Four schemas qualify exclusively through the Hslot lane (*perf-be+pp_within_NP*, *past-be+pp_out_NP*, *modal-be+pp_around_NP*, *modal-be+pp_outside_NP*), exhibiting productive slot-filling patterns that would be rejected under NPMI-only criteria due to sample size constraints on G^2 .

Instance counts range from $n = 4$ to $n = 20$, with a mean of 8.4 instances per schema. The *modal-be+pp_up_NP* schema exhibits the highest frequency ($n = 20$) and strongest statistical significance ($G^2 = 55.28$). NPMI values range from 0.037 to 0.588 (mean = 0.227), with the highest values in *perf-be+pp_within_NP* (0.588) and *past-be+pp_out_NP* (0.551), indicating very strong TAM-complement associations. Hslot values range from 1.02 to 3.75 bits (mean = 2.10), with the highest lexical diversity in *modal-be+pp_over_NP* (3.75 bits, corresponding to approximately 13 equiprobable verb types).

To ground the abstract schema notation in actual learner production, representative examples from the corpus are provided. The most frequent schema, *modal-be+pp_up_NP* ($n = 20$), is instantiated in phrases such as “*should be left up to*” and “*should be promoted up to*,” where diverse modal auxiliaries combine with the directional preposition *up* in argumentative contexts. The schema *modal-be+pp_above_NP* ($n = 7$) appears in constructions such as “*they must be married above all*” and “*man can be put above any other species*,” while *past-be+pp_out_NP*

($n = 5$) includes “*Her daughter was born out of wedlock*” and “*I was chucked out of the classroom*.” These examples illustrate the directional and spatial semantics shared across all validated schemas.

4.3 Validation results

4.3.1 Permutation testing

Permutation testing employed two complementary designs to validate schema robustness and test different hypotheses about the nature of identified patterns. The results provide convergent evidence that the LC Framework identifies genuine distributional regularities operating at the structural TAM $\times$ complement level rather than statistical artifacts or lexically-specific memorization.

The TAM-shuffling design tested whether observed TAM $\times$ complement patterns represent systematic co-occurrence beyond chance. TAM categories were randomly reassigned across the 183 formulaic instances while preserving complement distributions. Across 1,000 permutations, the mean Jaccard similarity between observed and permuted schema sets was 0.112 (SD = 0.111), indicating that random TAM assignments rarely recovered the same schemas identified in the observed data. The observed Jaccard similarity was 0.244, and the permutation p -value was 0.023, computed as the proportion of permutations yielding Jaccard ≥ 0.244 . This result confirms that the identified schemas reflect genuine TAM $\times$ complement associations rather than artifacts of the dual-lane acceptance criteria.

The verb-shuffling design tested whether schema acceptance depends on specific lexical items or operates at the structural TAM $\times$ complement level. Verb lemmas were shuffled within TAM categories while preserving TAM $\times$ complement pattern assignments. Across 1,000 permutations, the mean Jaccard similarity was 0.985 (SD = 0.030), indicating that permuted data recovered nearly identical schema sets. The observed Jaccard similarity of 0.797 is lower than the permutation mean (0.985, SD = 0.030) but, by construction, almost all permutations yield Jaccard values greater than or equal to 0.797. Under the p -value definition in Section 3.4.2, this

Table 5: Permutation testing: TAM-shuffling versus verb-shuffling comparison.

Design	Observed Jaccard	Perm. mean (SD)	p	Interpretation
TAM-shuffling	0.244	0.112 (0.111)	0.023*	Non-random structure
Verb-shuffling	0.797	0.985 (0.030)	1.000	Structural determination

Both designs employed 1,000 permutation iterations. Observed Jaccard = similarity between observed schema set and permuted sets. Perm. Mean = average Jaccard across 1,000 permutations. * $p < 0.05$.

corresponds to $p \approx 1.0$, so the data provide no evidence that schema acceptance depends on the identity of particular verbs. The verb-shuffling test is therefore consistent with a structurally, rather than lexically, defined schema inventory: shuffling verbs within TAM categories does not systematically disrupt which TAM $\times$ complement patterns are accepted. Table 5 presents the comparison of permutation designs and their theoretical implications.

The convergence of these findings validates the dual-lane architecture’s theoretical rationale. TAM-shuffling confirms that identified schemas are not artifacts of the acceptance criteria applied to arbitrary TAM $\times$ complement combinations – randomizing TAM assignments disrupts the patterns. Verb-shuffling confirms that the patterns operate at the structural level – the metrics correctly identify TAM $\times$ complement frames with systematic organization rather than lexically-specific chunks.

Figure 1 visualizes the permutation distributions for both designs, showing the stark contrast between TAM-shuffling (low Jaccard, confirming non-randomness) and verb-shuffling (high Jaccard, confirming structural determination).

4.3.2 Sensitivity analysis

Sensitivity analysis (see Figure 2 for a summary heatmap) tested the robustness of each schema across threshold variations. For each of the five critical thresholds (ATP, $\Delta P_{\text{backward}}$, H_r , NPMI, Hslot), values were systematically varied around their default

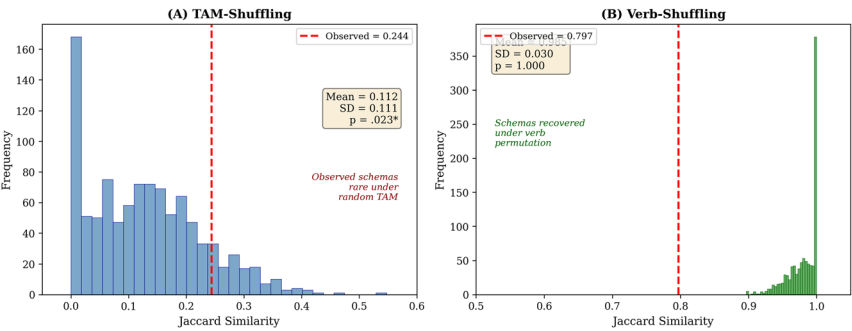


Figure 1: Permutation testing: TAM-shuffling versus verb-shuffling distributions.

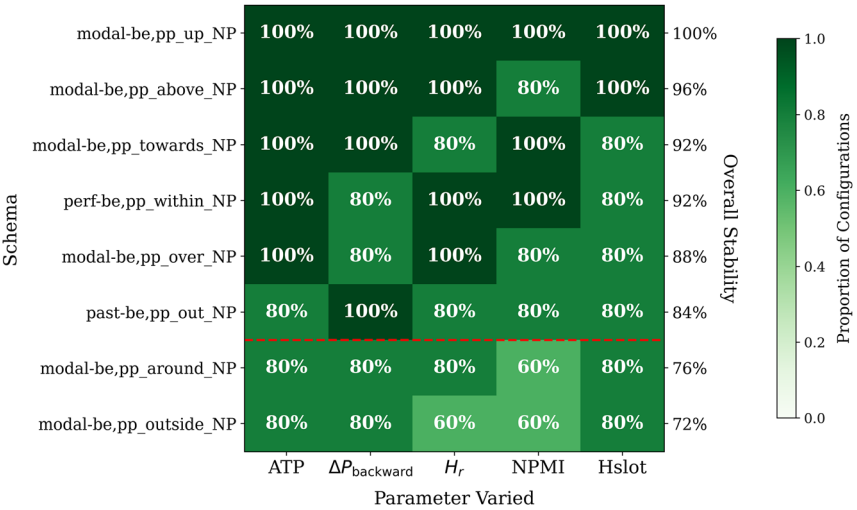


Figure 2: Sensitivity analysis: schema stability across parameter variations. Schemas above the red dashed line are globally stable ($\geq 80\%$ retention); schemas below are parameter-sensitive.

settings while holding other thresholds constant, using a one-factor-at-a-time (OFAT) design.

Six of eight schemas (75 %) demonstrated global stability, appearing in more than 80 % of parameter configurations tested. The globally stable schemas are *modal-be+pp_up_NP* (100 % stability, appearing in all 25 configurations), *modal-be+pp_above_NP* (96 %), *modal-be+pp_towards_NP* (92 %), *perf-be+pp_within_NP* (92 %), *modal-be+pp_over_NP* (88 %), and *past-be+pp_out_NP* (84 %). These schemas exhibit robust distributional signatures that persist across substantial threshold variation.

Two schemas exhibited parameter sensitivity, appearing in 50–80 % of configurations. The schema *modal-be+pp_around_NP* appeared in 76 % of configurations (19 of 25), showing sensitivity primarily to NPMI threshold variations. The schema *modal-be+pp_outside_NP* appeared in 72 % of configurations (18 of 25), showing sensitivity to both H_r and NPMI thresholds. Both parameter-sensitive schemas exhibit borderline values on multiple metrics and low instance counts ($n = 4\text{--}5$), explaining their threshold dependency.

Figure 2 presents a heatmap showing schema stability across parameter variations, distinguishing globally stable schemas (above the dividing line) from parameter-sensitive schemas (below). The global stability of 75 % of schemas (6 of 8) provides strong evidence that the LC Framework identifies robust distributional regularities rather than parameter-dependent artifacts.

The presence of two parameter-sensitive schemas does not undermine this conclusion; rather, it demonstrates that sensitivity analysis correctly identifies patterns whose acceptance depends on threshold choices, providing transparency about which schemas warrant greater interpretive caution.

4.3.3 Ablation analysis

Ablation analysis quantified Layer 1's contribution to false positive control by comparing schema acceptance with versus without instance-level prefiltering. Without Layer 1 filters (ATP, $\Delta P_{\text{backward}}$, H_r), all 57,090 valid constructions aggregated into TAM $\times$ complement patterns and proceeded directly to Layer 2 dual-lane evaluation.

The results demonstrate Layer 1's critical role in precision. Without instance prefiltering, 148 patterns passed Layer 2 criteria; with the full pipeline, 8 schemas were accepted. This represents a 94.6 % reduction in accepted patterns (140 of 148 filtered), indicating that Layer 1 eliminates the vast majority of patterns that would otherwise pass schema-level criteria despite lacking instance-level formulaic signatures.

Manual inspection of a random sample of filtered patterns ($n = 30$) revealed two primary categories of false positives. First, many patterns reflect topical clustering rather than formulaic cohesion. For example, *present-be+pp_despite_NP* passes Hslot criteria due to diverse verbs appearing in argumentative contexts (*be seen despite*, *be considered despite*, *be found despite*) without constituting a productive passive schema. Second, many patterns involve grammatically necessary associations rather than formulaic preferences. For example, *modal-be+to_VP* (modal passive + infinitival complement) achieves high frequency and passes association thresholds because modal passives grammatically require infinitival complementation in certain contexts, not because learners have entrenched this as a formulaic unit.

These findings validate the three-layer architecture as essential for distinguishing formulaic patterns (high ATP, reliable ΔP , low H_r , passing schema-level criteria) from spurious schema-level associations arising from topical or grammatical confounds. The ablation result speaks to the necessity of multi-metric integration, as schema-level metrics alone (NPMI, Hslot) capture association and productivity but cannot distinguish formulaic from non-formulaic usage without the instance-level prefilters that assess transitional probability and boundary predictability.

5 Discussion

The results address the three research questions in turn. Section 5.1 examines how the multi-metric LC Framework models stable patterns in learner passive production (RQ1). Section 5.2 presents the eight TAM $\times$ complement configurations that exhibit statistically robust stability across the validation procedures (RQ2). Section 5.3 discusses what these configurations reveal about the organization of learner passive usage in argumentative writing (RQ3), with Section 5.4 elaborating the frequency-formulaicity dissociation as the most distinctive interpretive finding.

5.1 Dual-lane architecture: theoretical and empirical validation (RQ1)

The dual-lane acceptance system captures two valid pathways to formulaic status, reflecting theoretical distinctions between lexically-specific constructions and abstract schematic templates discussed in Section 2.1. The NPMI lane identifies patterns where specific TAM-complement combinations exhibit strong association and conventional statistical significance, targeting fixed collocations where particular TAM categories systematically co-occur with particular complement types. The Hslot lane identifies patterns where TAM-complement frames accommodate diverse verbs systematically with positive association and statistical consistency, targeting productive templates where lexical variation occurs within stable structural frames. Section 3.3 details the threshold configurations for each lane.

The empirical validation demonstrates that these two pathways are not merely methodological conveniences but reflect genuine distinctions in formulaic organization. Ablation analysis confirms that Hslot-exclusive acceptance rescues schemas that association-only approaches would miss. Four schemas qualify exclusively through the Hslot lane (*perf-be+pp_within_NP*, *past-be+pp_out_NP*, *modal-be+pp_around_NP*, *modal-be+pp_outside_NP*), exhibiting high lexical productivity and positive association but insufficient sample sizes to reach conventional statistical significance ($G^2 < 3.84$). Rejecting these patterns due to sample size limitations would constitute a systematic Type II error, excluding genuine productive templates not due to weak formulaicity but due to mathematical constraints on G^2 for low-frequency patterns.

High Hslot alone does not guarantee productivity – Shannon entropy measures uncertainty, which could reflect systematic accommodation of diverse verbs serving similar discourse functions, accidental heterogeneity where verbs co-occur without systematic constraint, or functional polysemy where different sub-functions within a

schema license different verbs. The Hslot lane criteria jointly ensure that qualifying schemas exhibit systematic lexical diversity rather than arbitrary variation. The NPMI floor (≥ 0) excludes negatively-associated patterns representing grammatical necessity rather than formulaic preference. The G^2 threshold (≥ 1.0) requires statistical consistency, filtering patterns where observed-expected discrepancies are minimal.

The verb-shuffling permutation result (Section 4.3.1) provides convergent validation for this interpretation. If Hslot-exclusive schemas represented accidental heterogeneity or spurious associations, verb-shuffling should disrupt their acceptance by reassigning verbs randomly within TAM categories. Instead, Jaccard similarity between observed and permuted schema sets remained high, with the observed value falling within the permutation distribution. This stability confirms that schema acceptance is determined by structural TAM $\times$ complement patterns rather than specific verb-pattern associations, supporting the interpretation that Hslot-exclusive schemas represent genuine productive templates that accommodate diverse verbs systematically within stable structural frames.

The asymmetric G^2 thresholds employed in the dual-lane architecture require careful theoretical justification, as they represent a departure from conventional hypothesis testing practices where uniform significance thresholds are standard. The justification rests on three principles. First, different constructs warrant different thresholds. Association measures (NPMI) target specific TAM-complement pairings where conventional significance is appropriate. Productivity measures (Hslot), by contrast, target schematic frames where lexical variation is the defining property and sample sizes are inherently constrained by token distribution across verb types. Second, G^2 is proportional to sample size such that low-frequency patterns cannot achieve $G^2 \geq 3.84$ regardless of association strength, creating systematic bias against rare but validated patterns. Third, empirical validation through permutation testing confirms that the asymmetric configuration yields a schema set that differs significantly from randomized baselines. Because the four schemas retained under symmetric thresholds are a subset of this inventory, they can likewise be interpreted as non-random patterns, even under more conservative G^2 requirements.

5.2 The validated schema inventory (RQ2)

The empirical distribution of schemas across the two lanes illuminates the distinction between lexically-cohesive and productive systematic patterning. One schema qualifies exclusively through the NPMI lane (modal-be+pp_up_NP), exhibiting strong TAM-complement association (NPMI = 0.183, $G^2 = 55.28$) but lower lexical productivity

(Hslot = 1.02). This represents a lexically-cohesive pattern approaching the formulaic sequence prototype, a recurrent TAM-complement pairing with relatively constrained verb inventory.

Four schemas qualify exclusively through the Hslot lane (perf-be+pp_within_NP, past-be+pp_out_NP, modal-be+pp_around_NP, modal-be+pp_outside_NP), exhibiting high lexical diversity (Hslot ≥ 1.50) but weaker association relative to conventional significance thresholds ($G^2 < 3.84$). These represent productive templates, schematic TAM-complement frames that systematically accommodate diverse verbs serving similar discourse functions.

Three schemas qualify through both lanes (modal-be+pp_above_NP, modal-be+pp_towards_NP, modal-be+pp_over_NP), exhibiting both strong association ($G^2 \geq 4.69$) and meaningful lexical productivity (Hslot ≥ 1.92), representing intermediate positions on the specificity-schematicity continuum (Section 2.1) where systematic TAM-complement pairing co-exists with lexical variation.

This distribution validates the dual-lane architecture as capturing theoretically meaningful distinctions rather than arbitrary methodological divisions. The schemas identified represent genuine variation in how systematic patterning manifests in learner production. Some achieve this through lexical cohesion (relatively fixed verb inventories within strong TAM-complement associations), others through productive schematicity (diverse verbs systematically instantiating stable structural frames), and still others through both simultaneously. All eight schemas exhibit non-random distributional organization validated through permutation testing (Section 4.3.1), but they differ in whether this organization emerges through lexical fixedness, structural productivity, or both.

5.3 Organization of learner passive usage in argumentative writing (RQ3)

The validated inventory demonstrates that learner passive production in argumentative writing exhibits systematic distributional organization beyond raw frequency. The eight schemas represent a highly selective subset (9.4 %) of the 85 attested TAM $\times$ complement patterns that passed instance-level formulaicity criteria, and an even smaller proportion relative to the 57,090 valid passive tokens in the corpus (8 schematic types for 57,090 tokens, ≈ 0.014 %). This selectivity demonstrates that formulaic organization in advanced learner writing reflects convergent evidence across the multiple distributional dimensions operationalized in Section 3 rather than simple frequency accumulation.

The schema inventory exhibits two striking characteristics. First, TAM concentration is evident, with six of eight schemas (75 %) involving modal-*be* auxiliaries,

despite modal-*be* accounting for only 12.4 % of valid passive constructions. This frequency-formulaicity dissociation suggests that systematic patterning emerges through constraint and predictability rather than through repetition alone. Second, complement specialization is apparent, with all eight schemas involving prepositional complements with directional, spatial, or boundary semantics (*up, over, above, towards, around, within, out, outside*). Canonical *by*-phrase passives, despite their grammatical centrality and corpus frequency, do not emerge as validated schemas. This absence suggests that learners deploy *by*-phrases productively across varied contexts and TAM categories without developing stable TAM-specific routines.

These findings should be interpreted within their register context. Argumentative essays impose specific discourse demands – stance-marking, agent suppression, epistemic hedging, information management – that may privilege modal passives with directional prepositions for encoding abstract argumentative relationships. The concentration of validated schemas around modal-*be* + directional PP may reflect register-specific adaptations rather than general principles of learner passive organization. Whether the same patterns emerge in narrative registers (which might favor eventive past-*be* passives), scientific registers (which might favor resultative perfect-*be* constructions), or spoken discourse (which might show different TAM distributions entirely) requires systematic investigation across genre boundaries. As noted in Section 3.1, the ICLE corpus includes 364 literary essays (3.8 %) and 200 texts classified as “other” (2.1 %). All validated schema examples originate from argumentative texts, confirming that the findings pertain to the argumentative register.

5.4 The modal passive puzzle: frequency-formulaicity dissociation (RQ3, continued)

The most striking finding is the frequency-formulaicity dissociation reported in Section 4.1. Modal-*be* passives account for only 12.4 % of valid constructions but generate 75 % of validated schemas (6 of 8), while present-*be* passives dominate the corpus (48.2 %) but yield no validated schemas. This inversion challenges simplistic frequency-based accounts of formulaic language acquisition and suggests that systematic patterning requires constraint and predictability rather than exposure frequency alone.

Several factors consistent with usage-based accounts may explain this dissociation. Present-*be* passives, encountered frequently across diverse contexts in both input and learner production, may remain productive and flexible rather than consolidating into fixed schemas. High-frequency exposure across varied functions (describing states, reporting claims, establishing facts, expressing evaluations) may

prevent the formation of context-specific routines, as each present-be passive instance contributes to a broad, flexible template rather than to specialized sub-templates. This aligns with exemplar-theoretic proposals where forms encountered across diverse contexts build broad exemplar clusters encoding contextual variety, supporting productive generalization; by contrast, forms concentrated in constrained contexts consolidate into formulaic chunks through repetition (Bybee 2010).

Alternatively, the pattern may reflect input distributions in instructional materials rather than inherent properties of learner grammar development. If pedagogical materials emphasize modal passives with directional prepositions as exemplars of academic writing style, learners may reproduce these patterns systematically, creating distributional biases in production that reflect instructional emphasis rather than usage-based consolidation from naturalistic input. Testing this hypothesis requires systematic comparison of identified schemas with constructions emphasized in writing instruction and model texts.

According to usage-based accounts (Section 2.1), this frequency-formulaicity dissociation supports the position that constructions exhibit graded entrenchment (Hilpert and Diessel 2017) and that formulaicity emerges through constraint and systematicity rather than frequency alone. Frequency provides opportunity for pattern extraction, but systematic organization requires convergence across the metrics operationalized in the LC Framework. High-frequency patterns lacking this convergence remain productive but not formulaic; lower-frequency patterns exhibiting convergence consolidate into formulaic schemas.

5.5 Theoretical and methodological implications

The LC Framework demonstrates that formulaic patterns in learner language emerge through multidimensional distributional convergence rather than frequency alone, addressing a fundamental gap between usage-based acquisition theories (which emphasize frequency-driven learning) and corpus-based observations (which show frequency–formulaicity dissociations). The validated schemas exhibit systematic properties across the five metrics operationalized in Section 3, and this multi-metric integration captures convergent distributional signatures that single-metric approaches cannot identify.

The dual-lane architecture provides a principled solution to the lexical-specificity versus schematicity tension in Construction Grammar. Rather than imposing uniform criteria that privilege either lexically-fixed collocations or abstract templates, the framework recognizes both as valid manifestations of formulaic organization, admitting patterns through complementary pathways with asymmetric thresholds accommodating their different statistical properties. This

methodological pluralism aligns with Construction Grammar's theoretical position that constructions vary along a lexical-specificity continuum and that formulaicity manifests at different levels of abstraction (Section 2.1).

Methodologically, the framework provides a replicable template for construction mining that balances statistical conservatism with theoretical interpretability. The validation procedures (Section 4.3) demonstrate that rigorous statistical validation – permutation testing confirming non-random structure, ablation analysis quantifying false positive control, sensitivity analysis establishing parameter robustness – can coexist with theoretical flexibility in recognizing different types of formulaic organization. This addresses longstanding concerns in corpus linguistics about balancing Type I error control with Type II error avoidance (Gries 2015, 2024).

The theoretical contribution extends to debates about corpus evidence and cognitive interpretation in linguistics broadly (Gries 2024; Stefanowitsch 2013). The framework demonstrates that corpus linguistics can provide rigorous distributional characterization – identifying patterns through multi-metric convergence, validating non-randomness through permutation, establishing parameter robustness through sensitivity, controlling false positives through ablation – without overstepping into cognitive claims requiring independent psycholinguistic evidence. Corpus distributions identify candidates for formulaic constructions; cognitive status requires converging evidence from processing studies, priming experiments, and elicited production. This methodological position preserves corpus linguistics' empirical rigor while avoiding overinterpretation of corpus patterns as direct evidence of cognitive representation.

Whether the identified schemas reflect cognitive representation (entrenched TAM-complement mappings in learner grammars), processing efficiency (formulaic sequences retrieved holistically during production; cf. Conklin and Schmitt 2012), or pedagogical artifacts (patterns emphasized in instruction and reproduced systematically) remains an open question requiring independent psycholinguistic and instructional investigation. If schemas reflect processing routines, they should exhibit retrieval characteristics – rapid production with minimal disfluency, resistance to online modification, chunked prosody – testable through spoken production tasks, self-paced production with time pressure, or eye-tracking during reading and writing. What the present findings establish with confidence is that advanced learner passive production in argumentative writing exhibits systematic distributional organization, and that this organization is captured through multidimensional convergence measured by the LC Framework.

The pipeline's modular architecture supports extension beyond passive constructions to other grammatical domains exhibiting TAM variation and complement diversity. Candidate constructions include ditransitive alternations, complement clause patterns, and aspectual periphrases, each presenting analogous

combinatorial complexity amenable to multi-metric analysis. The dual-lane logic generalizes to any construction type where lexically-cohesive and productive schematic patterns coexist. Adaptation requires only substitution of the structural variables (replacing TAM and complement categories with construction-appropriate dimensions) while retaining the three-layer filtering architecture and validation procedures.

5.6 Limitations

Several limitations constrain interpretation of the present findings. First, register specificity is a concern, as the analysis targeted ICLE texts (predominantly argumentative), and as noted in Section 5.1, the concentration of validated schemas around modal-be + directional PP may reflect register-specific properties of argumentative discourse. Whether these patterns generalize to other registers requires systematic cross-register comparison.

Second, the corpus represents advanced learners (CEFR B2-C1) from diverse L1 backgrounds writing under varied task conditions (approximately 40 % timed, with varied access to reference tools). Whether the identified schemas characterize specific proficiency levels, L1 groups, or task conditions requires investigation with more fine-grained corpus partitioning and developmental sampling.

Third, the framework focuses on TAM $\times$ complement patterns, treating verb lemmas as variable within schemas (high Hslot indicates productivity) or as anchoring specific schemas (low Hslot with high NPMI). This design cannot capture verb-specific formulaic sequences where particular verbs systematically co-occur with particular TAM-complement configurations (e.g., *can be seen as, is believed to*). Extending the framework to verb $\times$ TAM $\times$ complement triples would enable detection of such lexically-anchored schemas but would require larger datasets to achieve adequate cell counts.

Fourth, distributional evidence cannot establish cognitive representation directly. The schemas identified exhibit robust corpus-level systematicity, but whether this systematicity reflects entrenched mental representations, processing routines, or artifacts of instruction requires independent psycholinguistic and pedagogical investigation. The LC Framework provides principled distributional characterization that can motivate and inform such investigation, but corpus evidence alone cannot resolve questions of cognitive status.

6 Conclusions

6.1 Summary of findings

This study introduced the Linguistic Clusters Framework for identifying stable distributional patterns in learner passive production through multi-metric integration (Section 3). Applied to 9,529 ICLE texts (predominantly argumentative essays; 81,138 passive constructions, 57,090 valid after basic filtering), the three-layer filtering architecture identified eight statistically validated schemas from 183 formulaic instances distributed across 85 unique TAM $\times$ complement patterns.

The validation procedures detailed in Section 4.3 confirmed non-random structure through complementary permutation designs, demonstrated parameter robustness through sensitivity analysis (75 % of schemas exhibiting global stability), and quantified false positive control through ablation analysis (94.6 % reduction in accepted patterns with Layer 1 filtering).

The validated schemas cluster predominantly around modal passives with directional prepositional complements (*up, over, above, towards, around, within, out, outside*), encoding recurrent discourse functions in argumentative writing. As reported in Section 4.1, the schema inventory exhibits a striking frequency-formulaicity dissociation, with modal-*be* passives generating 75 % of validated schemas despite constituting only 12.4 % of corpus constructions. This inversion shows that formulaic status aligns more closely with constraint and contextual predictability than with raw token counts.

The dual-lane architecture (Section 3.3) validated two complementary pathways to formulaic status, with the NPMI lane capturing fixed collocations and the Hslot lane capturing productive templates. Ablation analysis confirmed that Hslot-exclusive acceptance rescues 4 schemas (50 % of the validated inventory) that association-only approaches would miss, demonstrating the necessity of accommodating productive templates alongside fixed collocations in formulaic pattern identification.

6.2 Future directions

Cross-register comparison represents the most critical next step. The application of the LC Framework to narrative, scientific, and spoken learner corpora can establish whether the modal + directional PP cluster reflects argumentative discourse demands or more general patterns of L2 passive organization. If the pattern persists

across registers, it suggests learner-specific distributional regularities independent of genre; if it varies by register, it confirms genre-specific formulaic adaptation.

Experimental validation through psycholinguistic methods would test whether schemas identified through corpus analysis predict processing advantages, production preferences, or structural priming effects. If validated schemas reflect entrenched constructions, they should facilitate comprehension (faster reaction times), prime production (structural priming within schemas but not across schemas), and resist online modification (chunked retrieval). Converging evidence from corpus distributions and experimental methods would establish cognitive validity of distributional patterns.

Finally, instructional analysis comparing identified schemas with pedagogical materials, model texts, and explicit instruction would test pedagogical influence hypotheses. If modal + directional PP schemas appear as formulaic exemplars in academic writing guides or model essays, this would support pedagogical chunking as a contributing factor. Combined with native speaker corpus comparison and experimental validation, instructional analysis would help adjudicate among entrenchment, processing, and pedagogical interpretations of identified schemas. As a preliminary step, the LC Framework has been applied to argumentative essays from the LOCNESS native speaker corpus (results available in the repository); full cross-register comparison is in preparation.

Acknowledgments: We thank two anonymous reviewers for their careful reading and constructive suggestions.

Research funding: This research received no external funding.

Data availability: The complete analytical pipeline, including Python scripts for metric computation, validation procedures, and statistical testing, is available on OSF and GitHub. The ICLE corpus is available under license from the Centre for English Corpus Linguistics, UCLouvain (Granger et al. 2020). Filtered construction data and validated schema inventories are included in the repository. Document-level identifiers for all corpus examples cited in the text are provided in the repository.

References

- Appel, Randy & Pavel Trofimovich. 2017. Transitional probability predicts native and non-native use of formulaic sequences. *International Journal of Applied Linguistics* 27(1). 24–43.
- Aslin, Richard N. & Elaine L. Newport. 2012. Statistical learning: From acquiring specific items to forming general rules. *Current Directions in Psychological Science* 21(3). 170–176.
- Bentz, Christian & Dimitrios Alikaniotis. 2016. The word entropy of natural languages. arXiv preprint. <https://arxiv.org/abs/1606.06996> (accessed 18 December 2025).
- Biber, Douglas & Susan Conrad. 2009. *Register, genre, and style*. Cambridge: Cambridge University Press.

- Bouma, Gerlof. 2009. Normalized (pointwise) mutual information in collocation extraction. In Sam Featherston & Susanne Winkler (eds.), *The fruits of empirical linguistics*, Vol. 1, 31–44. Berlin: Mouton de Gruyter.
- Bybee, Joan. 2010. *Language, usage, and cognition*. Cambridge: Cambridge University Press.
- Church, Kenneth W. & Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1). 22–29.
- Conklin, Kathy & Norbert Schmitt. 2012. The processing of formulaic language. *Annual Review of Applied Linguistics* 32. 45–61.
- Degaetano-Ortlieb, Stefania & Elke Teich. 2018. Using relative entropy for detection and analysis of periods of diachronic linguistic change. In *Proceedings of the 2nd joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature*, 22–33. Stroudsburg, PA: Association for Computational Linguistics.
- Degaetano-Ortlieb, Stefania & Elke Teich. 2019. Toward an optimal code for communication: The case of scientific English. *Corpus Linguistics and Linguistic Theory* 18(1). 175–207.
- Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1). 61–74.
- Ellis, Nick C. & Teresa Cadierno. 2009. Constructing a second language: Introduction to the special section. *Annual Review of Cognitive Linguistics* 7(1). 111–139.
- Ellis, Nick C. & Fernando Ferreira-Junior. 2009. Construction learning as a function of frequency, frequency distribution, and function. *The Modern Language Journal* 93(3). 370–385.
- Ellis, Nick C. 2006. Selective attention and transfer phenomena in SLA: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics* 27(2). 164–194.
- Forsyth, Richard & Łukasz Grabowski. 2015. Is there a formula for formulaic language? *Poznań Studies in Contemporary Linguistics* 51(4). 511–549.
- Gilquin, Gaëtanelle & Magali Paquot. 2008. Too chatty: Learner academic writing and register variation. *Corpora* 3(2). 125–155.
- Gilquin, Gaëtanelle. 2025. Transfer of collocations: The case of causative constructions. *Corpus Linguistics and Linguistic Theory* 21(3). 665–685.
- Goldberg, Adele E. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.
- Goldberg, Adele E. 2006. *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.
- Granger, Sylviane, Maité Dupont, Fanny Meunier, Hubert Naets & Magali Paquot (eds.). 2020. *International Corpus of Learner English*, version 3. Louvain-la-Neuve: Presses universitaires de Louvain.
- Granger, Sylviane. 2009. The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In Karin Aijmer (ed.), *Corpora and language teaching*, 13–32. Amsterdam & Philadelphia: John Benjamins.
- Gries, Stefan Th. & Anatol Stefanowitsch. 2004. Extending collocation analysis: A corpus-based perspective on ‘alternations’. *International Journal of Corpus Linguistics* 9(1). 97–129.
- Gries, Stefan Th. & Stefanie Wulff. 2025. Introduction to the special issue on collocations. *Corpus Linguistics and Linguistic Theory* 21(3). 465–474.
- Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4). 403–437.
- Gries, Stefan Th. 2013. 50-something years of work on collocations: What is or should be next.... *International Journal of Corpus Linguistics* 18(1). 137–166.

- Gries, Stefan Th. 2015. The most under-used statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora* 10(1). 95–125.
- Gries, Stefan Th. 2024. *Frequency, dispersion, association, and keyness: Revising and tupleizing corpus-linguistic measures*. Amsterdam & Philadelphia: John Benjamins.
- Hilpert, Martin & Holger Diessel. 2017. Entrenchment in construction grammar. In Hans-Jörg Schmid (ed.), *Entrenchment and the psychology of language learning: How we reorganize and adapt linguistic knowledge*, 57–74. Washington, DC: American Psychological Association.
- Hilpert, Martin. 2014. *Construction grammar and its application to English*. Edinburgh: Edinburgh University Press.
- Honnibal, Matthew, Ines Montani, Sofie Van Landeghem & Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python. <https://spacy.io> (accessed 18 December 2025).
- Larsson, Tove, Magali Paquot & Biber Douglas. 2021. On the importance of register in learner writing: A multi-dimensional approach. In Elena Seoane & Douglas Biber (eds.), *Corpus-based approaches to register variation*, 235–258. Amsterdam & Philadelphia: John Benjamins.
- Lijffijt, Jeffrey, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamäki & Heikki Mannila. 2016. Significance testing of word frequencies in corpora. *Digital Scholarship in the Humanities* 31(2). 374–397.
- Mansfield, John. 2021. The word as a unit of internal predictability. *Linguistics* 59(6). 1427–1472.
- Maziarz, Marek, Łukasz Grabowski, Tadeusz Piotrowski, Ewa Rudnicka & Maciej Piasecki. 2023. Lexicalisation of Polish and English word combinations: An empirical study. *Poznań Studies in Contemporary Linguistics* 59(2). 381–406.
- Rastelli, Stefano. 2020. Contingency learning and perfective morpheme productivity in L2 Italian: A study on lexeme–morpheme associations with ΔP. *Corpus Linguistics and Linguistic Theory* 16(3). 459–486.
- Rastelli, Stefano. 2022. Intra-language: The study of L2 morpheme productivity as within-item variance. *International Review of Applied Linguistics in Language Teaching* 60(4). 1143–1171.
- Saffran, Jenny R., Richard N. Aslin & Elaine L. Newport. 1996. Statistical learning by 8-month-old infants. *Science* 274(5294). 1926–1928.
- Shannon, Claude E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27(3). 379–423.
- Stefanowitsch, Anatol & Stefan Th. Gries. 2003. Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8(2). 209–243.
- Stefanowitsch, Anatol & Stefan Th. Gries. 2005. Covarying collexemes. *Corpus Linguistics and Linguistic Theory* 1(1). 1–43.
- Stefanowitsch, Anatol. 2013. Collostructional analysis. In Thomas Hoffmann & Graeme Trousdale (eds.), *The Oxford handbook of construction grammar*, 290–306. Oxford: Oxford University Press.
- Sun, Kun & Rong Wang. 2021. Using the relative entropy of linguistic complexity to assess L2 language proficiency development. *Entropy* 23(8). 1080.
- Sung, Hakyung & Kristopher Kyle. 2024. Leveraging pre-trained language models for linguistic analysis: A case of argument structure constructions. In .Proceedings of the 2024 conference on empirical methods in natural language processing, 6200–6214. Stroudsburg, PA: Association for Computational Linguistics.
- Suttle, Laura & Adele E. Goldberg. 2011. The partial productivity of constructions as induction. *Linguistics* 49(6). 1237–1269.
- Wray, Alison & Michael R. Perkins. 2000. The functions of formulaic language: An integrated model. *Language & Communication* 20(1). 1–28.
- Wray, Alison. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

- Wulff, Stefanie & Nick C. Ellis. 2018. Usage-based approaches to second language acquisition. In David Miller, Fatih Bayram, Jason Rothman & Ludovica Serratrice (eds.), *Bilingual cognition and language: The state of the science across its subfields*, 37–56. Amsterdam & Philadelphia: John Benjamins.
- Zhong, Yu, Simon Todd, Stefan Th. Gries & Laurel Brehm. 2025. Metacognition of frequency, directional association strength, and dispersion of MWEs in first and second language speakers. In *Studies in second Language Acquisition*. Advance online publication .